The Condensation Phenomenon of Deep Neural Networks

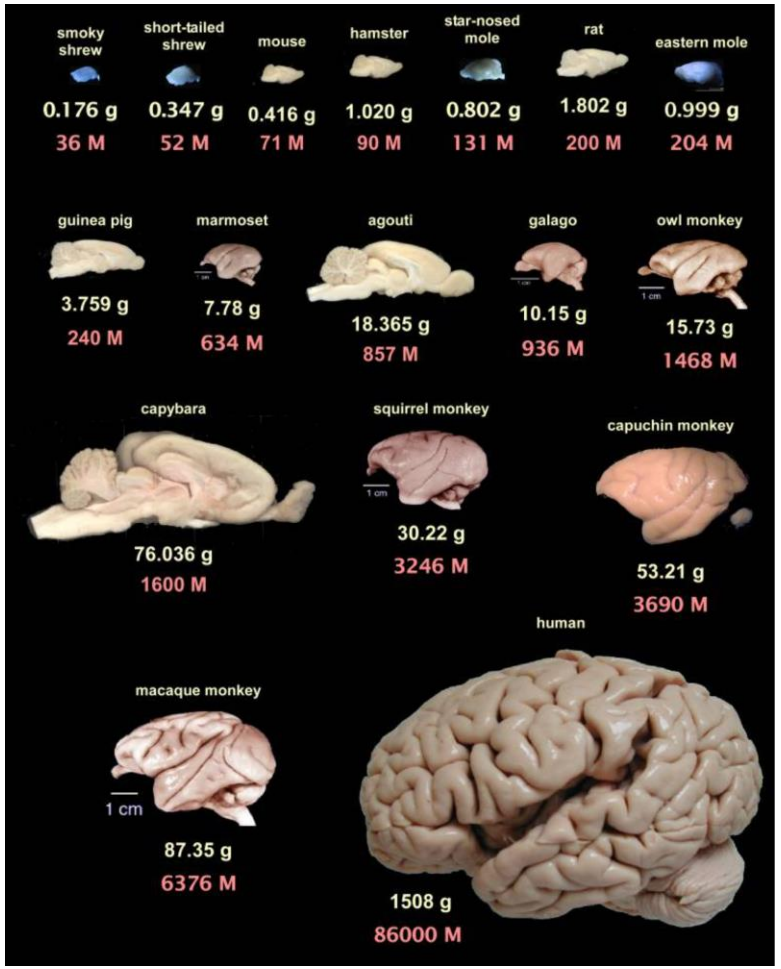
Yaoyu Zhang

Institute of Natural Sciences & School of Mathematical Sciences

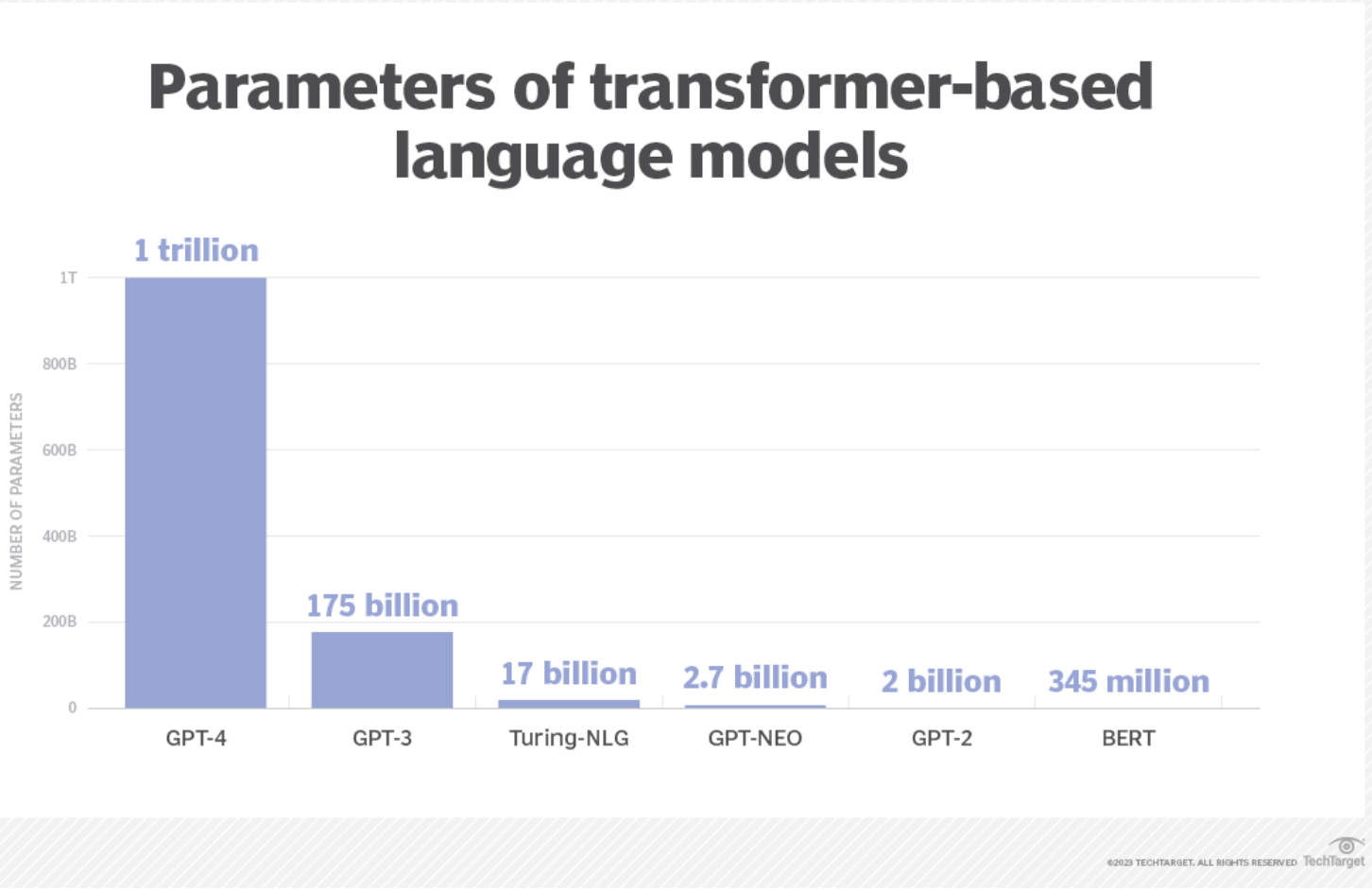
Shanghai Jiao Tong University

MaD Seminar, New York University

Learning systems with increasingly large size

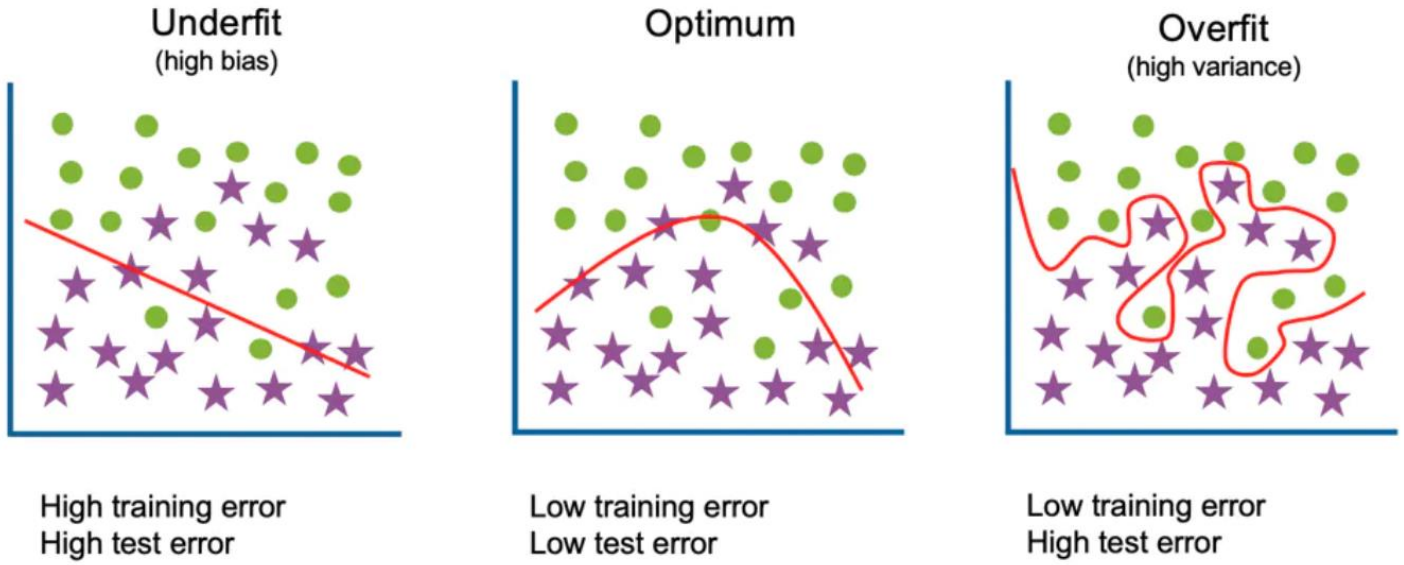


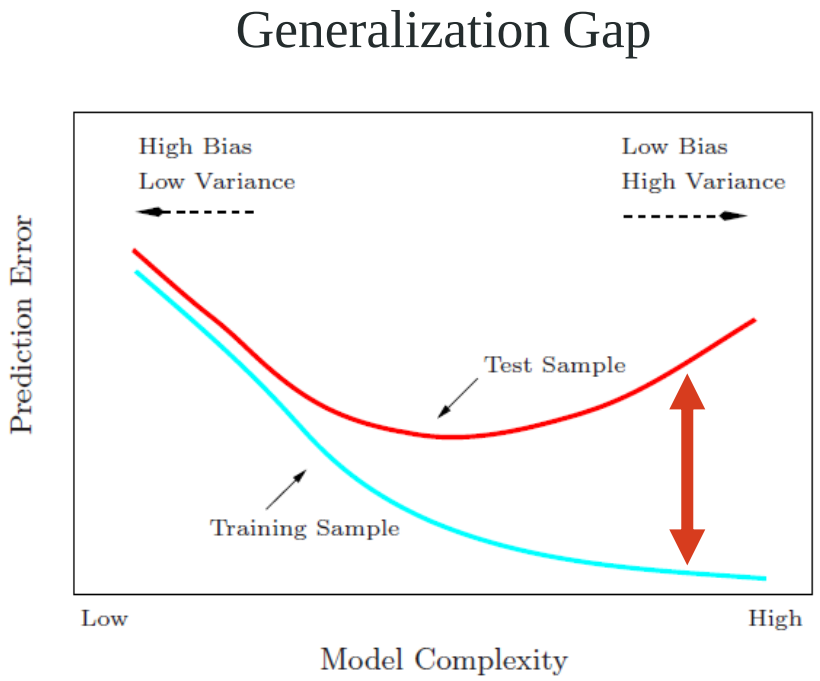
Suzana Herculano-Houzel, 2009



Failure of traditional wisdom

Large complexity → Large generalization gap





Traditional wisdom: complex models easily overfit

Long-standing problems

1995

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

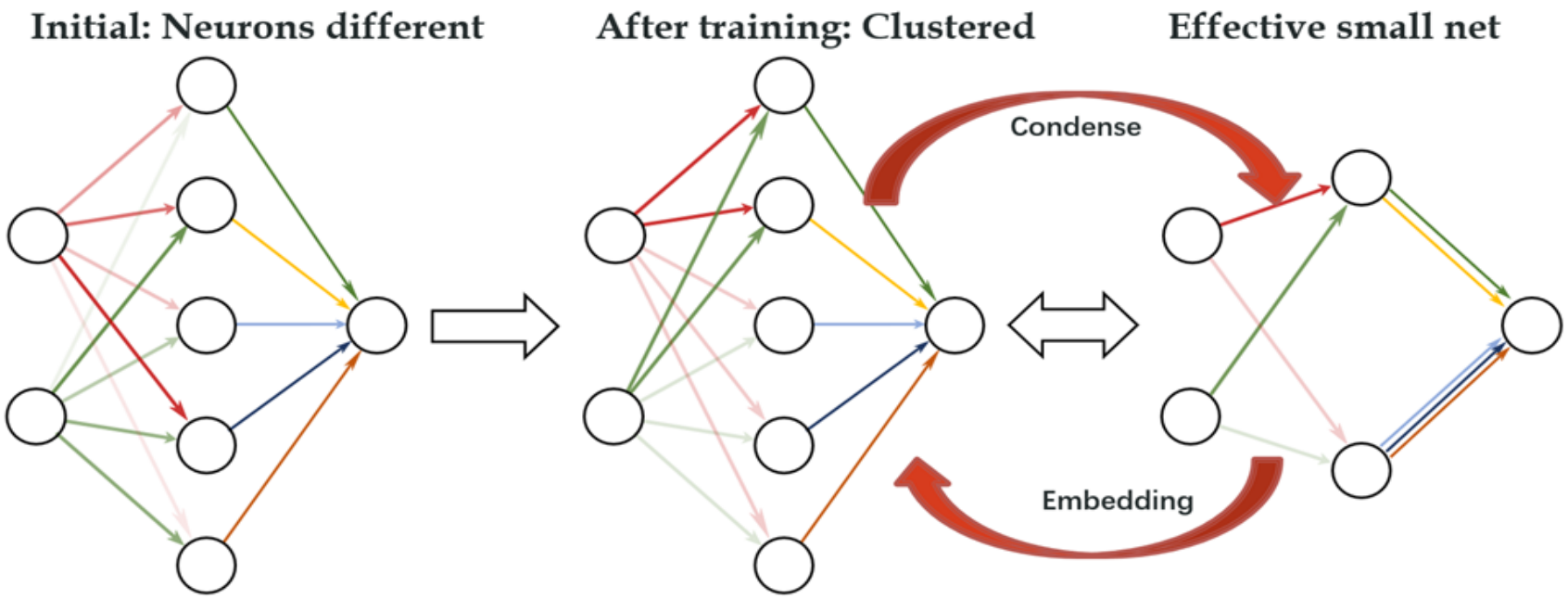
Reflections After Refereeing Papers for NIPS

- 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?

How (overparameterized) neural networks control the complexity of output function during **nonlinear** training?

