Optimistic Sample Size Estimate for Deep Neural Networks

Yaoyu Zhang

Shanghai Jiao Tong University

Institute of Natural Sciences&School of Mathematical Sciences

One World Seminar Series on the Mathematics of Machine Learning

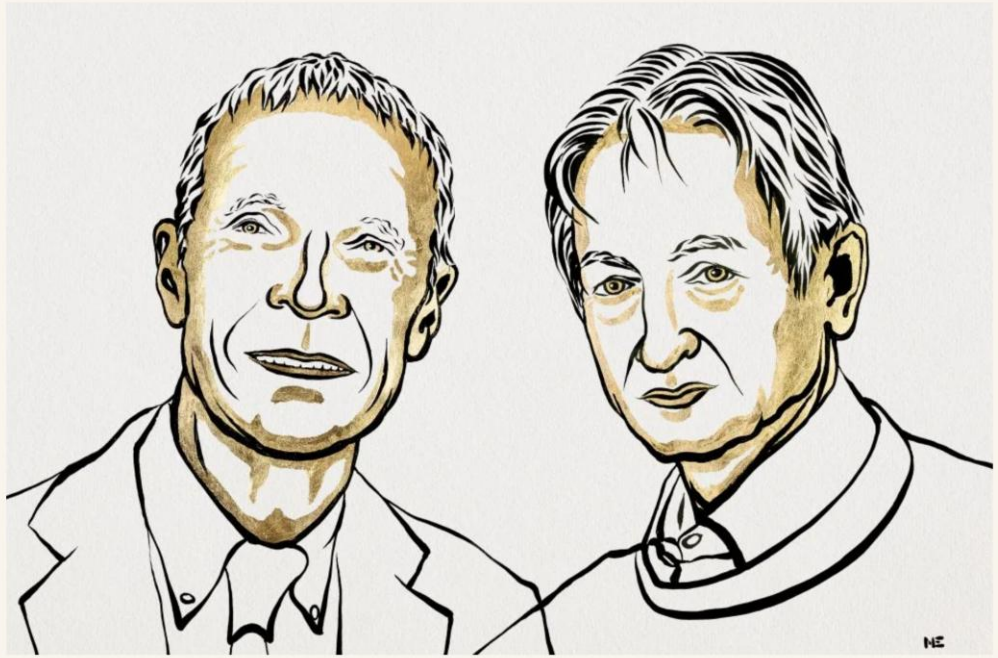
Success of theory-inspired models

Nobel Prize in Physics

The 2024 physics laureates

The Nobel Prize in Physics 2024 was awarded to John J. Hopfield and Geoffrey E. Hinton "for foundational discoveries and inventions that enable machine learning with artificial neural networks."

Hopfield created a structure that can store and reconstruct information. Hinton invented a method that can independently discover properties in data and which has become important for the large neural networks now in use.



Ill. Niklas Elmehed © Nobel Prize Outreach

(Neurons) Their basic structure has close similarities with spin models in statistical physics applied to magnetism or alloy theory. This year's Nobel Prize in Physics recognizes research exploiting this connection to make breakthrough methodological advances in the field of ANN.

What theory may inspire next-gen model?

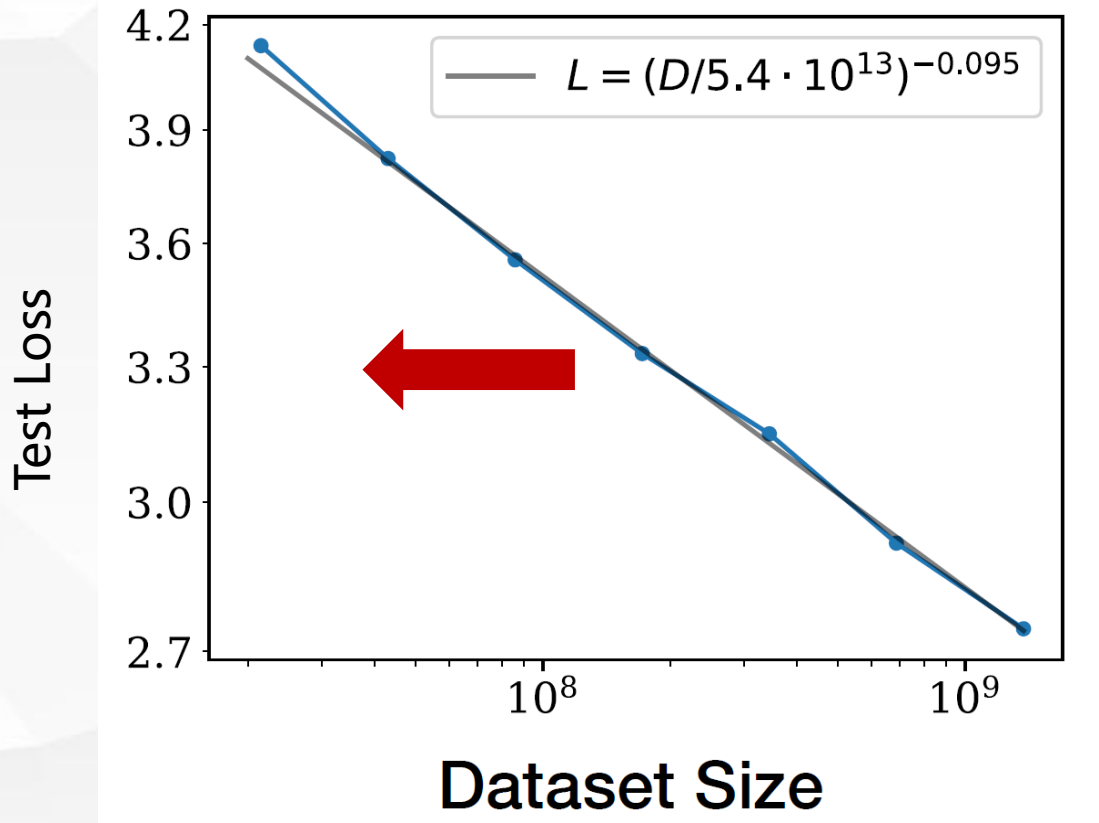
Spin theory -> Hopfield network, Boltzmann machine

? theory -> next-gen model?

Let’s build up theory for current ANNs first!

Estimate sample efficiency is important

Sample efficiency: sample size required for certain performance



Scaling Laws for Neural Language Models. OpenAI 2020

Why optimistic?

Limitation of worst-case based classic estimate

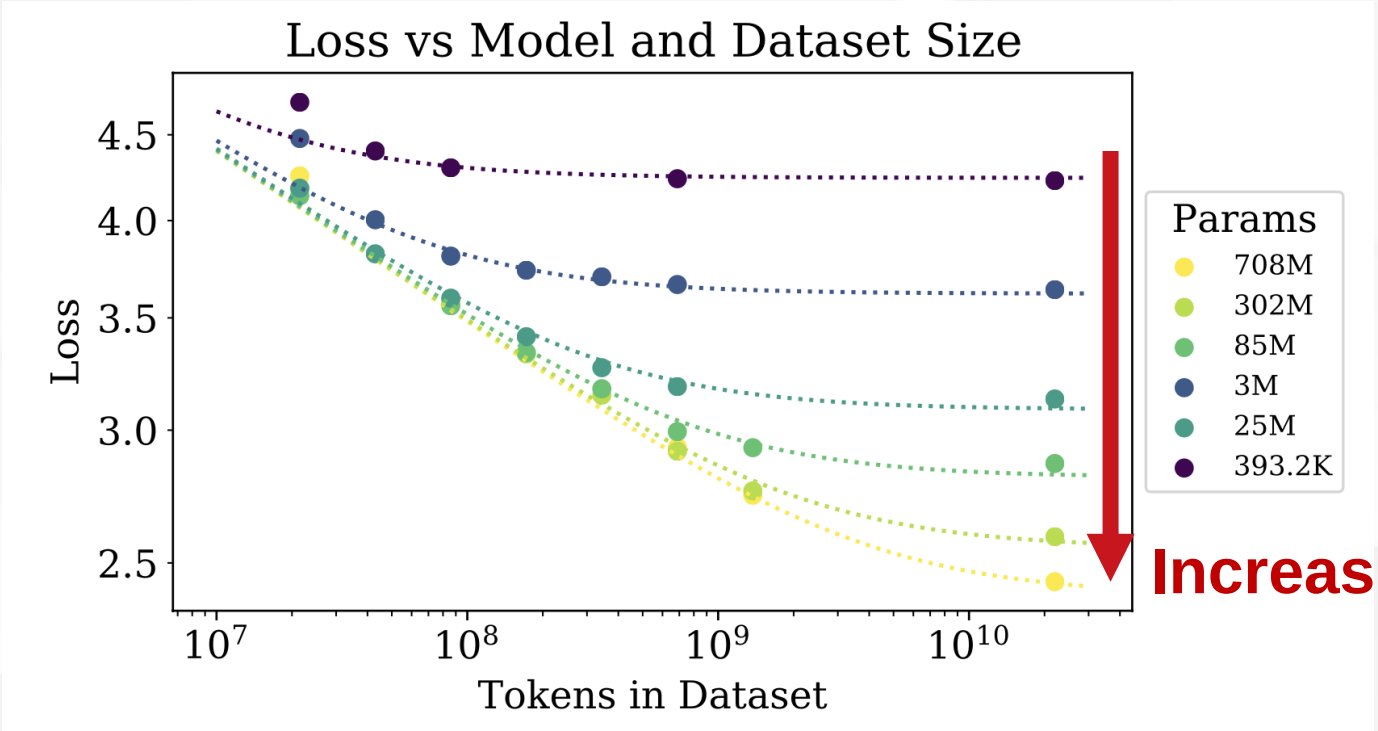
"Parameters" should not be multiplied unnecessarily

Example:

target	**model**	**sample size needed**
1-st order polynomial	$(M-1)$-th polynomial	M

$$a_0 + \sum_{i=1}^{M-1} a_i x^i$$

Increasing parameters benefits sample efficiency



Scaling Laws for Neural Language Models. OpenAI 2020

Long standing problems in deep learning theory

Why don't overparameterized NNs reduce sample efficiency?

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?

What is the sample size required for fitting

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

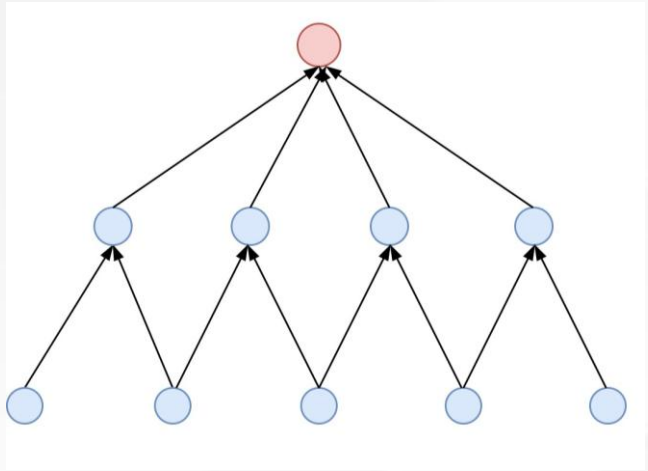
Reflections After Refereeing Papers for NIPS

1995

Optimistic estimate

At the best-possible scenario，what is the sample size required for fitting?

How to increase parameter size?

