

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**



Friedrich-Alexander-Universität
Research Center for
Mathematics of Data | MoD

FAU MoD Course



Towards a mathematical foundation of Deep Learning: From phenomena to theory

Yaoyu Zhang

SHANGHAI JIAO TONG UNIVERSITY



WWW.MOD.FAU.EU
#FAUMoDCourse

WHEN
Fri.-Thu. May 2-8, 2025
10:00H (Berlin time)

WHERE
On-site / Online

Friedrich-Alexander-Universität
Erlangen-Nürnberg (FAU)
Room H11 / H16
Felix-Klein building
Cauerstraße 11, 91058
Erlangen, Bavaria, Germany

Live-streaming:
www.fau.tv/fau-mod-livestream-2025

*Check room/day on website

Establishing a mathematical foundation for deep learning is a significant and challenging endeavor in mathematics. Recent theoretical advancements are transforming deep learning from a black box into a more transparent and understandable framework. This course offers an in-depth exploration of these developments, emphasizing a promising phenomenological approach. It is designed for those seeking an intuitive understanding of how neural networks learn from data, as well as an appreciation of their theoretical underpinnings. (...)

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

Overall, this course serves as a gateway to the vibrant field of deep learning theory, inspiring participants to contribute fresh perspectives to its advancement and application.



Generalization advantage of condensation

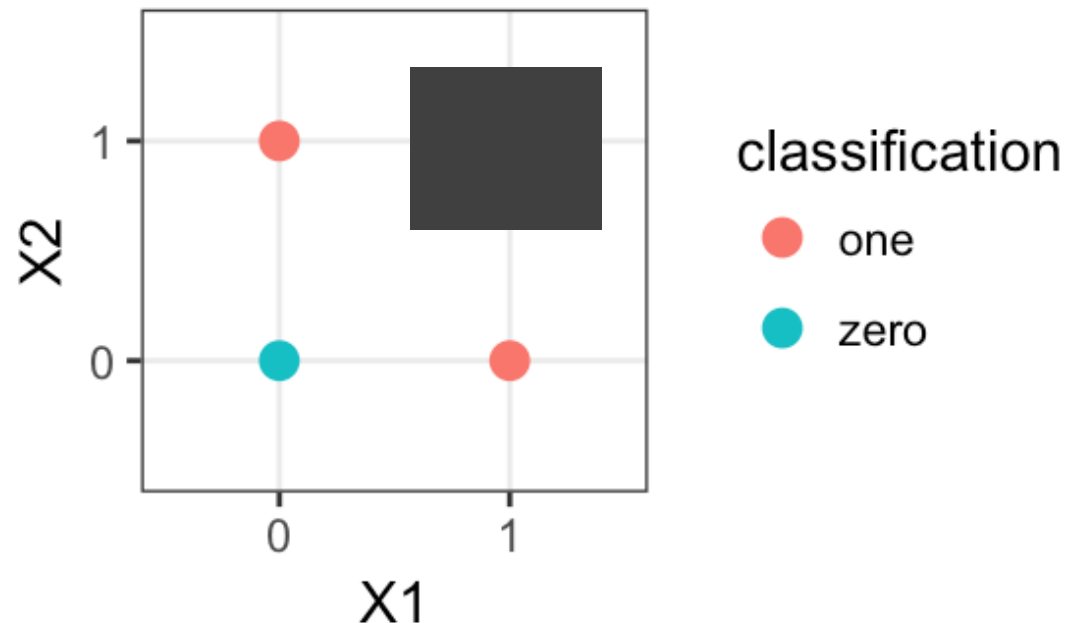
1. Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).
2. Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Optimistic Estimate Uncovers the 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 $f : V \rightarrow S$ 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.



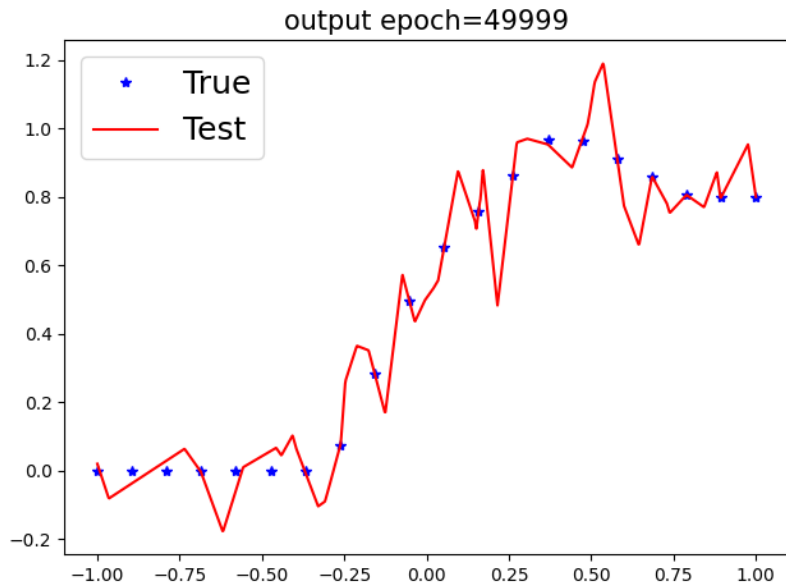
How to infer the missing spot?



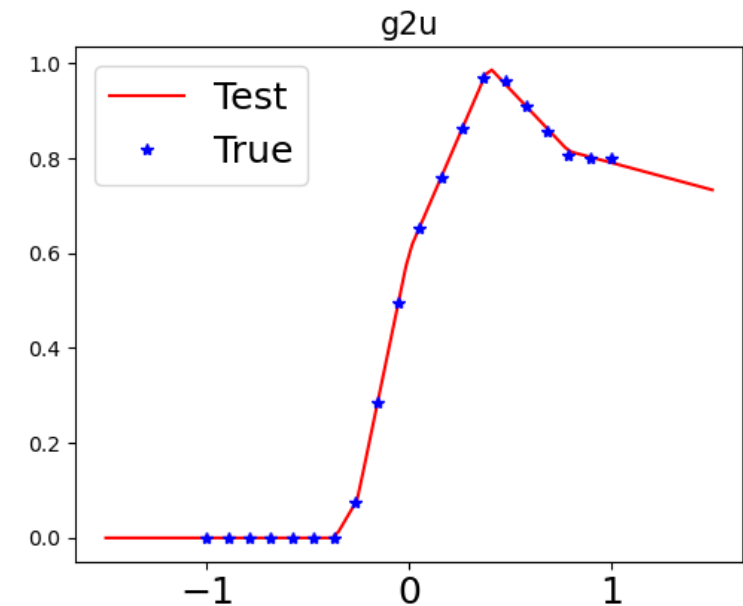


Generalization consequence of condensation

Large initialization
(no condensation)



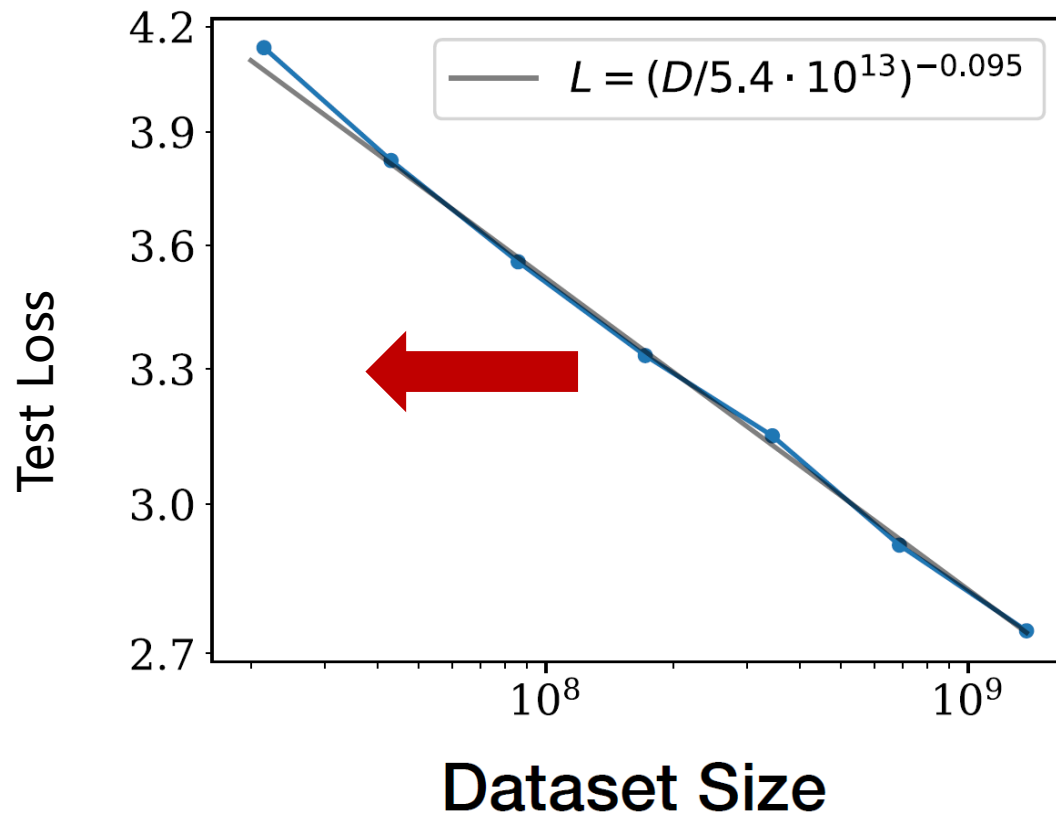
Small initialization
(Strong condensation)





