V.From condensation to generalization theory

Yaoyu Zhang

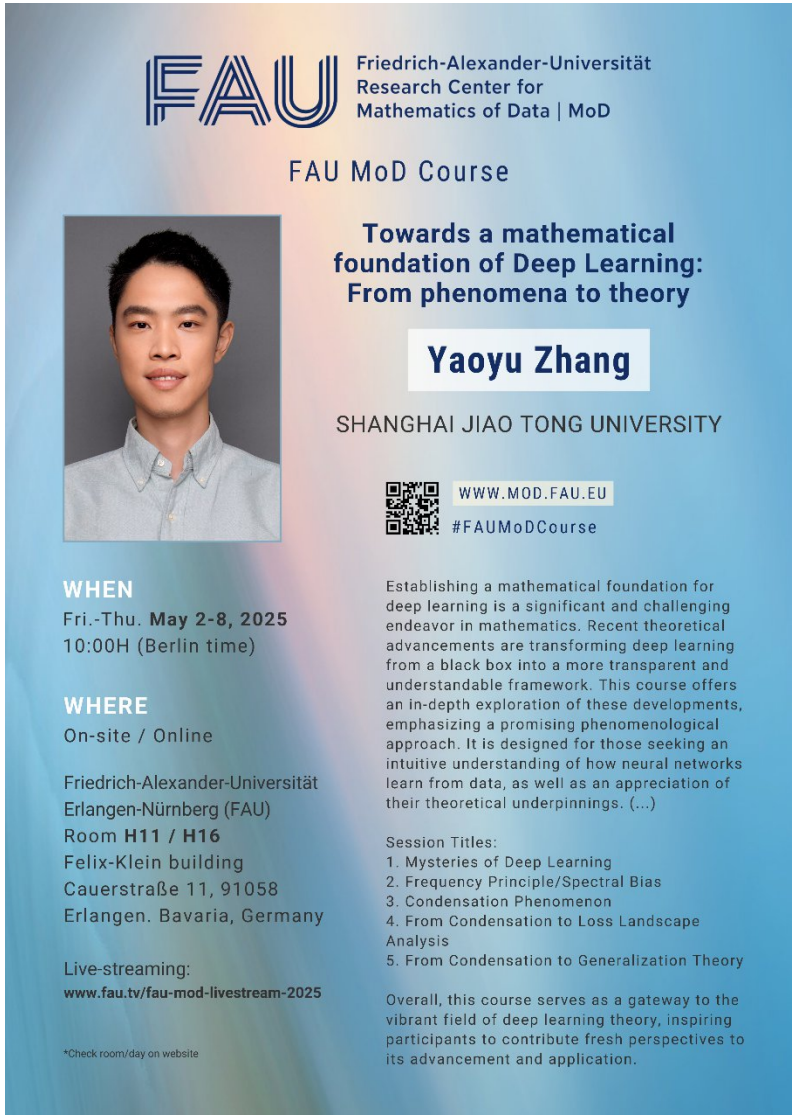
Institute of Natural Sciences & School of Mathematical Sciences

Shanghai Jiao Tong University

FAU MoD Course

饮 水 思 源 • 爱 国 荣 校

Deep learning is no longer a black-box



Towards a Mathematical Foundation of Deep Learning: From Phenomena to Theory

Date

Fri. – Thu. May 2 – 8, 2025

Session Titles

1. Mysteries of Deep Learning
2. Frequency Principle/Spectral Bias
3. Condensation Phenomenon
4. From Condensation to Loss Landscape Analysis
5. **From Condensation to Generalization Theory**

Generalization advantage of condensation

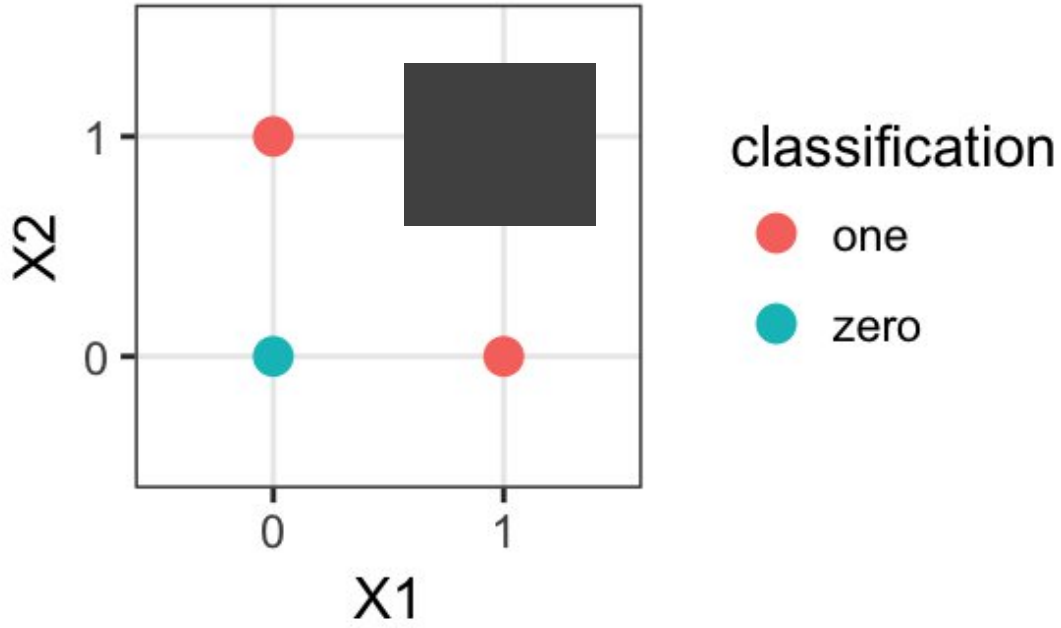
1. Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).

2. 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).

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

No Free Lunch Theorem (Wolpert and Macready)

Theorem—Given a finite set V and a finite set S of real numbers, assume that $f : V \to S$ is chosen at random according to uniform distribution on the set S^V of all possible functions from V to S. For the problem of optimizing f over the set V, then no algorithm performs better than blind search.

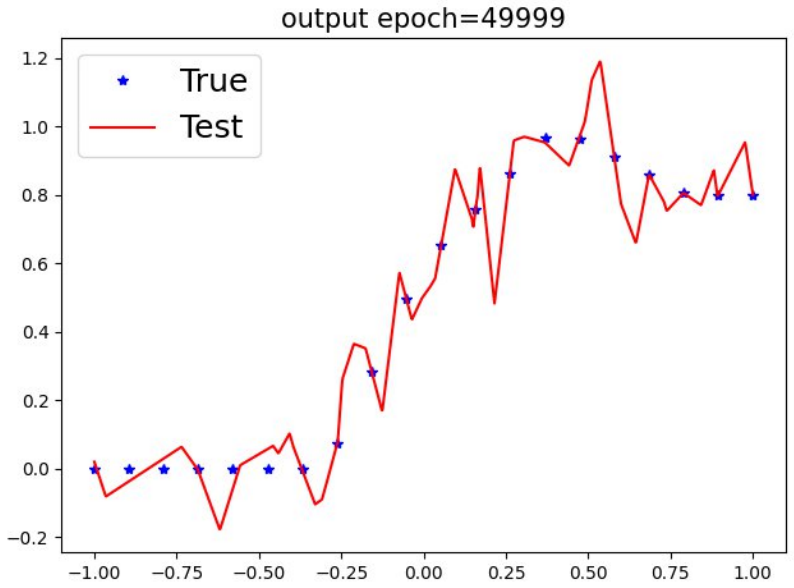


How to infer the missing spot?

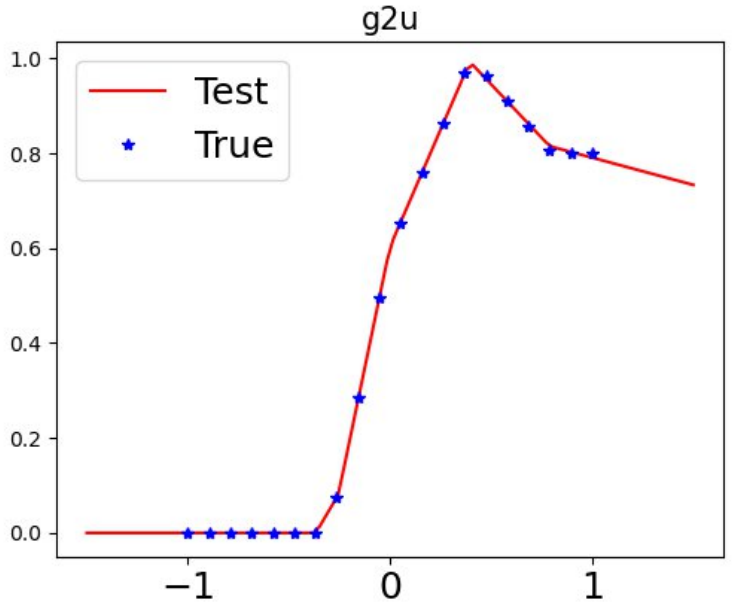
https://en.wikipedia.org/wiki/No_free_lunch_theorem

Generalization consequence of condensation

Large initialization
(no condensation)

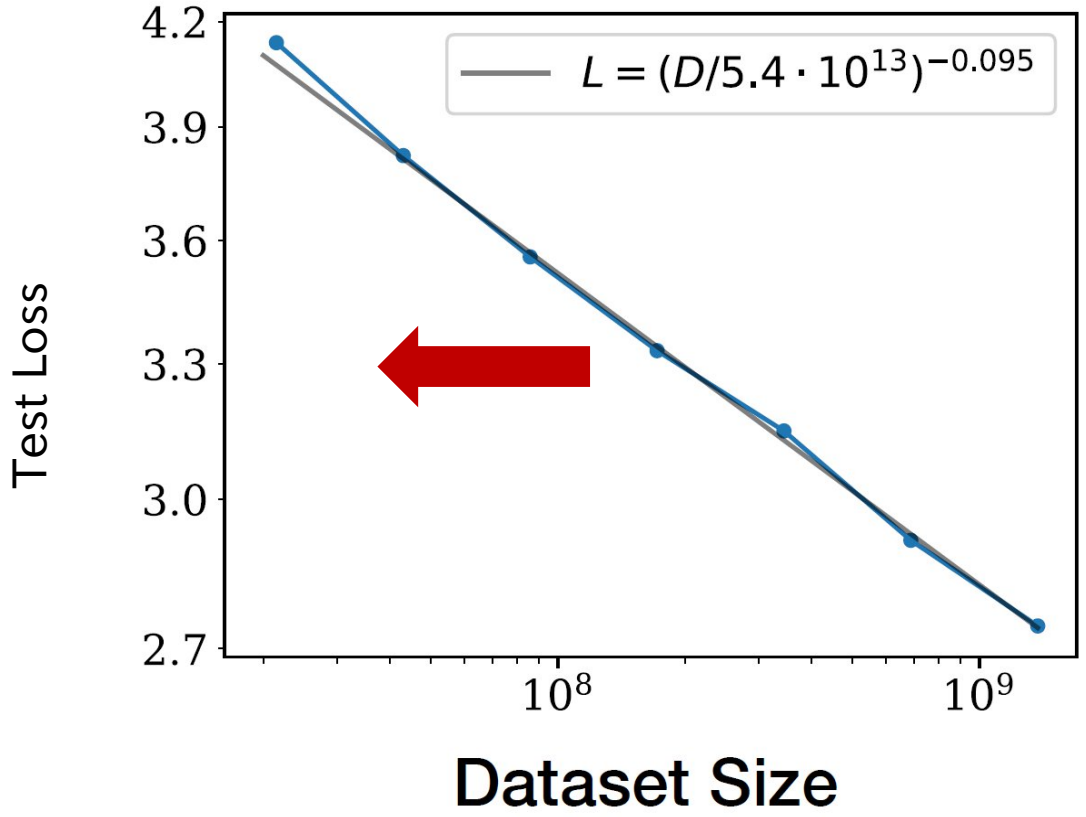


Small initialization
(Strong condensation)



Estimate sample efficiency is important

Sample efficiency: sample size required for certain performance

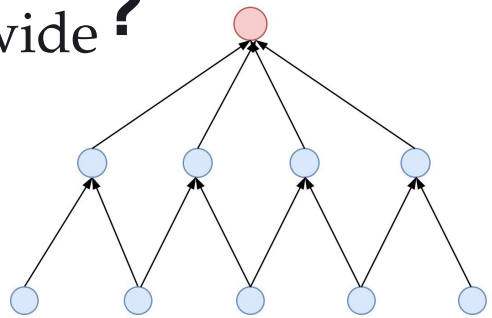


Scaling Laws for Neural Language Models. OpenAI 2020

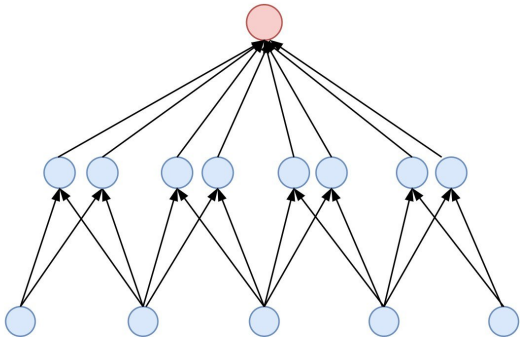
Intuition: condensation improves sample efficiency

How many samples are required to recover f^* by $\mathrm{NN}_{\mathrm{wide}}$?