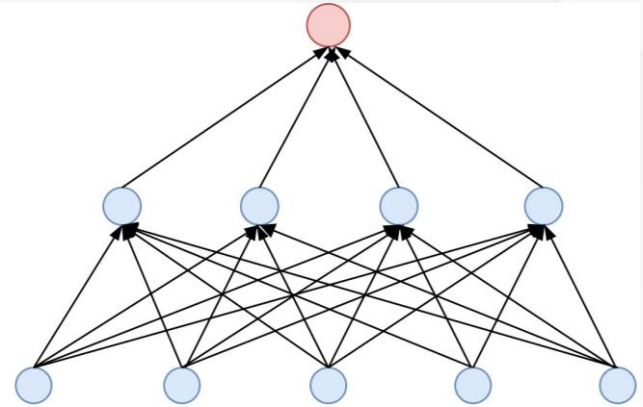


12

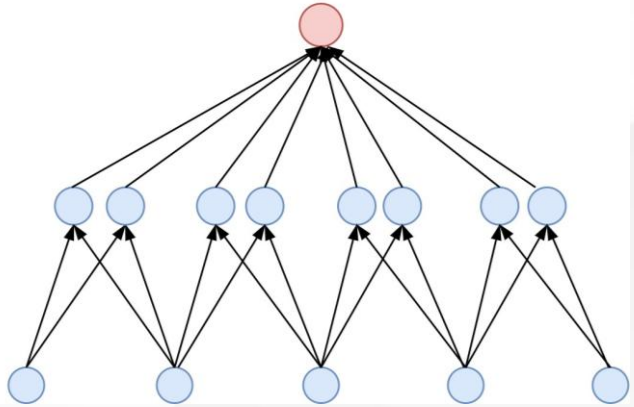
Increase connection

or

Increase width

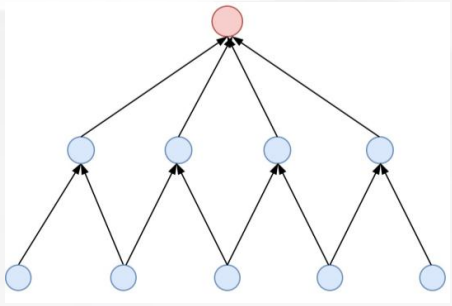


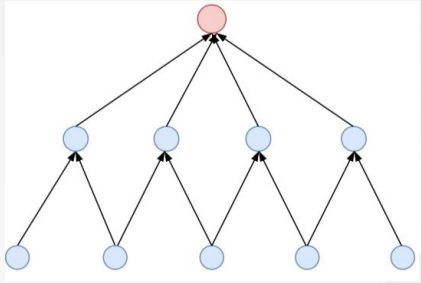
24

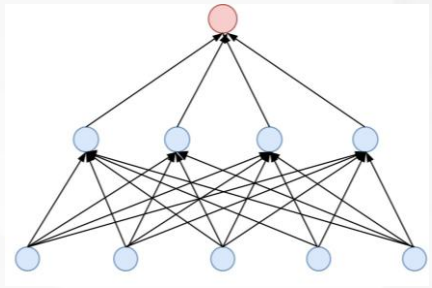


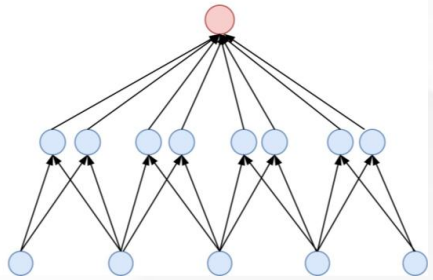
24

Failure of classic estimate based on worst cases

$f^* =$ 

$\mathrm{NN_A}$ 

$\mathrm{NN_B}$ 

$\mathrm{NN_C}$ 

classic estimate

12

24

~~24~~ 12

reduce sample efficiency

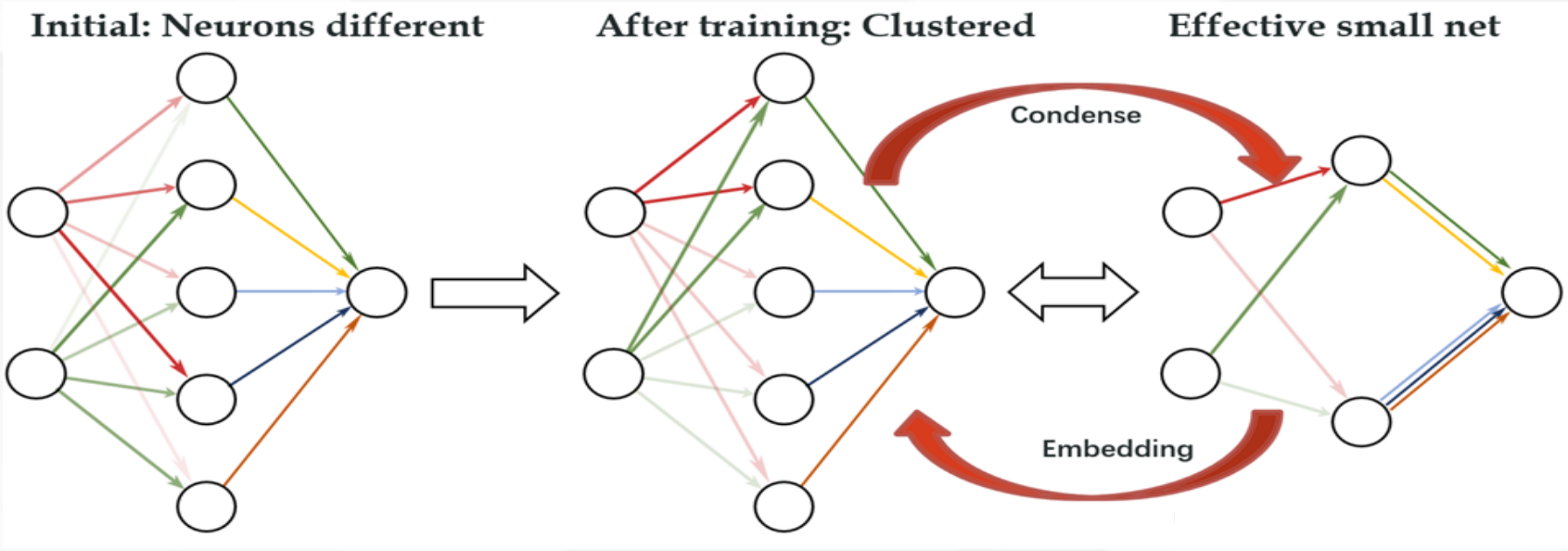
optimistic!

Classic estimate:

sample size for fitting = parameter size

Condensation——the clustering behavior of neurons

Condensation improves sample efficiency



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

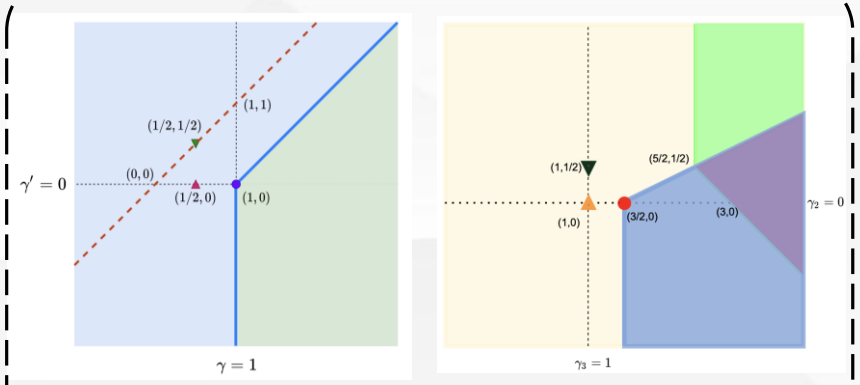
$$\boldsymbol{w}_1 = \boldsymbol{w}_2,$$
$$\boldsymbol{w}_3 = \boldsymbol{w}_4 = \boldsymbol{w}_5$$

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

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 (JMLR) 22(71):1−47, (2021)

Series works on condensation

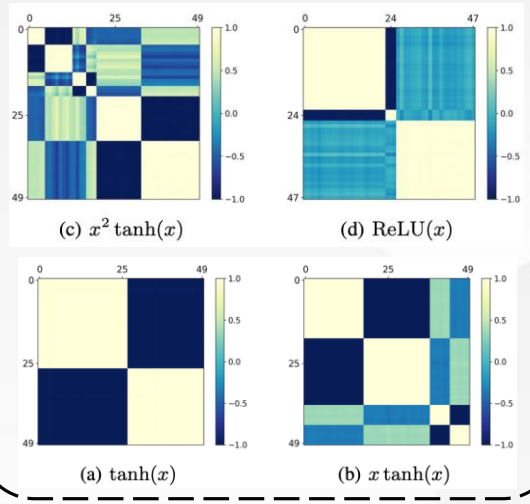


2-layer phase-diagram

3-layer phase-diagram

Condition of condensation [1, 2]

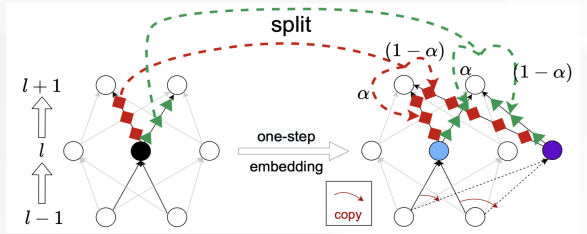
[1] Luo, Xu, Ma, Zhang, JMLR, 2021,
[2] Zhou, Zhou, Jin, Luo, Zhang, Xu, NeurIPS 2022.



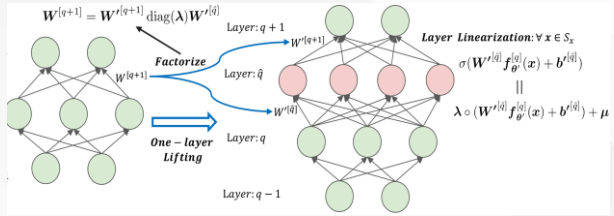
Initial condensation [3]

[3] Zhi-Qin John Xu, Hanxu Zhou, Tao Luo, Yaoyu Zhang NeurIPS 2022

Embedding Principle (width)

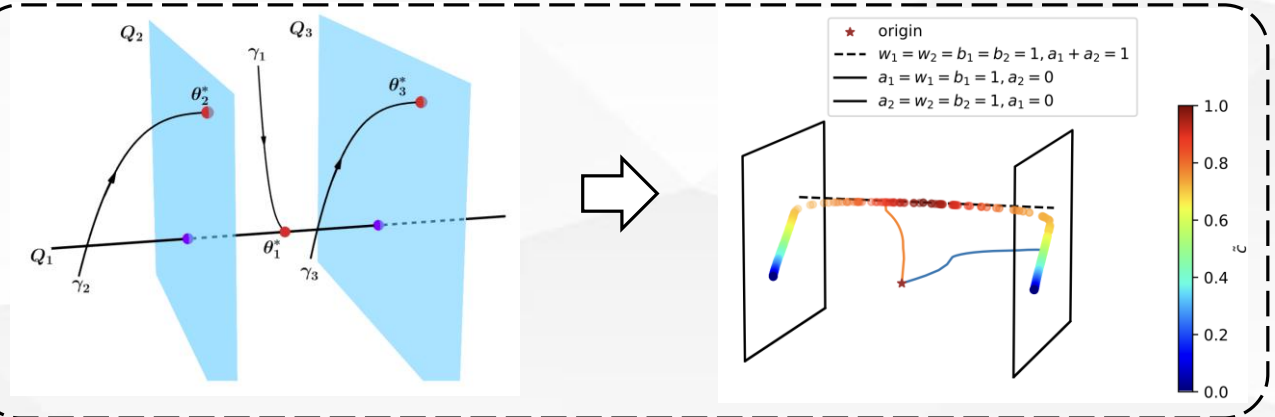


Embedding Principle (depth)



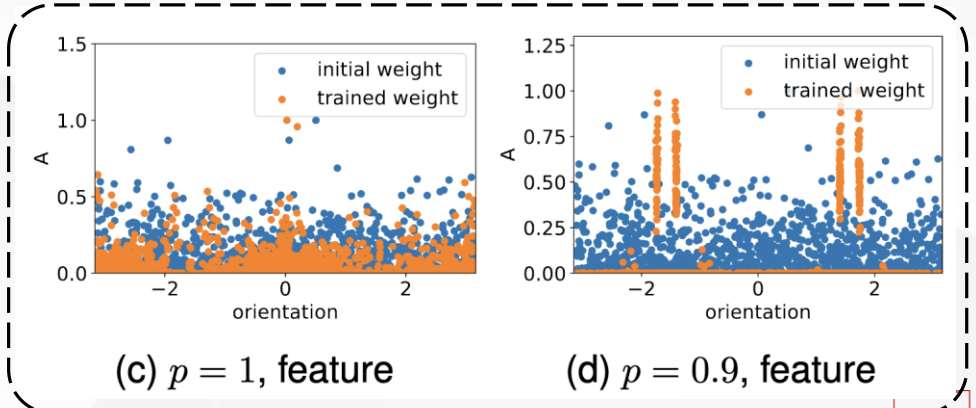
Condensed critical points [4, 5, 6]

[4] Zhang, Zhang, Luo, Xu, NeurIPS 2021 Spotlight
[5] Zhang, Li, Zhang, Luo, Xu, JML 2022
[6] Bai, Luo, Xu, Zhang, CSIAM 2024.