Estimate sample efficiency is important

Sample efficiency: sample size required for certain performance

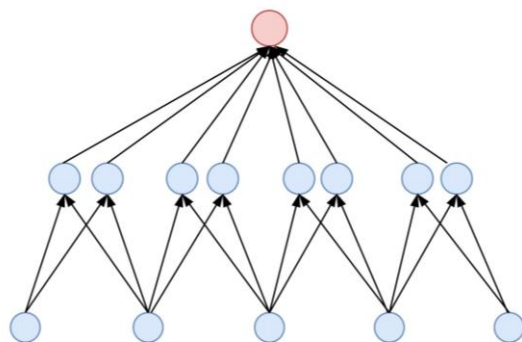
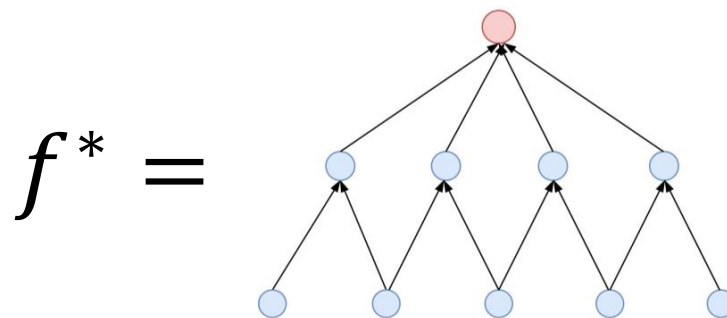




Intuition: condensation improves sample efficiency

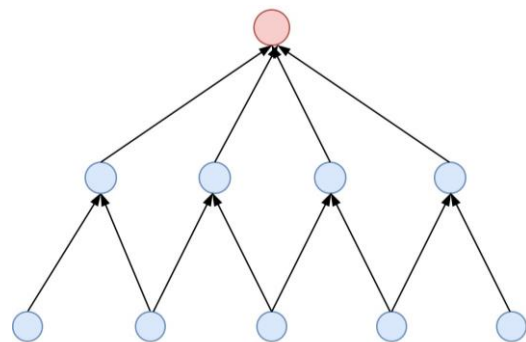


How many samples are required to recover f^* by NN_{wide} ?



NN_{wide}

condense



NN_{narr} (equiv)

parameter count

12

optimistic
sample size

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

566 m

95 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:

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

Model:

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

Optimization:

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

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

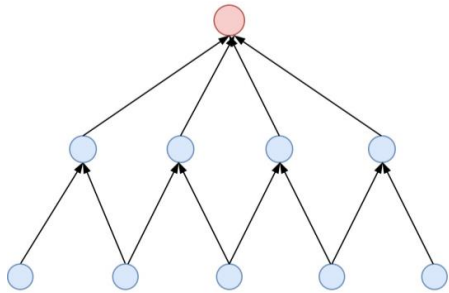
High dimensional, nonlinear, nonconvex!



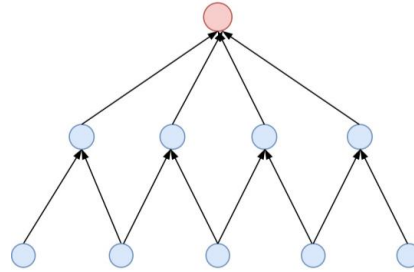


Failure of classic estimate

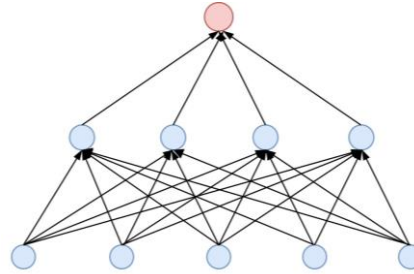
$f^* =$



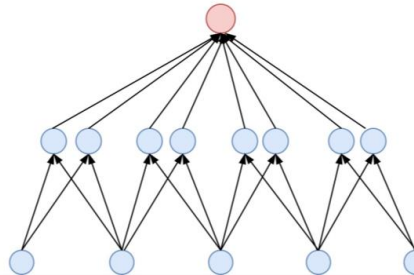
NN_A



NN_B



NN_C



classic estimate

12

24

~~24~~ 12

reduce
sample
efficiency

optimistic!

Classic estimate:

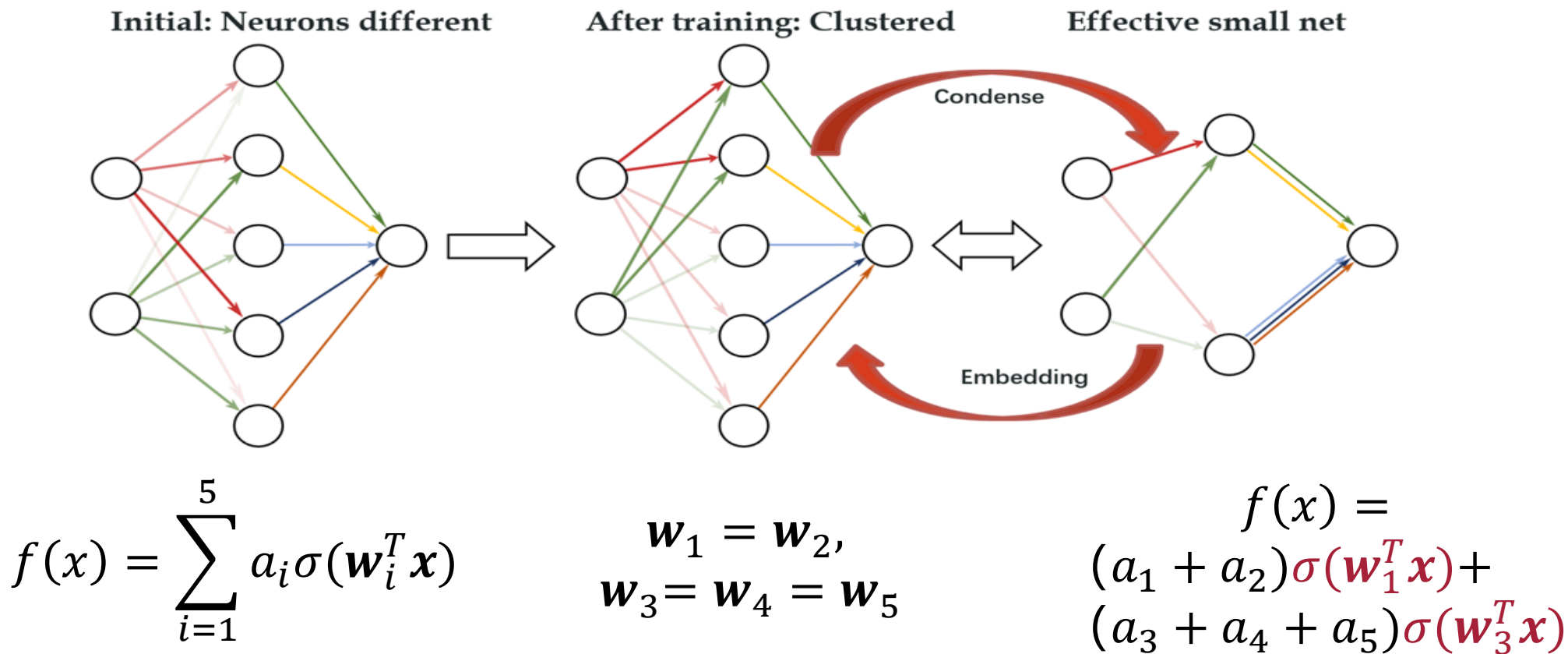
sample size for fitting = parameter size





Condensation—the clustering behavior of neurons

Condensation improves sample efficiency



Initial: **random**

Training: **condense**

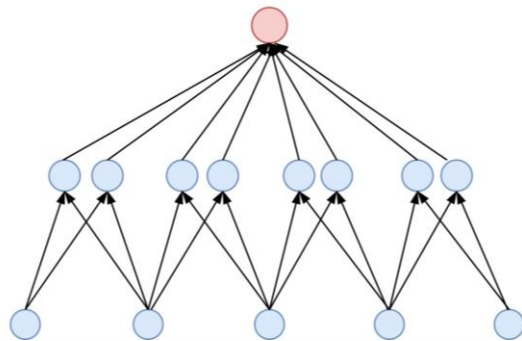
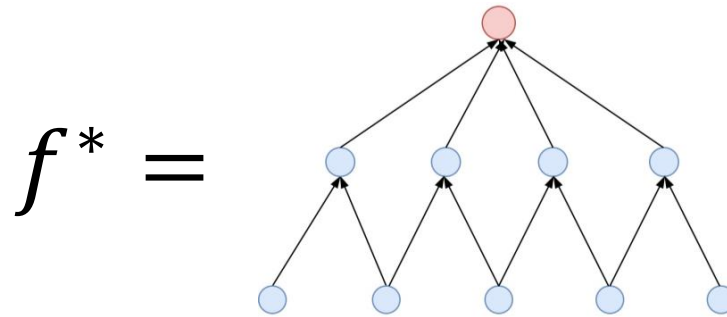
Effect: **equiv to small net**





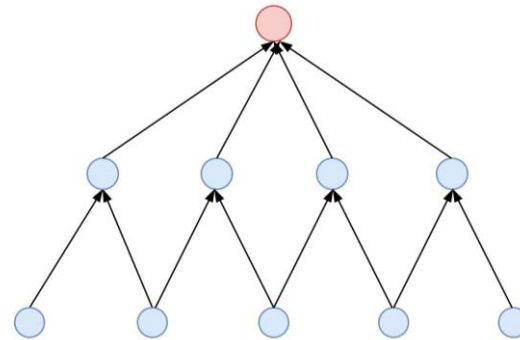
Condensation improves sample efficiency

How many samples are required to fit f^* ?