$f^* =$

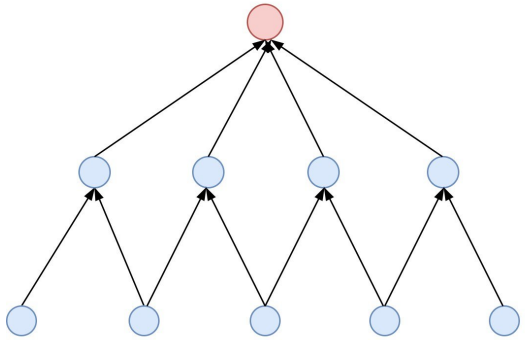


optimistic sample size



$\mathrm{NN}_{\mathrm{wide}}$

condense



$\mathrm{NN}_{\mathrm{narr}}$ (equiv)

parameter count

12

Optimistic estimate

In the best-possible scenario, what is the sample size required for fitting $f^* \in \mathcal{F}$?

How much time required from Nuremberg to Frankfurt?

Best | 2h 33m | 2h 22m | 48 hr | 13 hr | —

Nuremberg

Frankfurt am Main

Add destination

Options

Send directions to iPhone | Copy link

via B8 — 48 hr — 210 km

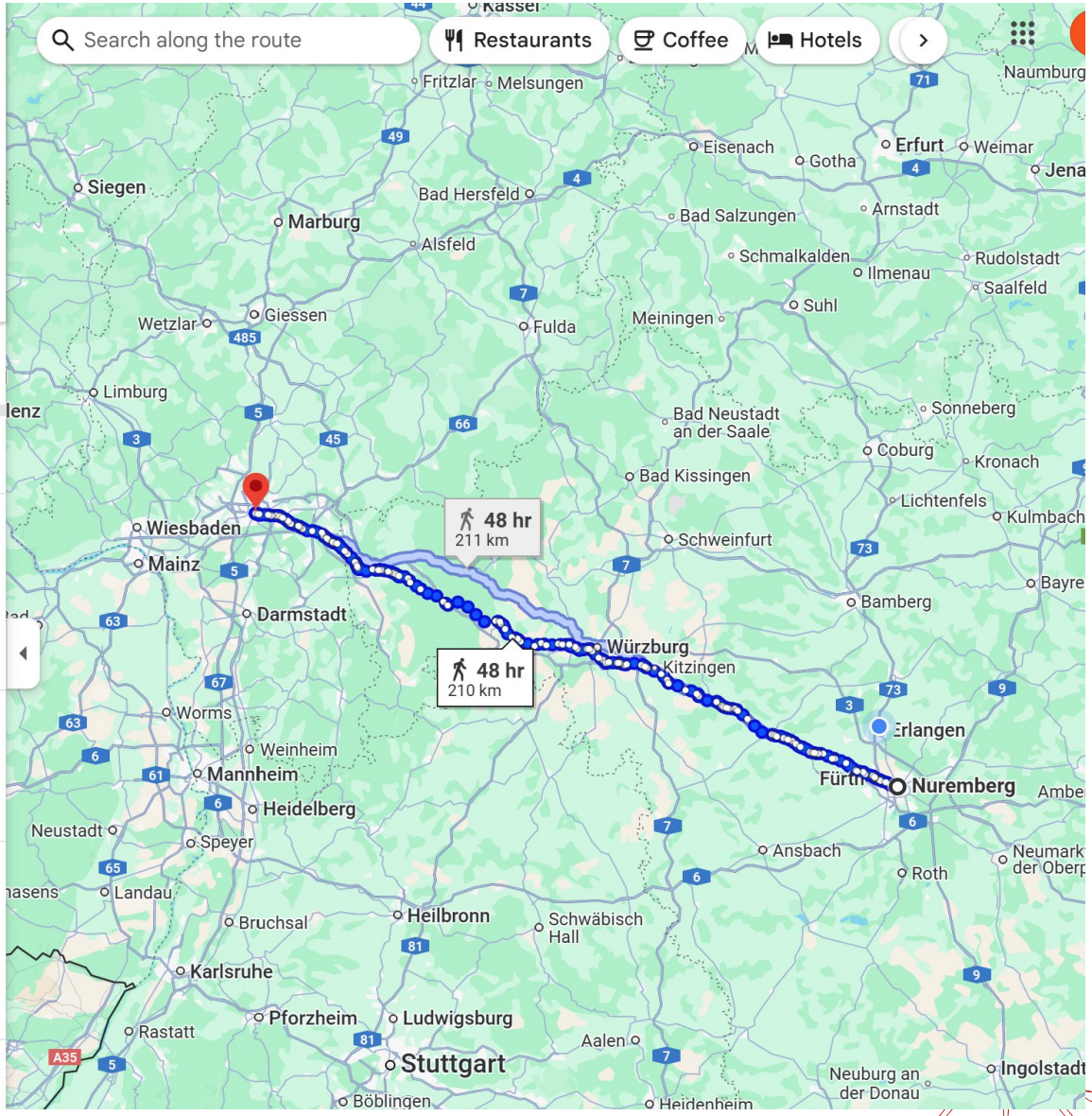
⚠ This route has restricted usage or private roads.

Details

via B26 and B8 — 48 hr — 211 km

⚠ This route has restricted usage or private roads.

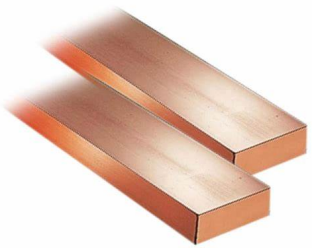
↑ 1,434 m · ↓ 1,633 m



Idea of optimistic estimate

Optimistic estimate:

Estimating the best-possible performance.



Optimistic estimate for chemical reaction:

Gold cannot be obtained from copper

Graphite has the potential to be converted to diamond

Sample size estimation--simplest setup

Data:

$$S=\left\{\left(x_i\in\mathbb{R}^d, f^*(x_i)\in\mathbb{R}\right)\right\}_{i=1}^{n}$$

Model:

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

Optimization:

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

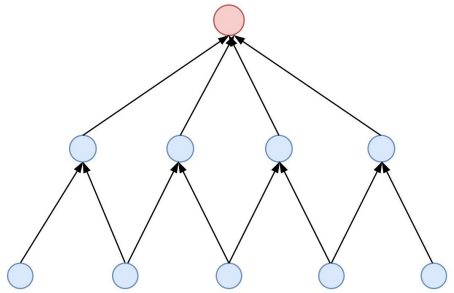
$$\dot{\theta}=-\nabla R_S(\theta)$$

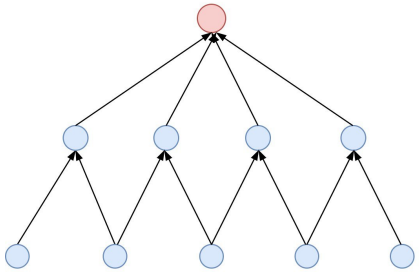
Problem: Given $f^*\in\mathscr{F}$, how many samples are required to recover f^* in the **best-possible scenario**?

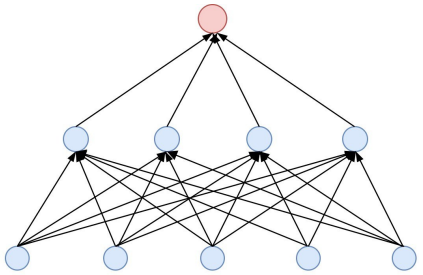
High dimensional, nonlinear, nonconvex!

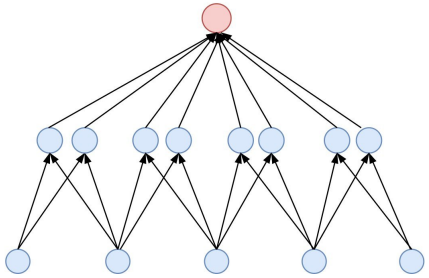
Yaoyu Zhang*, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Failure of classic estimate

$f^* =$ 

NN_A 

NN_B 

NN_C 

classic estimate

12

24

~~24~~412

reduce sample efficiency

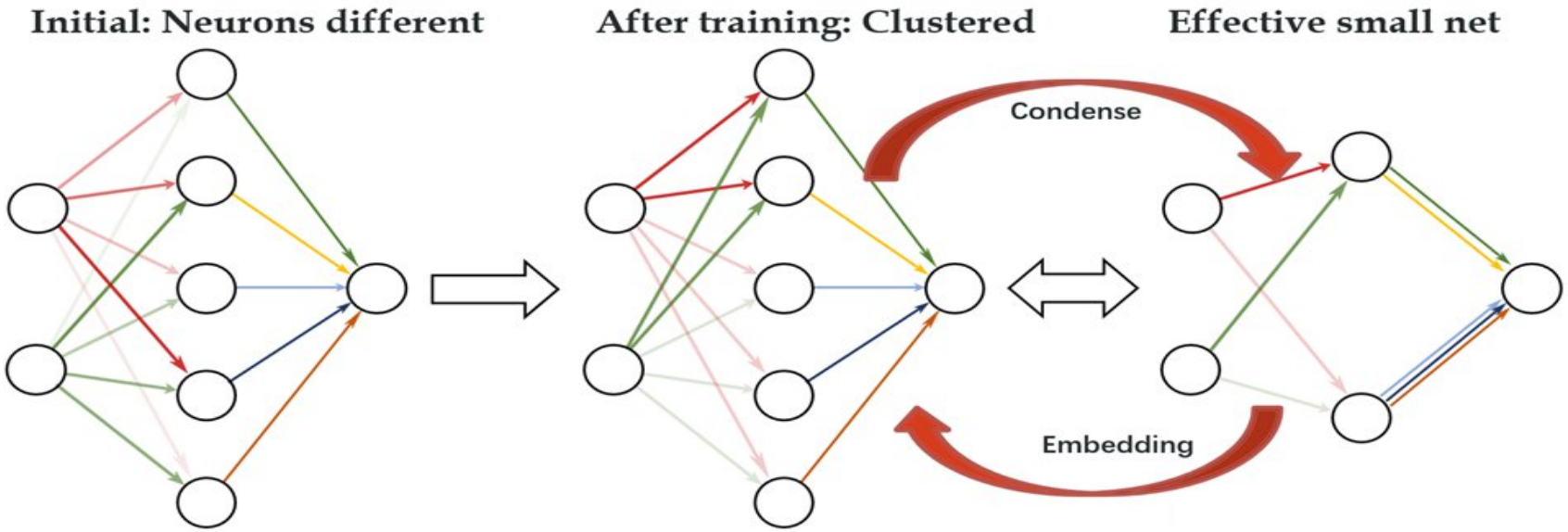
optimistic!