Final state condensation [7, 8]

[7] Leyang Zhang, Yaoyu Zhang, Tao Luo, arXiv (2023)
[8] Jiajie Zhao, Zhiwei Bai, Yaoyu Zhang, arXiv (2024)

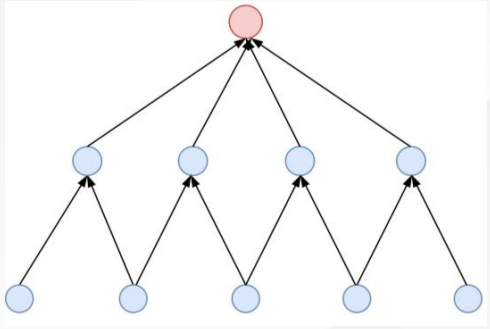


Condensation facilitation (dropout) [9, 10]

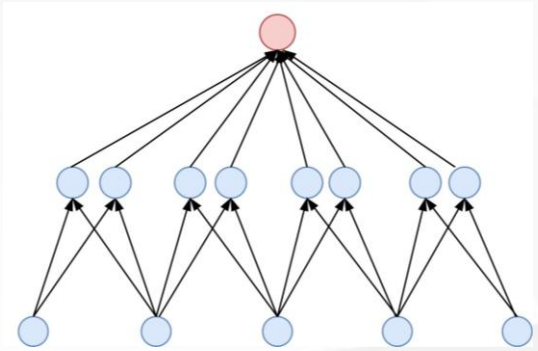
[9] Zhongwang Zhang, Zhi-Qin John Xu,(TPAMI), 2024
[10] Zhongwang Zhang, Yuqing Li*, Tao Luo*, Zhi-Qin John Xu*,ICLR,2024

Condensation improves sample efficiency

How many samples are required to fit f^*?

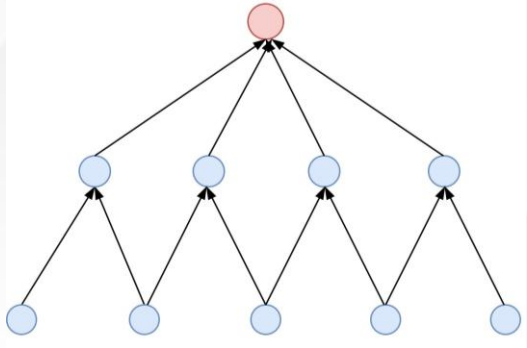
$f^* =$ 

optimistic estimate



$\mathrm{NN_C}$

condense



$\mathrm{NN_A}$ (equiv)

classic estimate

12

Model rank

Model:

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

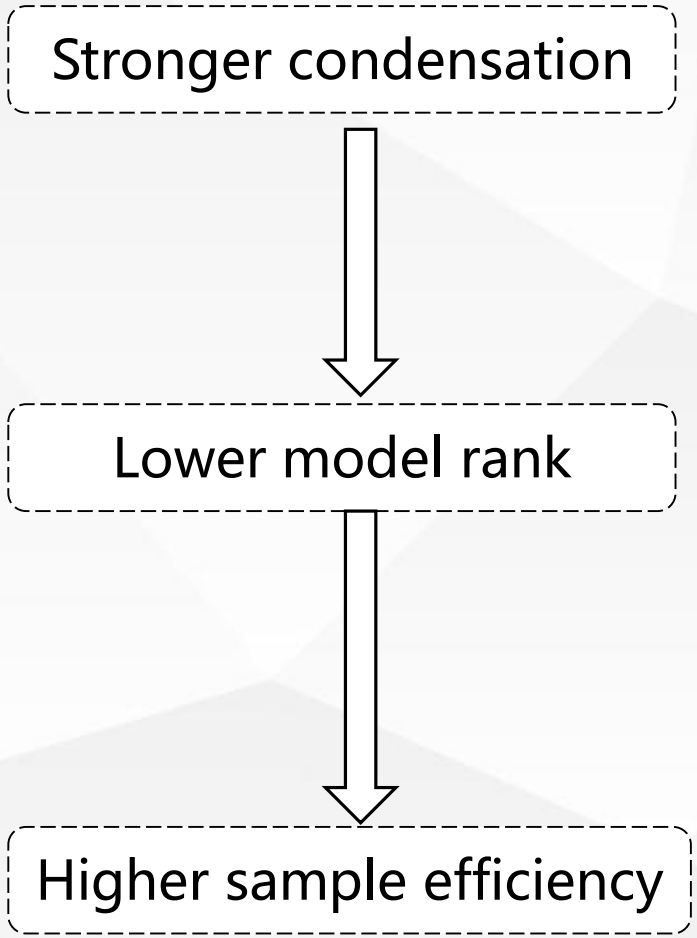
Model rank:

$$r_{\boldsymbol{\theta}} := \operatorname{rank} DF(\boldsymbol{\theta}) = \dim \operatorname{Im}\big(DF(\boldsymbol{\theta})\big)$$

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

Intuition: effective degrees of freedom at $\boldsymbol{\theta}$

$$F(\boldsymbol{\theta} + \boldsymbol{\delta})(\cdot) \approx F(\boldsymbol{\theta})(\cdot) + \sum_{i=1}^{M} \partial_{\theta_i} F(\boldsymbol{\theta})(\cdot)\, \delta_i$$



Yaoyu Zhang*, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu*, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).

Condensation means lower model rank

Example:

$$F(\boldsymbol{\theta})(x) = a_1 \tanh(w_1 x) + a_2 \tanh(w_2 x)$$

Model rank:

$$\dim \operatorname{span}\{\tanh(w_1 x), a_1 \tanh'(w_1 x)x, \tanh(w_2 x), a_2 \tanh'(w_2 x)x\}$$

- **Condensed($w_1 = \pm w_2$):**

$$r_{\boldsymbol{\theta}} \leq 2$$

- **Not condensed($w_1 \neq \pm w_2 \neq 0, a_1 \neq 0, a_2 \neq 0$):**

$$r_{\boldsymbol{\theta}} = 4$$

Yaoyu Zhang*, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu*, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).

Basic setup

Data:

$$S = \{(\boldsymbol{x}_i \in \mathbb{R}^d, f^*(\boldsymbol{x}_i) \in \mathbb{R})\}_{i=1}^n$$

Model:

$$F: \mathbb{R}^M \to \mathcal{F}$$

Ex: two-layer NN $F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{j=1}^m a_j \tanh(\boldsymbol{w}_j^T \boldsymbol{x})$, $\boldsymbol{\theta} = (a_j, \boldsymbol{w}_j)_{j=1}^m$

Optimization:

$$R_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \left(F(\boldsymbol{\theta})(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right)^2$$

$$\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta})$$

Problem: how many samples are required for fitting f^*?

Optimistic sample size estimate

Model:

$$F: \mathbb{R}^M \to \boldsymbol{\mathcal{F}}$$

Model rank:

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

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

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

$F^{-1}(f^*)$: 目标集(零泛化误差)

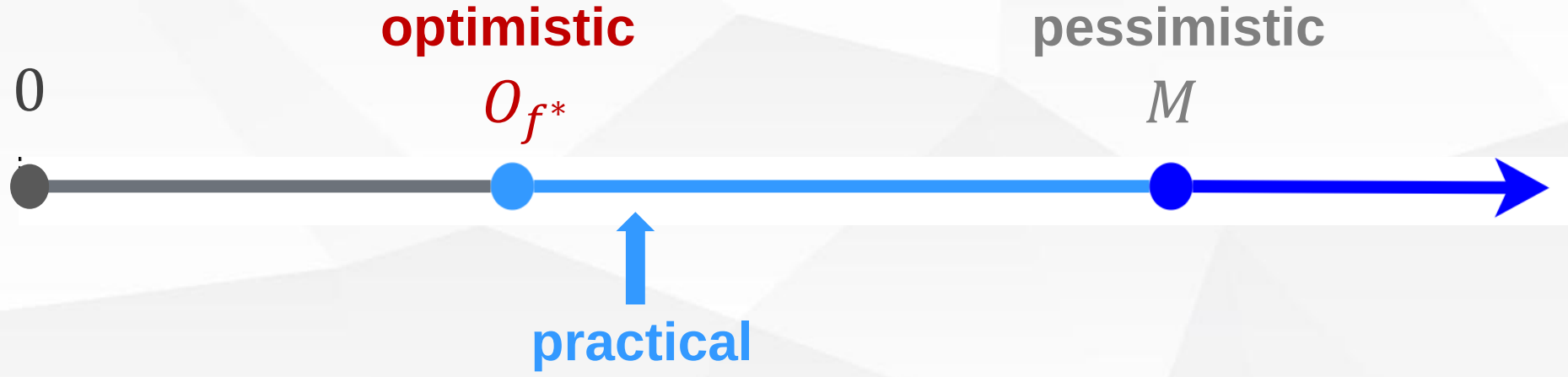
Intuitive procedure:

Given target f^* ⇒ find $\theta^* \in F^{-1}(f^*)$ with minimum rank ⇒ $O_{f^*} = r_{\theta^*}$

Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Properties of optimistic estimate

- **Optimistic:** lower bound sample size for fitting;
- **Target-adaptive:** less samples for simpler target;
- **Tune-for-realization:** realization requires strong condensation.



Width vs. sample efficiency

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*

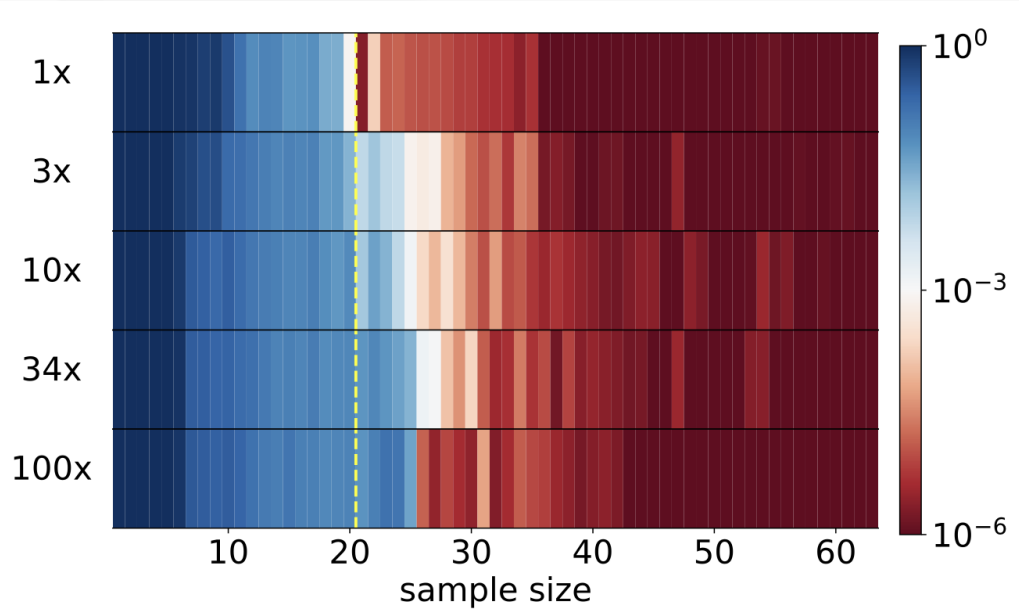
$$O_{f_{\boldsymbol{\theta}}}(f^*) = k(d+1).$$

→ **No m!**

f^*	generic bound (Cor. 2)	$O_{f_{\boldsymbol{\theta}}}(f^*)$ (Thm. 5)
$\{0(\cdot)\}$		0
$\mathcal{F}_1^{\mathrm{NN}} \backslash \{0(\cdot)\}$		$d+1$
$\vdots$		$\vdots$
$\mathcal{F}_k^{\mathrm{NN}} \backslash \mathcal{F}_{k-1}^{\mathrm{NN}}$	$m(d+1)$	$k(d+1)$
$\vdots$		$\vdots$
$\mathcal{F}_m^{\mathrm{NN}} \backslash \mathcal{F}_{m-1}^{\mathrm{NN}}$		$m(d+1)$