Condensation Phenomenon

Illustration of Condensation



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

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

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

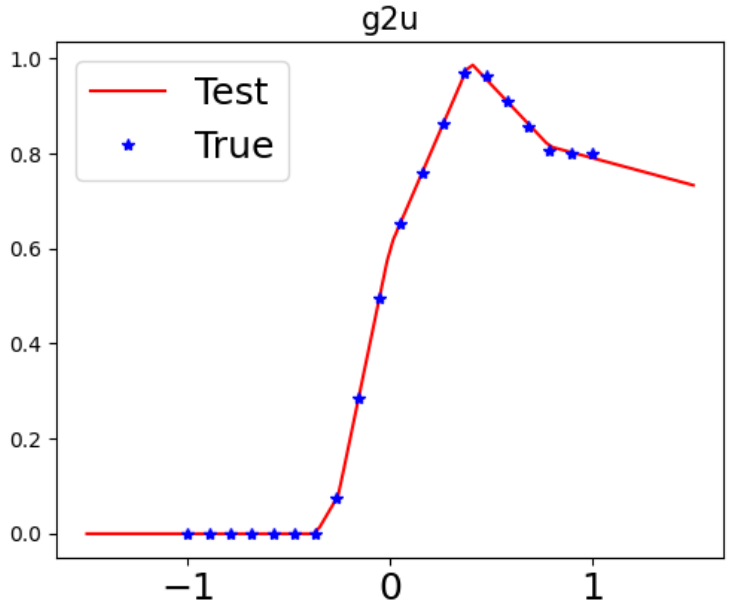
Initial: random **Training: condense** **Effect: equiv to small net**

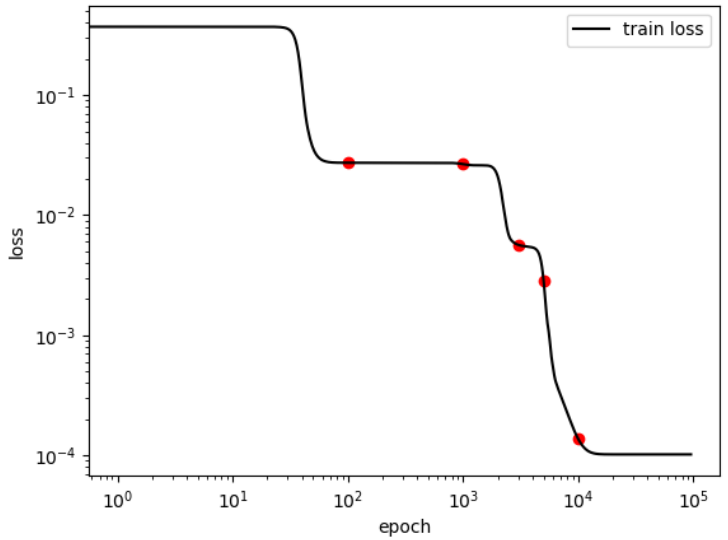
1d example: condensation with small initialization

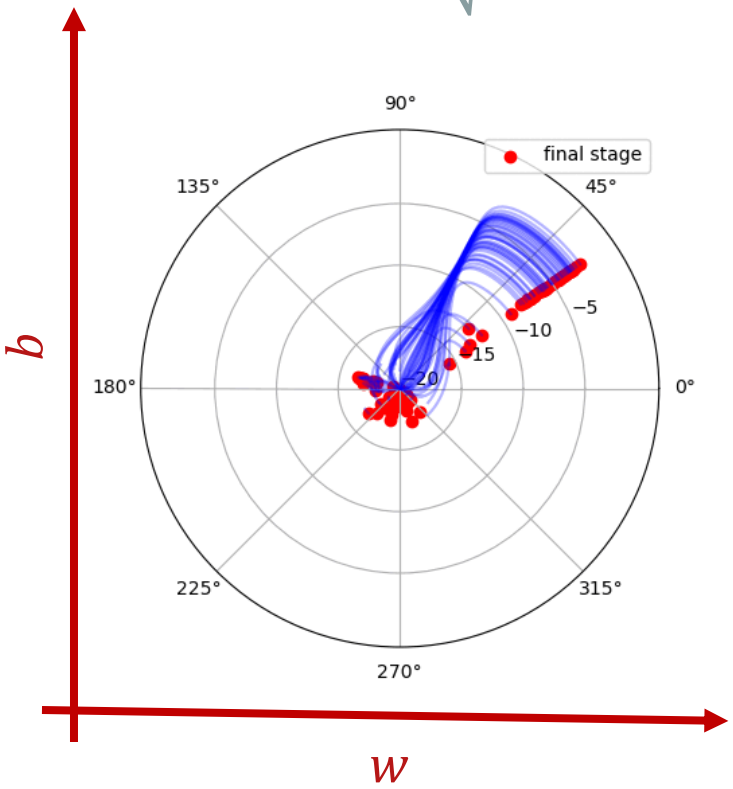
$$f_\theta(x) = \sum_{j=1}^{m} a_j \,\mathrm{relu}(w_j x + b_j)$$

$$\mathrm{relu}(z) = \max(0, z)$$

$$A_j = |a_j|\sqrt{w_j^2 + b_j^2}$$

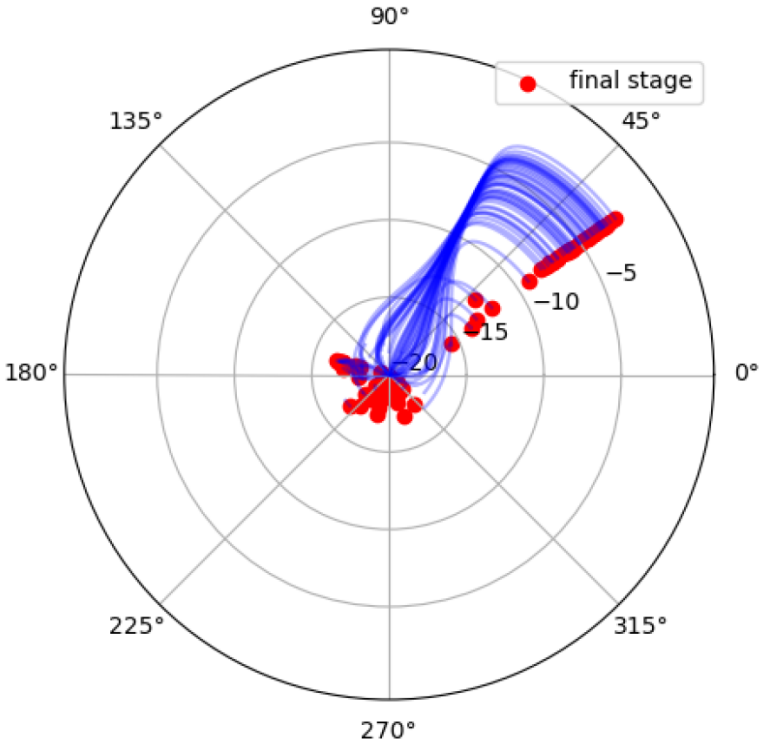




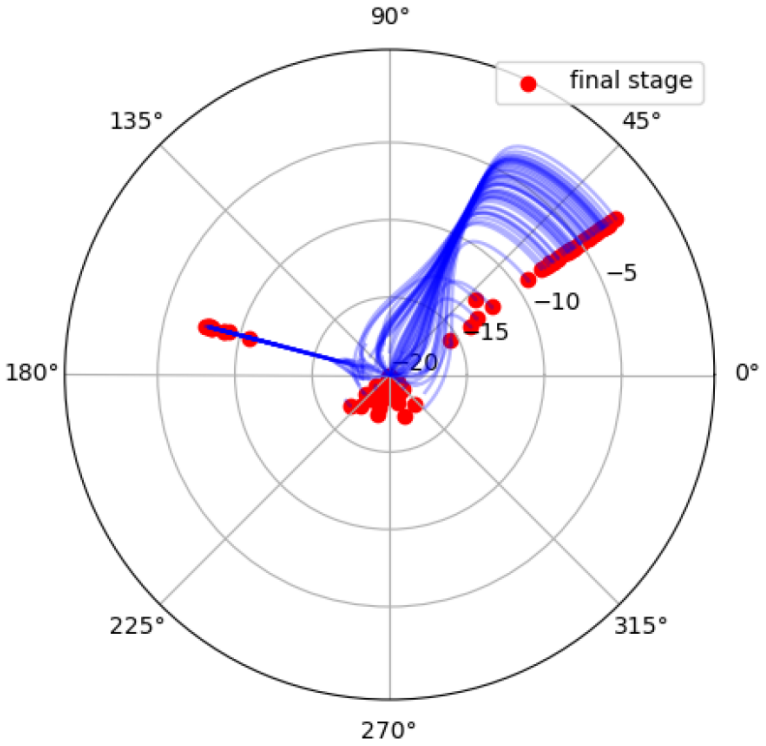


Small initialization: $a_j(0), w_j(0), b_j(0) \sim N(0, \sigma^2)$ with small σ

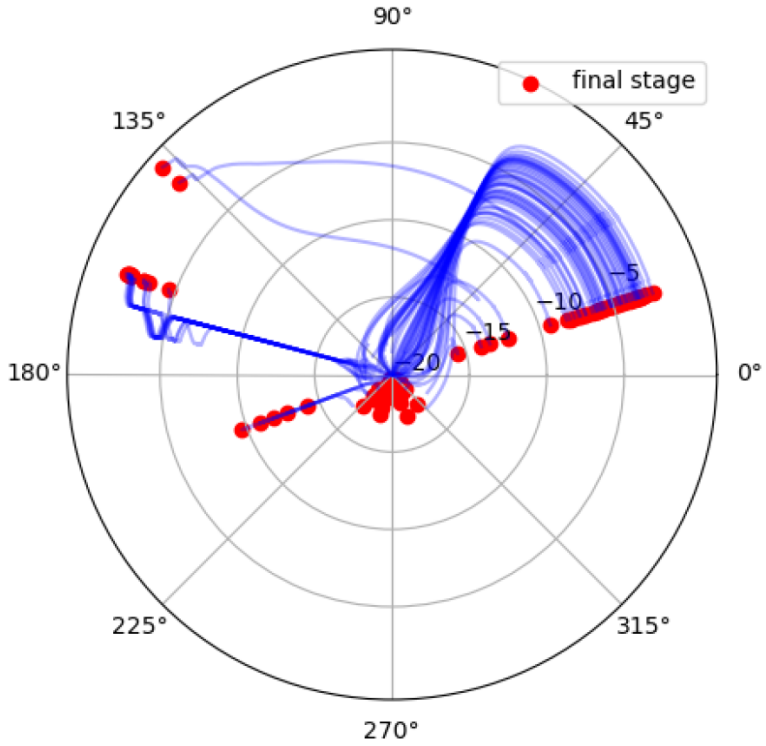
Evolution trajectory: change significantly



(a) epoch=100

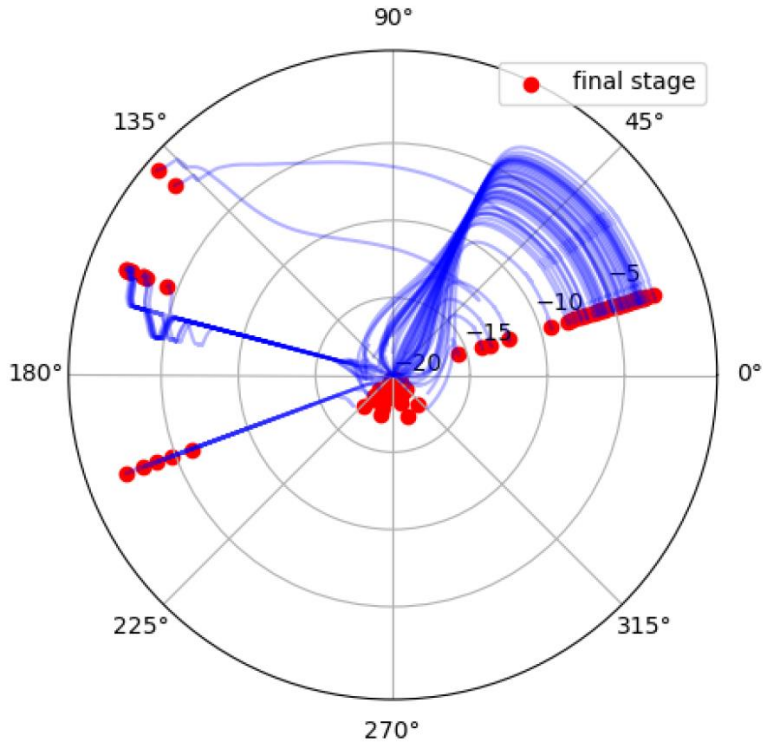


(b) epoch=1000

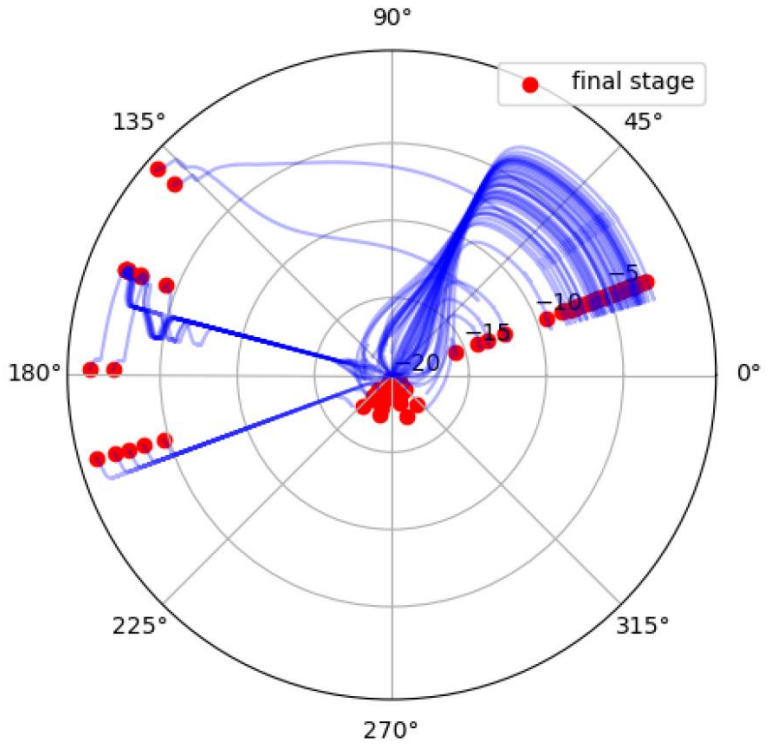


(c) epoch=3000

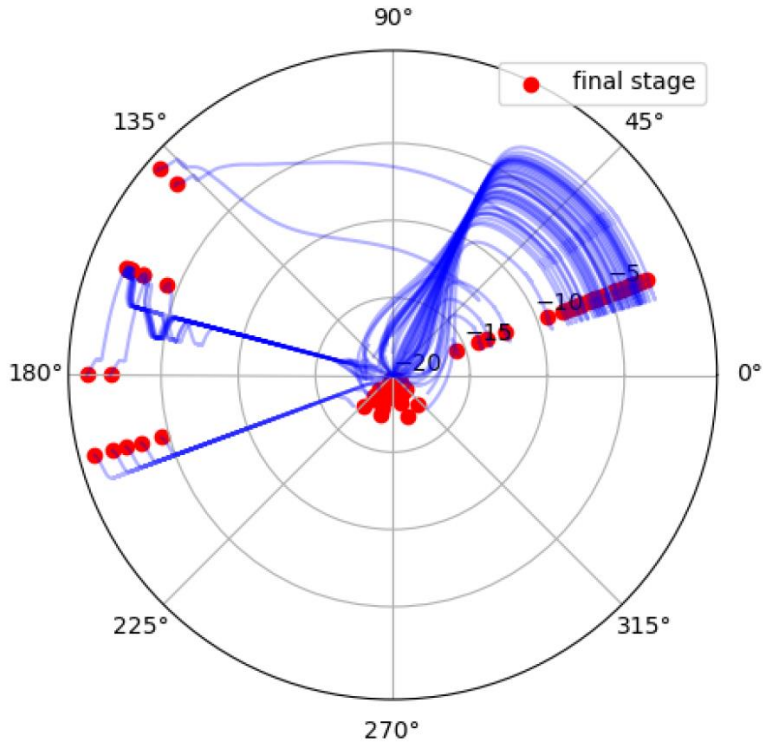
Evolution trajectory: change significantly