Classic estimate:

sample size for fitting = parameter size

Condensation—the clustering behavior of neurons

Condensation improves sample efficiency



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

$$w_1 = w_2,$$
$$w_3 = w_4 = w_5$$

$$f(x) = (a_1 + a_2)\sigma(w_1^T x) + (a_3 + a_4 + a_5)\sigma(w_3^T x)$$

Initial: random

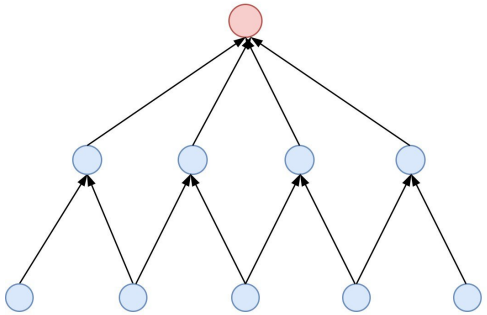
Training: condense

Effect: equiv to small net

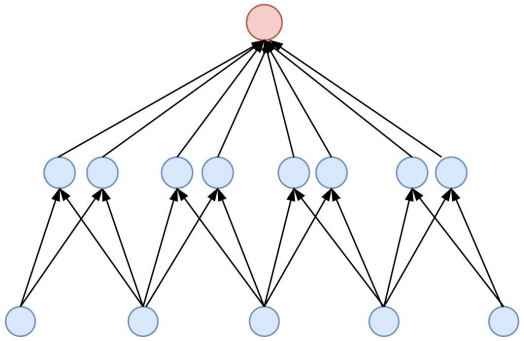
Tao Luo, Zhi-Qin John Xu, Zheng Ma, **Yaoyu Zhang*,** Phase diagram for two-layer ReLU neural networks at infinite-width limit, Journal of Machine Learning Research (JMLR) 22(71):1−47, (2021)

Condensation improves sample efficiency

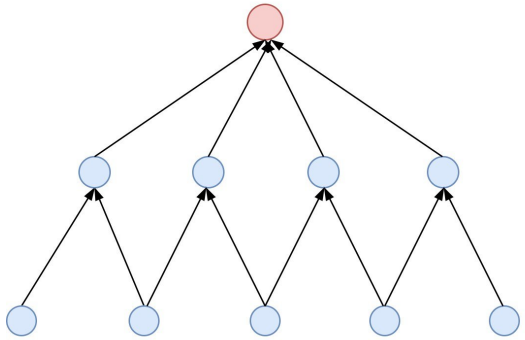
How many samples are required to fit f^*?

$f^* =$ 

optimistic estimate



condense



classic estimate

12

$\mathrm{NN_C}$

$\mathrm{NN_A}$ (equiv)

Model rank

Model:

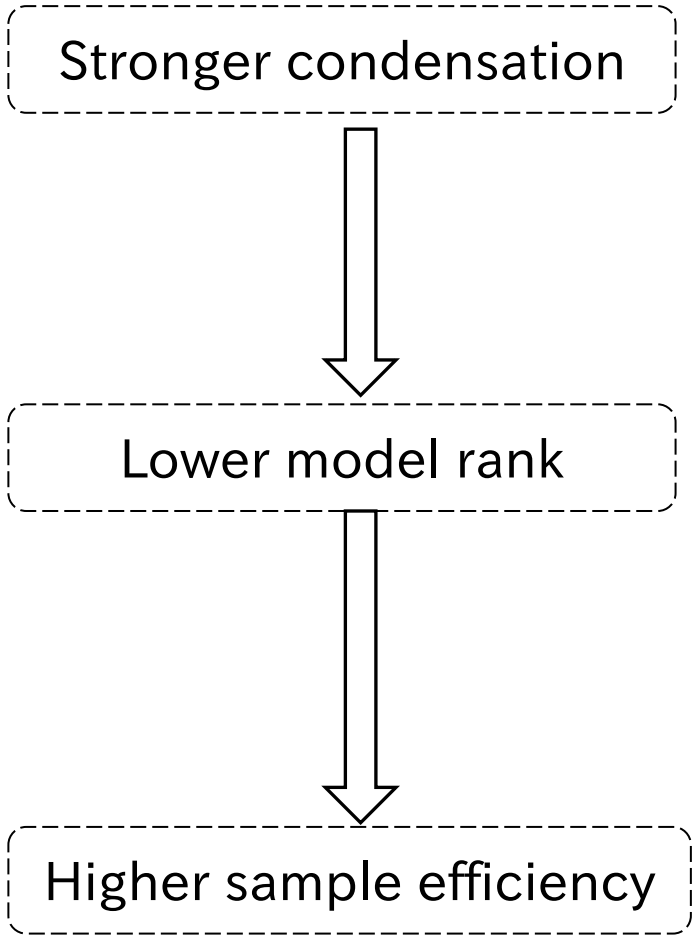
$$F:\mathbb{R}^M \to \mathscr{F} \subseteq C(\mathbb{R}^d)$$

Model rank:

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

Intuition: effective degrees of freedom at θ

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



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:

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

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

$$r_\theta \leq 2$$

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

$$r_\theta = 4$$

Optimistic sample size estimate

Model:

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

Model rank:

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

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

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

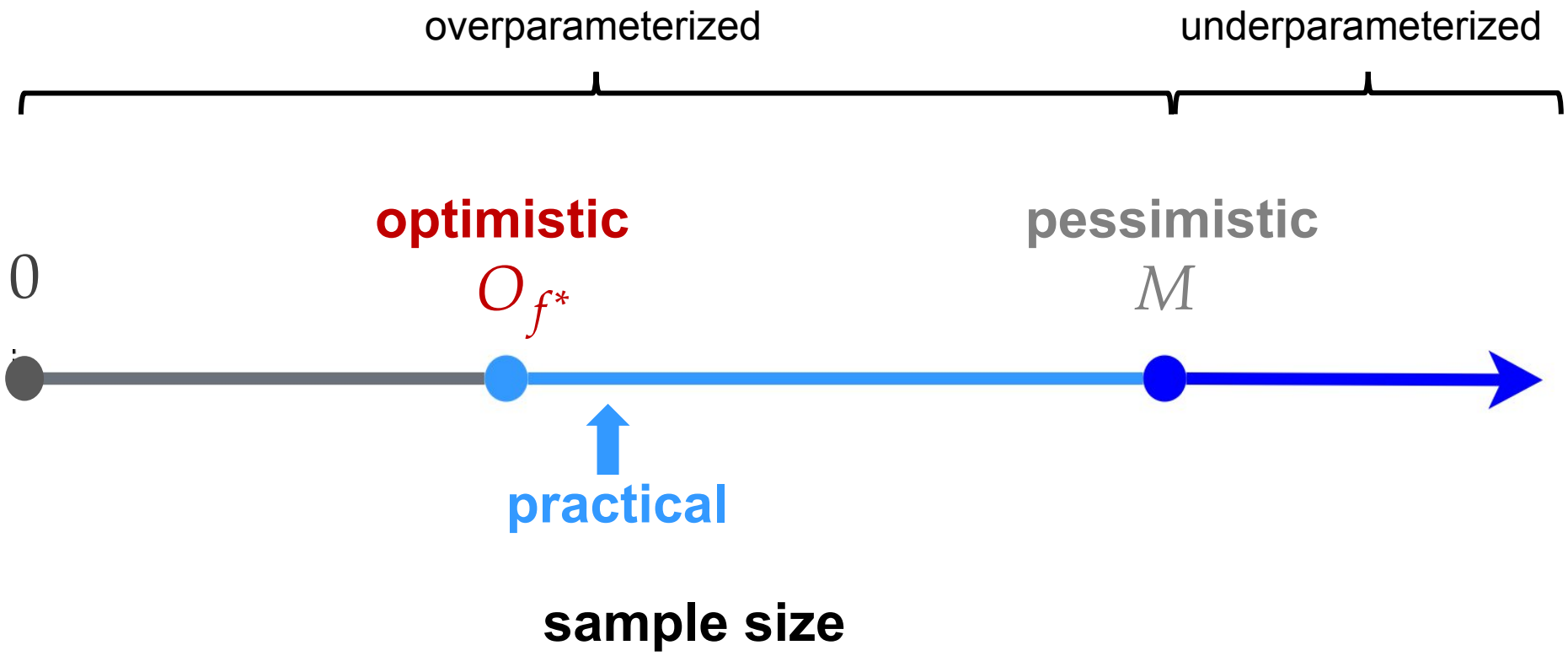
$F^{-1}(f^*)$：目标集(零泛化误差)

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).

Picture of sample size requirement to recover $f^* \in \mathcal{F}$

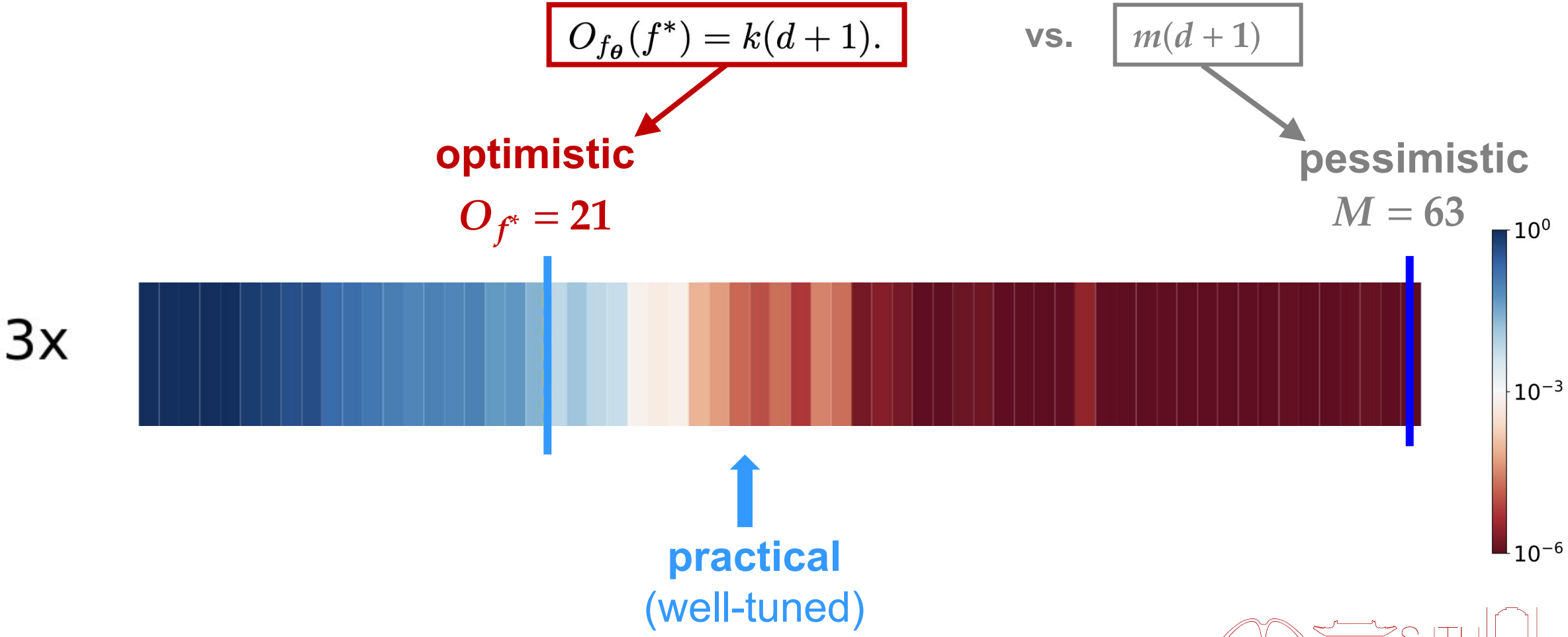


Optimistic sample size estimate vs practice

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 sample size estimate vs 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*

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

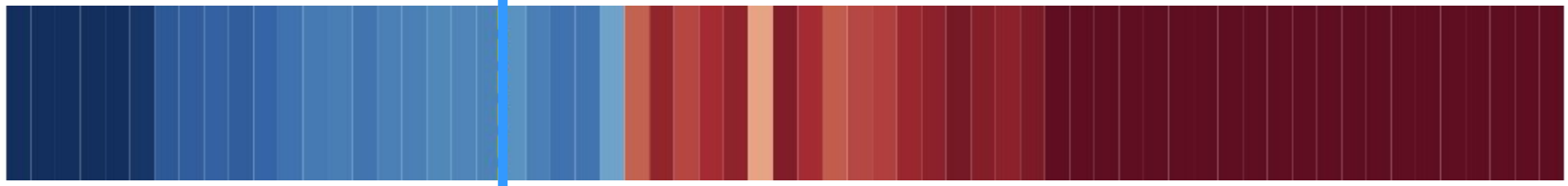
vs. $m(d+1)$

optimistic

$O_{f^*} = 21$

pessimistic

$M = 2100$



100x

...