NN_C

condense



NN_A (equiv)

classic
estimate

optimistic
estimate

12





Model:

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

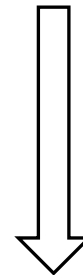
Model rank:

$$\begin{aligned} r_{\theta} &:= \text{rank } DF(\theta) = \dim \text{Im}(DF(\theta)) \\ &= \text{dim span } \{\partial_{\theta_i} F(\theta)(\cdot)\}_{i=1}^M \end{aligned}$$

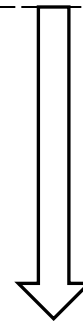
Intuition: effective degrees of freedom at θ

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

Stronger condensation



Lower model rank



Higher sample efficiency





Condensation means lower model rank

Example:

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

Model rank:

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

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

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

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

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





Model:

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

Model rank:

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

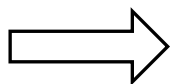
Optimistic sample size ($f^* \in \mathcal{F}$) :

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

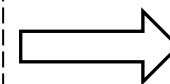
$F^{-1}(f^*)$: 目标集(零泛化误差)

Intuitive procedure:

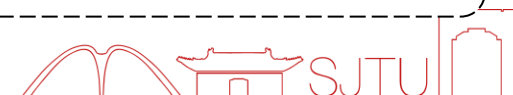
Given target f^*



find $\boldsymbol{\theta}^* \in F^{-1}(f^*)$ with minimum rank

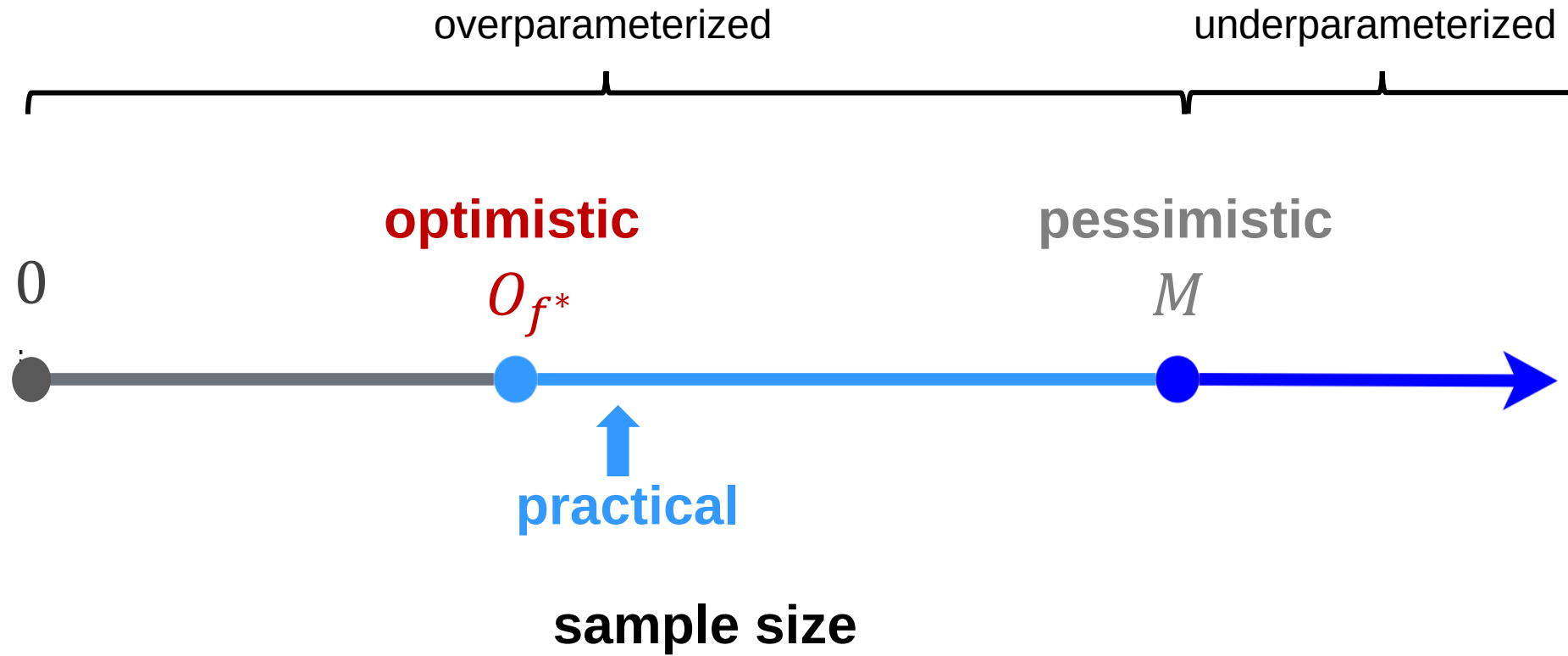


$$O_{f^*} = r_{\boldsymbol{\theta}^*}$$





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 $f_{\theta}(\mathbf{x}) = \sum_{i=1}^m a_i \tanh(\mathbf{w}_i^T \mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, $\theta = (a_i, \mathbf{w}_i)_{i=1}^m$, for any target function $f^* \in \mathcal{F}_k^{\text{NN}} \setminus \mathcal{F}_{k-1}^{\text{NN}}$ with $0 \leq k \leq m$, the optimistic sample size*

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

vs.

$$m(d+1)$$

optimistic

$$O_{f^*} = 21$$

pessimistic

$$M = 63$$

3x



practical
(well-tuned)





Optimistic sample size estimate vs practice



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

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

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 $O_{f_{\theta_{\text{wide}}}}(f^*) \leq O_{f_{\theta_{\text{narr}}}}(f^*) \leq M_{\text{narr}}$.*



wider network is sample efficient



Key tool for optimistic estimation: critical embedding



Criticality of optimistic sample size

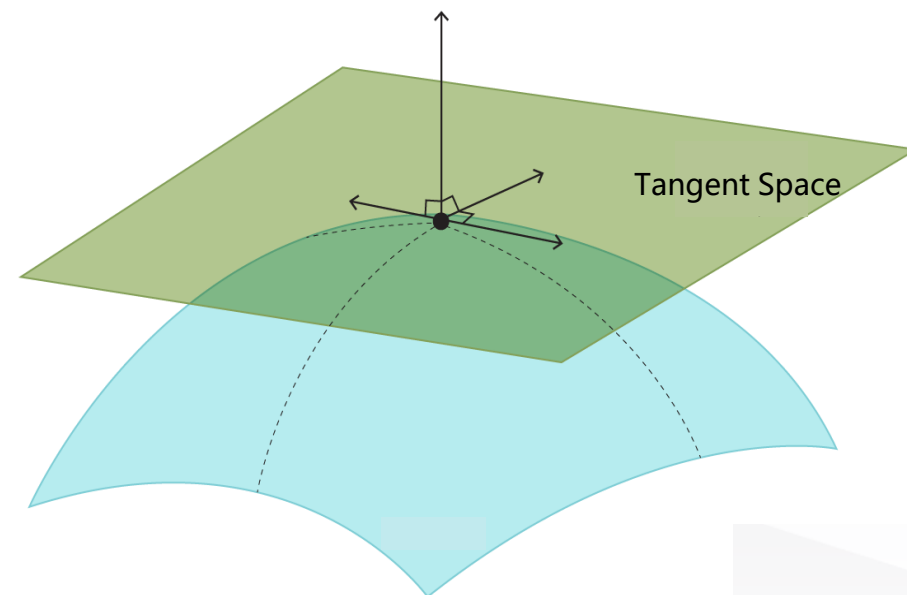
Theorem (phase transition of LLR-guarantee at a target point). For any $\theta' \in \mathcal{M}_{f^*}$

- 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 exists 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^*

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

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

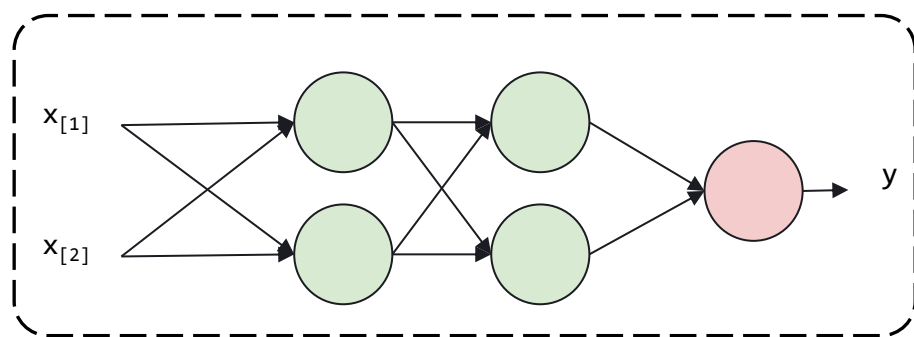




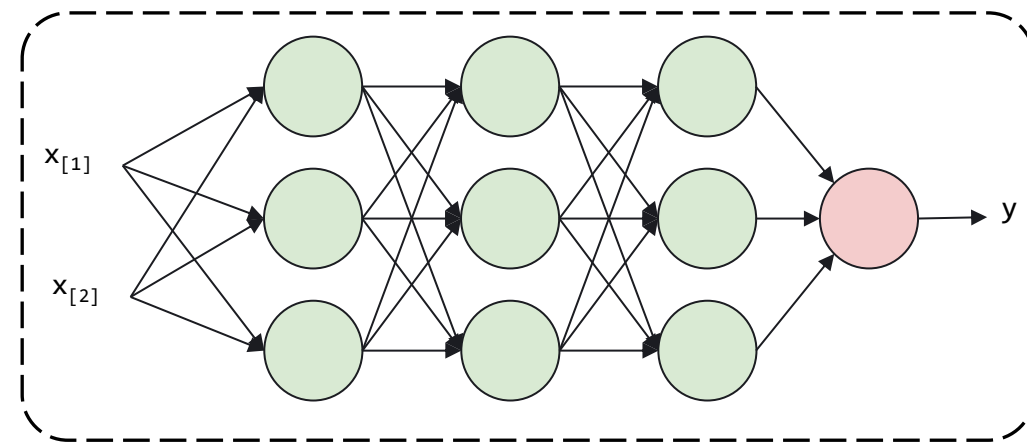
Optimistic sample size estimation for general DNNs

Key tool: critical Mapping

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



critical Mapping
 $\mathcal{P}: \mathbb{R}^{M_A} \rightarrow \mathbb{R}^{M_B}$



critical Mapping:

- (i) Output Preserving: $f_{\theta} = g_{\mathcal{P}(\theta)}$
- (ii) Criticality Preserving: if $\nabla_{\theta} R_S(f_{\theta}) = 0$, then $\nabla_{\theta} R_S(g_{\mathcal{P}(\theta)}) = 0$, 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 θ_{narr}^c of the narrower NN, $\theta_{\text{wide}}^c := \mathcal{T}(\theta_{\text{narr}}^c)$ is still a critical point of the K -neuron wider NN with the same output function, i.e., $f_{\theta_{\text{narr}}^c} = f_{\theta_{\text{wide}}^c}$.*

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

- (i) $f_{\theta'_c}(x) = f_{\theta_c}(x)$ for $x \in S_x$;
- (ii) $\nabla_{\theta'} R_S(\theta'_c) = \mathbf{0}$.