(d) epoch=5000

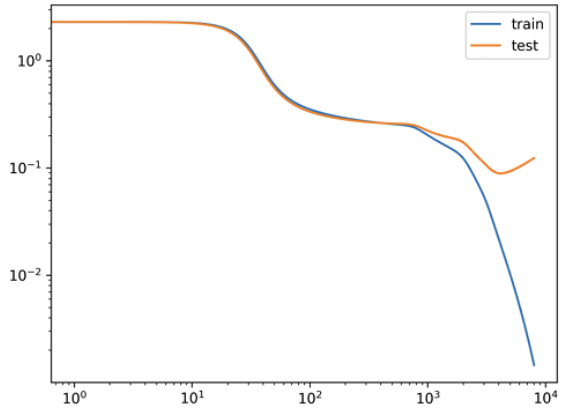


(e) epoch=10000

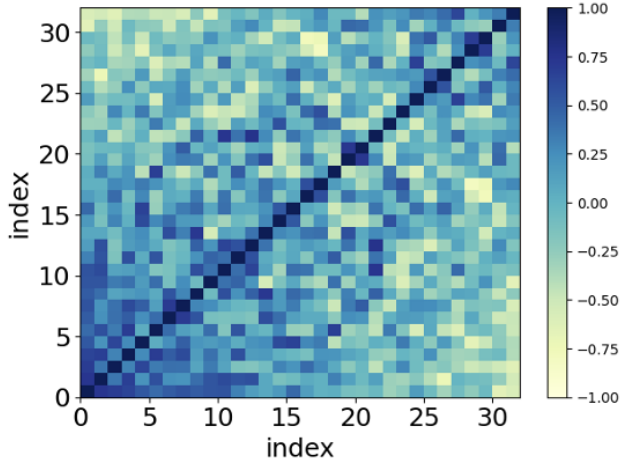


(f) epoch=100000

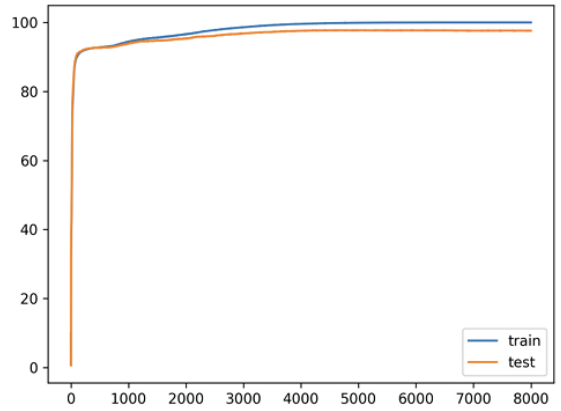
Condensation in CNN on MNIST



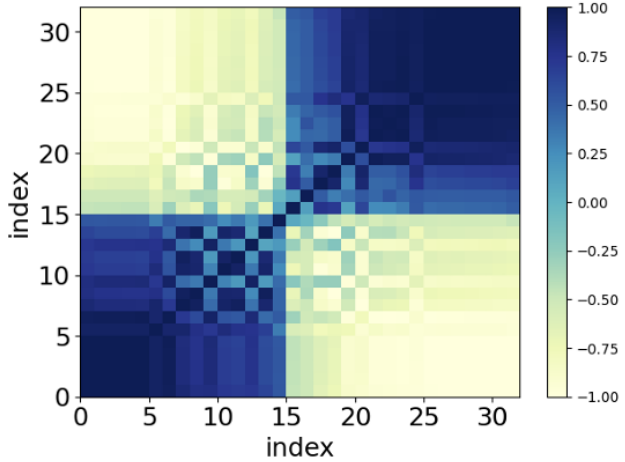
(a) Loss



(b) initial weight



(d) Accuracy



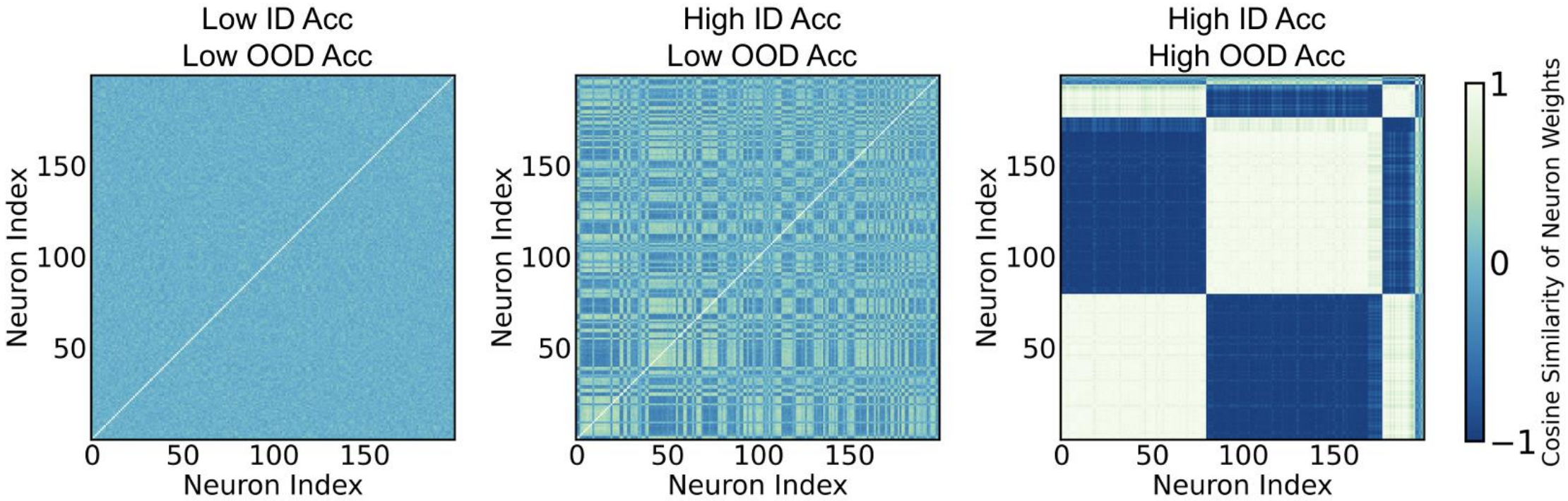
(e) final weight

Cosine similarity:

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

100% training and 97.62% test accuracy

Condensation in transformer



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

Regime of Condensation

1.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).

2.Hanxu Zhou, Qixuan Zhou, Zhenyuan Jin, Tao Luo, Yaoyu Zhang, Zhi-Qin John Xu, "Empirical Phase Diagram for Three-layer Neural Networks with Infinite Width," NeurIPS 2022.

Normalization and scaling parameters

Two layer ReLU network

$$f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}) = \frac{1}{\alpha}\sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x}) \qquad a_k^0 \sim N(0, \beta_1^2),\ \boldsymbol{w}_k^0 \sim N(0, \beta_2^2 \boldsymbol{I}_d)$$

$$x = [x^T, 1]^T \qquad w_k = [w_k^T, b_k]^T$$

Normalized gradient flow

$$\bar{a}_k = \beta_1^{-1} a_k, \quad \bar{\boldsymbol{w}}_k = \beta_2^{-1}\boldsymbol{w}_k, \quad \bar{t} = \frac{1}{\beta_1\beta_2}t,$$

$$\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}\bar{t}} = -\frac{1}{\kappa'}\frac{1}{n}\sum_{i=1}^{n} \kappa\sigma(\bar{\boldsymbol{w}}_k^{\mathsf{T}}\boldsymbol{x}_i)\left(\kappa\sum_{k'=1}^{m}\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^{\mathsf{T}}\boldsymbol{x}_i) - y_i\right),$$

$$\frac{\mathrm{d}\bar{\boldsymbol{w}}_k}{\mathrm{d}\bar{t}} = -\kappa'\frac{1}{n}\sum_{i=1}^{n} \kappa\bar{a}_k\sigma'(\bar{\boldsymbol{w}}_k^{\mathsf{T}}\boldsymbol{x}_i)\boldsymbol{x}_i\left(\kappa\sum_{k'=1}^{m}\bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^{\mathsf{T}}\boldsymbol{x}_i) - y_i\right).$$

$$m \to +\infty$$

$$\frac{\beta_1\beta_2}{\alpha} = m^{-\gamma}$$

$$\frac{\beta_1}{\beta_2} = m^{-\gamma'}$$

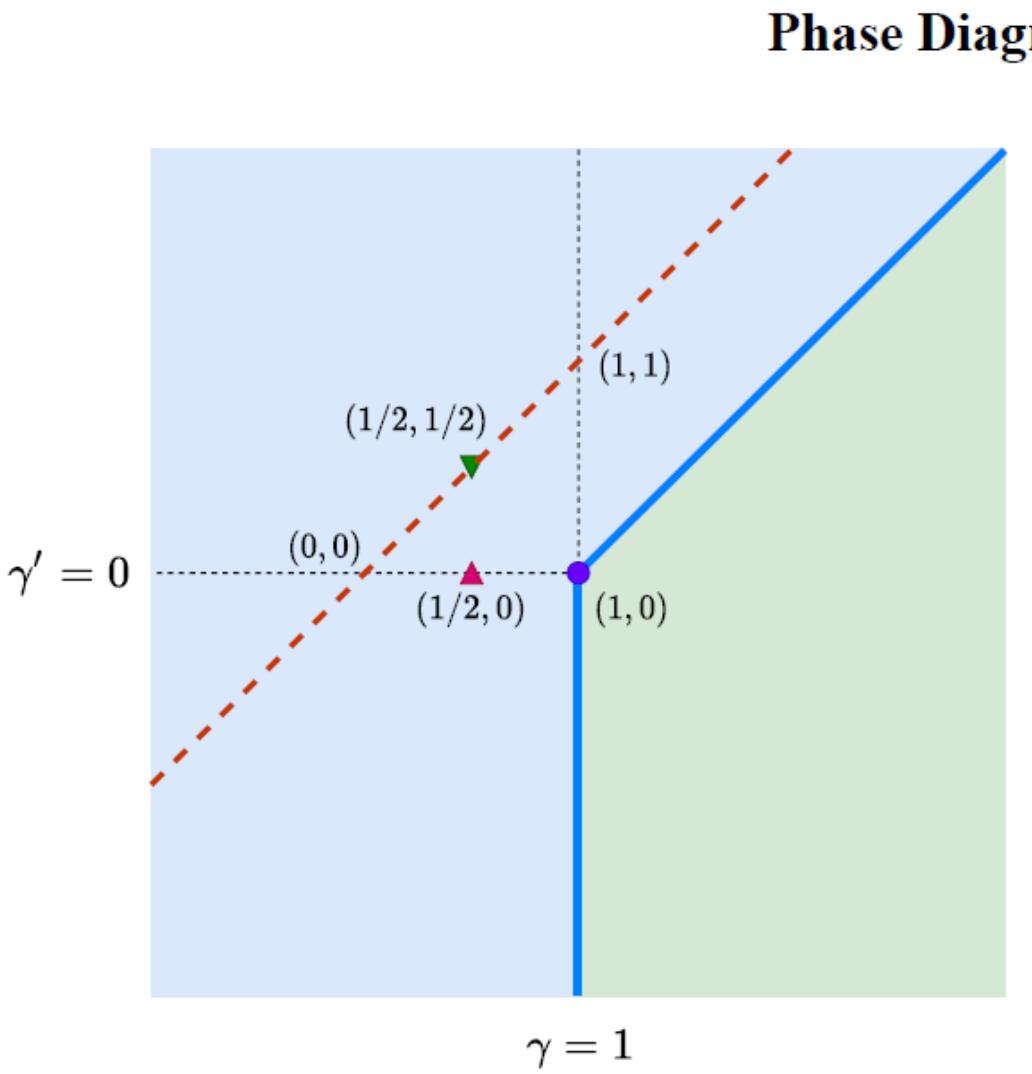
Scaling parameters and infinite-width limit

$$\kappa := \frac{\beta_1\beta_2}{\alpha}, \quad \kappa' := \frac{\beta_1}{\beta_2}, \qquad \gamma = \lim_{m\to\infty} -\frac{\log\kappa}{\log m}, \quad \gamma' = \lim_{m\to\infty} -\frac{\log\kappa'}{\log m},$$

Initialization scheme

Name (related works)	α	β_1	β_2	κ $\left(\frac{\beta_1\beta_2}{\alpha}\right)$	κ' $\left(\frac{\beta_1}{\beta_2}\right)$	γ $\left(\lim_{m\to\infty}\frac{\log 1/\kappa}{\log m}\right)$	γ' $\left(\lim_{m\to\infty}\frac{\log 1/\kappa'}{\log m}\right)$
LeCun (LeCun et al., 2012)	1	$\sqrt{\frac{1}{m}}$	$\sqrt{\frac{1}{d}}$	$\sqrt{\frac{1}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
He (He et al., 2015)	1	$\sqrt{\frac{2}{m}}$	$\sqrt{\frac{2}{d}}$	$\sqrt{\frac{4}{md}}$	$\sqrt{\frac{d}{m}}$	$\frac{1}{2}$	$\frac{1}{2}$
Xavier (Glorot and Bengio, 2010)	1	$\sqrt{\frac{2}{m+1}}$	$\sqrt{\frac{2}{m+d}}$	$\sqrt{\frac{4}{(m+1)(m+d)}}$	$\sqrt{\frac{m+d}{m+1}}$	1	0
NTK (Jacot et al., 2018)	$\sqrt{m}$	1	1	$\sqrt{\frac{1}{m}}$	1	$\frac{1}{2}$	0
Mean-field (Mei et al., 2018) (Sirignano and Spiliopoulos, 2020) (Rotskoff and Vanden-Eijnden, 2018)	m	1	1	$\frac{1}{m}$	1	1	0
E et al. (E et al., 2020)	1	β	1	β	β	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$