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

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

Key tool for optimistic estimation: critical embedding

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

Criticality of optimistic sample size

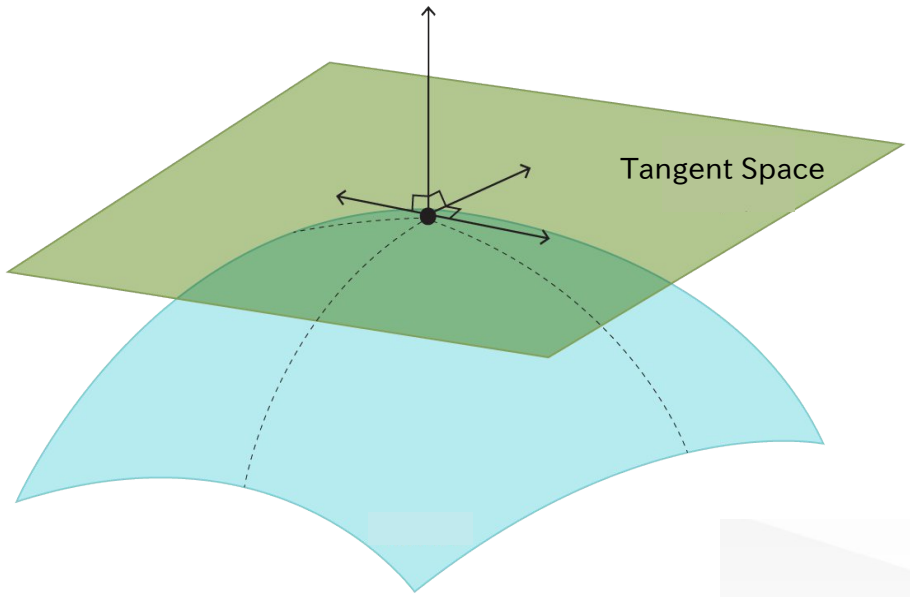
Theorem (phase transition of LLR-guarantee at a target point). For any $\theta' \in \mathscr{M}_{f^*}$

- 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^*

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

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

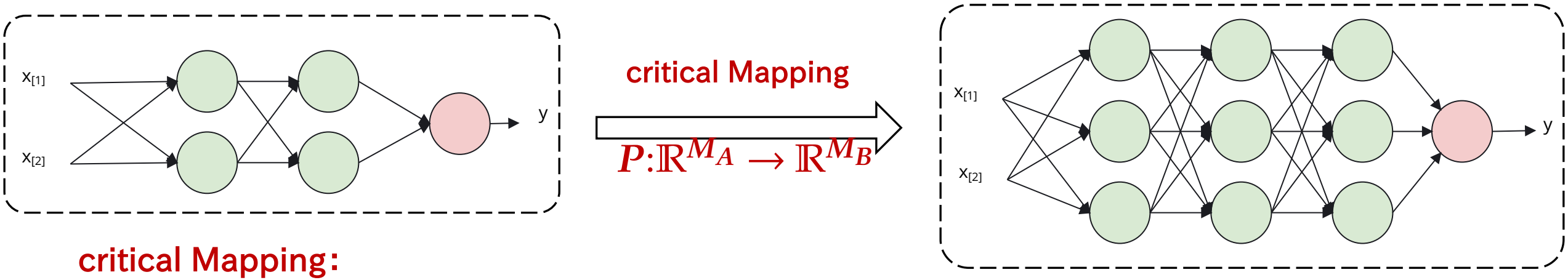


Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Optimistic sample size 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：

(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

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

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 $\boldsymbol{\theta}^{\mathrm{c}}_{\mathrm{narr}}$ of the narrower NN, $\boldsymbol{\theta}^{\mathrm{c}}_{\mathrm{wide}} := \mathcal{T}(\boldsymbol{\theta}^{\mathrm{c}}_{\mathrm{narr}})$ is still a critical point of the K-neuron wider NN with the same output function, i.e., $\boldsymbol{f}_{\boldsymbol{\theta}^{\mathrm{c}}_{\mathrm{narr}}} = \boldsymbol{f}_{\boldsymbol{\theta}^{\mathrm{c}}_{\mathrm{wide}}}$.*

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 $\boldsymbol{\theta}_{\mathrm{c}}$ of any shallower $\mathrm{NN}\left(\{m_l\}_{l=0}^{L}\right)$ satisfying $\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}_{\mathrm{c}}) = \mathbf{0}$, there exists parameter $\boldsymbol{\theta}'_{\mathrm{c}}$ in the loss landscape of $\mathrm{NN}'\left(\{m'_l\}_{l=0}^{L'}\right)$ satisfying the following conditions:*

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

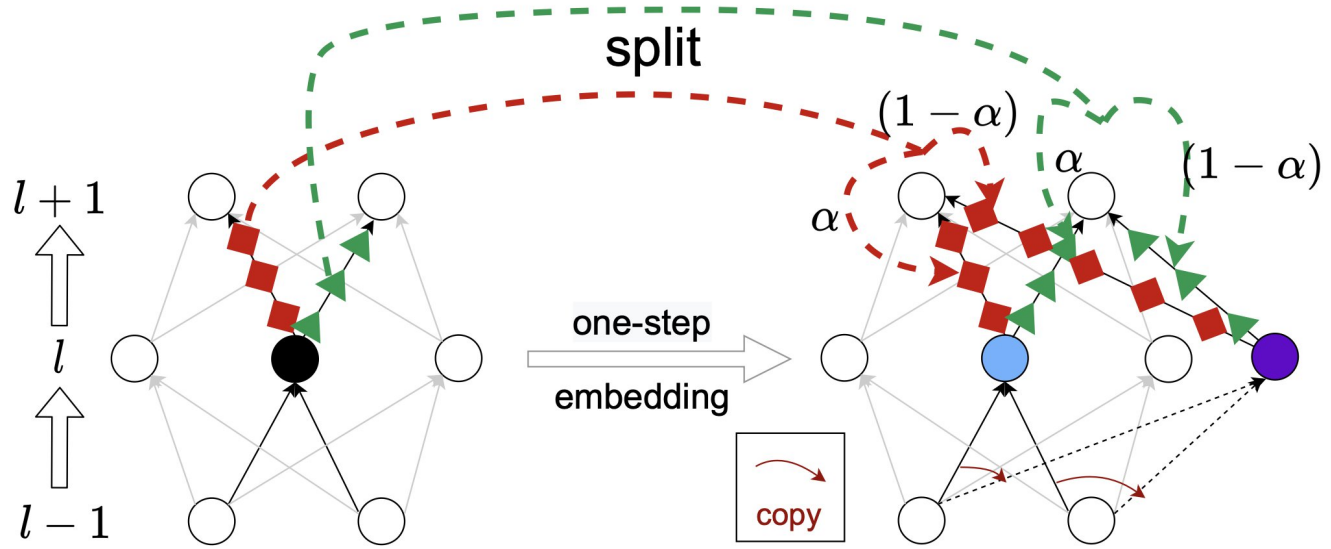
(ii) $\nabla_{\boldsymbol{\theta}'} R_S(\boldsymbol{\theta}'_{\mathrm{c}}) = \mathbf{0}$.

Embedding Principle (width/depth):
The loss landscape of a neural network “contains” all the critical points of narrower/shallower networks.

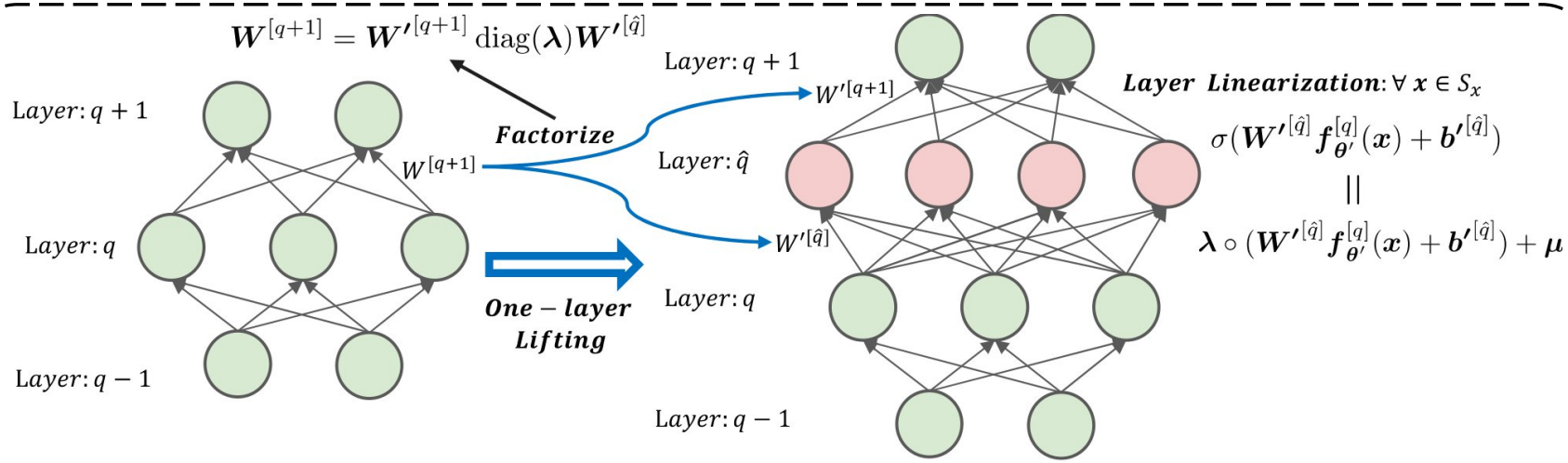
[1] Zhang, Zhang, Luo, Xu, Embedding Principle of Loss Landscape of Deep Neural Networks. NeurIPS 2021 Spotlight
[2] Zhang, Li, Zhang, Luo, Xu, Embedding Principle: a hierarchical structure of loss landscape of deep neural networks. JML 2022
[3] Bai, Luo, Xu, Zhang, Embedding Principle in Depth for the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

Embedding principle -> sample efficiency preserving

Embedding Principle (width)



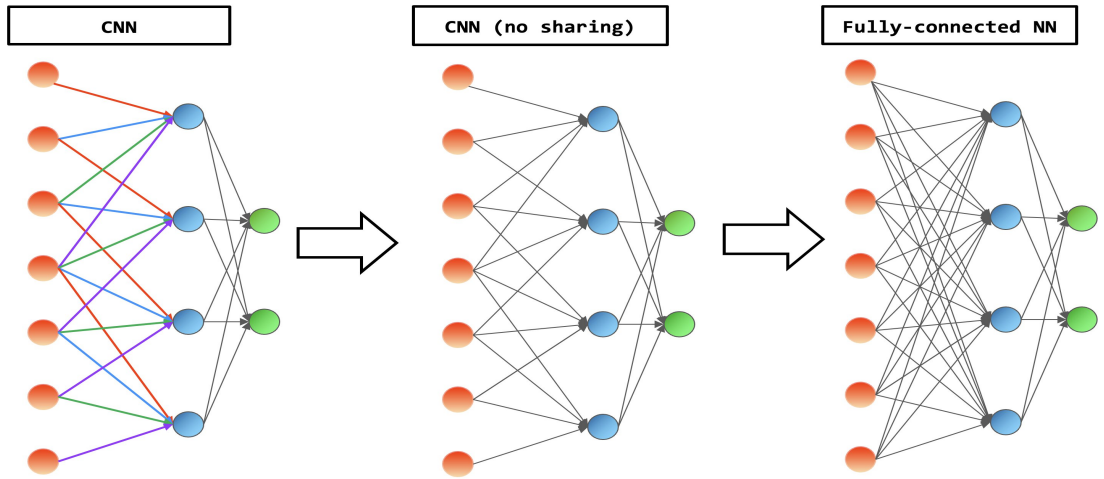
Embedding Principle (depth)



[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.