Pessimistic: efficiency worsens

Optimistic: efficiency preserves



Experiment: increasing width approximately preserves sample efficiency

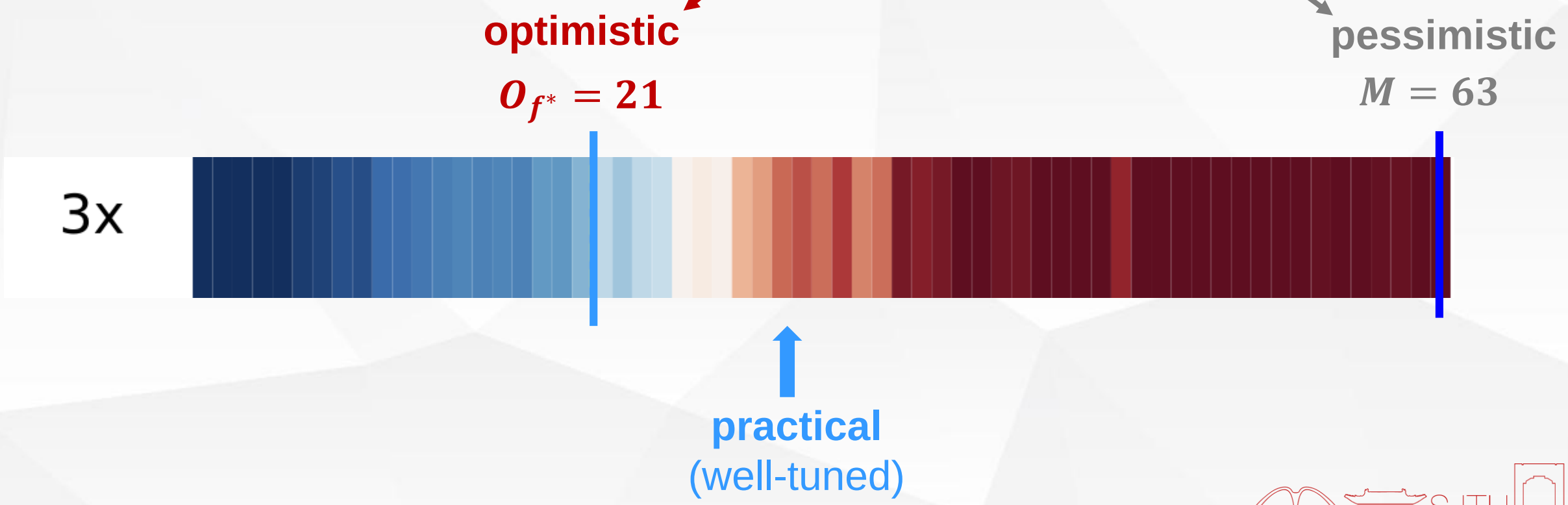
Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Optimistic sample size reflects practice

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*

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

vs. $m(d+1)$



Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Optimistic sample size reflects practice

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*

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

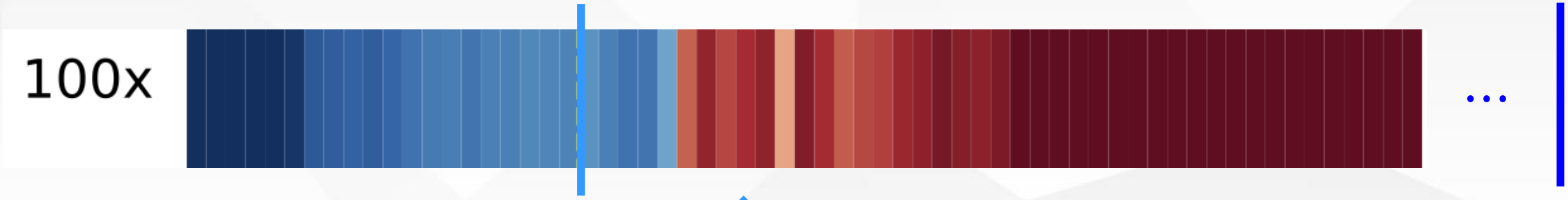
vs. $m(d+1)$

optimistic

$O_{f^*} = 21$

pessimistic

$M = 2100$



practical
(well-tuned)

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* $O_{f_{\boldsymbol{\theta}_{\text{wide}}}}(f^*) \leqslant O_{f_{\boldsymbol{\theta}_{\text{narr}}}}(f^*) \leqslant M_{\text{narr}}$.

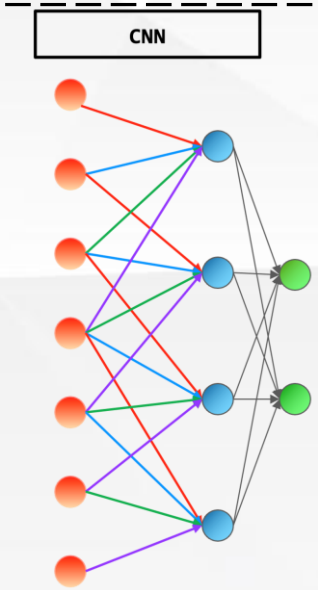
wider network is sample efficient

Corollary 6 (free expressiveness in width). *The optimistic sample size of a target function expressible by any DNN never increases as the DNN gets wider.*

increasing width preserves sample efficiency

Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

CNN vs. FNN



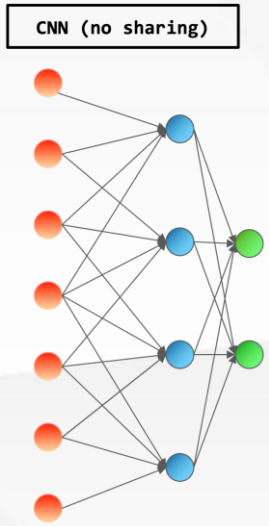
Theorem 6 (optimistic sample sizes for two-layer tanh-CNN). *Given a m-kernel two-layer CNN with weight sharing with 2-d input $I \in \mathbb{R}^{d\times d}$, $s \times s$ kernel and stride 1*

$$f_{\boldsymbol{\theta}}(\boldsymbol{I}) = \sum_{l=1}^{m} \sum_{i,j=1}^{d+1-s} a_{lij} \tanh\left(\sum_{\alpha,\beta} I_{i+s-\alpha,j+s-\beta} K_{l;\alpha,\beta} \right), \quad \boldsymbol{I} \in \mathbb{R}^{d\times d},$$

for any target function $f^* \in \mathcal{F}_k^{\mathrm{CNN}} \backslash \mathcal{F}_{k-1}^{\mathrm{CNN}}$ with $0 \leqslant k \leqslant m$, the optimistic sample size

$$O_{f_{\boldsymbol{\theta}}}(f^*) = k(s^2 + (d+1-s)^2).$$

Here $\mathcal{F}_k^{\mathrm{CNN}}$ indicates the function space of k-kernel CNN for $k \in \mathbb{N}^+$, $\mathcal{F}_0^{\mathrm{CNN}} := \{0(\cdot)\}$ and $\mathcal{F}_{-1}^{\mathrm{CNN}} := \emptyset$.



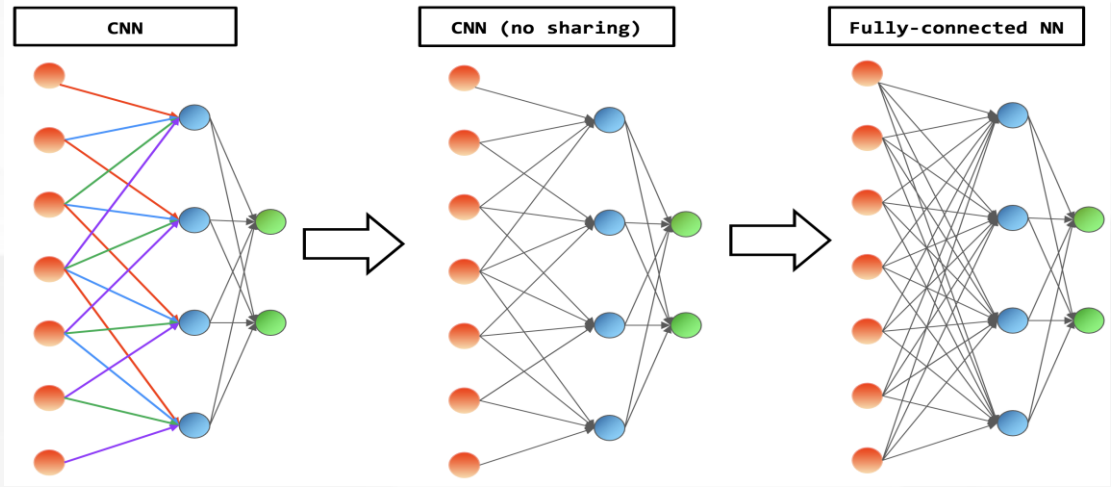
Theorem 7 (optimistic sample sizes of CNN functions in two-layer CNN without weight sharing). *We consider a m-kernel two-layer CNN without weight sharing with 2-d input $I \in \mathbb{R}^{d\times d}$, $s \times s$ kernel and stride 1*

$$f_{\boldsymbol{\theta}}(\boldsymbol{I}) = \sum_{l=1}^{m} \sum_{i,j=1}^{d+1-s} a_{lij} \tanh\left(\sum_{\alpha,\beta} I_{i+s-\alpha,j+s-\beta} K_{lij;\alpha,\beta} \right), \quad \boldsymbol{I} \in \mathbb{R}^{d\times d}.$$

For any target function expressible by a CNN (with sharing) $f^* \in \mathcal{F}_k^{\mathrm{CNN}} \backslash \mathcal{F}_{k-1}^{\mathrm{CNN}} \subset \mathcal{F}_k^{\mathrm{CNN-NS}}$ with $0 \leqslant k \leqslant m$, then the optimistic sample size

$$O_{f_{\boldsymbol{\theta}}}(f^*) = (s^2+1)(k(d+1-s)^2 - m_{\mathrm{null}}),$$

CNN vs. FNN



f^*	CNN	CNN (no sharing)	Fully-connected NN
$\{0\}$	0	0	0
$\mathcal{F}_1^{\mathrm{CNN}} \backslash \{0\}$	$s^2+(d+1-s)^2$	$(s^2+1)\left((d+1-s)^2-m_{\mathrm{n}}\right)$	$(d^2+1)\left((d+1-s)^2-m_{\mathrm{n}}\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{F}_k^{\mathrm{CNN}} \backslash \mathcal{F}_{k-1}^{\mathrm{CNN}}$	$k(s^2+(d+1-s)^2)$	$(s^2+1)\left(k(d+1-s)^2-m_{\mathrm{n}}\right)$	$(d^2+1)\left(k(d+1-s)^2-m_{\mathrm{n}}\right)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{F}_m^{\mathrm{CNN}} \backslash \mathcal{F}_{m-1}^{\mathrm{CNN}}$	$m(s^2+(d+1-s)^2)$	$(s^2+1)\left(m(d+1-s)^2-m_{\mathrm{n}}\right)$	$(d^2+1)\left(m(d+1-s)^2-m_{\mathrm{n}}\right)$