When condensation happens (at infinite width limit)?



$$a_k^0 \sim N(0, \beta_1^2), \quad \mathbf{w}_k^0 \sim N(0, \beta_2^2 \mathbf{I}_d)$$

$$\gamma = \lim_{m\to\infty} -\frac{\log \beta_1 \beta_2 / \alpha}{\log m}, \quad \gamma' = \lim_{m\to\infty} -\frac{\log \beta_1 / \beta_2}{\log m}$$

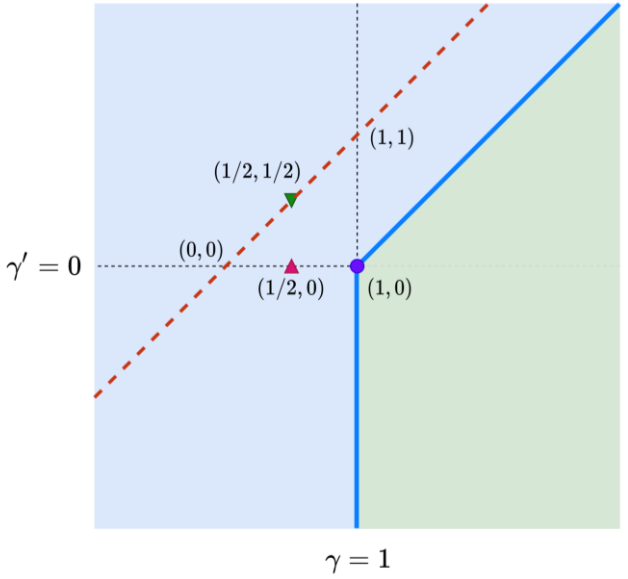
Regime separation -- theorems

Theorem 1*. *(Informal statement of Theorem 6) If $\gamma < 1$ or $\gamma' > \gamma - 1$, then with a high probability over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m \to +\infty} \sup_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = 0. \tag{20}$$

Theorem 2*. *(Informal statement of Theorem 8) If $\gamma > 1$ and $\gamma' < \gamma - 1$, then with a high probability over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{m \to +\infty} \sup_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = +\infty. \tag{21}$$



$$f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}) = \frac{1}{\alpha} \sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x})$$

$$\boldsymbol{\theta}_{\boldsymbol{w}} = \mathrm{vec}(\{\boldsymbol{w}_k\}_{k=1}^{m})$$

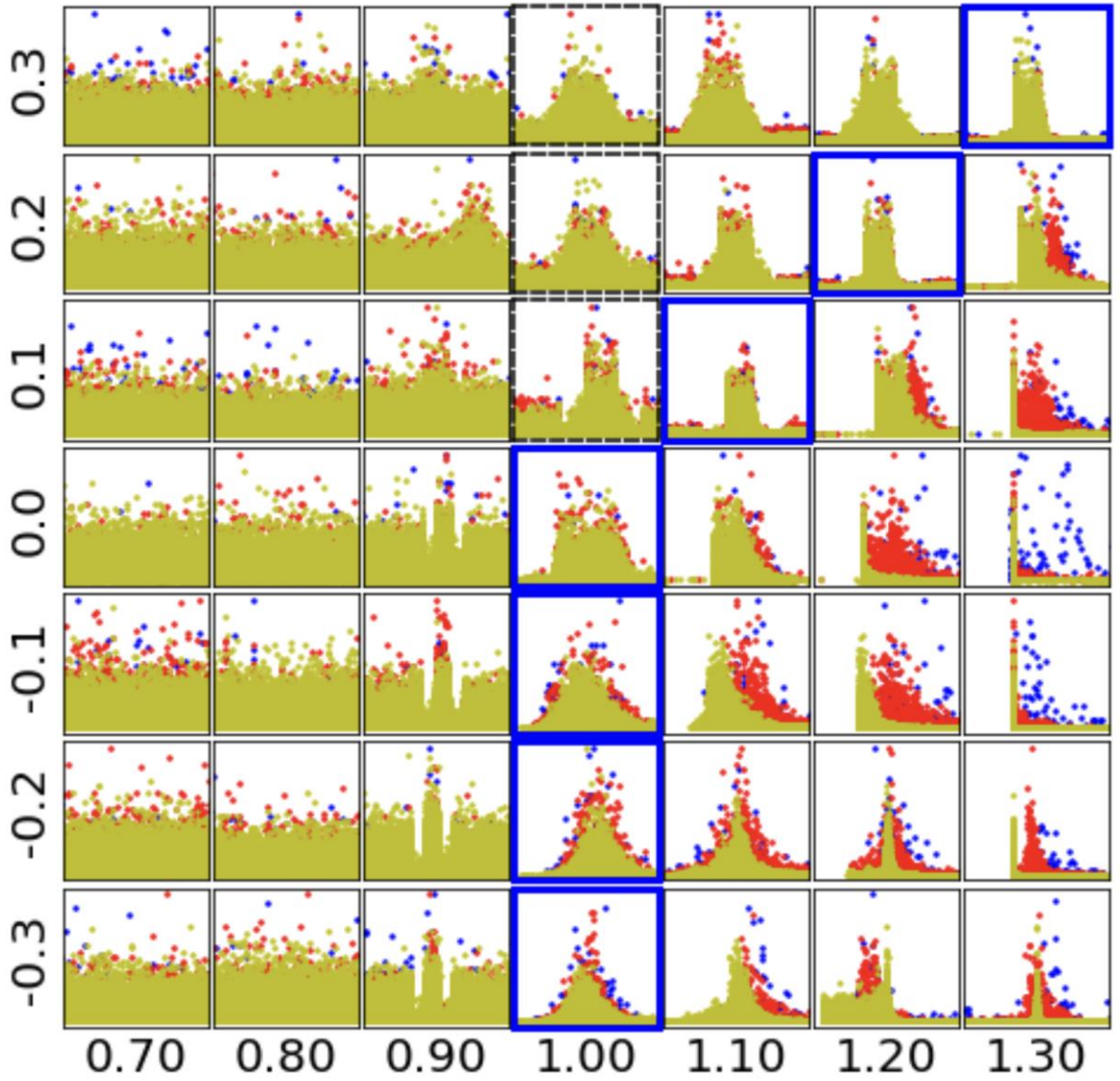
$$\mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = \frac{\|\theta_{\boldsymbol{w}}(t) - \theta_{\boldsymbol{w}}(0)\|_2}{\|\theta_{\boldsymbol{w}}(0)\|_2}.$$

Feature distribution across the phase diagram

$$\{(A_k, \hat{\boldsymbol{w}}_k)\}_{k=1}^m$$

$$A = |a| \|\boldsymbol{w}\|_2$$

$$f_{\boldsymbol{\theta}}^{\alpha}(\boldsymbol{x}) = \frac{1}{\alpha} \sum_{k=1}^{m} a_k \sigma(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x})$$



Blue: $m = 10^3$

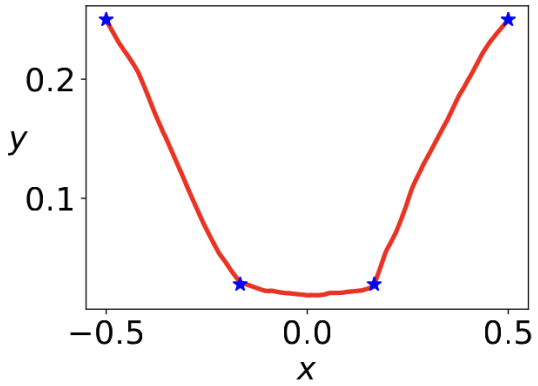
Red: $m = 10^4$

Yellow: $m = 10^6$

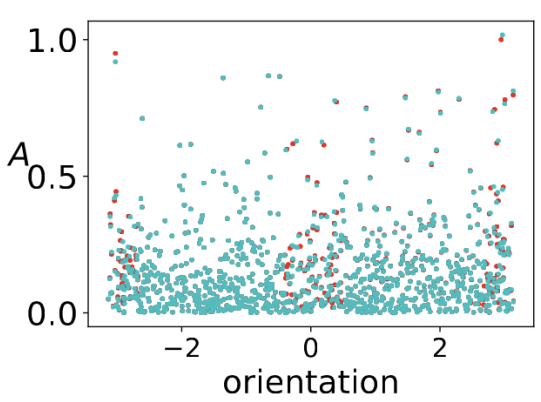
Typical cases across the phase diagram

$\gamma' = 0$

Linear regime

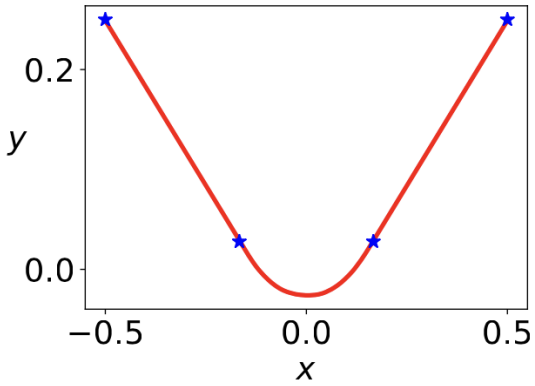


(a) $\gamma = 0.5$

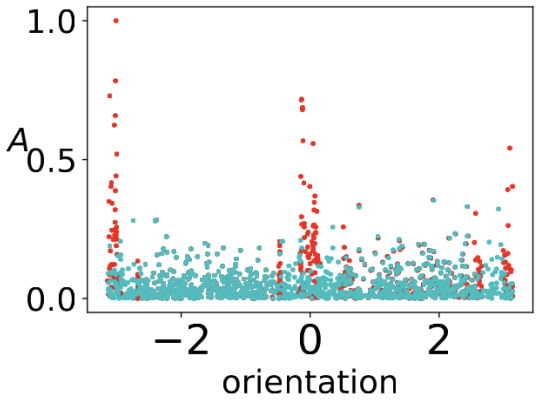


(d) $\gamma = 0.5$

critical regime

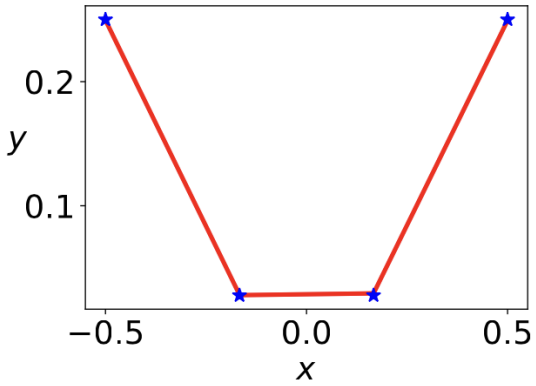


(b) $\gamma = 1$

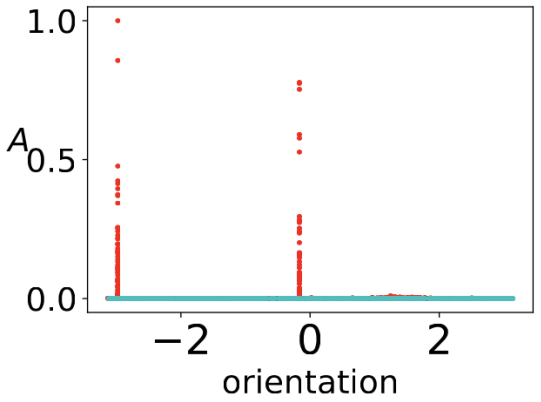


(e) $\gamma = 1$

condensed regime

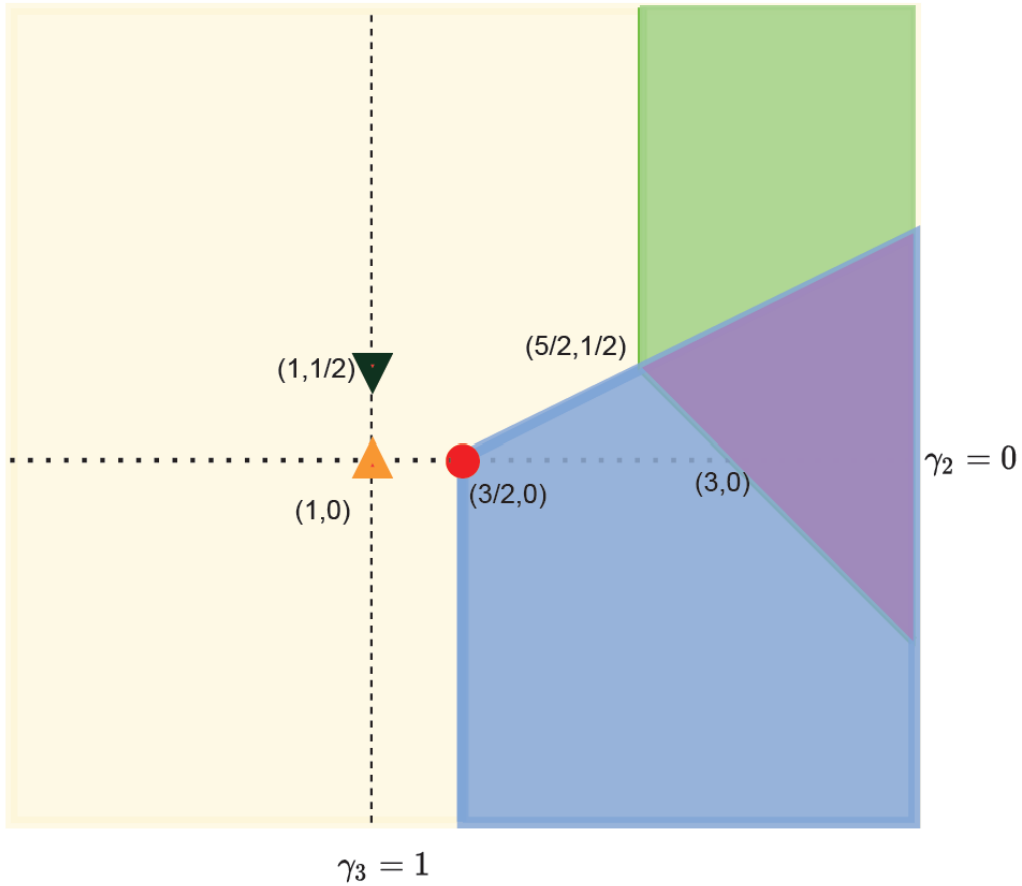


(c) $\gamma = 1.75$



(f) $\gamma = 1.75$

Phase diagram in three-layer ReLU NN



$CR\ for\ \mathbf{W}^{[2]},\ LR\ for\ \mathbf{W}^{[1]}$

$CR\ for\ \mathbf{W}^{[2]},\ CR\ for\ \mathbf{W}^{[1]}$

$LR\ for\ \mathbf{W}^{[2]},\ CR\ for\ \mathbf{W}^{[1]}$

$LR\ for\ \mathbf{W}^{[2]},\ LR\ for\ \mathbf{W}^{[1]}$

$Example:$

$Xavier$

NTK

$Lecun, He$

$$\mathbf{a}_k \sim \mathcal{N}(0, \beta_3^2),\ \mathbf{W}_{ij}^{[2]} \sim (0, \beta_2^2),\ \mathbf{W}_{ij}^{[1]} \sim (0, \beta_1^2)$$

$$\gamma_3 = \lim_{m\to\infty} -\frac{\log \beta_1\beta_2\beta_3/\alpha}{\log m},\ \gamma_2 = \lim_{m\to\infty} -\frac{\log \beta_3/\beta_1}{\log m}$$

$CR\ is\ short\ for\ Condensed\ Reigme,\ LR\ is\ short\ for\ Linear\ Regime$

Loss landscape structure underlying condensation

1.Yaoyu Zhang, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, "Embedding Principle of Loss Landscape of Deep Neural Networks," NeurIPS 2021 spotlight.