Impact of connection

Impact of connection



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

more connection lower sample efficiency

k intrinsic width

$s \times s$ conv ker size

$d \times d$ input dim

CNN vs. FNN

Adding (unnecessary) connections reduces sample efficiency

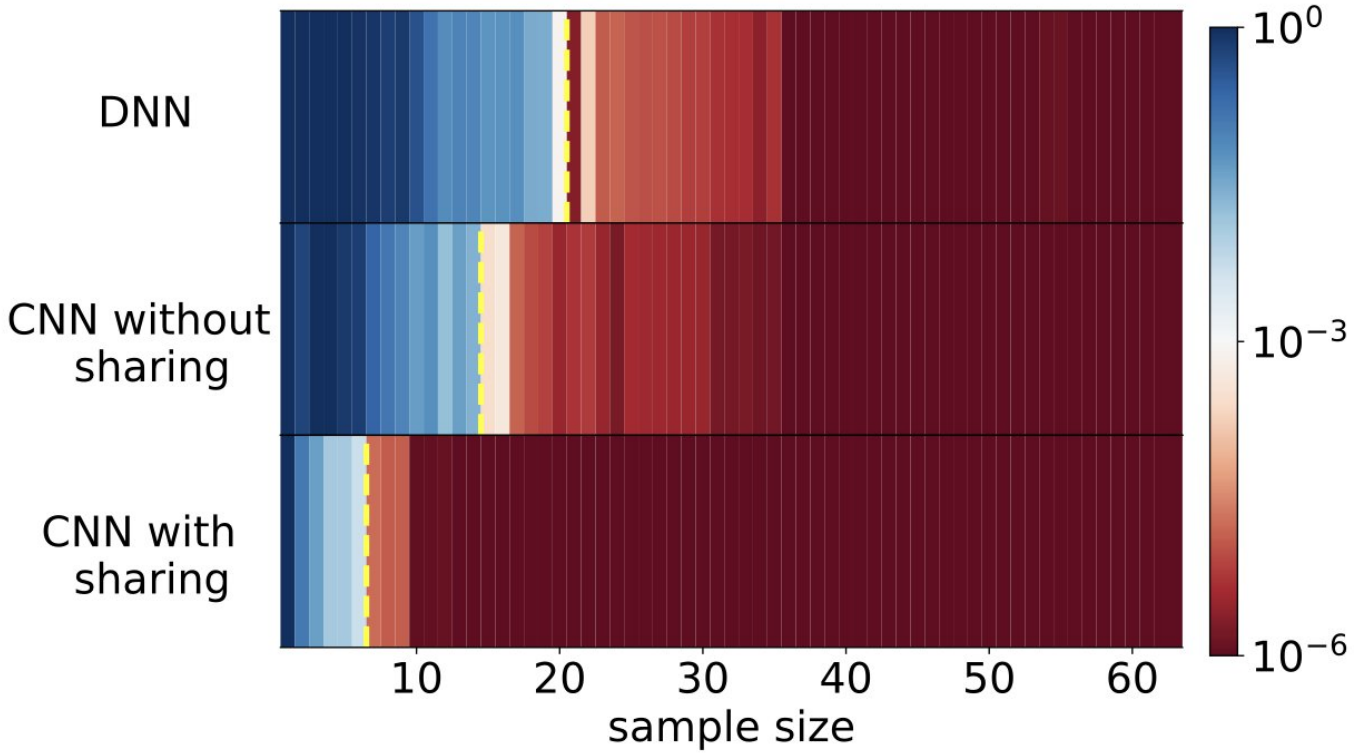
MNIST k-kernel, kernel size: 3x3

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

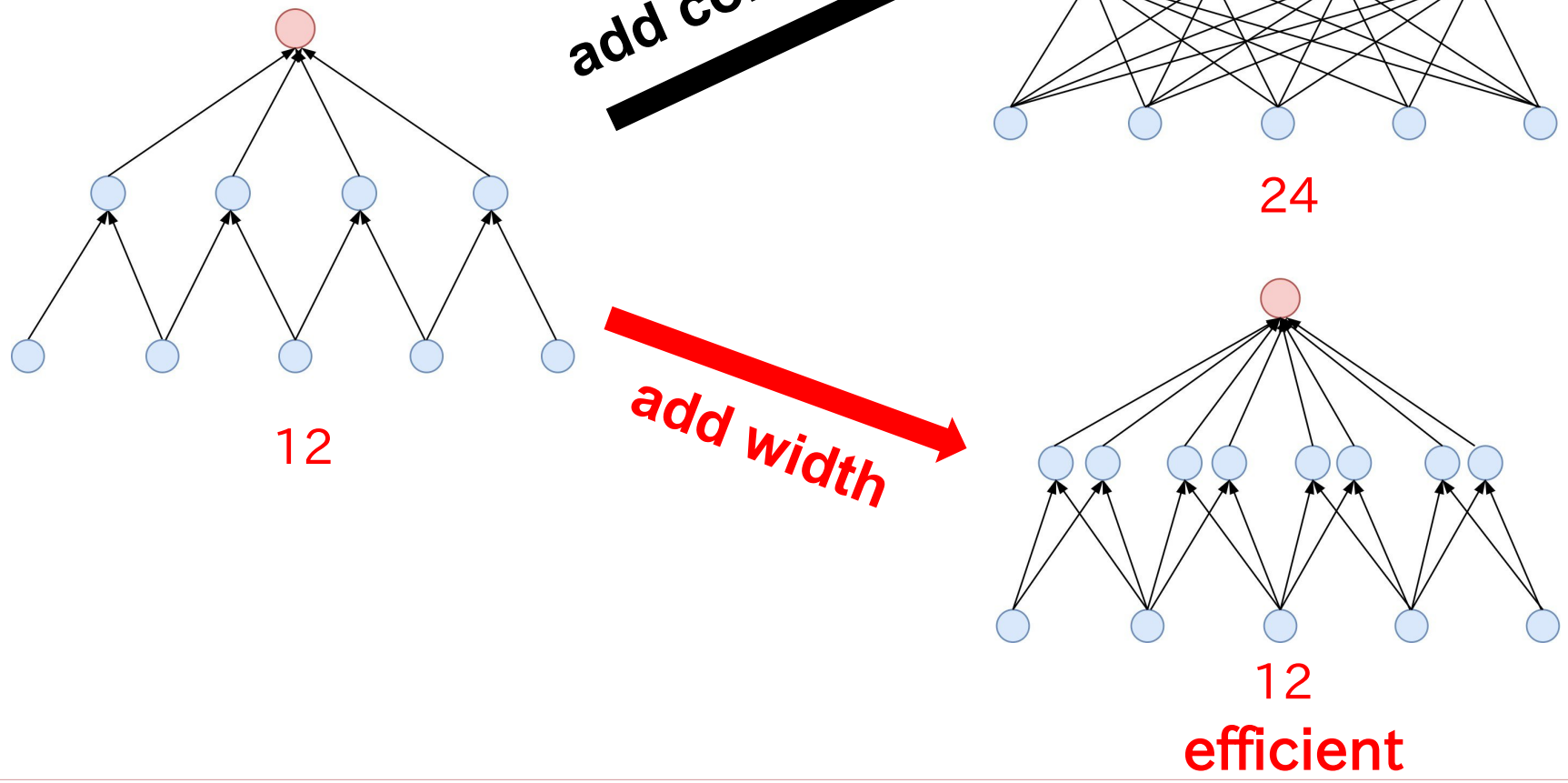
1000X worse!

Optimistic estimate vs. experiments

Experiment：adding connection reduces sample efficiency



Specialty of DNN models via optimistic estimate



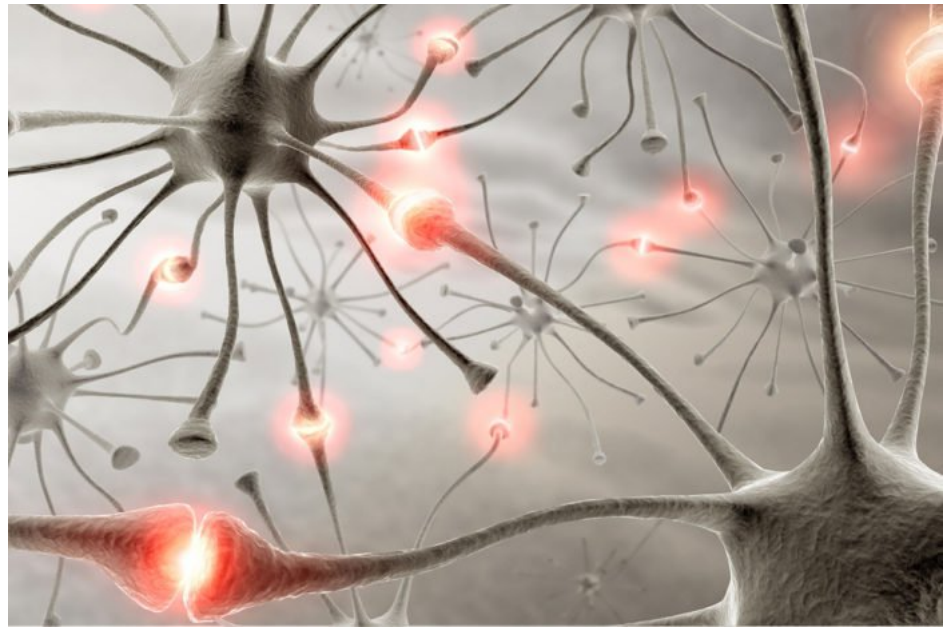
Sample inefficient:
Adding (unnecessary) connections worsens generalization

Sample efficient:
Increasing width doesn't harm generalization

Inspiration

Principle of model scaling

- Freely increase width (also depth)
- 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

Optimistic estimate for simpler models

Linear models

Model:

$$F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{m} a_i x_i$$

Optimistic sample size:

For any $f^* \in \mathscr{F}$,

$$O_{f^*} = m$$

Conclusion:

Linear models cannot recover targets under overparameterization.