k intrinsic width

$s \times s$ conv ker size

$d \times d$ input dim

more connection lower sample efficiency

Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

CNN vs. FNN

Adding (unnecessary) connections reduces sample efficiency

MNIST width-k, kernel size: 3x3

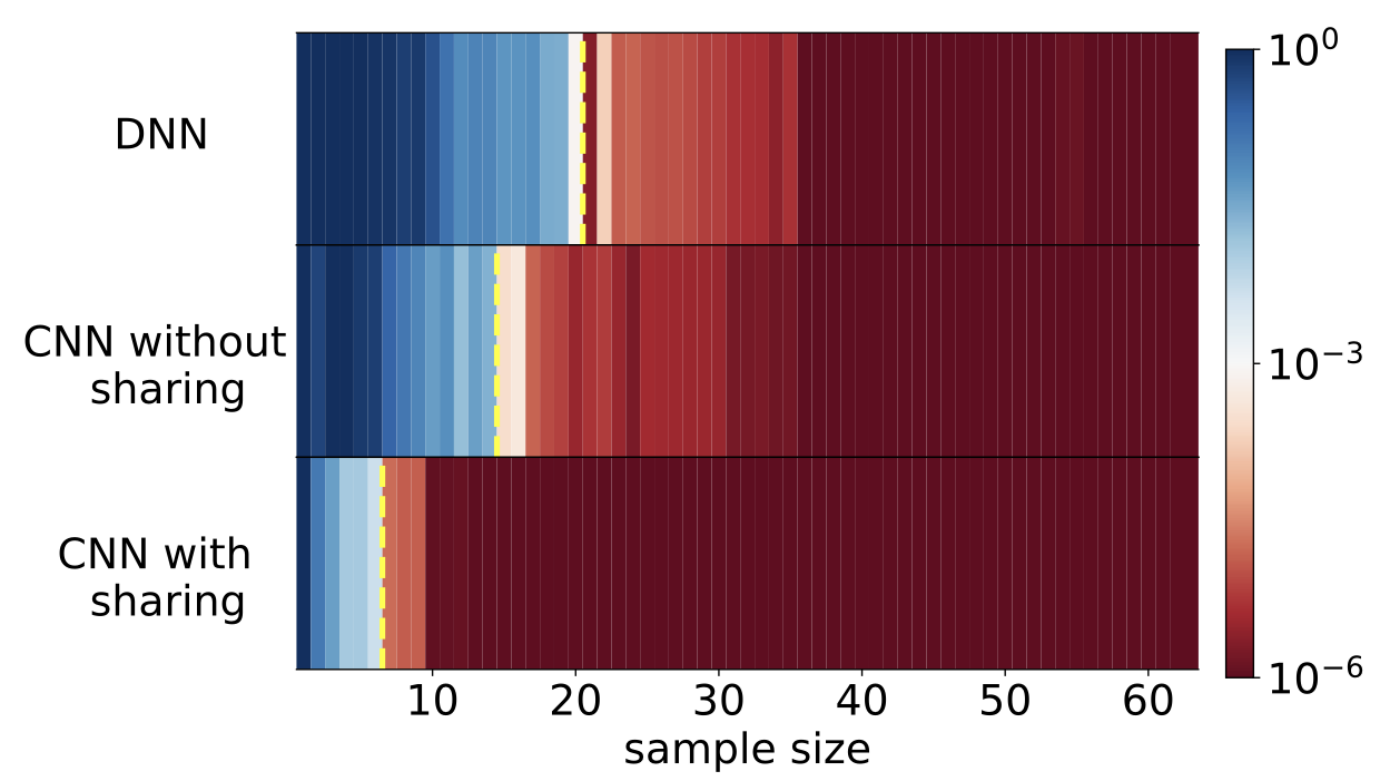
- CNN: $685k$
- CNN (no sharing) : $6760k$
- FNN: $530660k$

1000X worse!

Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

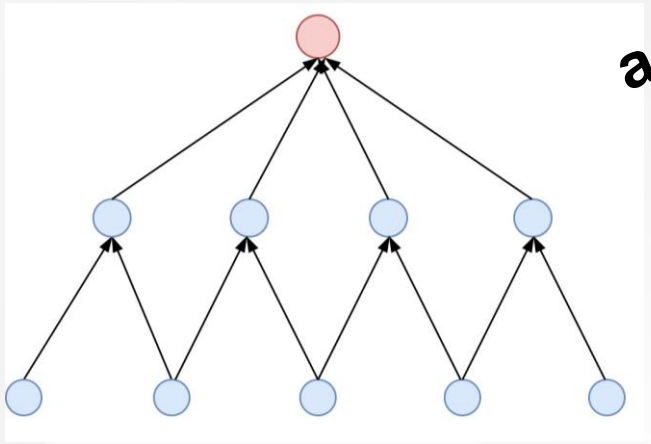
Optimistic Estimate of CNN

Experiment：adding connection reduces sample efficiency



Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

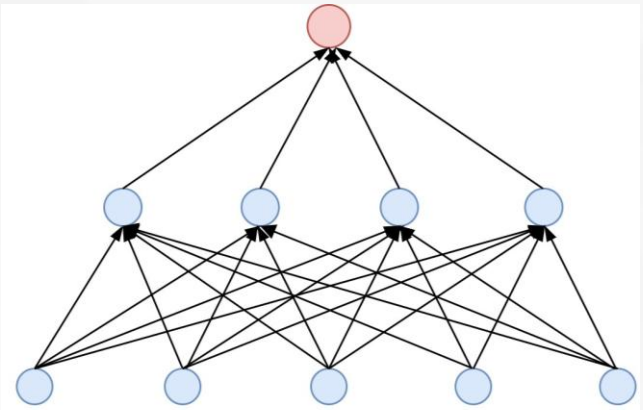
Specialty of DNN models via optimistic estimate



12

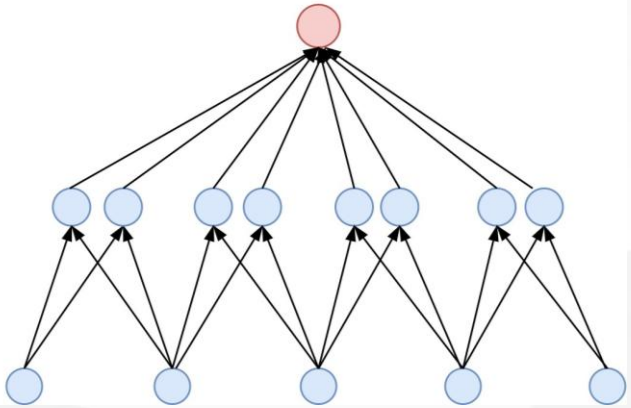
add connection

inefficient



24

add width



12

efficient

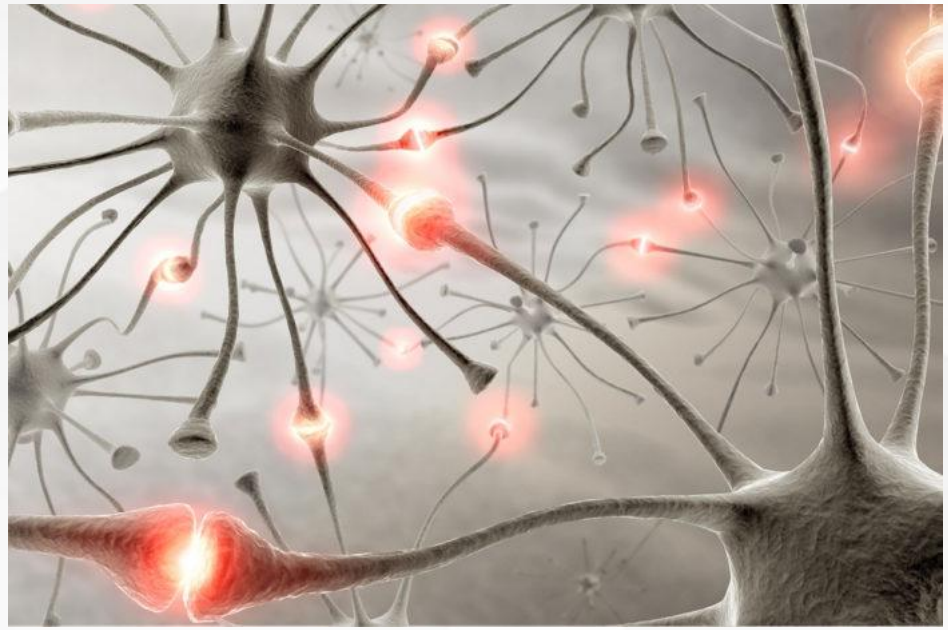
Sample inefficient:
Adding (unnecessary) connections worsens generalization

Sample efficient:
Increasing width doesn't harm generalization

Inspiration

Principle of model scaling

- Freely increase width
- Refrain from adding connection



vs. Scaling of brain

- mouse: $\sim 10^8$ neurons, $\sim 10^3 - 10^4$ connections/neuron
- human: $\sim 10^{11}$ neurons, $\sim 10^3 - 10^4$ connections/neuron

Permutation symmetry → sample efficiency preserving

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

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

Theorem(informal):

permutation-invariant manifolds are invariant manifolds of gradient flow.

e.g., $\left(\boldsymbol{W}_{,j}^{[l-1]}, \boldsymbol{W}_{j}^{[l-2]}, b_{j}^{[l-2]}\right) = \left(\boldsymbol{W}_{,j'}^{[l-1]}, \boldsymbol{W}_{j'}^{[l-2]}, b_{j'}^{[l-2]}\right)$

Permutation symmetry → invariant manifolds (equiv to smaller network)

→ optimistically as efficient as smaller networks

Permutation symmetry in Transformer

permutation symmetry -> optimistic sample efficiency preserving

Permutation symmetric:

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

→ **Scale up freely!**

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

KAN (Kolmogorov–Arnold Networks)

KAN:

$$f_\theta(\boldsymbol{x}) = f_\theta(x_1, \cdots, x_d) = \sum_{i=1}^{m} \Phi_i \left(\sum_{j=1}^{d} \phi_{i,j}(x_j) \right)$$

width： permutation symmetric

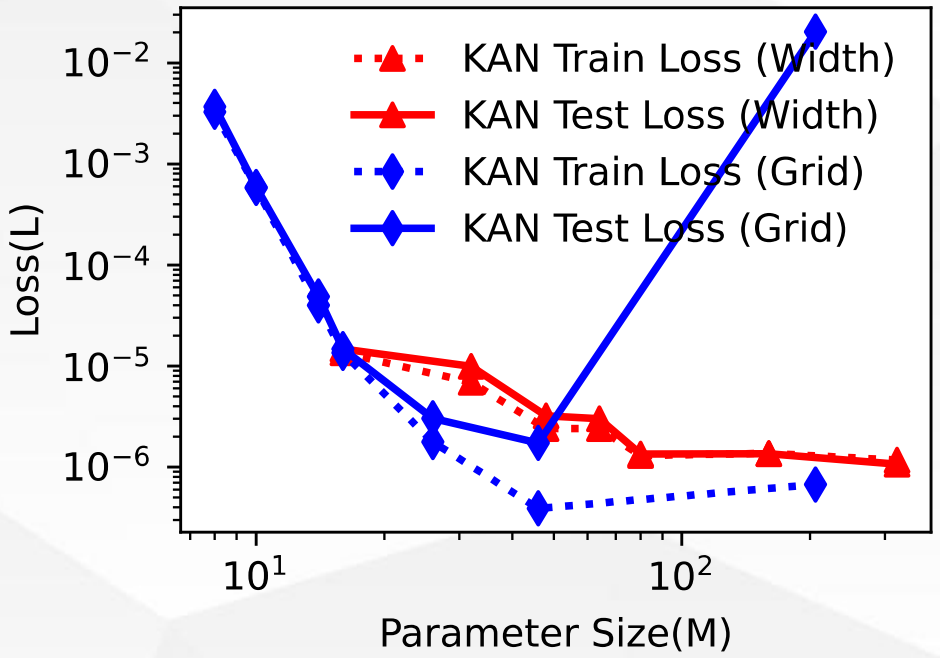
Increase m preserves sample efficiency

$$\Phi_i(x) \text{ or } \phi_{i,j}(x) = \sum_{i=1}^{G} c_i B_i(x), \theta = \{c_i\} \text{ is learnable}$$

grid: no symmetry

Increase G reduces sample efficiency

Experiment



Empirical estimation and application

Empirical estimation of optimistic sample size

Empirical estimation: **under best tuning of hyperparameters**, the minimal sample size for (close to) 100% test accuracy.