Embedding Principle (width/depth):

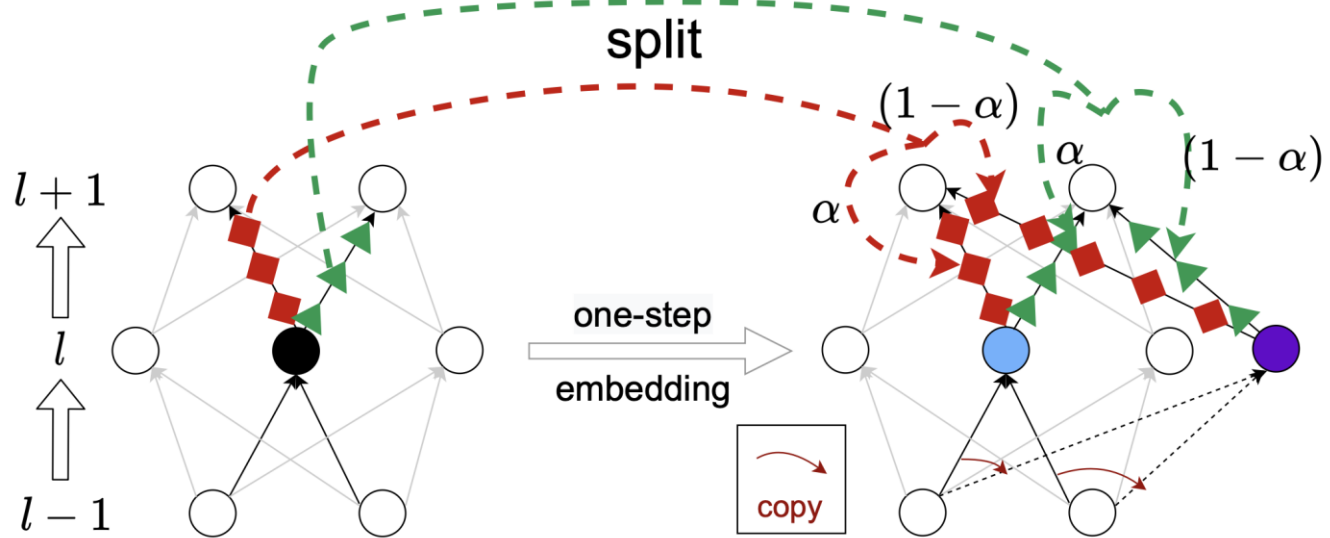
The loss landscape of a neural network “contains” all the critical points of narrower/shallower networks.



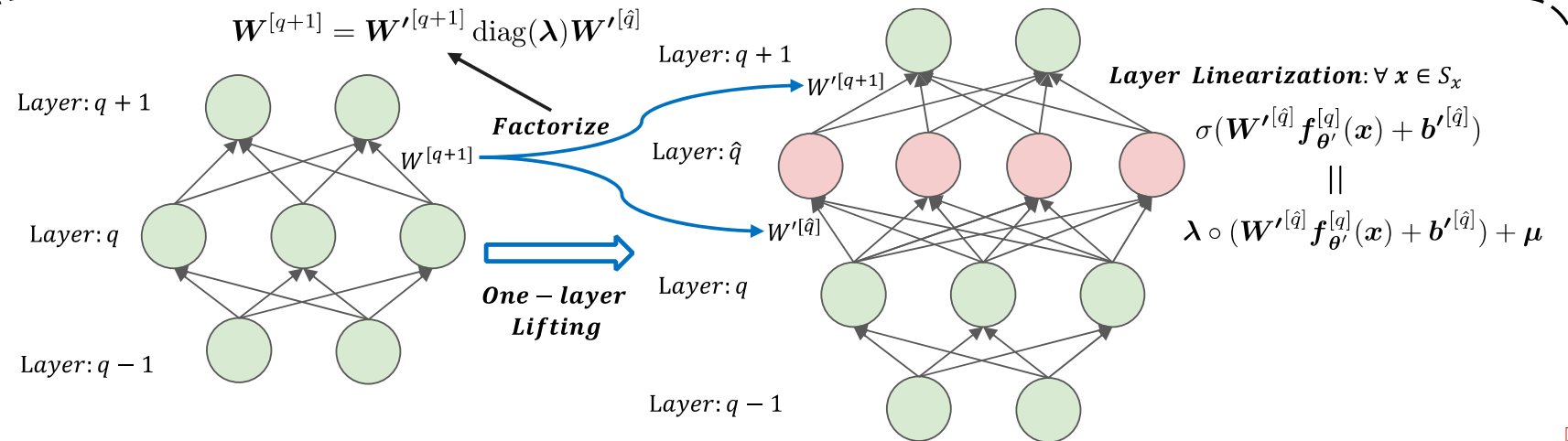


Embedding principle -> sample efficiency preserving

Embedding Principle (width)



Embedding Principle (depth)

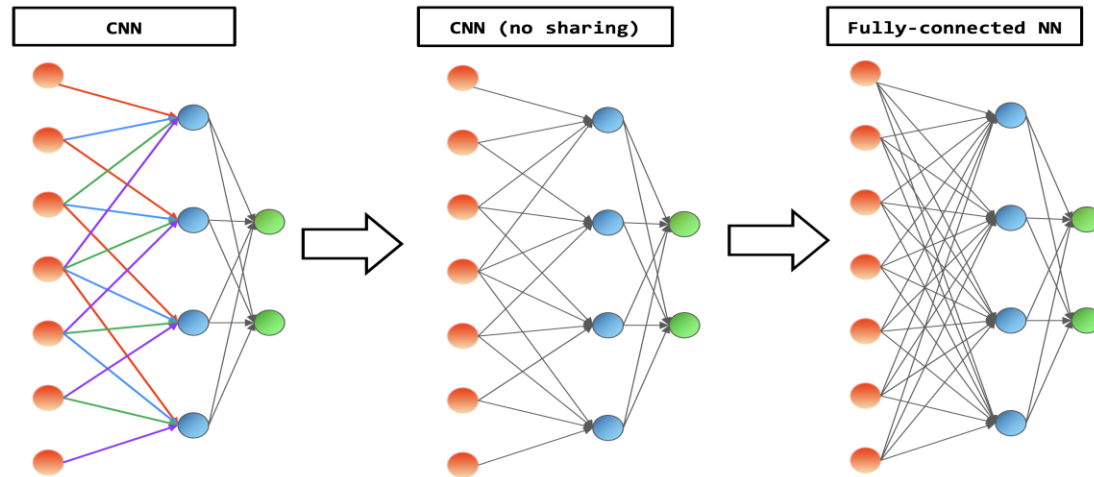


Impact of connection





Impact of connection



k intrinsic width
 $s \times s$ conv ker size
 $d \times d$ input dim

f^*	CNN	CNN (no sharing)	Fully-connected NN
$\{0\}$	0	0	0
$\mathcal{F}_1^{\text{CNN}} \setminus \{0\}$	$s^2 + (d+1-s)^2$	$(s^2 + 1) ((d+1-s)^2 - m_n)$	$(d^2 + 1) ((d+1-s)^2 - m_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{F}_k^{\text{CNN}} \setminus \mathcal{F}_{k-1}^{\text{CNN}}$	$k(s^2 + (d+1-s)^2)$	$(s^2 + 1) (k(d+1-s)^2 - m_n)$	$(d^2 + 1) (k(d+1-s)^2 - m_n)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{F}_m^{\text{CNN}} \setminus \mathcal{F}_{m-1}^{\text{CNN}}$	$m(s^2 + (d+1-s)^2)$	$(s^2 + 1) (m(d+1-s)^2 - m_n)$	$(d^2 + 1) (m(d+1-s)^2 - m_n)$

more connection lower
sample efficiency





Adding (unnecessary) connections reduces sample efficiency

MNIST k -kernel, kernel size: 3×3

1000X
worse!