Deep models

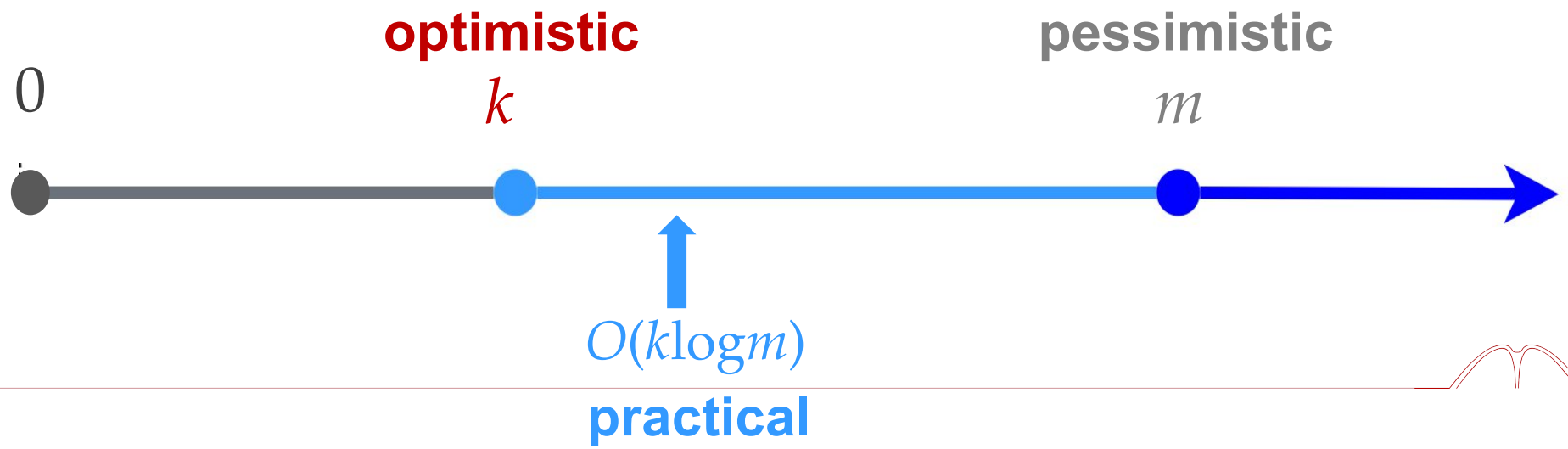
Model:

$$F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{m} (a_i^2 - b_i^2) x_i$$

Optimistic sample size:

For a k-sparse function f^*,

$$O_{f^*} = k$$



Deep models

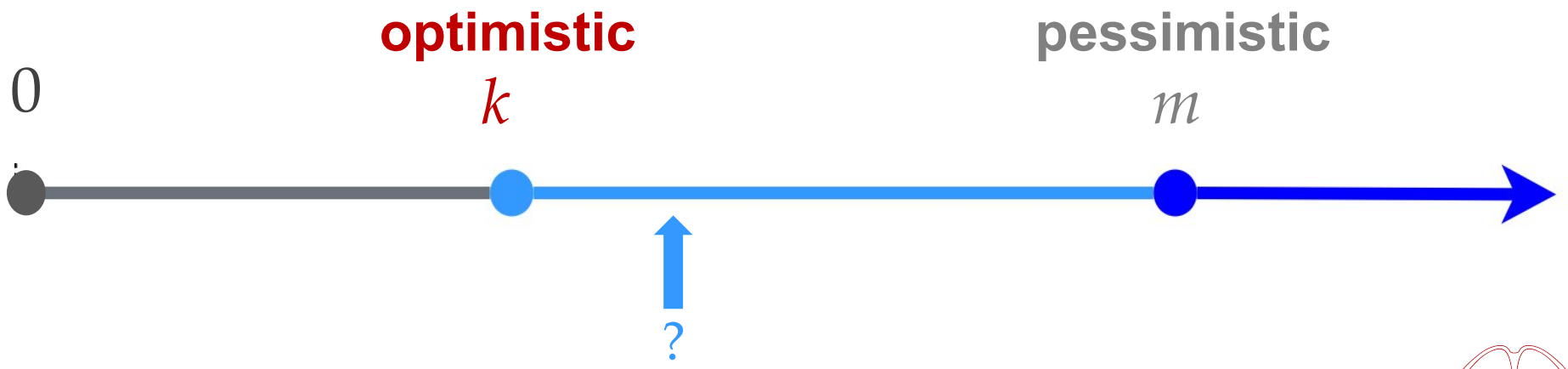
Model:

$$F(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{i=1}^{m} a_i^{[L]} \cdots a_i^{[1]} x_i$$

Optimistic sample size:

For a k-sparse function f^*,

$$O_{f^*} = k$$



Matrix factorization model

Model: $f_\theta = AB,\ \theta = (A,B),A,B \in \mathbb{R}^{d\times d}$

Loss: $L_S(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left([AB]_{j_i k_i} - M^*_{j_i k_i}\right)^2$

$$\begin{bmatrix} \boxed{12.1} & \boxed{17.3} & \boxed{24.1} & \boxed{-4.9} \\ 16.3 & \boxed{24.1} & 16.1 & 1.1 \\ 14.2 & 25.8 & \boxed{16.9} & 4.3 \\ 22.2 & 15.6 & 18.5 & \boxed{3.1} \end{bmatrix}$$

Matrix factorization model

Model: $f_\theta = AB,\ \theta = (A,B),A,B\in\mathbb{R}^{d\times d}$

Rank stratification over the parameter space:

$$d^2-(d-r_A)(d-r_B)$$

Optimistic estimation over the function space:

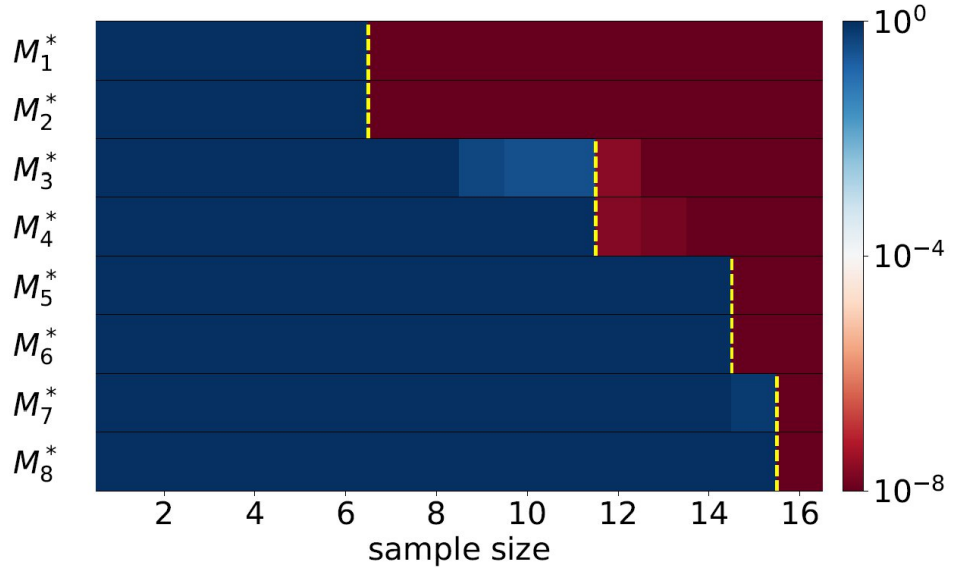
model	$f_\theta = AB, \theta = (A, B), A, B \in \mathbb{R}^{d\times d}$	
$\mathrm{rank}_{f_\theta}(f^*)$	f^*	$\arg\min_{\theta'\in\mathcal{M}_{f^*}} \mathrm{rank}_{f_\theta}(\theta')$
0	$\mathbf{0}_{d\times d}$	$A = B = \mathbf{0}_{d\times d}$
$2d-1$	$\mathrm{rank}(f^*) = 1$	$\mathrm{rank}(A) = \mathrm{rank}(B) = 1, AB = f^*$
$\vdots$	$\vdots$	$\vdots$
$2rd-r^2$	$\mathrm{rank}(f^*) = r$	$\mathrm{rank}(A) = \mathrm{rank}(B) = r, AB = f^*$
$\vdots$	$\vdots$	$\vdots$
d^2	$\mathrm{rank}(f^*) = d$	$\mathrm{rank}(A) = \mathrm{rank}(B) = d, AB = f^*$

Numerical experiments

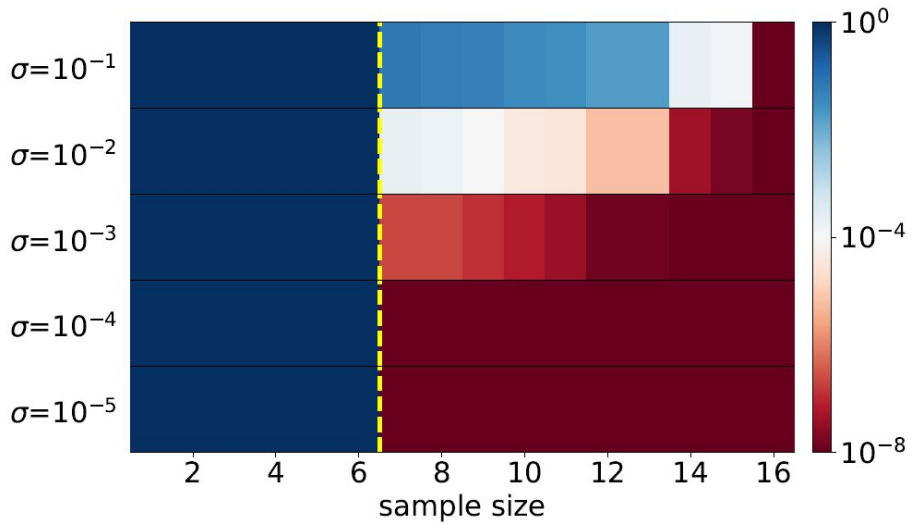
Matrix factorization model:

$$f_\theta = AB,\ \theta = (A,B),\ A,B \in \mathbb{R}^{4\times 4}$$

O_{f^*}	target matrix
0	$\mathbf{0}$
7	$\mathrm{rank}(f^*) = 1$
12	$\mathrm{rank}(f^*) = 2$
15	$\mathrm{rank}(f^*) = 3$
16	$\mathrm{rank}(f^*) = 4$



Effect of hyperparameter tuning



(b) matrix completion, rank=1

- Success of fitting depends on the initialization scale;
- Increasing data size beyond the optimistic sample size enhances tolerance on tuning.

Symmetry and optimistic estimate

Symmetry of a nonlinear model has profound impact on its optimistic sample sizes.

Permutation symmetry → sample efficiency preserving

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

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

Theorem(informal):

permutation-invariant manifolds are invariant manifolds of gradient flow.

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

Permutation symmetry → invariant manifolds (equiv to smaller network)

→ optimistic sample size no larger than 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:

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

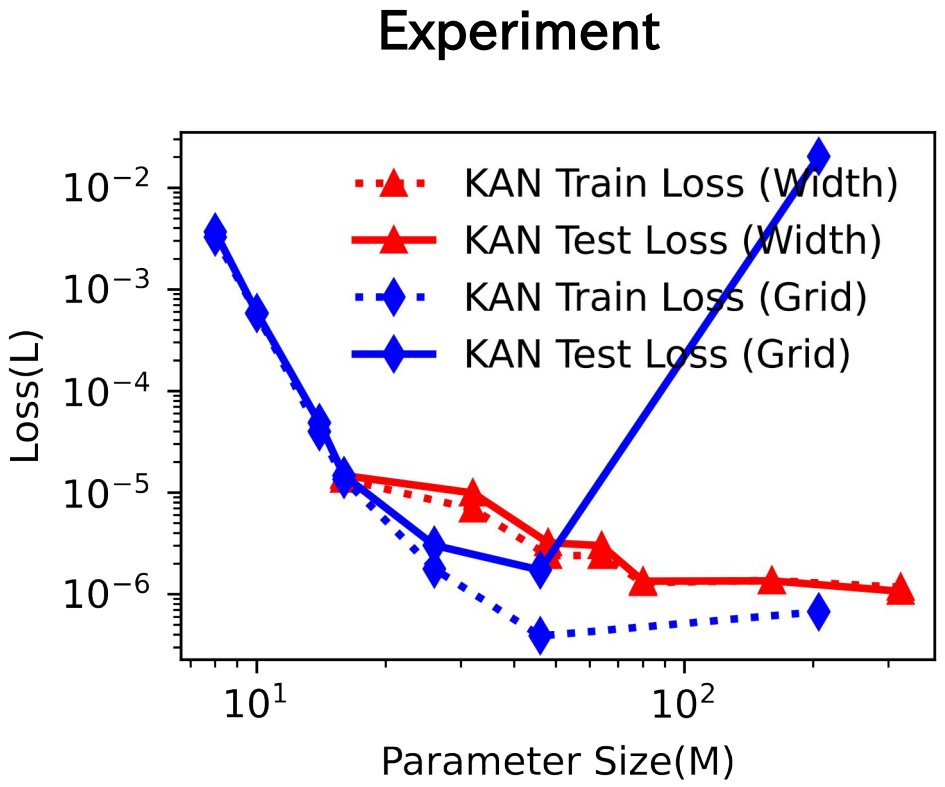
width : permutation symmetric

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

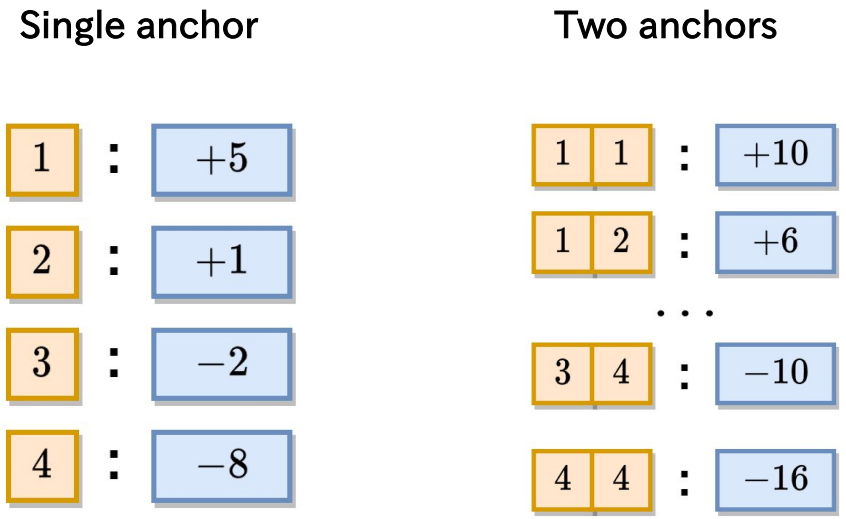
Increase G reduces sample efficiency



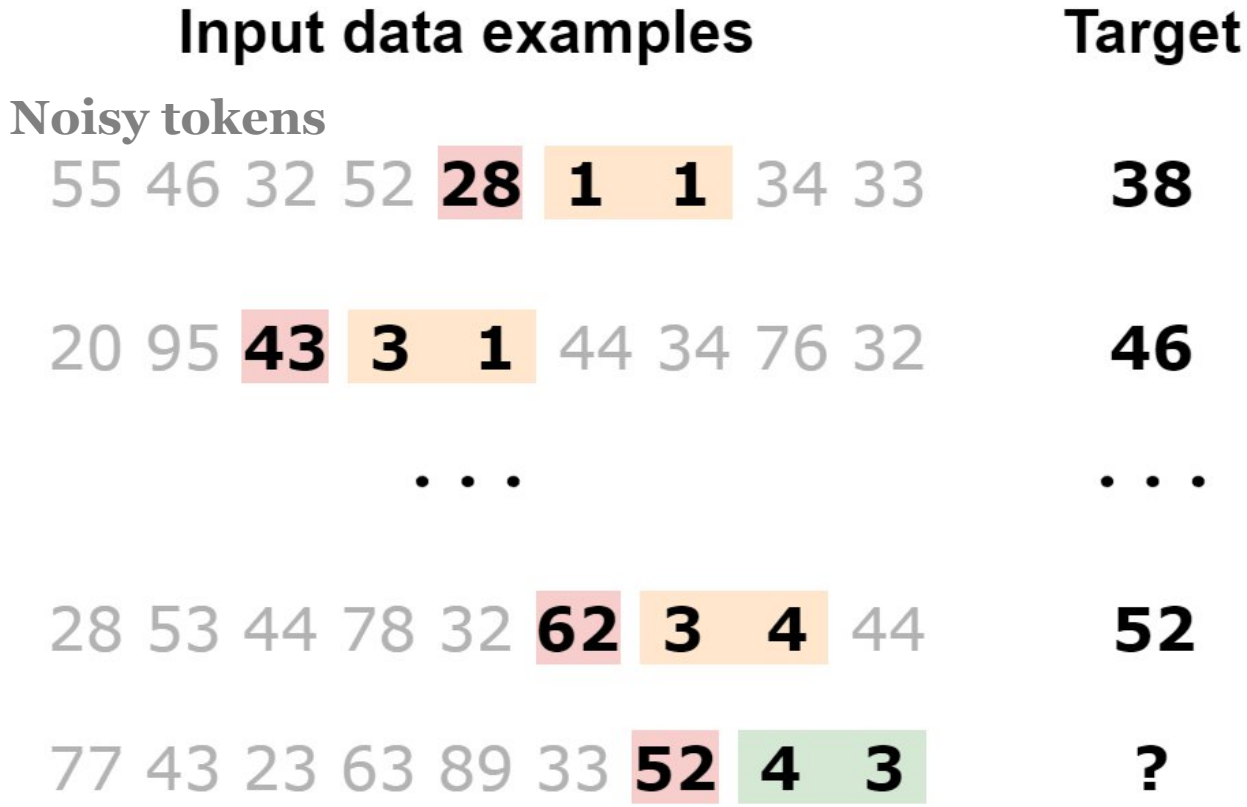
Empirical estimation and application

Empirical estimation of optimistic sample size

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

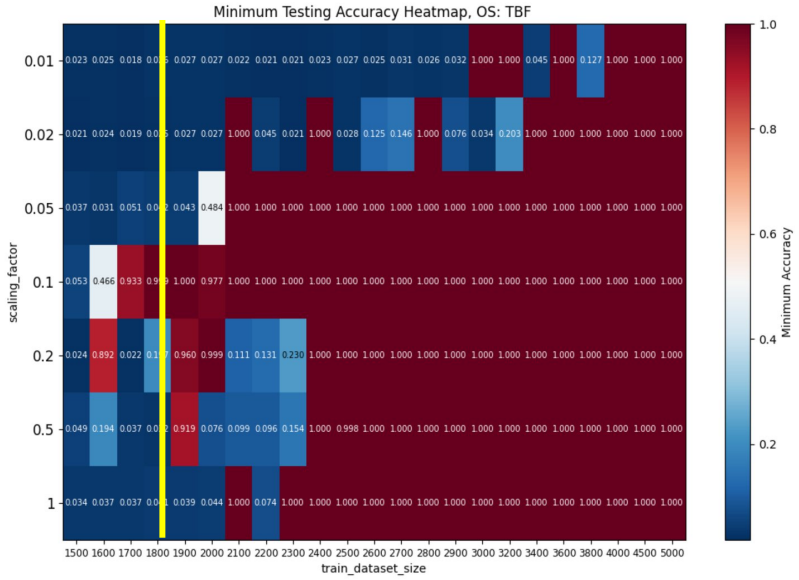


- Composite Anchor function
- Symmetry Anchor function
- Random Anchor function

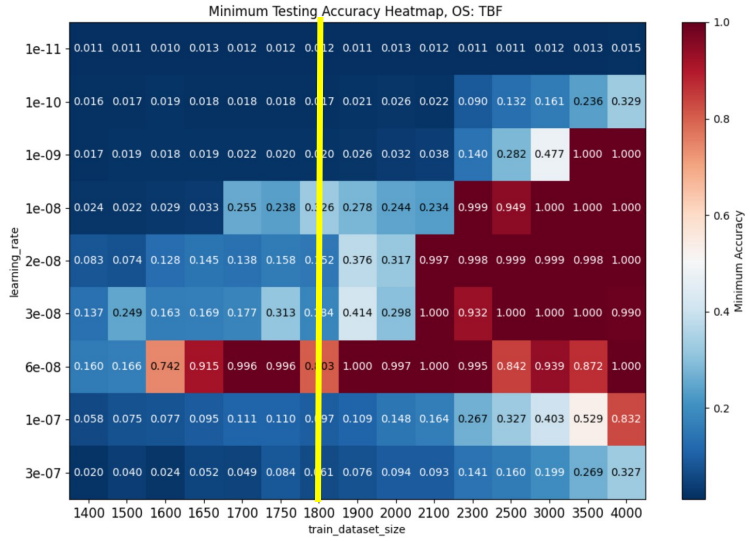


Empirical estimation of optimistic sample size

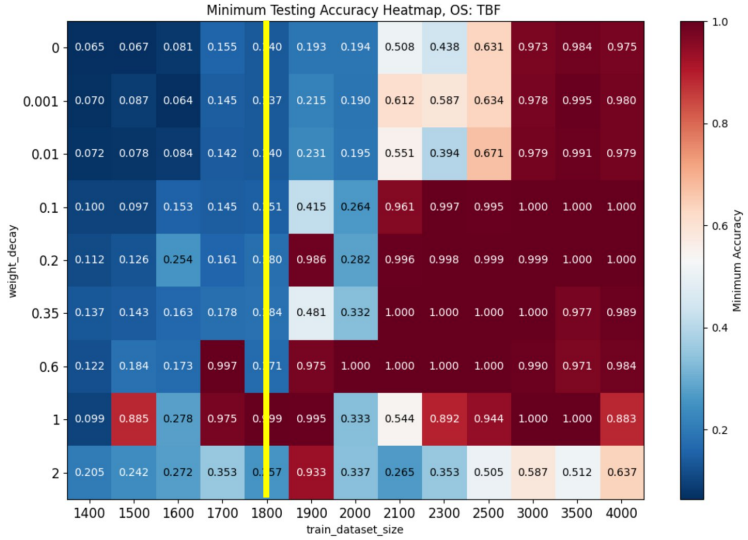
Identification of best tuning of hyperparameters



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

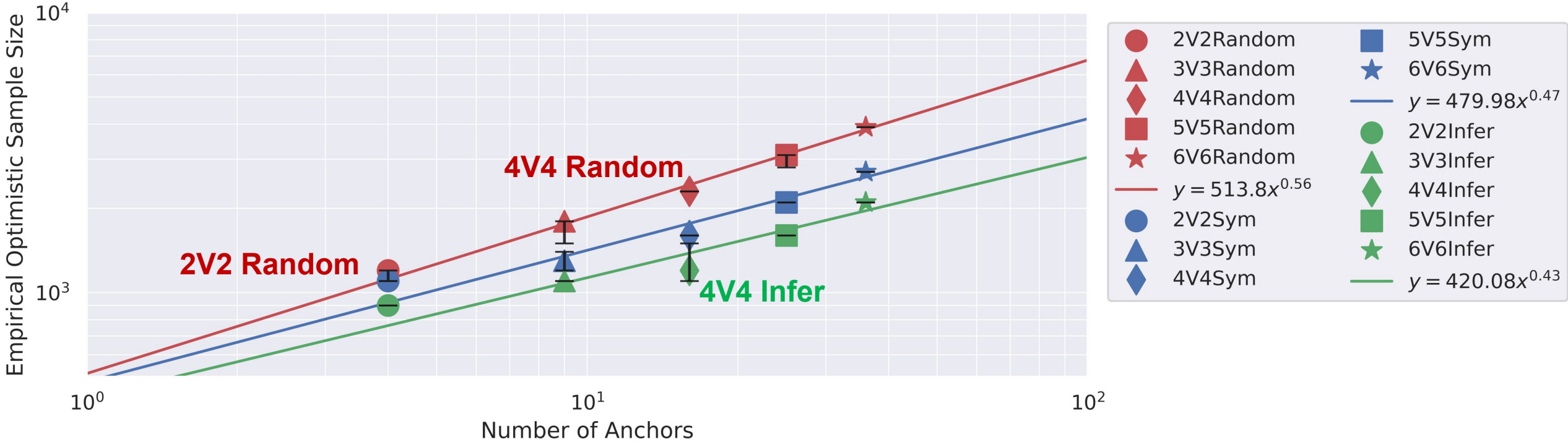
Can Transformer make inference?

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

(Optimistically) Memorization? No!