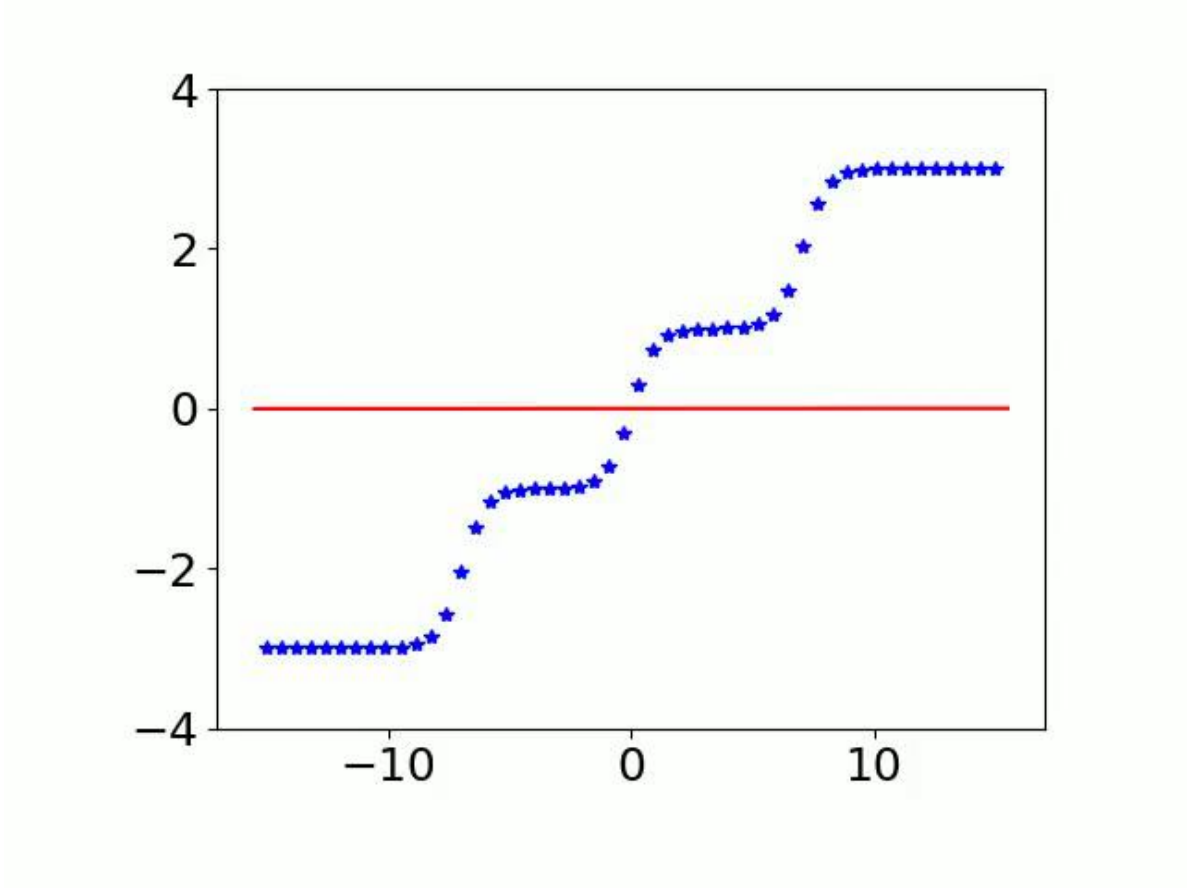
2.Yaoyu Zhang, Yuqing Li, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, "Embedding Principle: a hierarchical structure of loss landscape of deep neural networks," Journal of Machine Learning, 1(1), pp. 60-113, 2022.

3.Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, Zhi-Qin John Xu, "Towards Understanding the Condensation of Neural Networks at Initial Training," NeurIPS 2022.

4.Tao Luo, Leyang Zhang, Yaoyu Zhang, "Structure and Gradient Dynamics Near Global Minima of Two-layer Neural Networks," arXiv:2309.00508 (2023).

SJTU

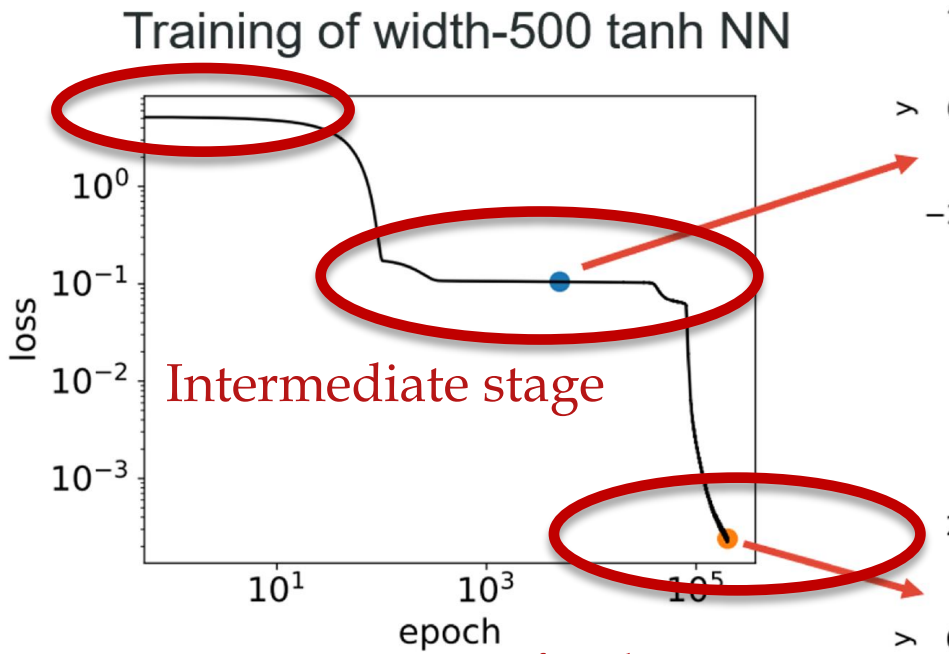
Typical training behavior with strong condensation

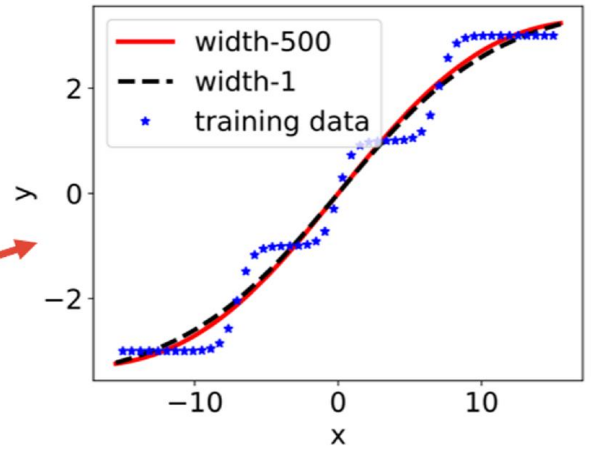


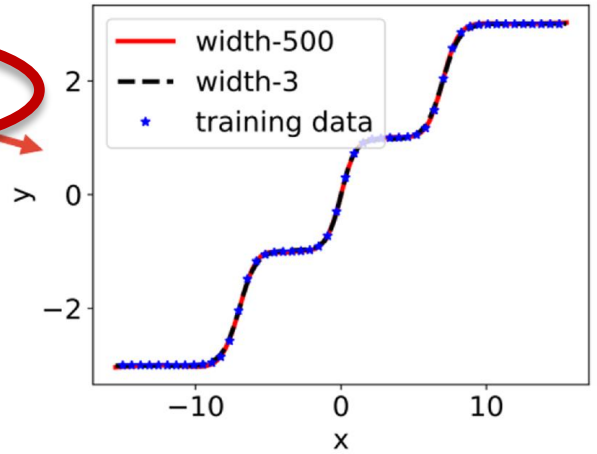
Width-500 tanh-NN (~1500 parameters)

Trajectory of training loss

Initial stage

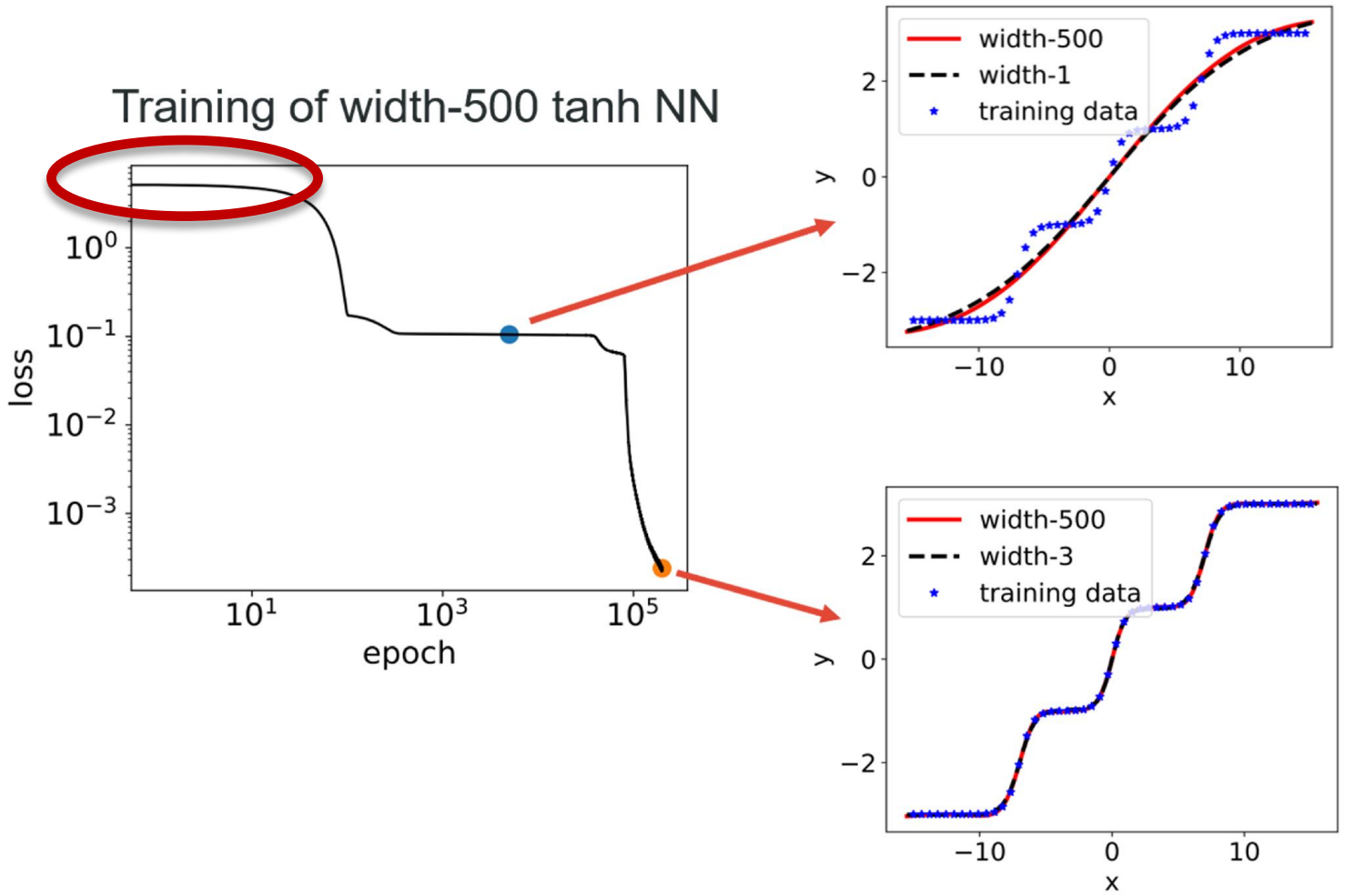






Initial condensation

Initial stage



Loss landscape around 0 and Initial condensation

$$\dot{\boldsymbol{w}}_j = \sum_{i=1}^{m} \big(y_i - f_{\boldsymbol{\theta}}(x_i)\big) a_j \sigma'\big(\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i\big) \boldsymbol{x}_i$$

When $\boldsymbol{\theta} \approx \mathbf{0}$, then $f_{\boldsymbol{\theta}}(\cdot) \approx 0(\cdot)$:

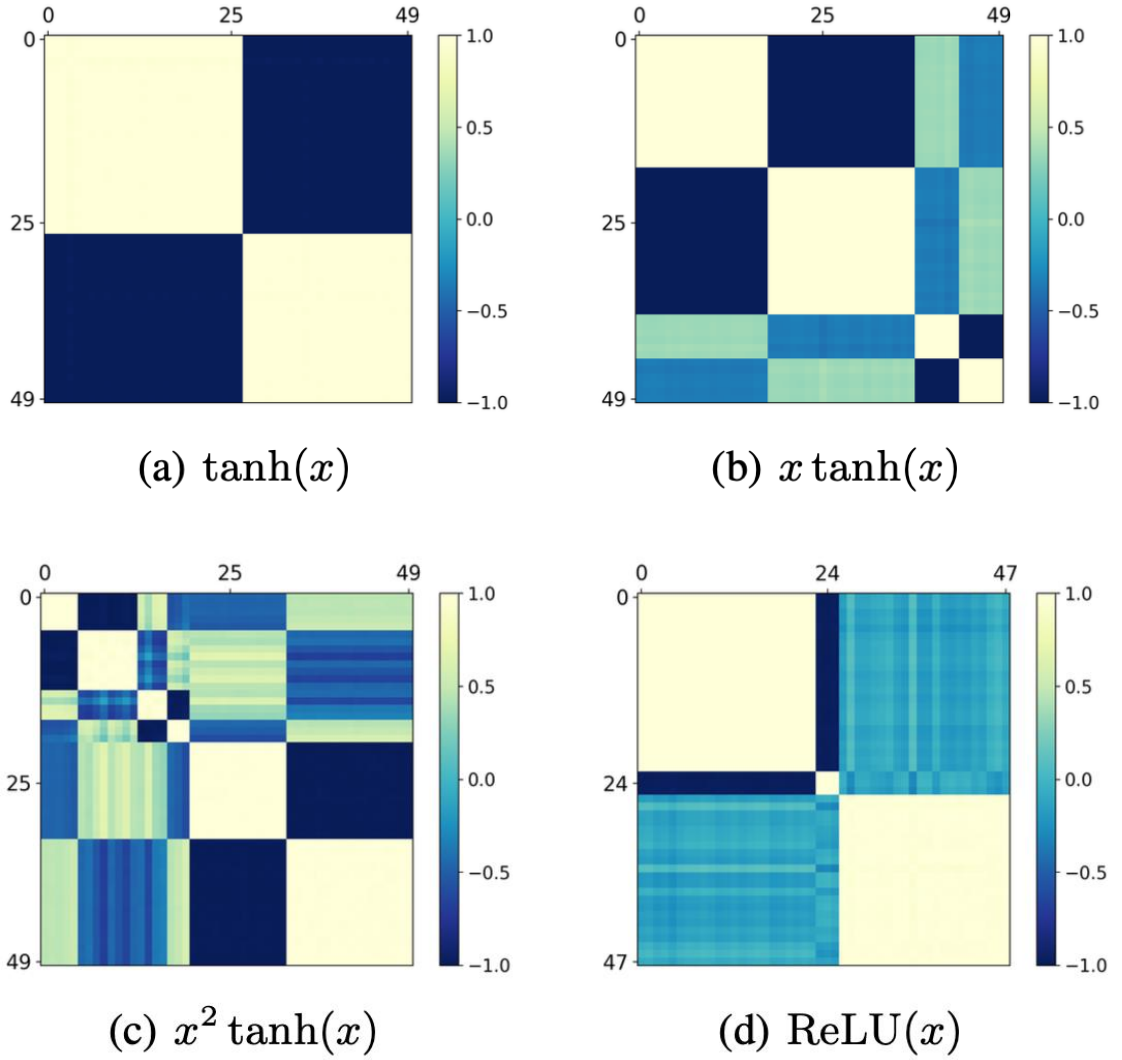
$$\dot{\boldsymbol{w}}_j \approx a_j \sum_{i=1}^{m} y_i \sigma'\big(\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i\big) \boldsymbol{x}_i$$

If $\sigma'(0) \neq 0$ (e.g. tanh, swish, gelu):

$$\dot{\boldsymbol{w}}_j \approx a_j \sigma'(0) \sum_{i=1}^{m} y_i \boldsymbol{x}_i$$

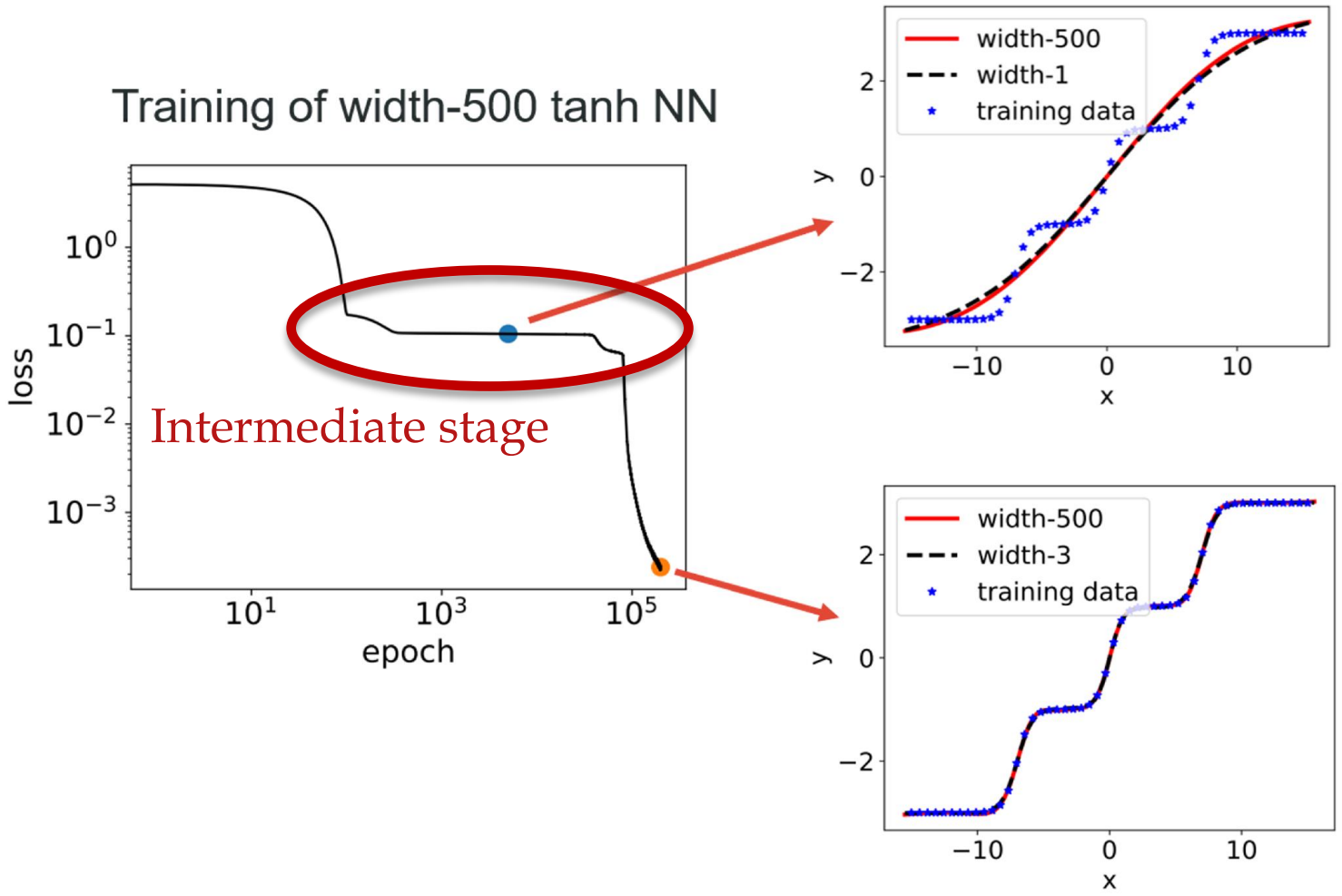
i. No coupling between w_j and $w_{j'}$!

ii. 2 limiting directions: $\pm \sum_{i=1}^{m} y_i \boldsymbol{x}_i$.



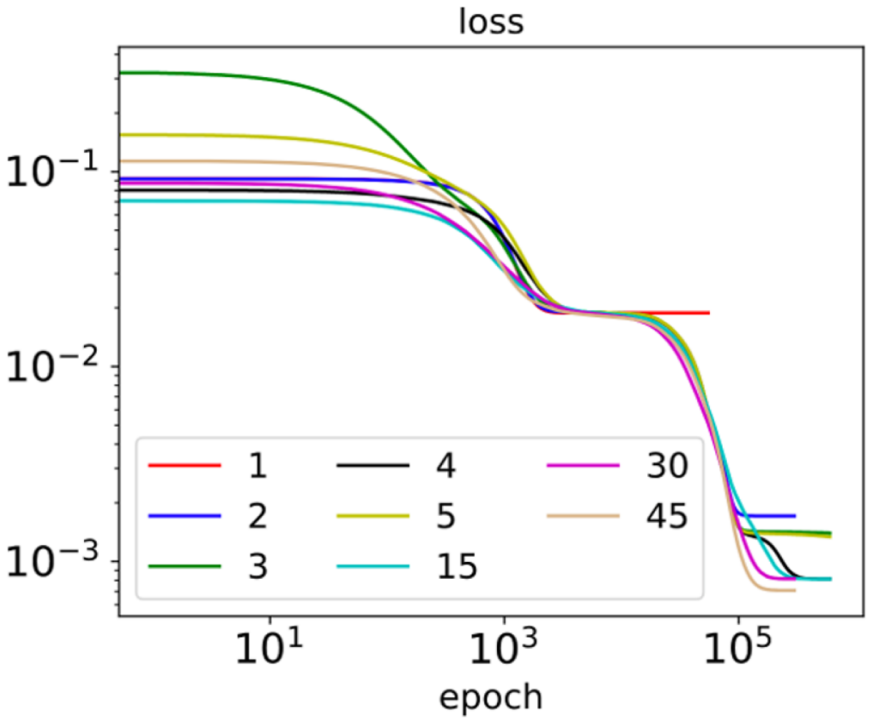
Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, Zhi-Qin John Xu,
Towards Understanding the Condensation of Neural Networks at Initial Training, NeurIPS 2022.

Intermediate condensation



Zhang, Zhang, Luo, Xu, NeurIPS 2021 spotlight.
Zhang, Li, Zhang, Luo, Xu, Journal of Machine Learning 2022.

Condensed critical points for intermediate stage



Observation:
Width similarity

Embedding Principle (informal Theorem)

The loss landscape of any network ``contains'' all critical points of all narrower networks.

Equivalent Statement

$$\mathcal{F}^{\mathbf{c}}_{\mathbf{narr}} \subseteq \mathcal{F}^{\mathbf{c}}_{\mathbf{wide}},$$

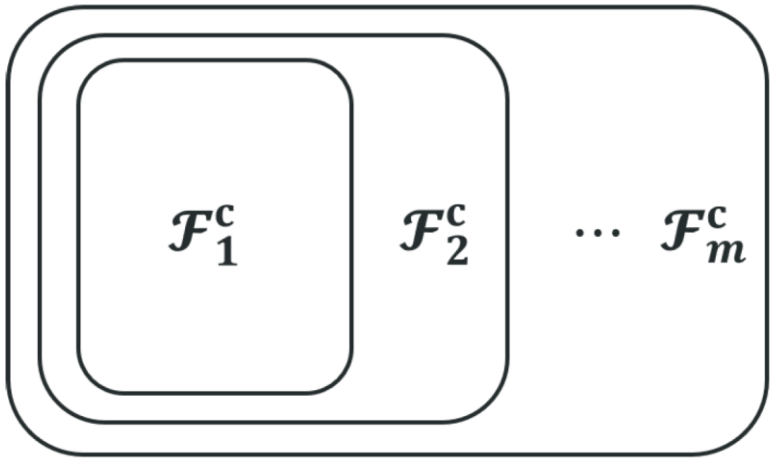
where $\mathcal{F}^{c} := \{f_\theta(\cdot) | \nabla R_S(\theta) = 0\}$.

Implication of theory:
simple condensed critical points are common

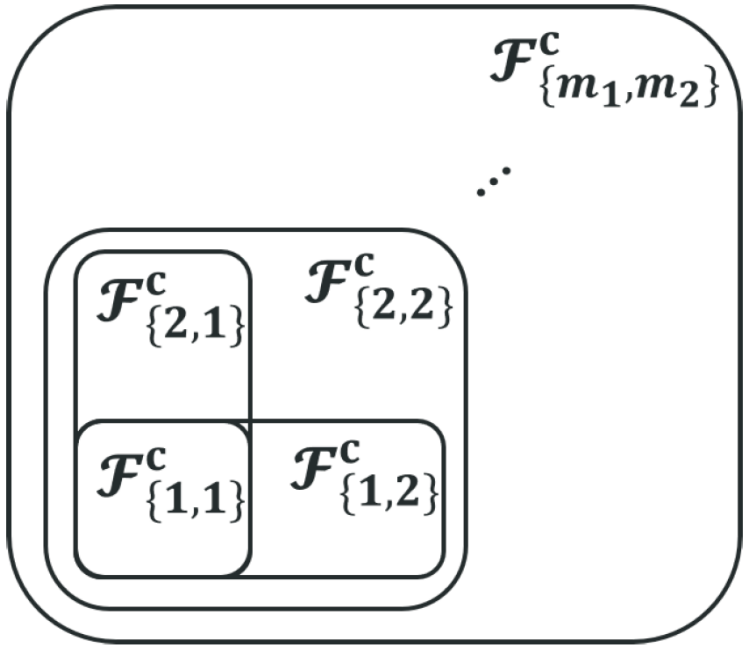
Zhang, Zhang, Luo, Xu, NeurIPS 2021 spotlight.
Zhang, Li, Zhang, Luo, Xu, Journal of Machine Learning 2022.