➤ CNN: $685k$

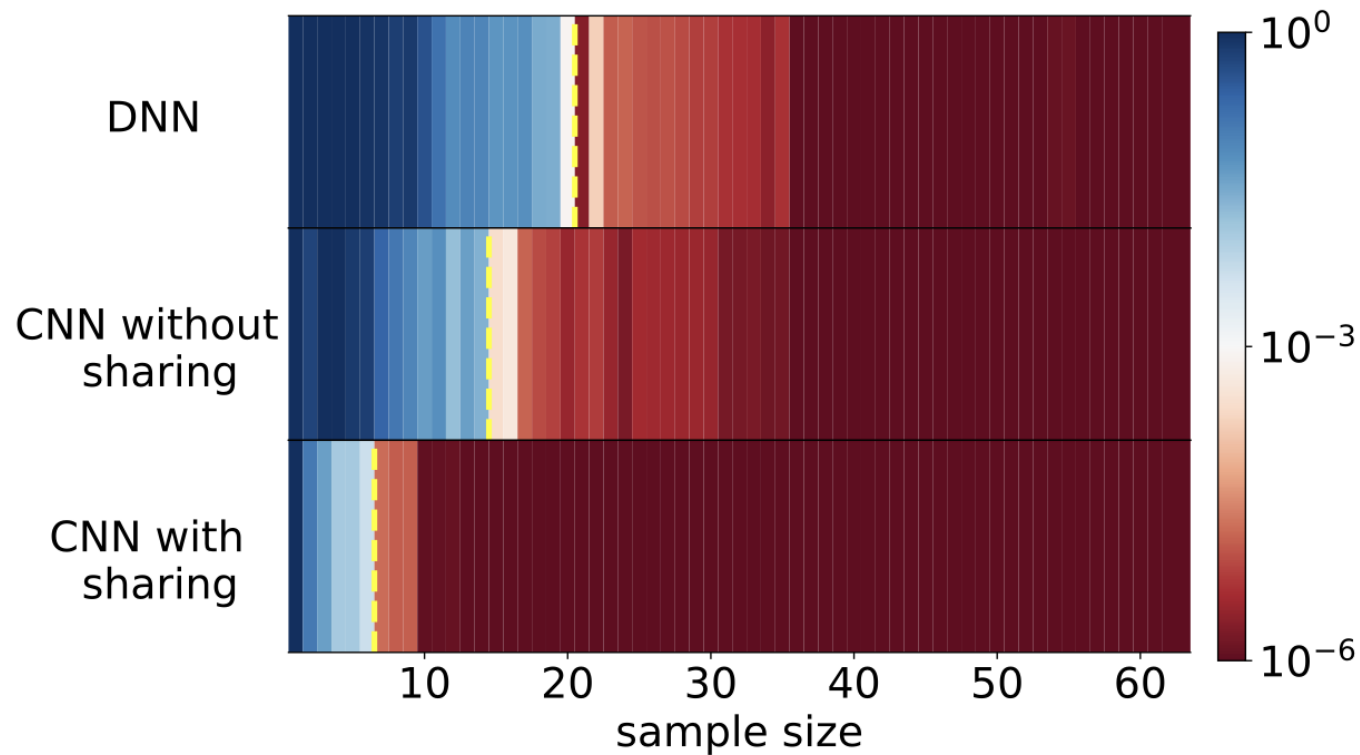
➤ CNN (no sharing) : $6760k$

➤ FNN: $530660k$



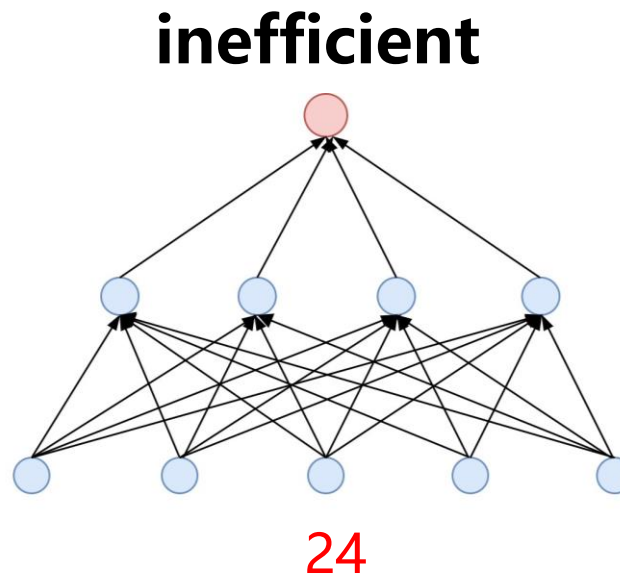
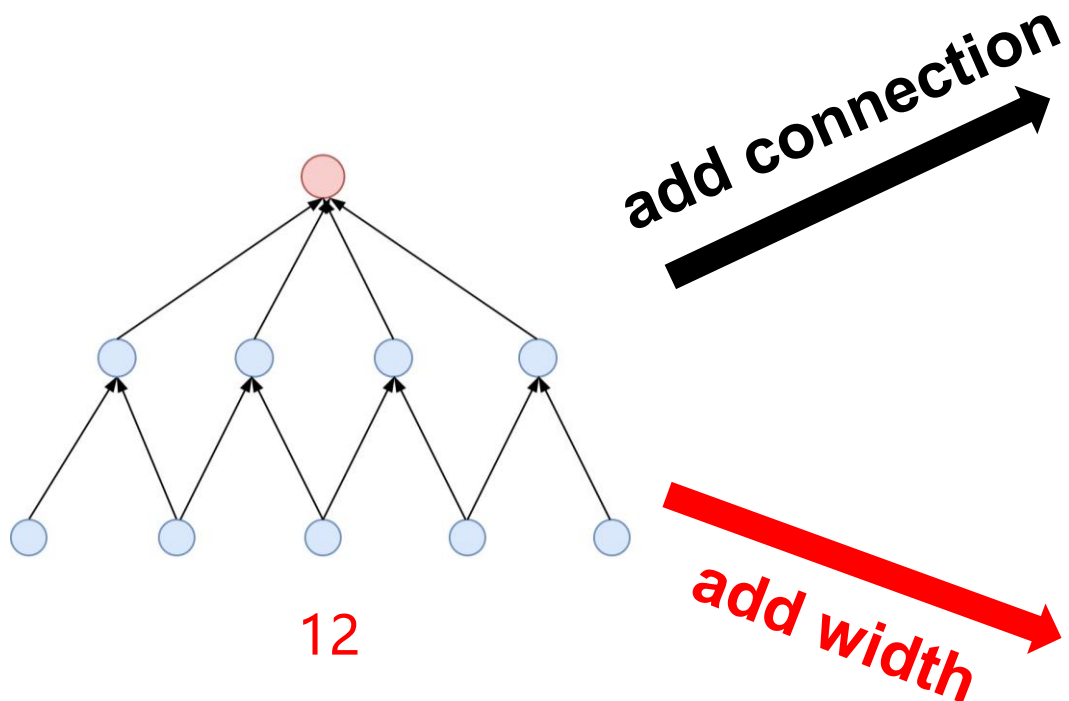
Optimistic estimate vs. experiments

Experiment: adding connection reduces sample efficiency

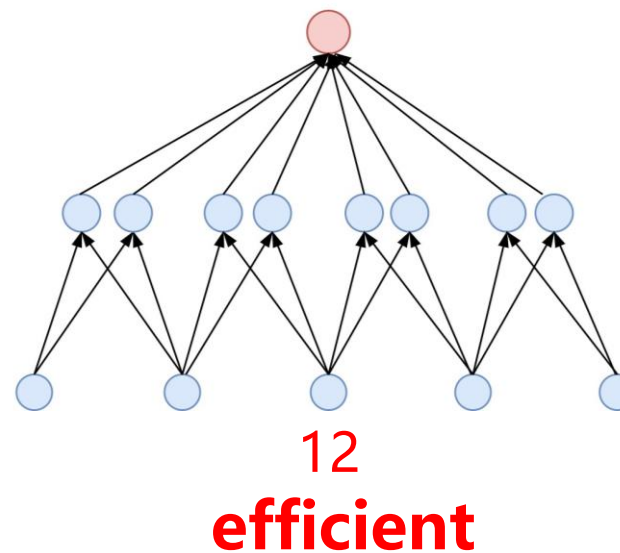




Specialty of DNN models via optimistic estimate



Sample inefficient:
Adding (unnecessary) connections worsens generalization



Sample efficient:
Increasing width doesn't harm generalization



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



Model:

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

Optimistic sample size:

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

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

Conclusion:

Linear models cannot recover targets under overparameterization.





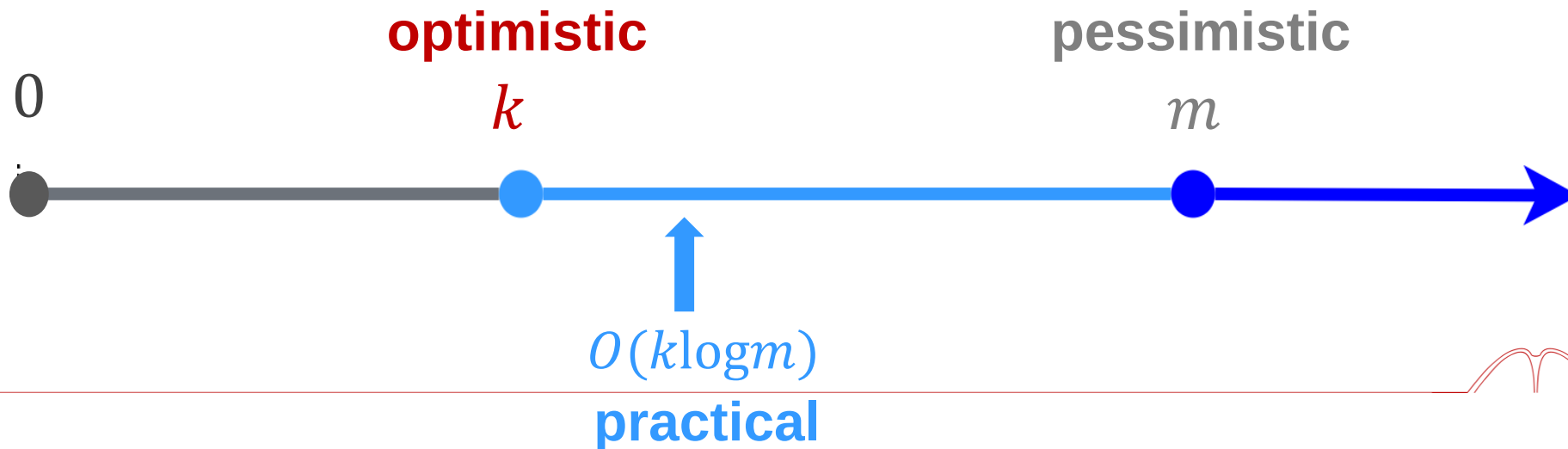
Model:

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

Optimistic sample size:

For a ***k*-sparse** function f^* ,

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





Deep models

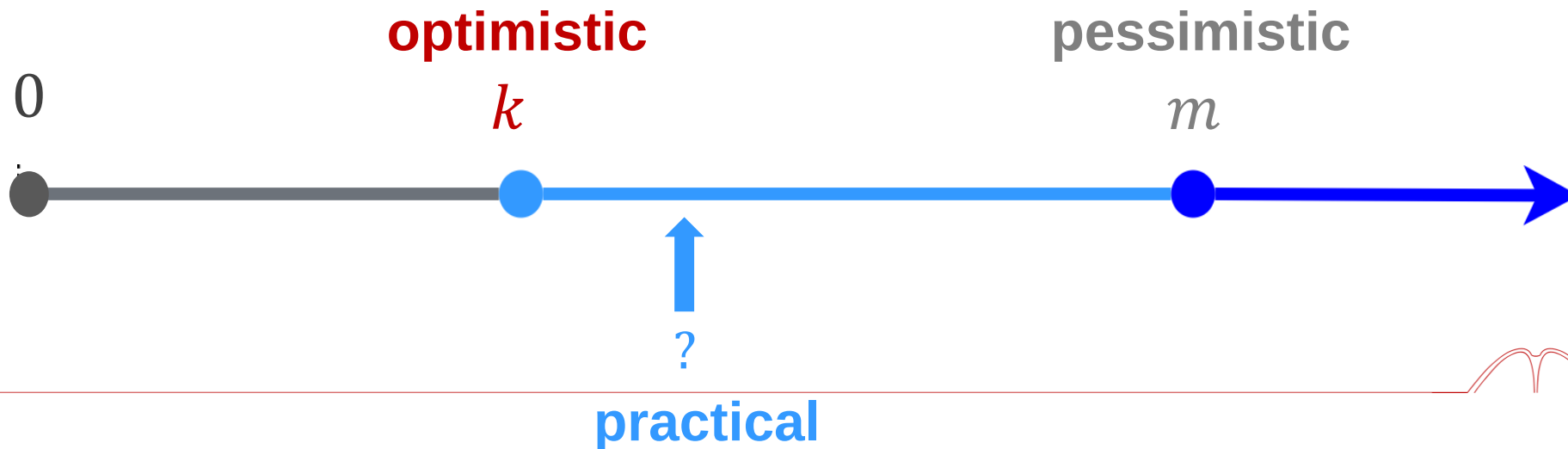
Model:

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

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

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



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:

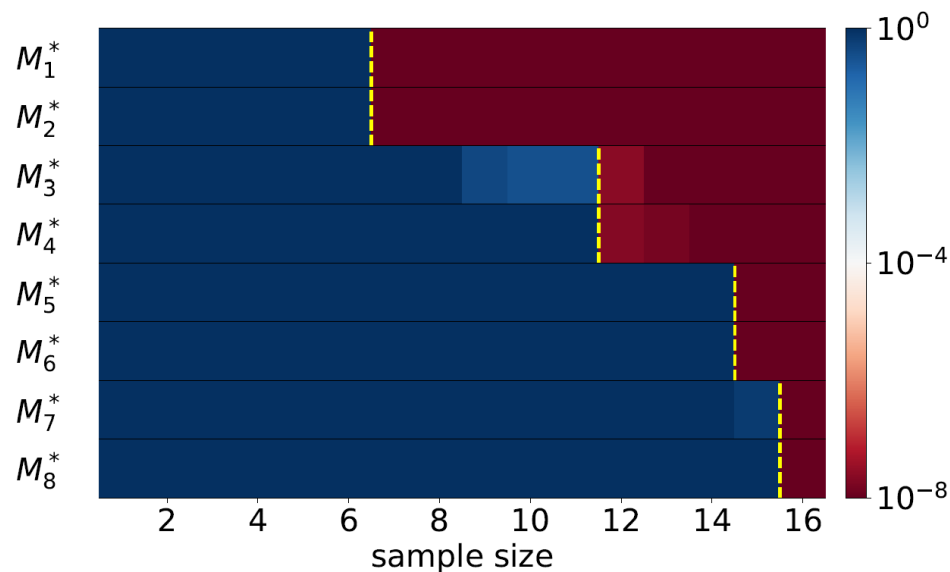
model	$f_{\theta} = AB, \theta = (A, B), A, B \in \mathbb{R}^{d \times d}$	
$\text{rank}_{f_{\theta}}(\mathbf{f}^*)$	$\mathbf{f}^*$	$\arg \min_{\theta' \in \mathcal{M}_{f^*}} \text{rank}_{f_{\theta}}(\theta')$
0	$\mathbf{0}_{d \times d}$	$\mathbf{A} = \mathbf{B} = \mathbf{0}_{d \times d}$
$2d - 1$	$\text{rank}(\mathbf{f}^*)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



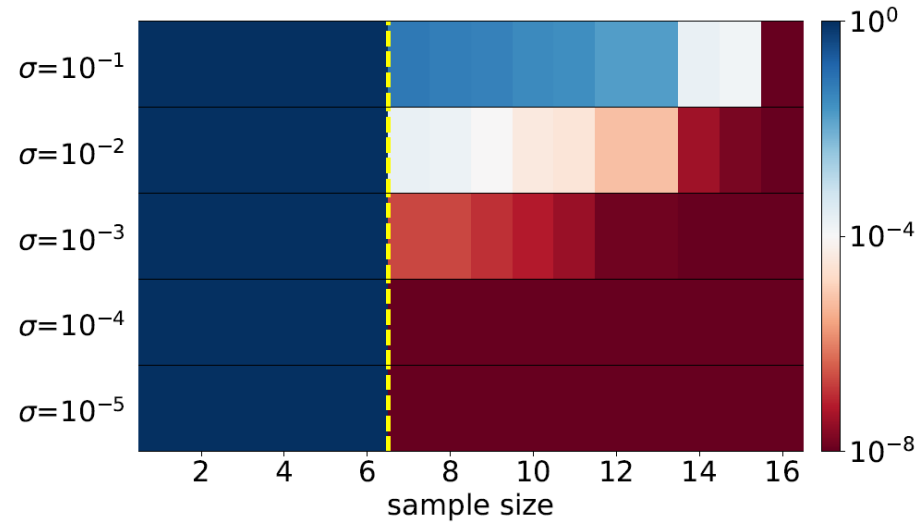
Matrix factorization model:

$$\mathbf{f}_\theta = \mathbf{A}\mathbf{B}, \quad \theta = (\mathbf{A}, \mathbf{B}), \quad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{4 \times 4}$$

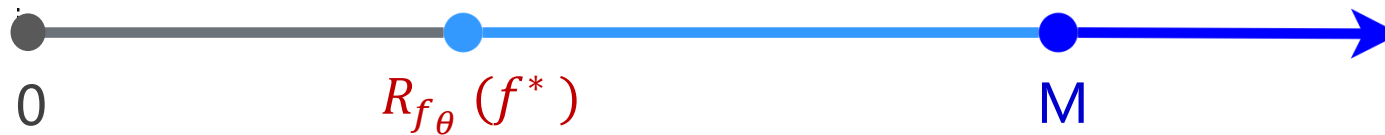
$O_{\mathbf{f}^*}$	target matrix
0	$\mathbf{0}$
7	$\text{rank}(\mathbf{f}^*) = 1$
12	$\text{rank}(\mathbf{f}^*) = 2$
15	$\text{rank}(\mathbf{f}^*) = 3$
16	$\text{rank}(\mathbf{f}^*) = 4$



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: 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.

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

Permutation symmetry $\rightarrow$ invariant manifolds (equiv to smaller network)

$\rightarrow$ 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!

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





KAN (Kolmogorov–Arnold Networks)

KAN:

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

width: permutation symmetric

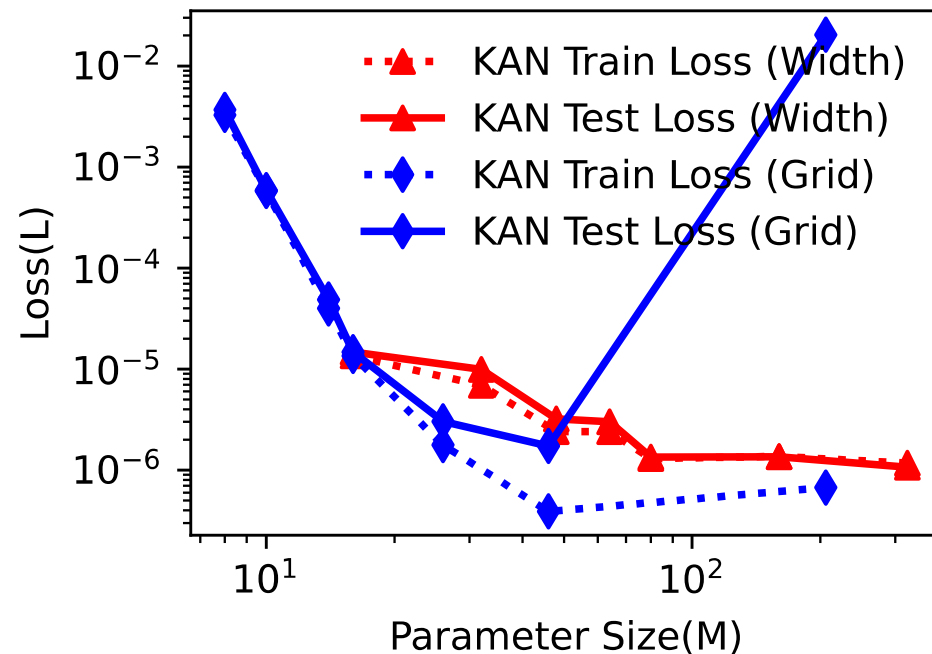
Increase m preserves sample efficiency

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

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.