Single anchor

$1 : +5$

$2 : +1$

$3 : -2$

$4 : -8$

Two anchors

$1\ 1 : +10$

$1\ 2 : +6$

$\cdots$

$3\ 4 : -10$

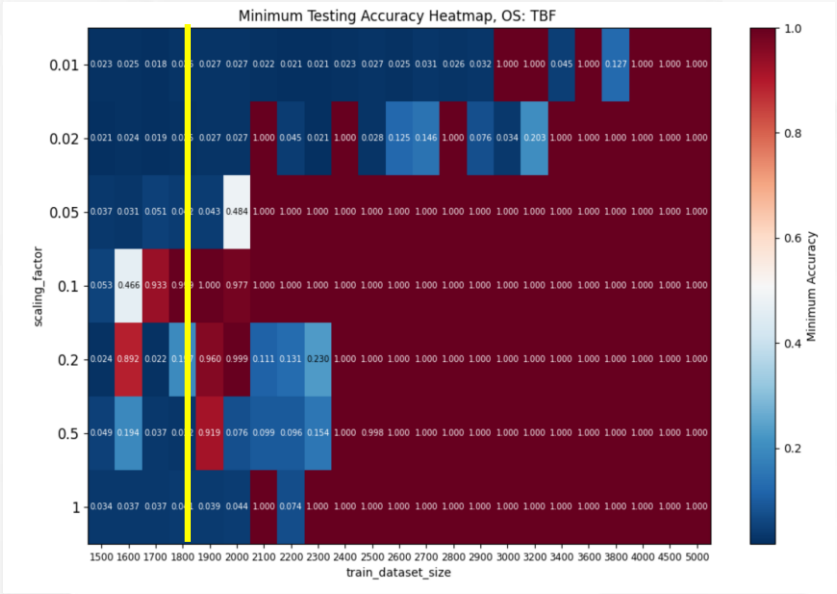
$4\ 4 : -16$

- Composite Anchor function
- Symmetry Anchor function
- Random Anchor function

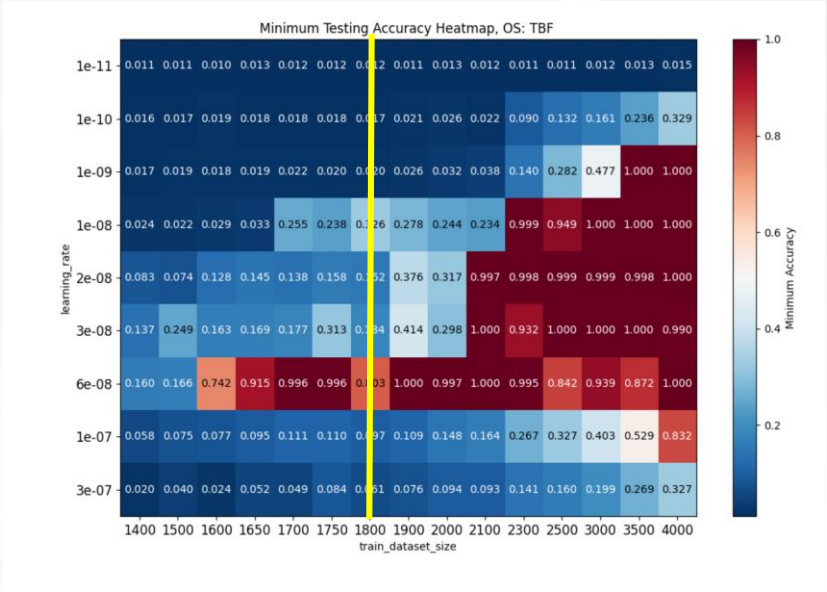
Input data examples	Target
Noisy tokens	
55 46 32 52 **28 1 1** 34 33	**38**
20 95 **43 3 1** 44 34 76 32	**46**
. . .	. . .
28 53 44 78 32 **62 3 4** 44	**52**
77 43 23 63 89 33 **52 4 3**	**?**

Empirical estimation of optimistic sample size

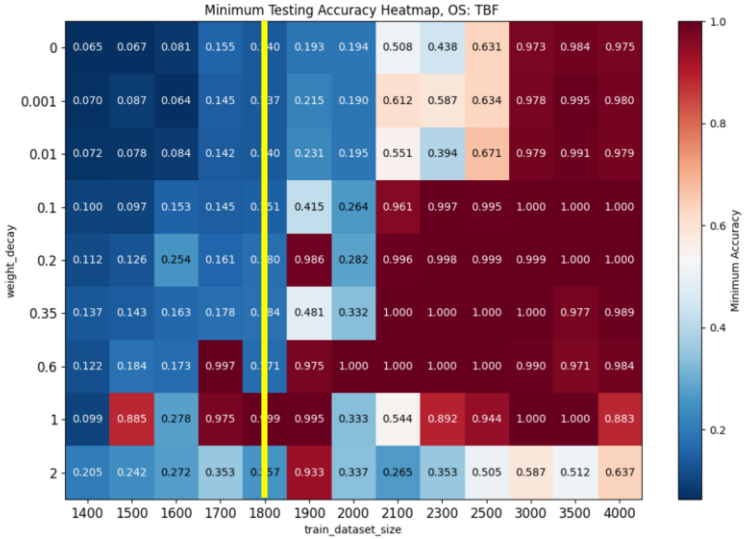
Identification of best tuning of hyperparameters



Initialization scale



Learning rate



Weight decay

Empicial estimate of optimistic sample size≈1800

Application——architecture analysis

- 2V2 Random

Anchor	1	2
1	+1	+5
2	-2	+7

- 4V4 Composite

Anchor	1	2	3	4
1	+2	+3	+4	+5
2	+3	+4	+5	+6
3	+4	+5	+6	+7
4	+5	+6	+7	+8

Memorization complexity	Inference complexity
4	4
16	4 1: +1 2: +2 3: +3 4: +4

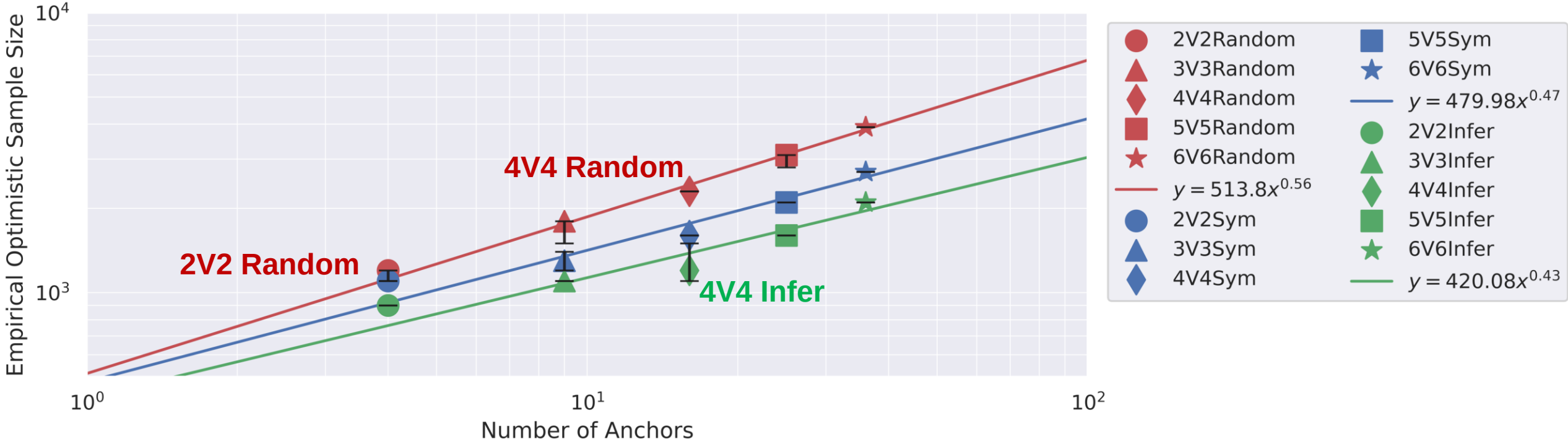
Can Transformer make inference?

Experiment:
Optimistic sample size scales with Memorization or Inference complexity?

(Optimistically) Memorization? No!

Transformer:

Memorization complexity poorly predicts optimistic sample size

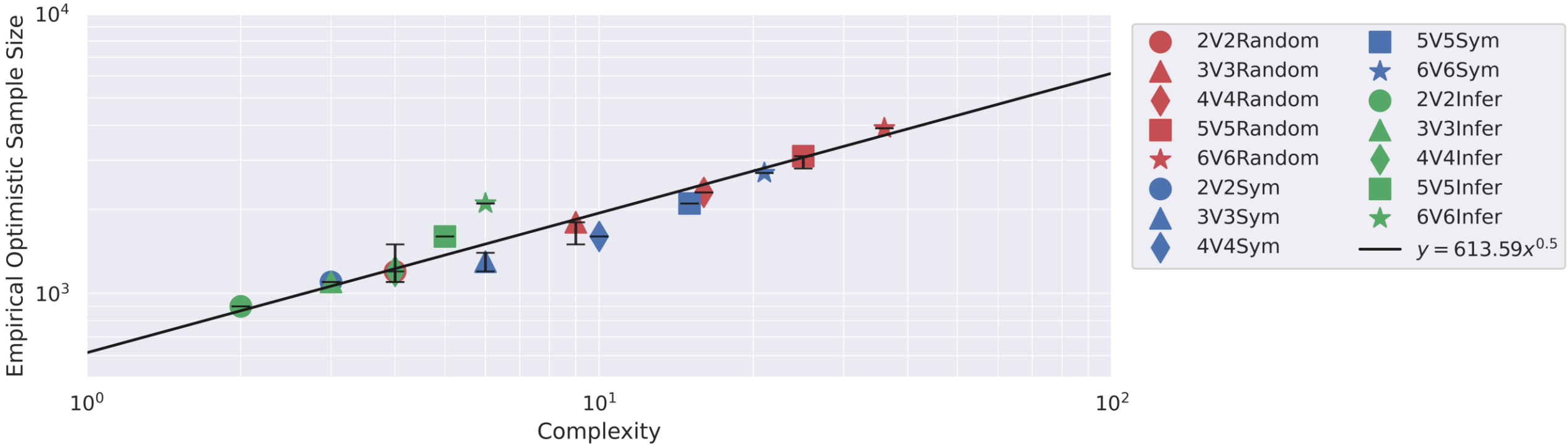


optimistic sample size vs. memorization complexity

(Optimistically) Inference? Yes!

Transformer:

Inference complexity well predicts optimistic sample size



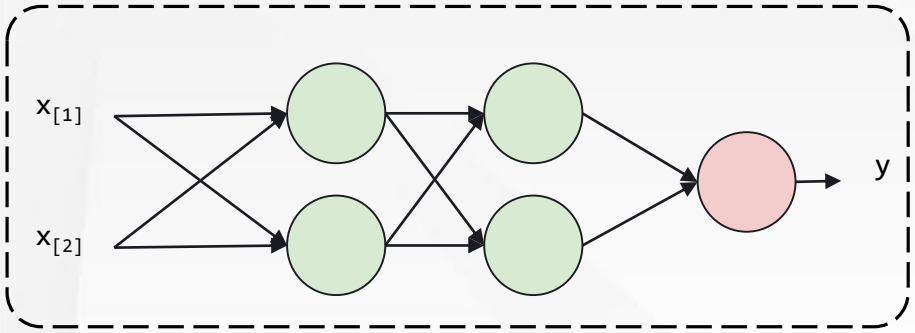
optimistic sample size vs. inference complexity

Theory of optimistic estimate

Optimistic sample size estimation for general DNNs

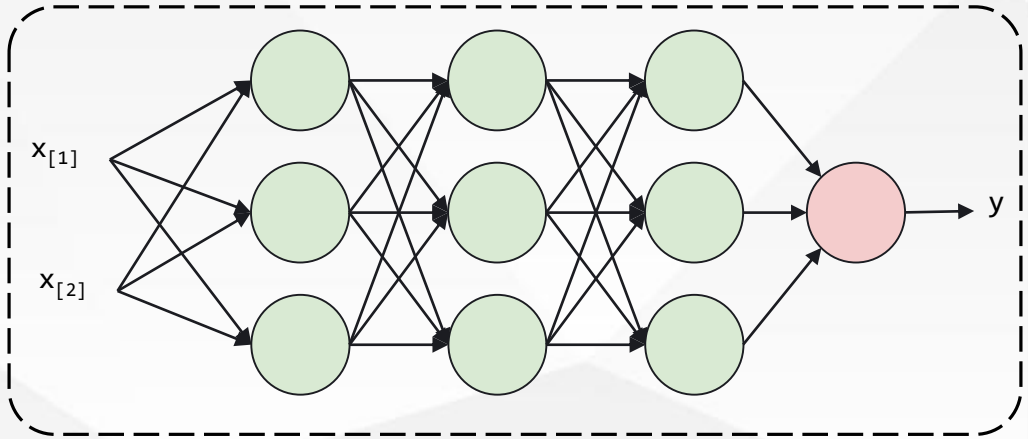
Key tool: critical Mapping

Lemma 12 (upper bound of optimistic sample size). *Given two models* $f_{\boldsymbol{\theta}_A} = f(\cdot;\boldsymbol{\theta}_A)$ *with* $\boldsymbol{\theta}_A \in \mathbb{R}^{M_A}$ *and* $g_{\boldsymbol{\theta}_B} = g(\cdot;\boldsymbol{\theta}_B)$ *with* $\boldsymbol{\theta}_B \in \mathbb{R}^{M_B}$, *if there exists a critical mapping* $\mathcal{P}$ *from model* A *to* B, *then the optimistic sample size* $O_g(f^*) \leqslant O_f(f^*) \leqslant M_A$ *for any* $f^* \in \mathcal{F}_A$.