Transformer:

Memorization complexity poorly predicts optimistic sample size

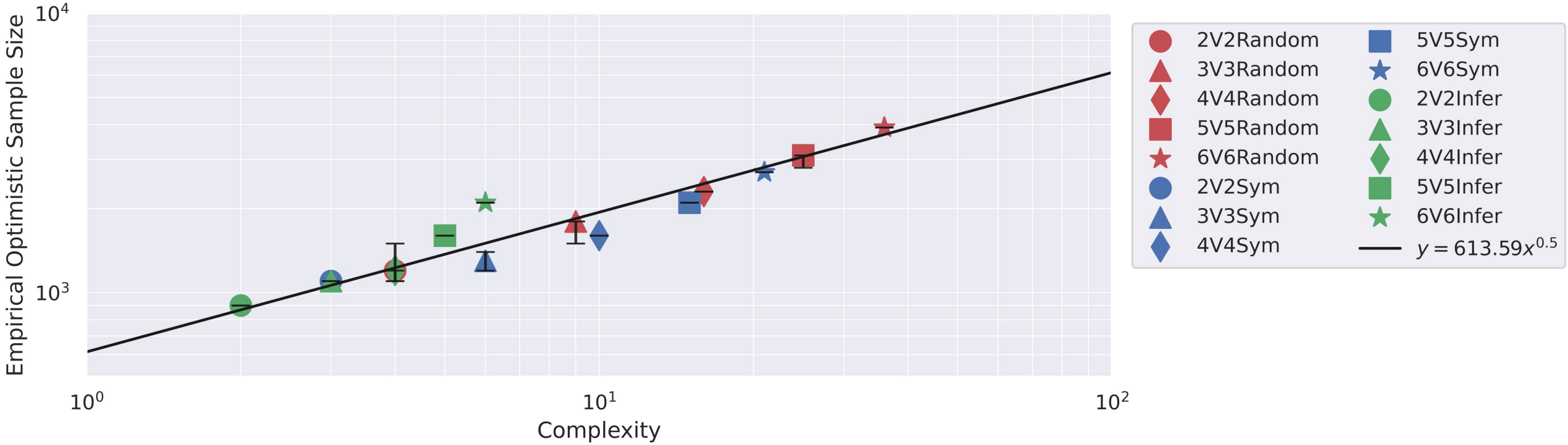


optimistic sample size vs. memorization complexity

(Optimistically) Inference? Yes!

Transformer:

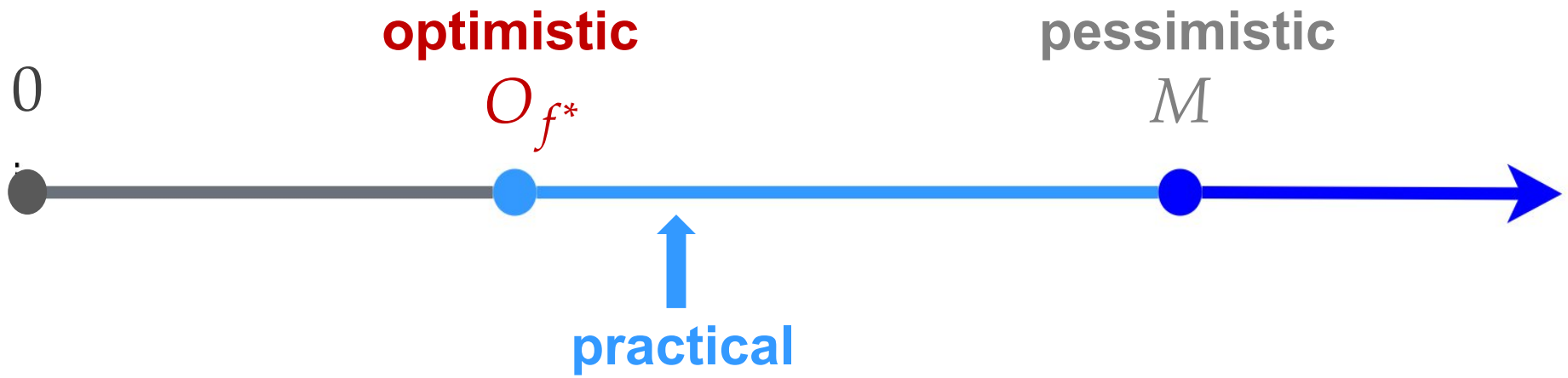
Inference complexity well predicts optimistic sample size



optimistic sample size vs. inference complexity

Messages

Picture of sample size requirement for nonlinear models

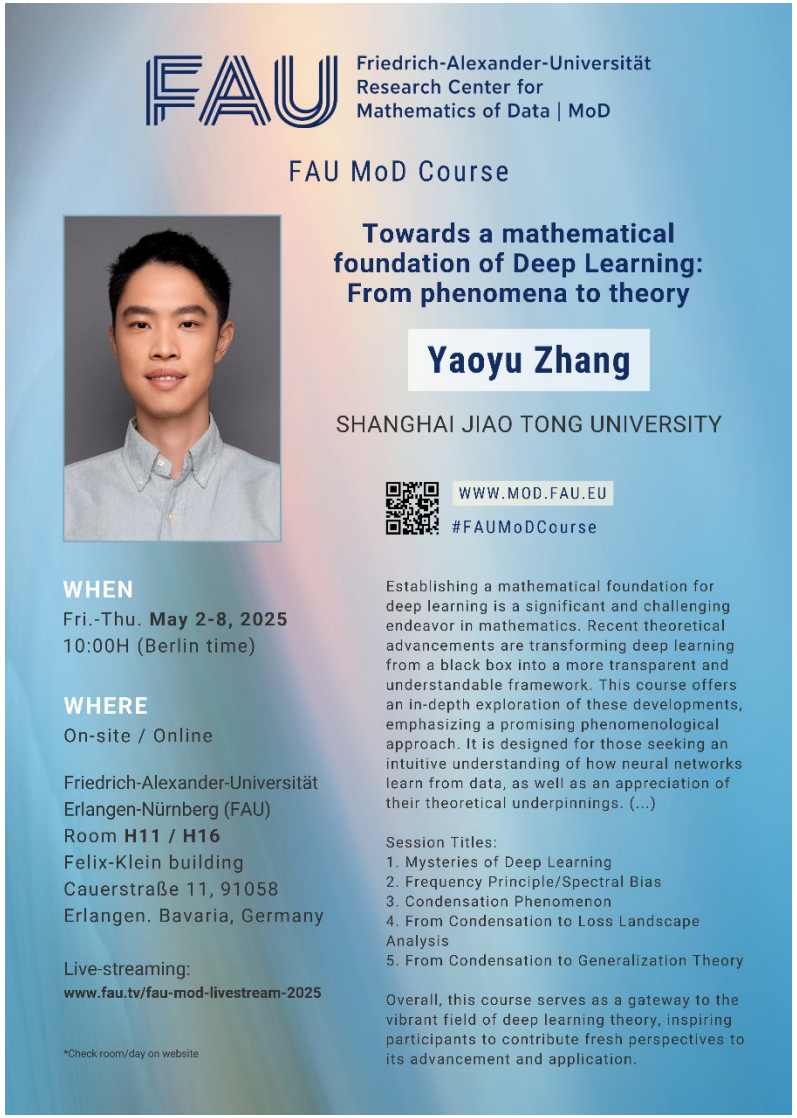


Ways to improve sample efficiency

- **Reduce O_{f^*}:** optimize the architecture (e.g., DNN to CNN)
- **Get close to O_{f^*} :** optimize the hyperparameters (e.g., smaller initialization scale, larger weight decay rate, larger learning rate)

Towards a mathematical foundation of deep learning

Deep learning is no longer a black-box



Towards a Mathematical Foundation of Deep Learning: From Phenomena to Theory

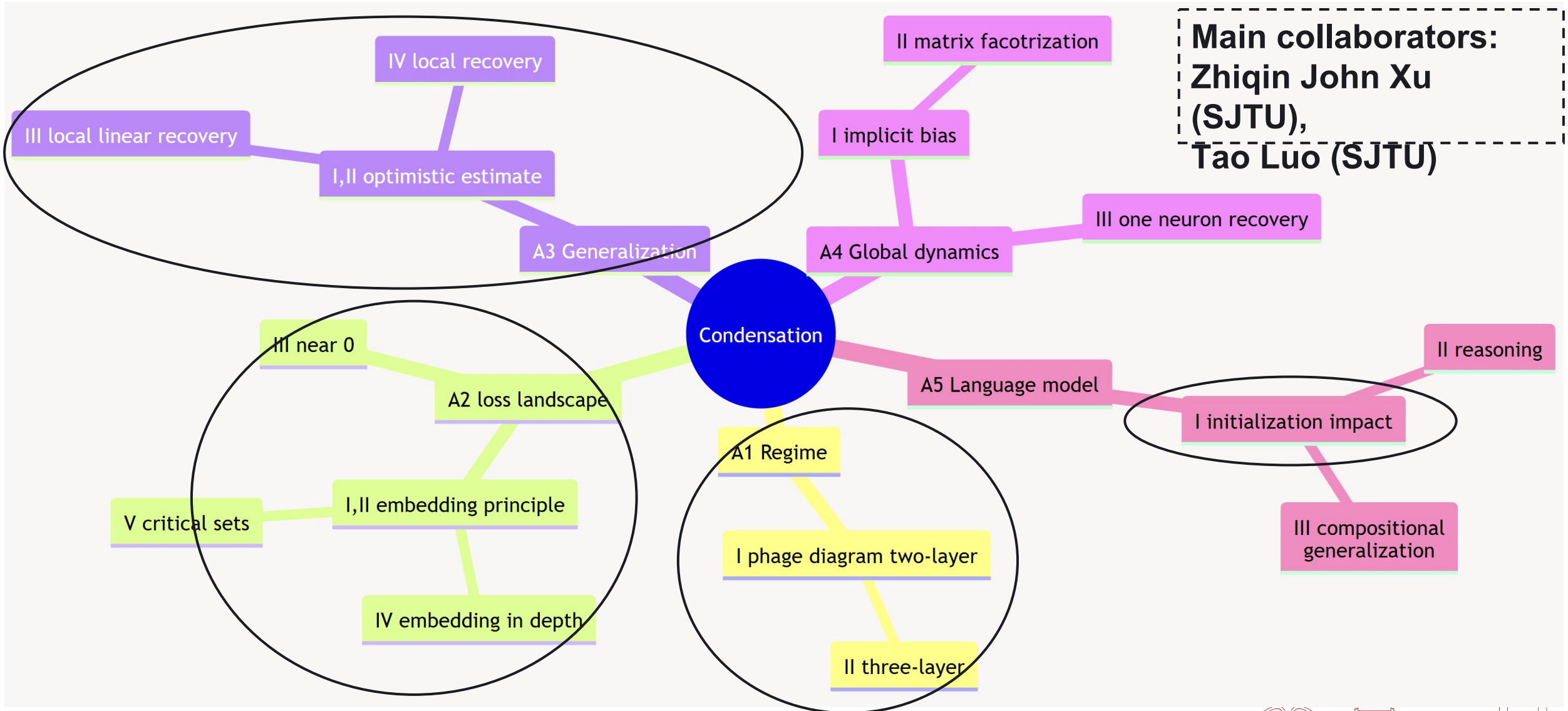
Date

Fri. – Thu. May 2 – 8, 2025

Session Titles

1. Mysteries of Deep Learning
2. Frequency Principle/Spectral Bias
3. Condensation Phenomenon
4. From Condensation to Loss Landscape Analysis
5. From Condensation to Generalization Theory

Condensation



See more works on my personal website: https://yaoyuzhang1.github.io/

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

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.

- **Suspend your framework, dive into the problem**
- **Experience before theorize**

Dawning of the mathematical foundation of AI

Frequency Principle

Condensation

Embedding Principle

Optimistic Estimate

Suspension

Cumulation

Emergence

Introduction to Deep Learning Phenomena

Personal web

PDF

深度学习现象导论

作者:

许志钦　　张耀宇

（笔画数大小排序）

参与学生:

王志伟　白志威　张众望

杭良慨　周章辰　赵佳杰　姚俊杰

（笔画数大小排序）

Thanks!

饮 水 思 源　　爱 国 荣 校