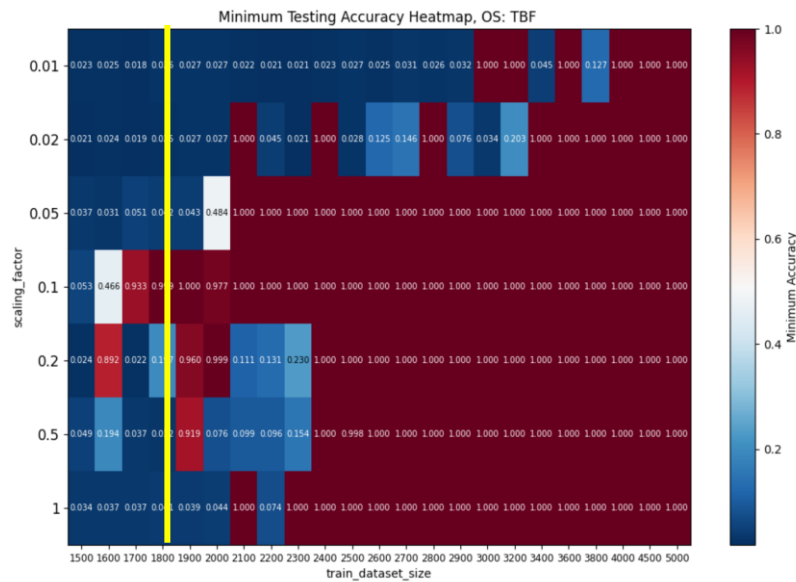
Single anchor	Two anchors	Input data examples	Target
1 : +5	1 1 : +10	Noisy tokens	
2 : +1	1 2 : +6	55 46 32 52 28 1 1 34 33	38
3 : -2	...	20 95 43 3 1 44 34 76 32	46
4 : -8	3 4 : -10	...	...
	4 4 : -16		
		28 53 44 78 32 62 3 4 44	52
		77 43 23 63 89 33 52 4 3	?

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

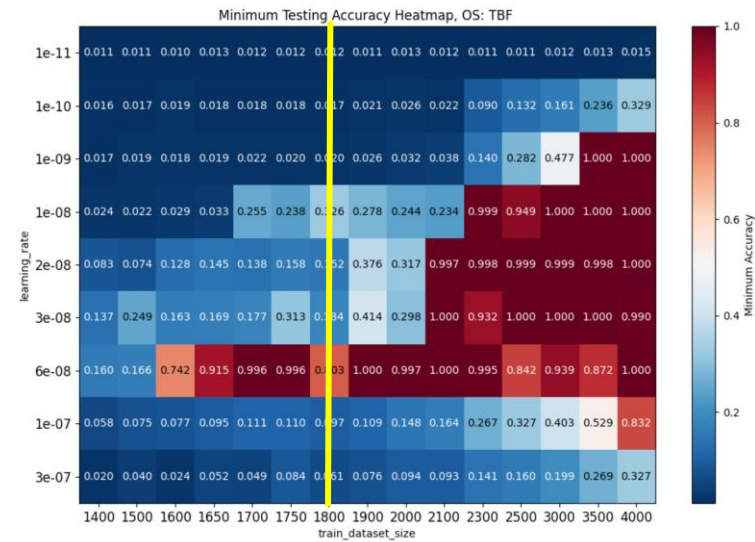


Empirical estimation of optimistic sample size

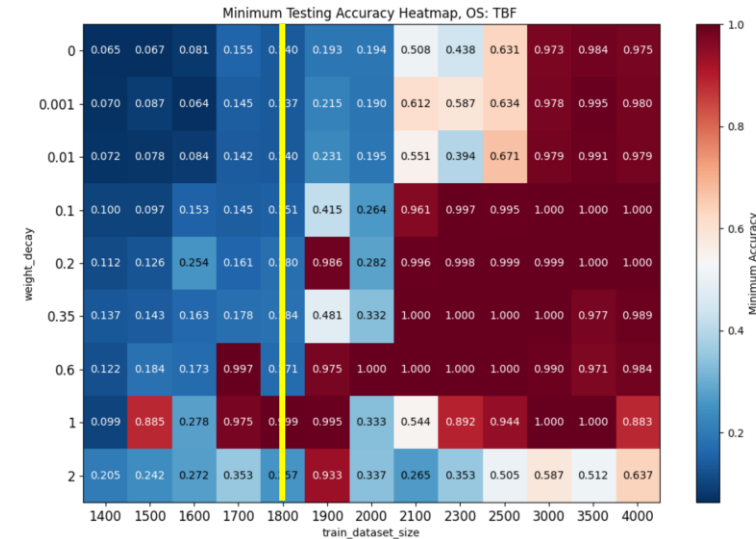
Identification of best tuning of hyperparameters



Initialization scale



Learning rate



Weight decay

Empirical estimate of optimistic sample size ≈ 1800





Application——architecture analysis

➤ 2V2 Random

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

➤ 4V4 Composite

Anc hor	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

4

Inference
complexity

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?**

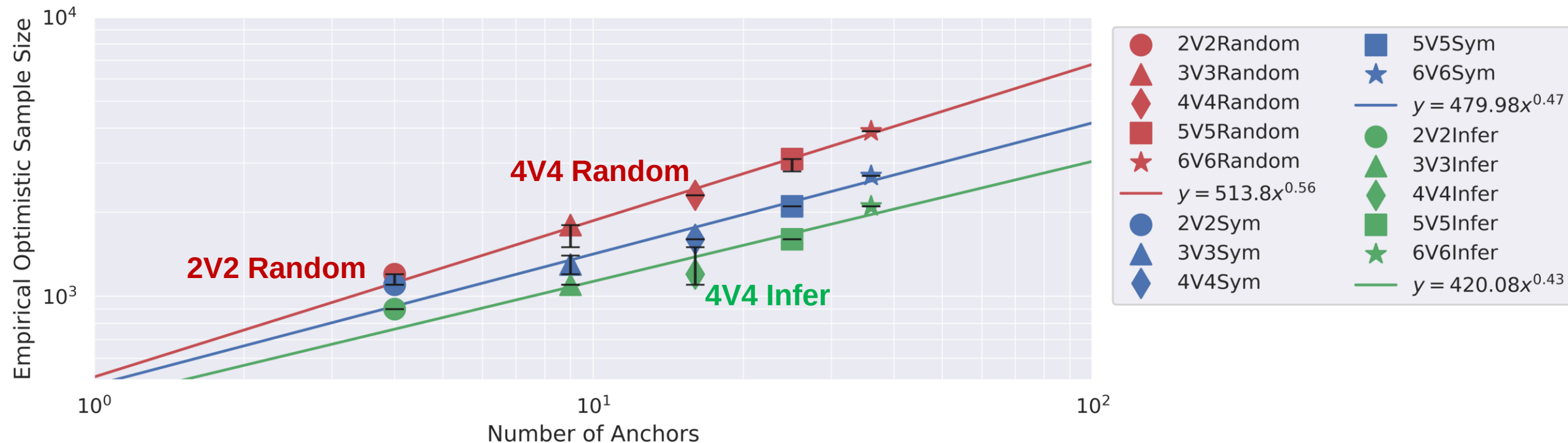




(Optimistically) Memorization? No!

Transformer:

Memorization complexity poorly predicts optimistic sample size



optimistic sample size vs. memorization complexity

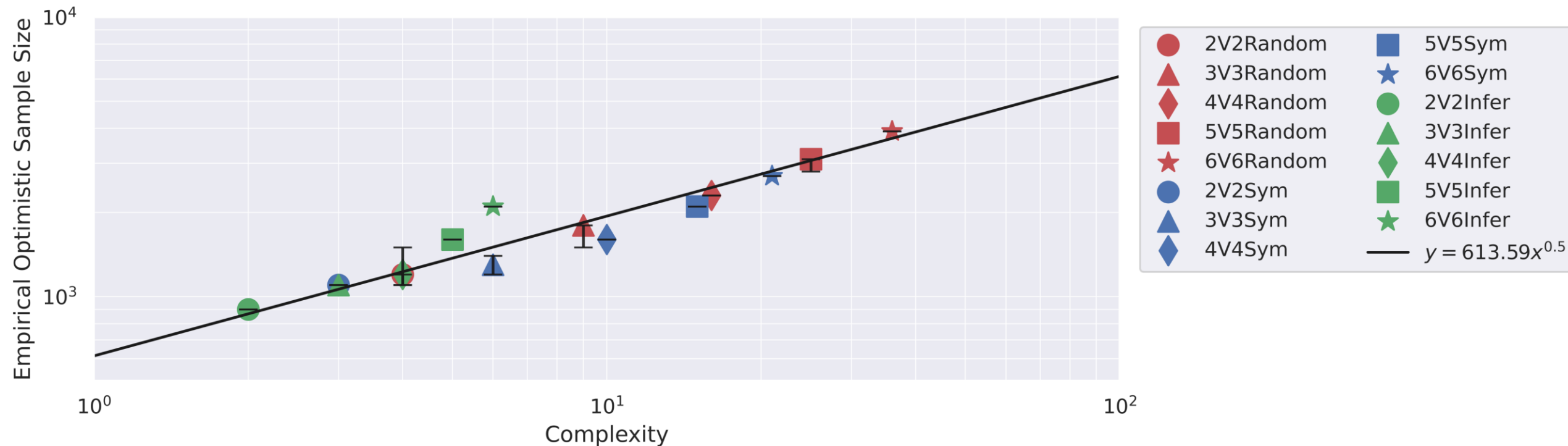




(Optimistically) Inference? Yes!