critical Mapping

$\boldsymbol{P}: \mathbb{R}^{\boldsymbol{M_A}} \rightarrow \mathbb{R}^{\boldsymbol{M_B}}$



critical Mapping:

(i) Output Preserving: $f_\theta = g_{P(\theta)}$

(ii) Criticality Preserving: if $\nabla_\theta R_S(f_\theta) = 0$, then $\nabla_\theta R_S\left(g_{P(\theta)}\right) = 0$, for any data S

Embedding principle: critical mapping exists

Theorem 10 (Embedding Principle) *Given any NN and any K-neuron wider NN, there exists a K-step composition embedding $\mathcal{T}$ satisfying that: For any given data S, loss function $\ell(\cdot,\cdot)$, activation function $\sigma(\cdot)$, given any critical point $\boldsymbol{\theta}^{c}_{\text{narr}}$ of the narrower NN, $\boldsymbol{\theta}^{c}_{\text{wide}} := \mathcal{T}(\boldsymbol{\theta}^{c}_{\text{narr}})$ is still a critical point of the K-neuron wider NN with the same output function, i.e., $\boldsymbol{f}_{\boldsymbol{\theta}^{c}_{\text{narr}}} = \boldsymbol{f}_{\boldsymbol{\theta}^{c}_{\text{wide}}}$.*

Theorem 4.1 (embedding principle in depth). *(see Appendix A: Thm. A.1 for proof) Given data S and an $\mathrm{NN}'\left(\{m'_l\}_{l=0}^{L'}\right)$, for any parameter $\boldsymbol{\theta}_c$ of any shallower $\mathrm{NN}\left(\{m_l\}_{l=0}^{L}\right)$ satisfying $\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}_c) = \mathbf{0}$, there exists parameter $\boldsymbol{\theta}'_c$ in the loss landscape of $\mathrm{NN}'\left(\{m'_l\}_{l=0}^{L'}\right)$ satisfying the following conditions:*

(i) $\boldsymbol{f}_{\boldsymbol{\theta}'_c}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{\theta}_c}(\boldsymbol{x})$ *for* $\boldsymbol{x} \in S_{\boldsymbol{x}}$;

(ii) $\nabla_{\boldsymbol{\theta}'} R_S(\boldsymbol{\theta}'_c) = \mathbf{0}$.

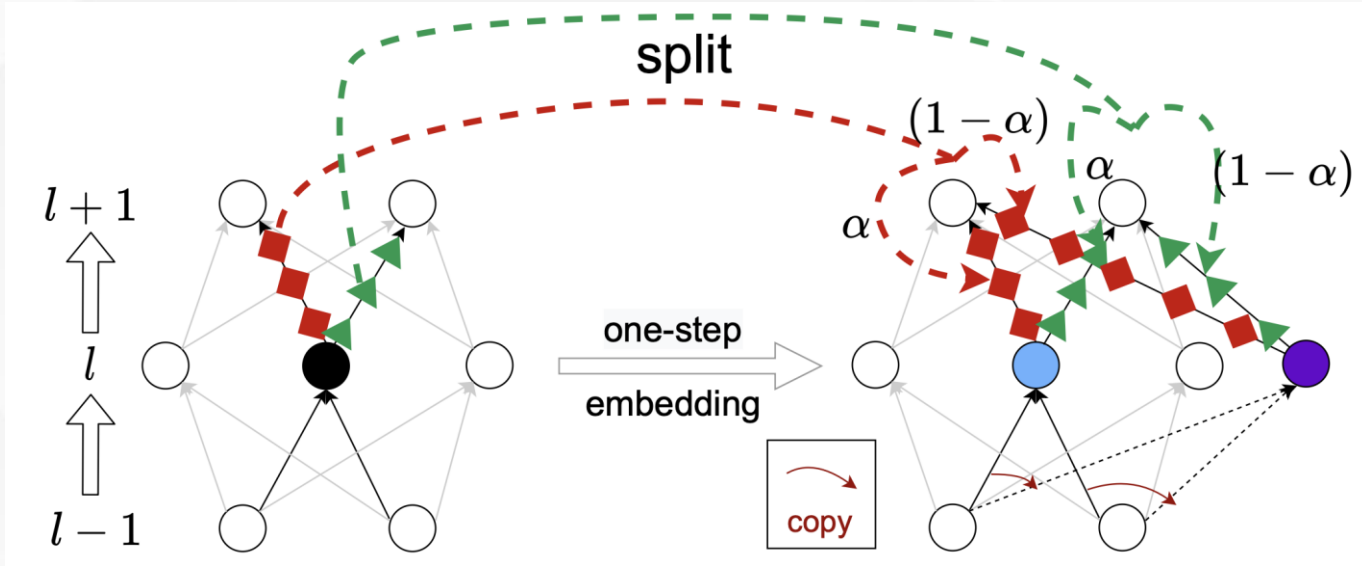
Embedding Principle (width/depth):

The loss landscape of a neural network “contains” all the critical points of narrower/shallower networks.

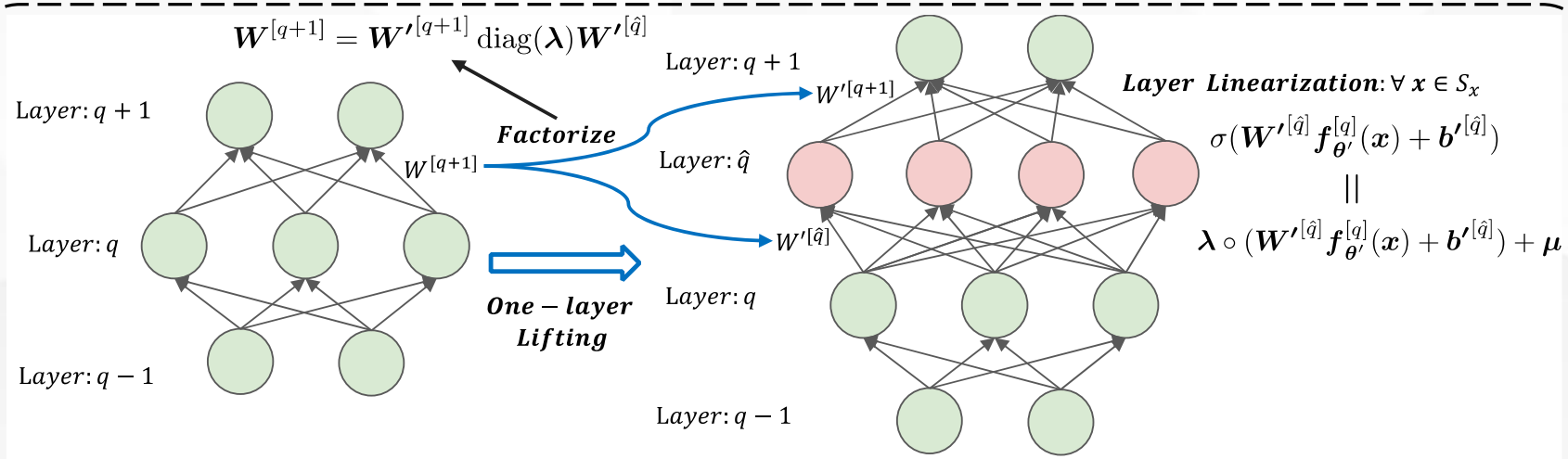
[1] Zhang, Zhang, Luo, Xu, Embedding Principle of Loss Landscape of Deep Neural Networks. NeurIPS 2021 Spotlight
[2] Zhang, Li, Zhang, Luo, Xu, Embedding Principle: a hierarchical structure of loss landscape of deep neural networks. JML 2022
[3] Bai, Luo, Xu, Zhang, Embedding Principle in Depth for the Loss Land- scape Analysis of Deep Neural Networks. CSIAM 2024.

Embedding principle -> sample efficiency preserving

Embedding Principle (width)



Embedding Principle (depth)



[1] Zhang, Zhang, Luo, Xu, Embedding Principle of Loss Landscape of Deep Neural Networks. NeurIPS 2021 Spotlight
[2] Zhang, Li, Zhang, Luo, Xu, Embedding Principle: a hierarchical structure of loss landscape of deep neural networks. JML 2022
[3] Bai, Luo, Xu, Zhang, Embedding Principle in Depth for the Loss Land- scape Analysis of Deep Neural Networks. CSIAM 2024.

Criticality of optimistic sample size—local theory

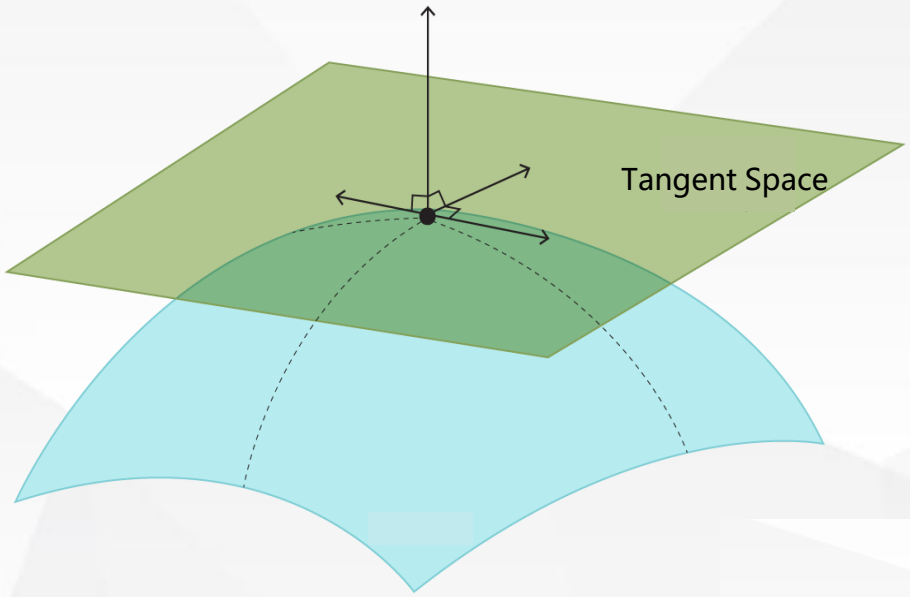
Theorem (phase transition of LLR-guarantee at a target point). For any $\theta' \in \mathcal{M}_{f^*}$

- If training data size $n < O_{f_\theta}(\theta')$, f^* has no local linear recovery guarantee at θ'
- If $n \geq O_{f_\theta}(\theta')$, f^* has n-sample LLR-guarantee, i.e., there exits an n-sample dataset $S' = \{(x_i, f^*(x_i))\}_{i=1}^n$ such that f^* has local linear recovery guarantee at θ'.