hierarchical structure of DNN loss landscape

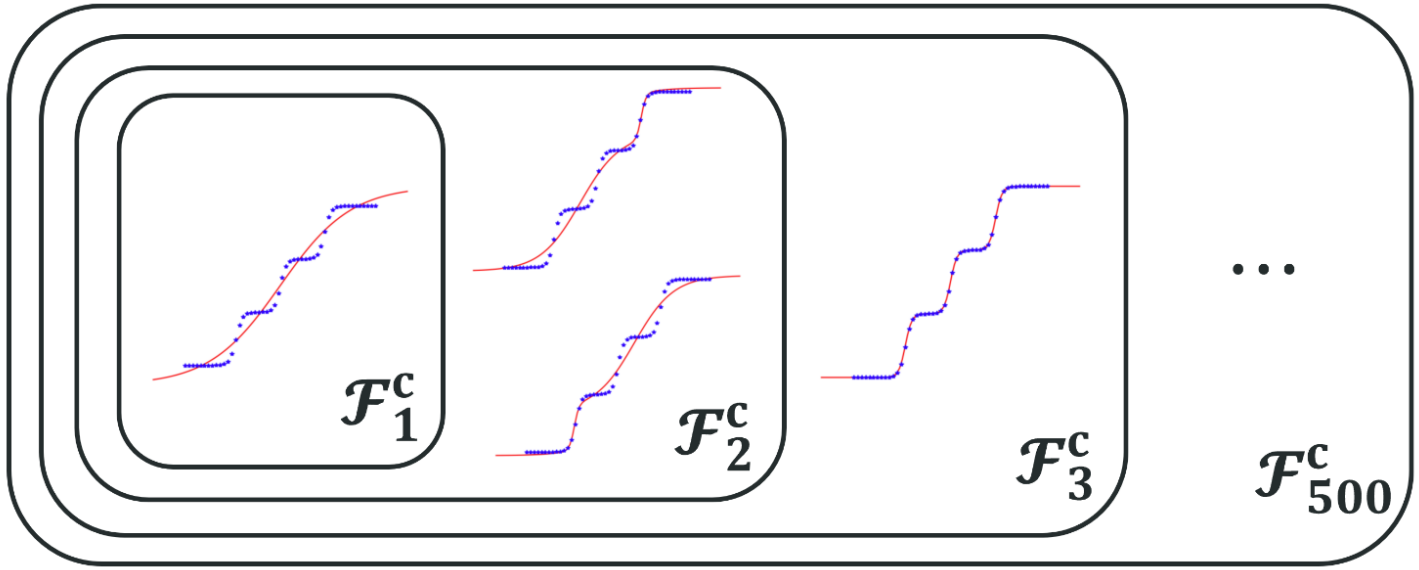
Simple → Complex



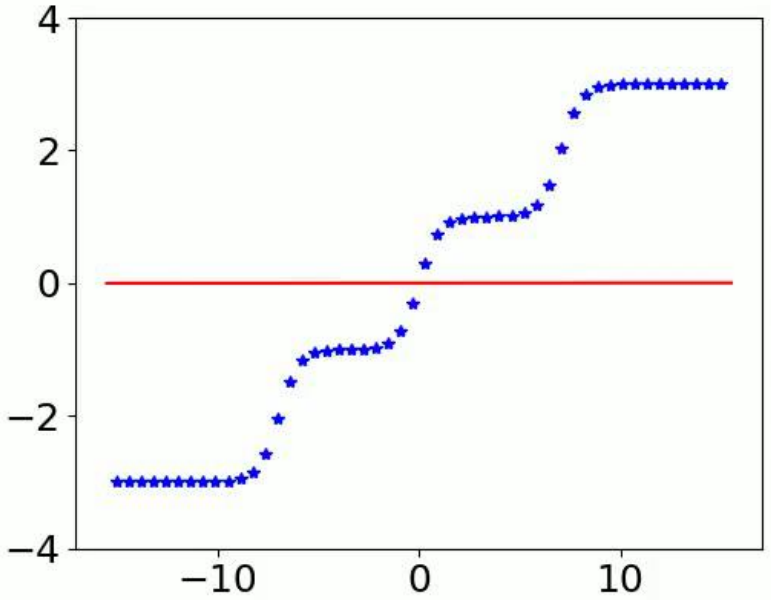
Simple → Complex



Example: identification of critical points and functions

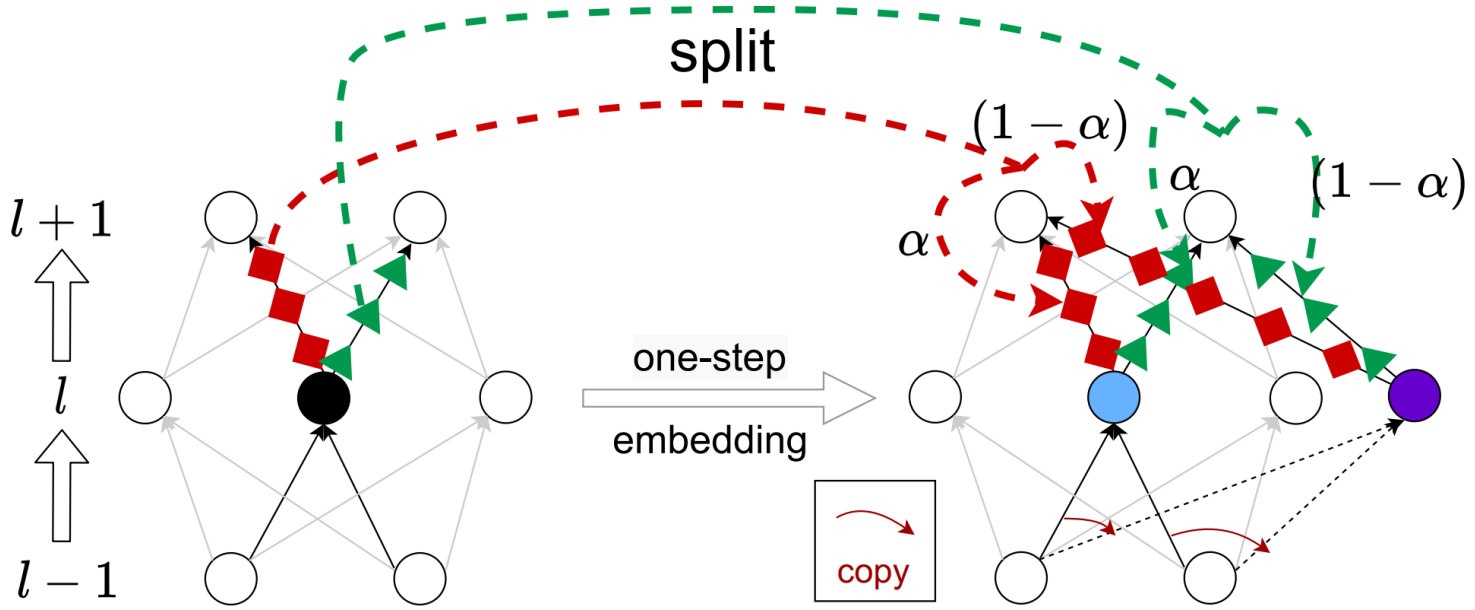


500 tanh neuron



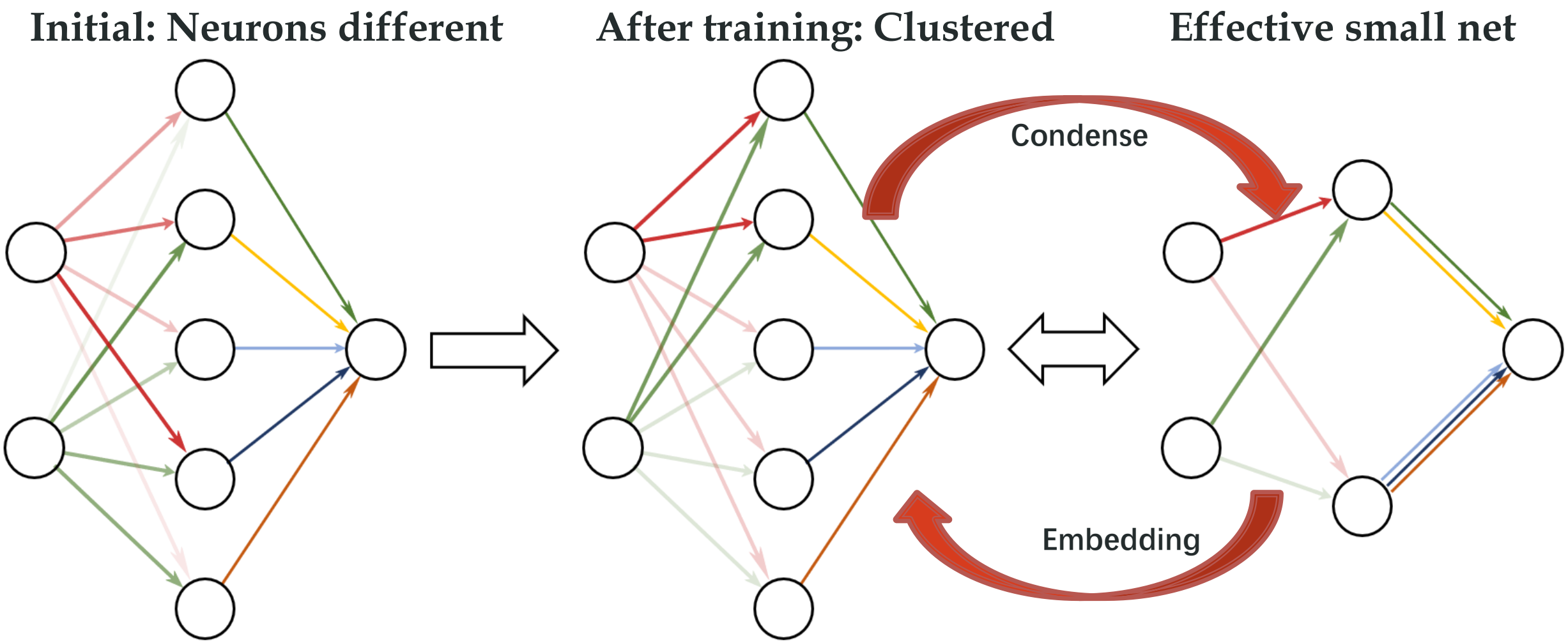
Embedding principle

One-step splitting embedding $T: \mathbb{R}^{M_{\text{narr}}} \to \mathbb{R}^{M_{\text{wide}}}$

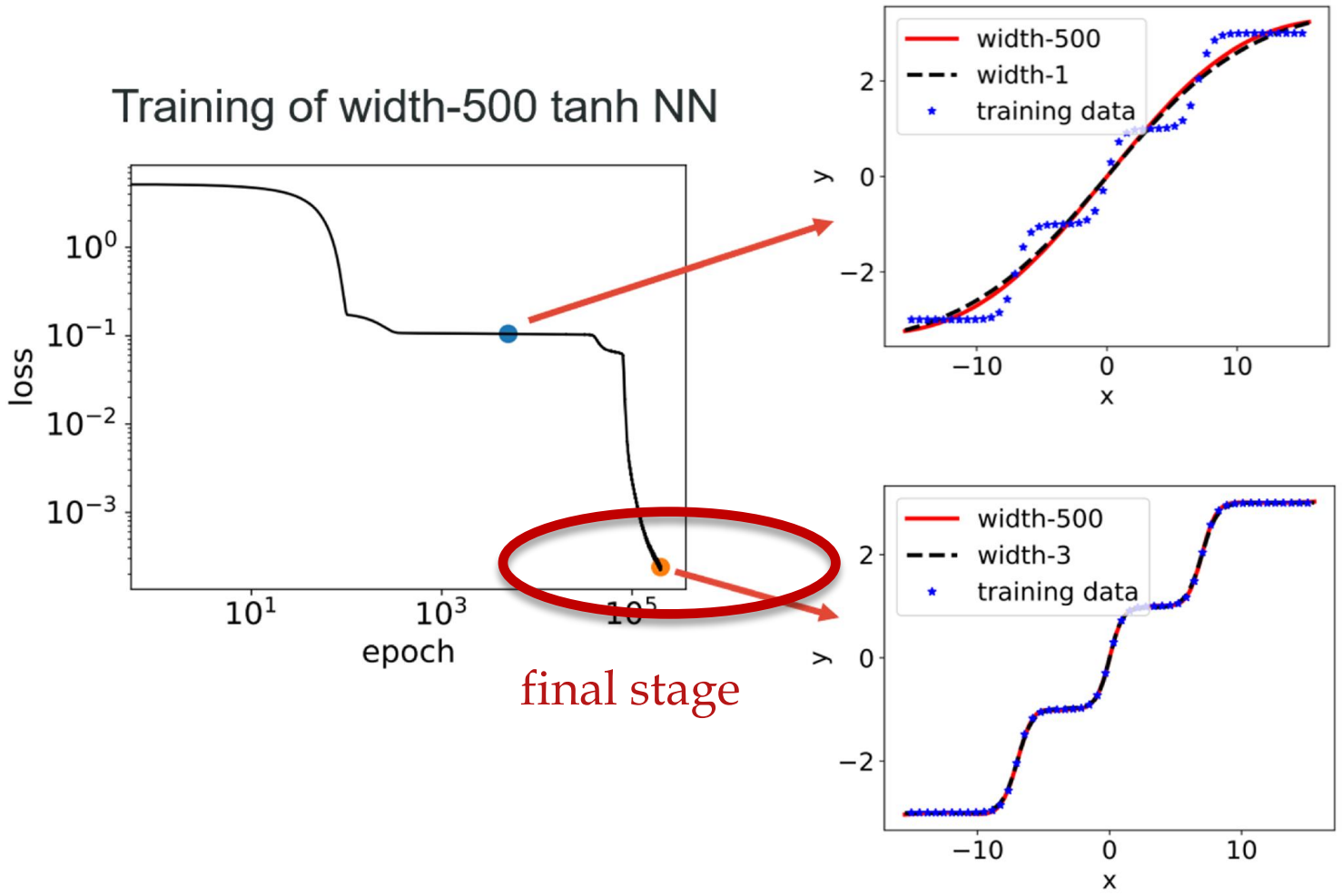


Theorem: One-step splitting embedding T with $\boldsymbol{\theta}_{\text{wide}} = T(\boldsymbol{\theta}_{\text{narr}})$ satisfies:
(i) **output preserving**: $f_{\boldsymbol{\theta}_{\text{narr}}}(x) = f_{\boldsymbol{\theta}_{\text{wide}}}(x)$;
(ii) **criticality preserving**: If $\nabla R_S(\boldsymbol{\theta}_{\text{narr}}) = \mathbf{0}$, then $\nabla R_S(\boldsymbol{\theta}_{\text{wide}}) = \mathbf{0}$.

Existance of condensed critical points---embedding principle



Final condensation



Tao Luo, Leyang Zhang, Yaoyu Zhang,
Structure and Gradient Dynamics Near Global Minima of Two-layer Neural Networks, arXiv:2309.00508 (2023)

Geometry of global-min: simpler f^*, higher-dim Q^*

- **Model:** $F(\theta)(x) = a_1\sigma\left(w_1^T x\right) + a_2\sigma\left(w_2^T x\right),\ x \in \mathbb{R}^2, \theta \in \mathbb{R}^6$
- **Target:** $f^* = \bar{a}\sigma\left(\bar{w}^{\mathrm{T}} x\right)$
- **Target Set** $Q^* = F^{-1}(f^*)$ **generally consists of three "branches" (sets)**

(a) $Q_1 = \left\{(a_k, w_k)_{k=1}^2 : w_1 = w_2 = \bar{w}, a_1 + a_2 = \bar{a}\right\}.$

(b) $Q_2 = \left\{(a_k, w_k)_{k=1}^2 : w_1 = \bar{w}, a_1 = \bar{a}, a_2 = 0\right\}.$

(c) $Q_3 = \left\{(a_k, w_k)_{k=1}^2 : w_2 = \bar{w}, a_2 = \bar{a}, a_1 = 0\right\}.$

As sample size n increases, how global min $\boldsymbol{L_S^{-1}(0)}$ **shrinks to** $\boldsymbol{Q^*}$?