Transformer:

Inference complexity well predicts optimistic sample size

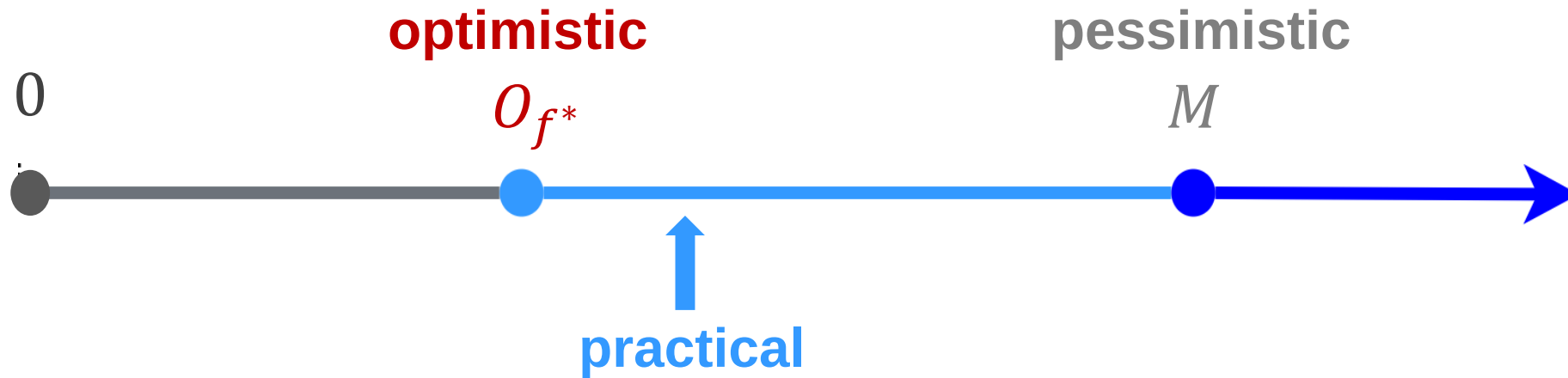


optimistic sample size vs. inference complexity





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



Friedrich-Alexander-Universität
Research Center for
Mathematics of Data | MoD

FAU MoD Course



**Towards a mathematical foundation of Deep Learning:
From phenomena to theory**

Yaoyu Zhang

SHANGHAI JIAO TONG UNIVERSITY



WWW.MOD.FAU.EU
#FAUMoDCourse

WHEN
Fri.-Thu. May 2-8, 2025
10:00H (Berlin time)

WHERE
On-site / Online

Friedrich-Alexander-Universität
Erlangen-Nürnberg (FAU)
Room H11 / H16
Felix-Klein building
Cauerstraße 11, 91058
Erlangen, Bavaria, Germany

Live-streaming:
www.fau.tv/fau-mod-livestream-2025

*Check room/day on website

Establishing a mathematical foundation for deep learning is a significant and challenging endeavor in mathematics. Recent theoretical advancements are transforming deep learning from a black box into a more transparent and understandable framework. This course offers an in-depth exploration of these developments, emphasizing a promising phenomenological approach. It is designed for those seeking an intuitive understanding of how neural networks learn from data, as well as an appreciation of their theoretical underpinnings. (...)

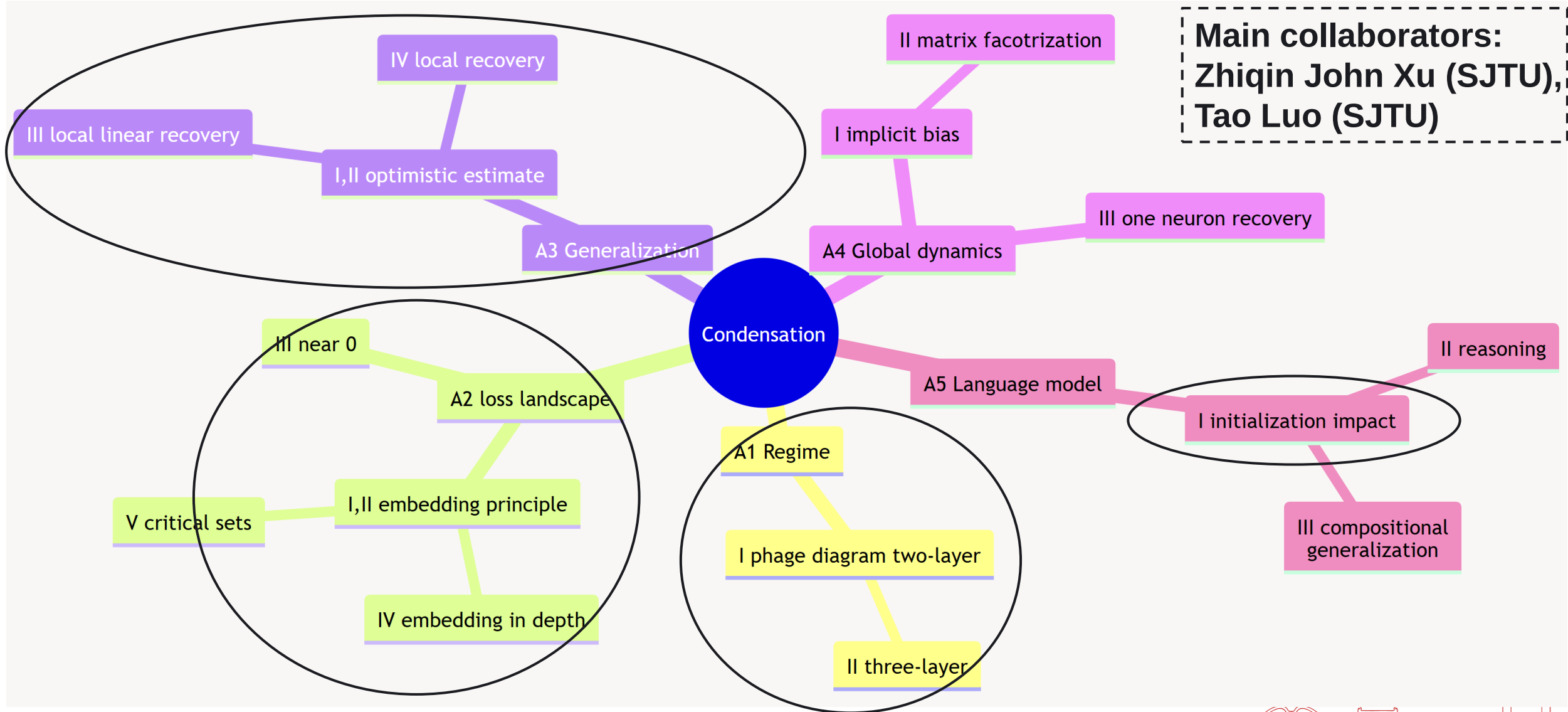
Session Titles:
1. Mysteries of Deep Learning
2. Frequency Principle/Spectral Bias
3. Condensation Phenomenon
4. From Condensation to Loss Landscape Analysis
5. From Condensation to Generalization Theory

Overall, this course serves as a gateway to the vibrant field of deep learning theory, inspiring participants to contribute fresh perspectives to its advancement and application.





Condensation





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



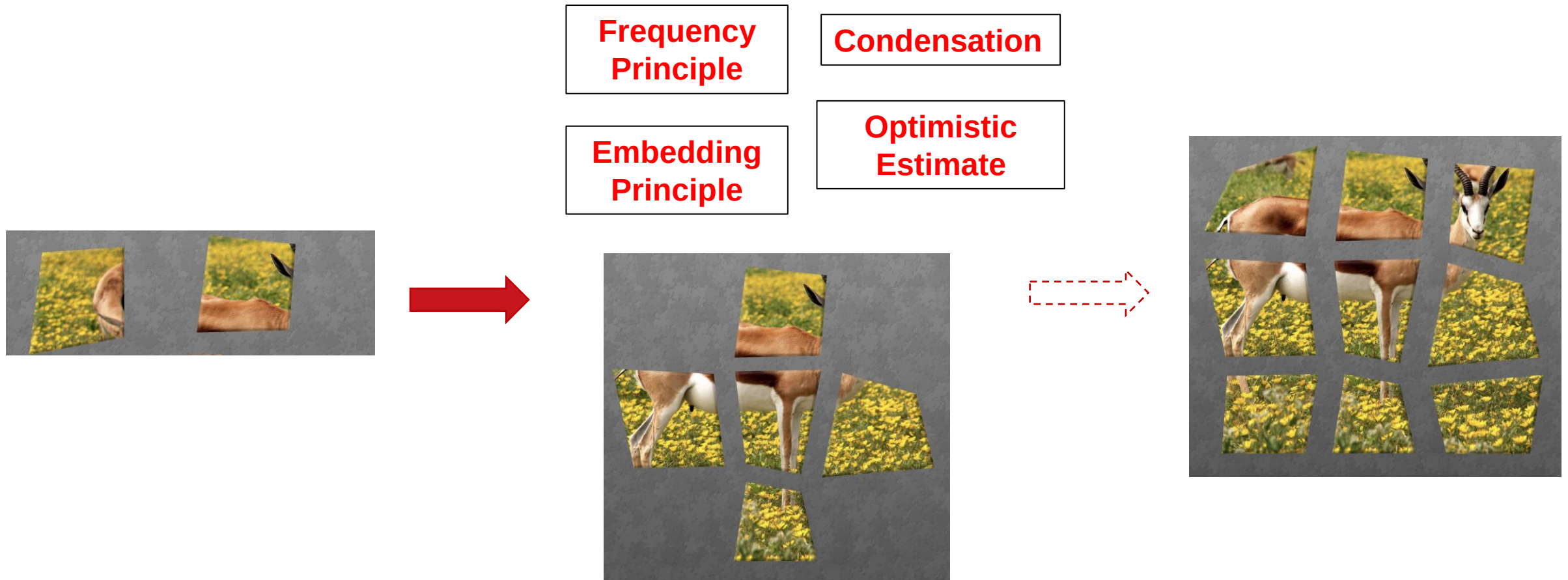
Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

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





Dawning of the mathematical foundation of AI



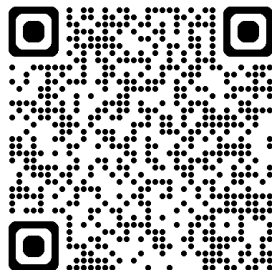


Personal web



深度学习现象导论

PDF



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