O_{f^*} is critical for the local linear recovery of f^*

$$f^* = \operatorname{argmin}_{g \in \widetilde{\mathcal{T}}_{\theta'}} \frac{1}{n} \sum_{i=1}^{n} \ell(g(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)),$$

$$\widetilde{\mathcal{T}}_{\theta'} = \{f(\cdot;\boldsymbol{\theta}') + \boldsymbol{a}^T \nabla_{\boldsymbol{\theta}} f(\cdot;\boldsymbol{\theta}') | \boldsymbol{a} \in \mathbb{R}^M\}$$



Criticality of optimistic sample Size—local theory

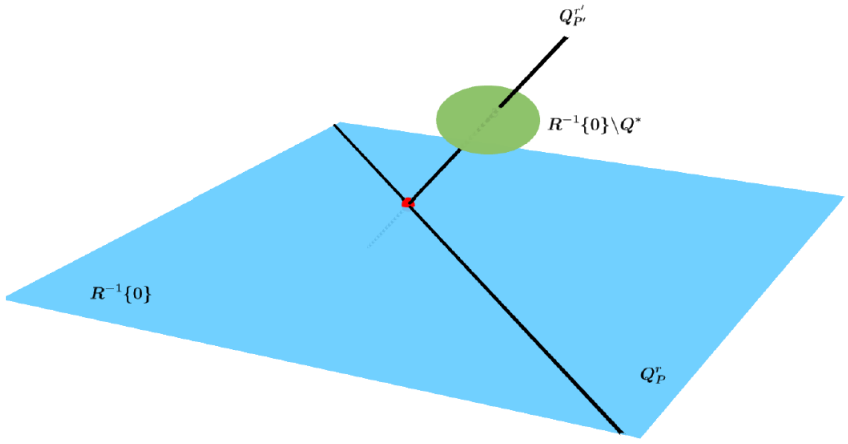
Theorem 2.2 (gradient flow near global minima). *Following the hypotheses and notations in Theorems 2.1, any gradient flow sufficiently close to $R^{-1}\{0\}$ converges. On the other hand, any point in $R^{-1}\{0\}$ is the limit of some gradient flow. The following results hold.*

(a) Whenever $t \leq N'$ and sample size $n \geq N_t$, a generic gradient flow sufficiently close to Q_t converges to a point $\theta^* \in Q_t$. The convergence does not have linear rate and the curve is "biased towards" $\ker \mathrm{Hess} R(\theta^*)$. Moreover, any small perturbation of it still converges to Q_t.

(b) Given $t > N'$. When sample size $n \geq N_t$, any gradient flow sufficiently close to Q_t converges to points Q_t at linear rate. Similar to (a), any small perturbation of it still converges to Q_t.

No local recovery with $\leq O_{f^*}$ samples

Sample size	Separation of Q^*
$\leq (d+1)m_0$	Nowhere separated
$\geq m + m_0 d$	$Q^m_{P,\pi}$ separated[1]
$\vdots$	$\vdots$
$\geq r + (m + m_0 - r)d$	$Q^r_{P,\pi}$ separated[1]
$\vdots$	$\vdots$
$\geq m_0 + md$	$Q^{m_0}_{P,\pi}$ separated[1]
$\geq (d+1)m$	Q^* is separated[2]



Tao Luo, Leyang Zhang, **Yaoyu Zhang**, Geometry and Local Recovery of Global Minima of Two-layer Neural Networks at Overparameterization, arXiv:2309.00508, (2024).

Series works

Optimistic Estimate

- Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).
- Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Optimistic Estimate Uncovers the Potential of Nonlinear Models. arXiv:2307.08921, (2023).
- Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).
- Tao Luo, Leyang Zhang, Yaoyu Zhang, Geometry and Local Recovery of Global Minima of Two-layer Neural Networks at Overparameterization, arXiv:2309.00508, (2024).

Embedding Principle

- Yaoyu Zhang, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, Embedding Principle of Loss Landscape of Deep Neural Networks. NeurIPS 2021 spotlight.
- Yaoyu Zhang, Yuqing Li, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, Embedding Principle: a hierarchical structure of loss landscape of deep neural networks. Journal of Machine Learning, 1(1), pp. 60-113, 2022.
- Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Yaoyu Zhang, Embedding Principle in Depth for the Loss Landscape Analysis of Deep Neural Networks, CSIAM Trans. Appl. Math., 5 (2024), pp. 350-389.

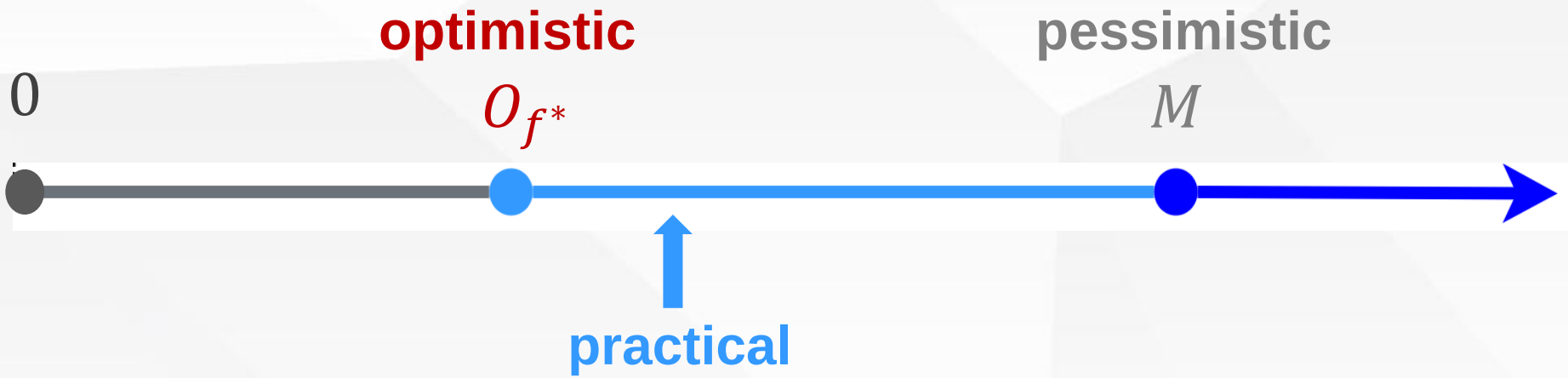
Condensation

- 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 (JMLR) 22(71):1−47, (2021)
- Zhi-Qin John Xu*, Hanxu Zhou, Tao Luo, Yaoyu Zhang, Towards Understanding the Condensation of Two-layer Neural Networks at Initial Training. NeurIPS 2022

Take Home Messages

homepage

Picture of sample size requirement for nonlinear models



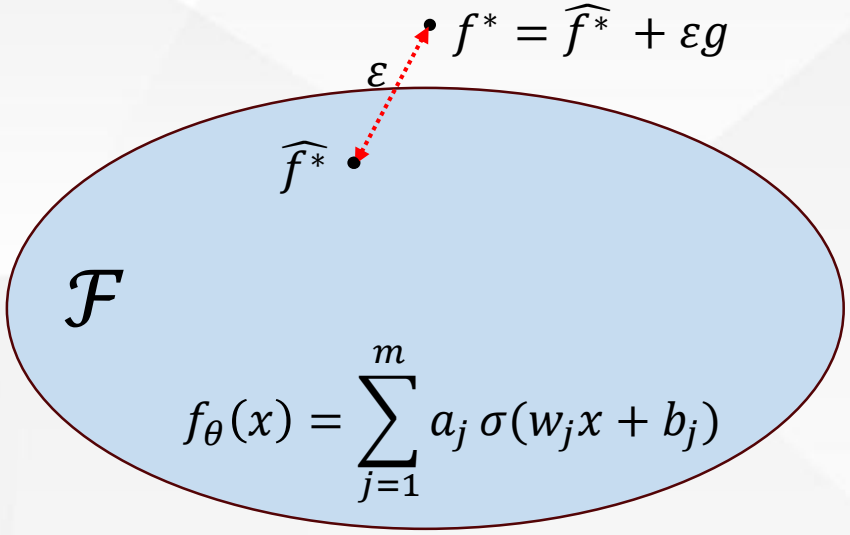
Ways to improve sample efficiency

- **Improve the potential:** optimize the architecture
- **Unleash the potential:** optimize the hyperparameters (e.g., initialization scale, weight decay rate, learning rate)

Thanks!

饮 水 思 源　爱 国 荣 校

乐观估计的推广——ε 乐观样本量

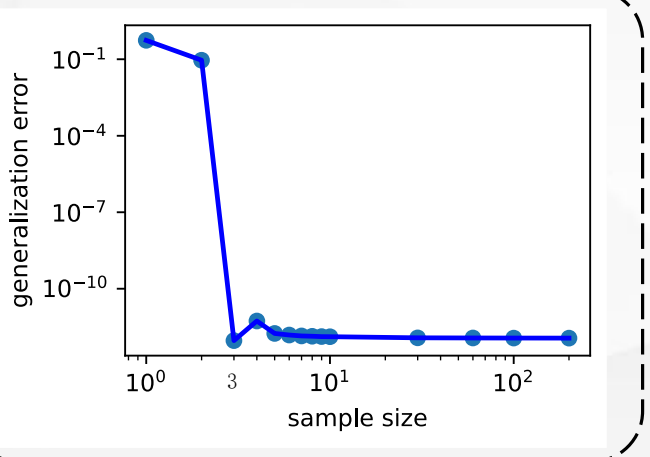


若 $f^* \notin \mathcal{F}$, 但是其在假设空间某个函数的小邻域内:

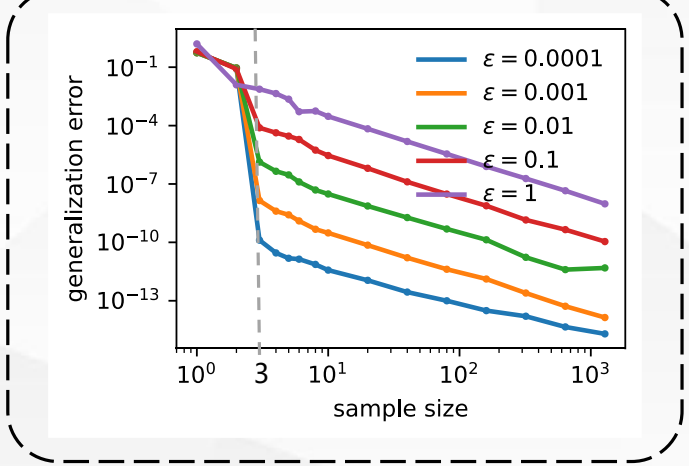
$$f^* = \widehat{f^*} + \varepsilon g$$

$\widehat{f^*}$: 乐观估计 $\Rightarrow$ 泛化误差的相变行为

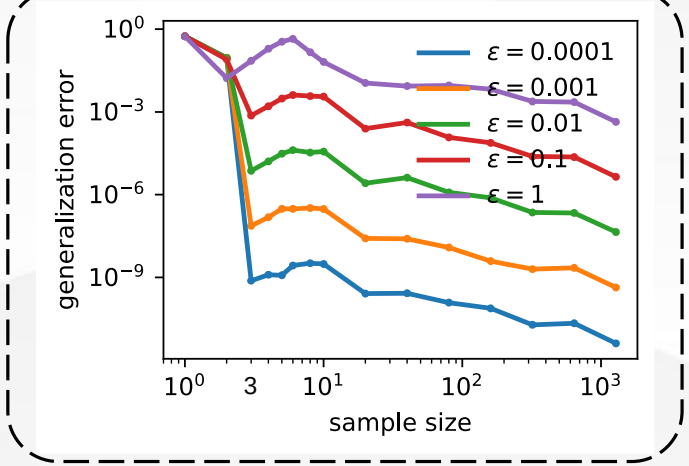
f^*: 泛化误差的相变行为



$\varepsilon = 0, f^* = \sigma(x+1) \in \mathcal{F}$



$f^* = \sigma(x+1) + \varepsilon \sin(x)$



带 ε 噪音采样

Revisit Leo Breiman's problems (1995)

- 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?

condensation、optimistic estimate、embedding principle、frequency principle

Time to make AI a science!

《深度学习导论与理解》

Take Home Messages

Why don't overparameterized NNs reduce sample efficiency?

Condensation!

What is the sample size required for fitting?

$$O_{f^*} = \min_{\boldsymbol{\theta} \in F^{-1}(f^*)} \dim \operatorname{span} \left\{ \partial_{\theta_i} F(\boldsymbol{\theta})(\cdot) \right\}_{i=1}^{M} \text{ (Optimistically)}$$

How to increase sample efficiency?

- **Improve the potential:** Optimize the architecture
- **Unleash the potential:** Optimize the hyperparameters (e.g., initialization scale, weight decay rate, learning rate)

Take Home Messages

- **Picture of sample size requirement for nonlinear models**