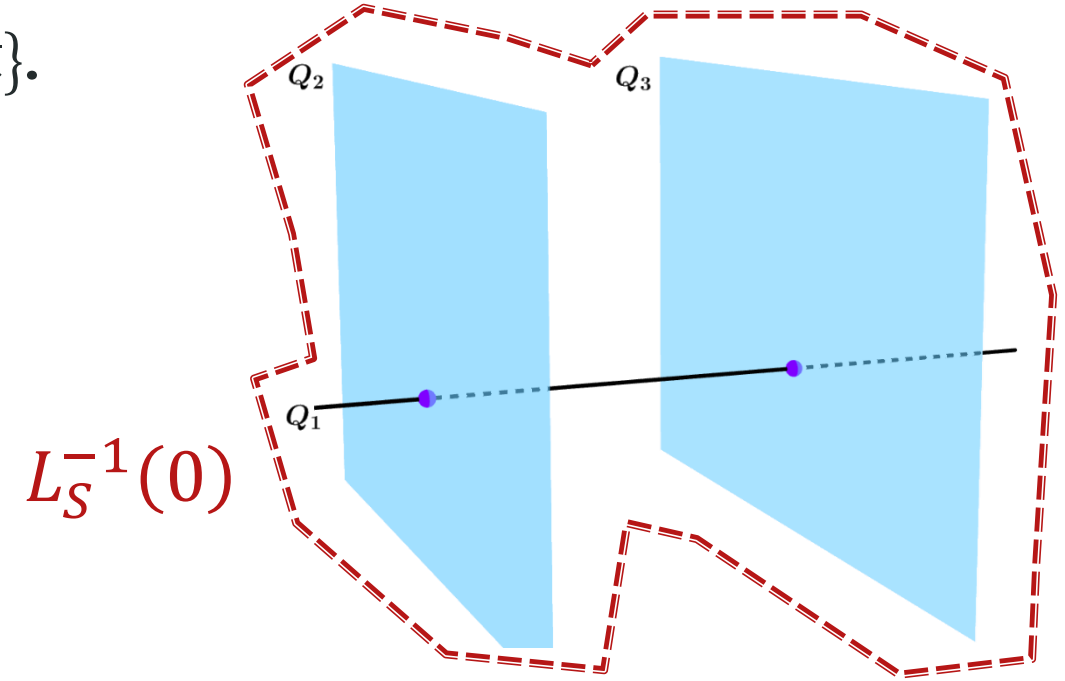
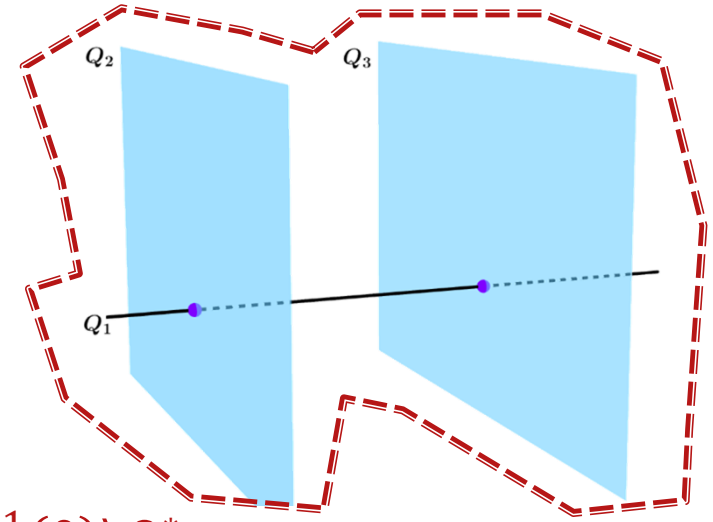


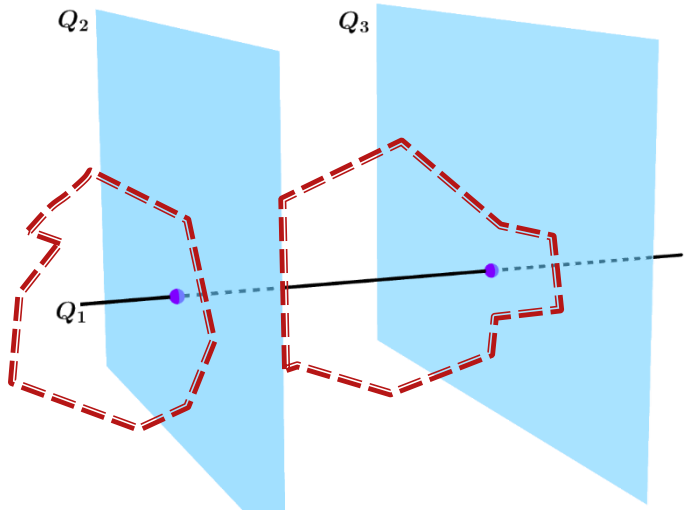
Illustration of Q^1, Q^2, Q^3

Geometry of global minima for final condensation

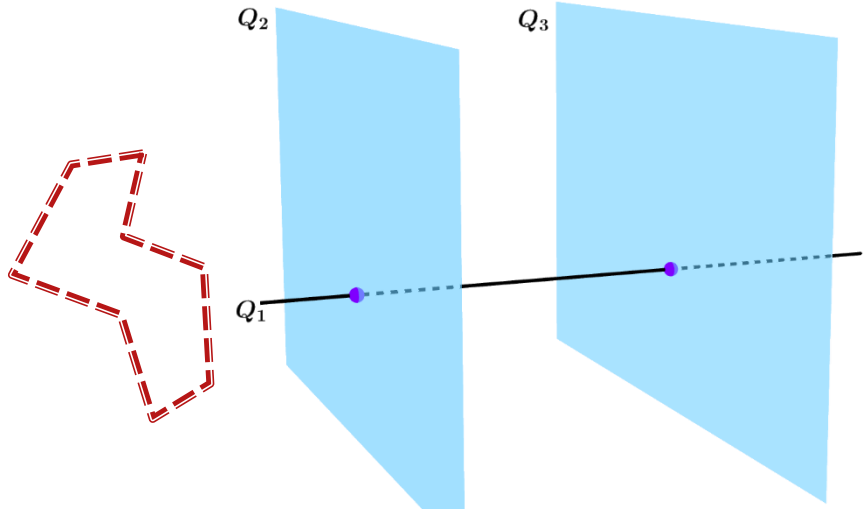


$L_S^{-1}(0)\backslash \mathrm{Q}^*$

$n < 4$

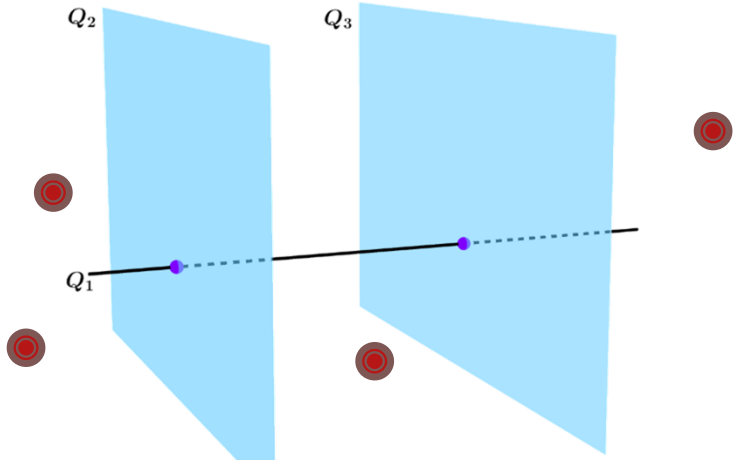


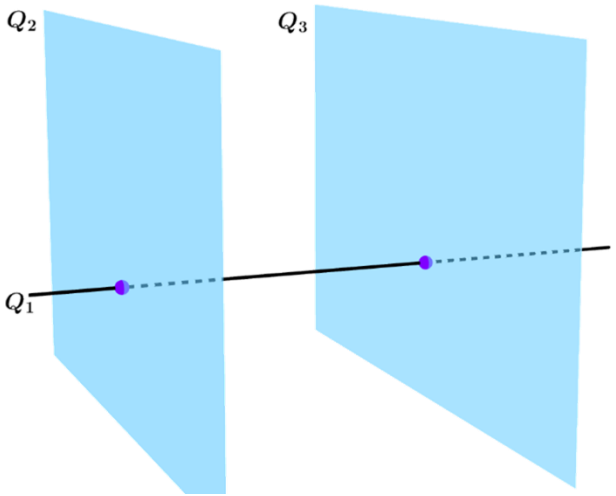
$n = 4$



$n = 5$

$n = 6$





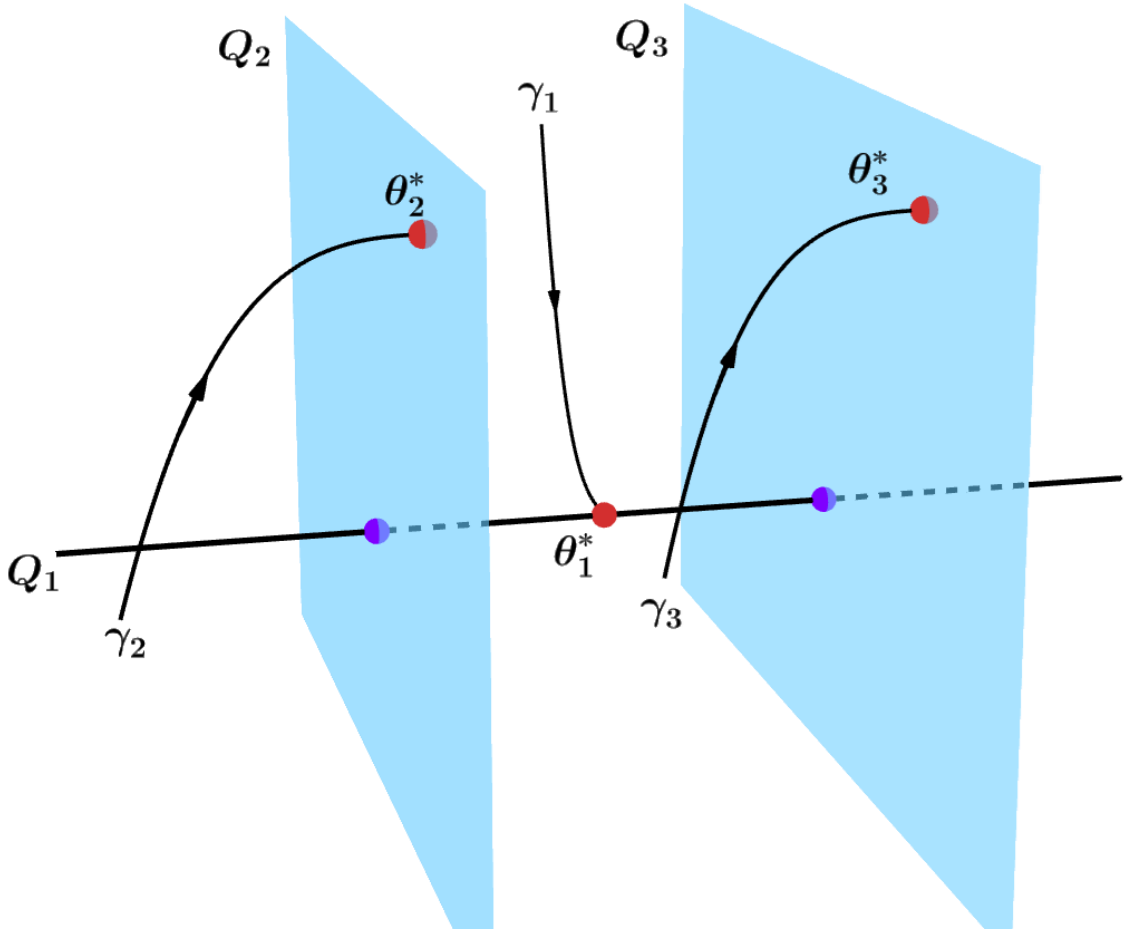
$n > 6$

Typical convergence rate for final condensation

Gradient flows near Q^*:

γ_1 : sublinear rate;

γ_2, γ_3: linear rate.



Stability of target branches underlies final condensation

Theorem 5.4 (recovery stability). *Given $m_0 \le r \le m$, partition P and permutation π and separating inputs $\{x_i\}_{i=1}^n$. Then no point in $Q^r_{P,\pi}$ is recovery stable when $n \le r + (r-l)d$ (l is the deficient number of P), and almost all points in $Q^r_{P,\pi}$ are recovery stable when $n \ge r + (m + m_0 - r)d$. Moreover, all points in Q^* are recovery stable when $n > (d+1)m$, namely, Q^* is recovery stable.*

Sample size/Branches	Q^{m_0}	...	Q^r	...	Q^m
$\le (d+1)m_0$	✗	...	✗	...	✗
$\ge m + m_0 d$					✓
⋮				...	⋮
$\ge r + (m + m_0 - r)d$			✓	...	✓
⋮		...	⋮	...	⋮
$\ge m_0 + md$	✓	...	✓	...	✓
$> (d+1)m$			$\checkmark^*$		
$\checkmark^*$: any point in Q^* is recovery stable					

Generalization advantage of condensation

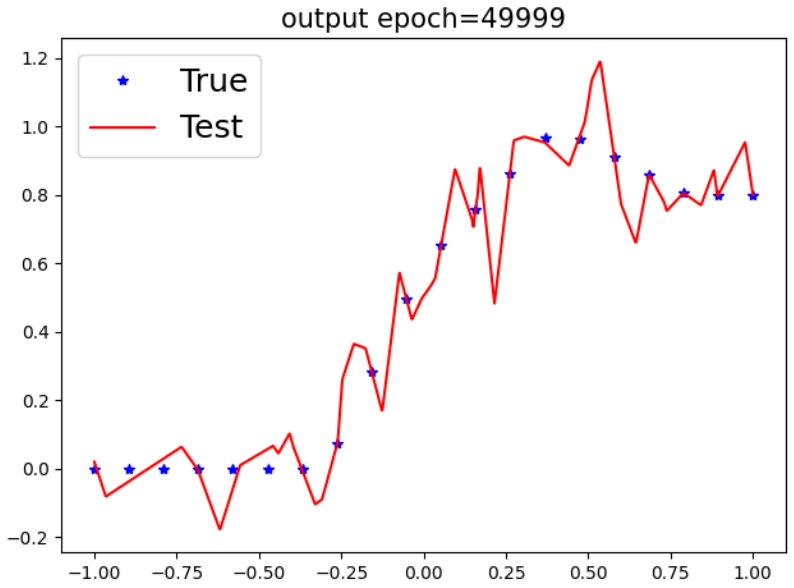
1. Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, Linear Stability Hypothesis and Rank Stratification for Nonlinear Models. arXiv:2211.11623, (2022).

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).

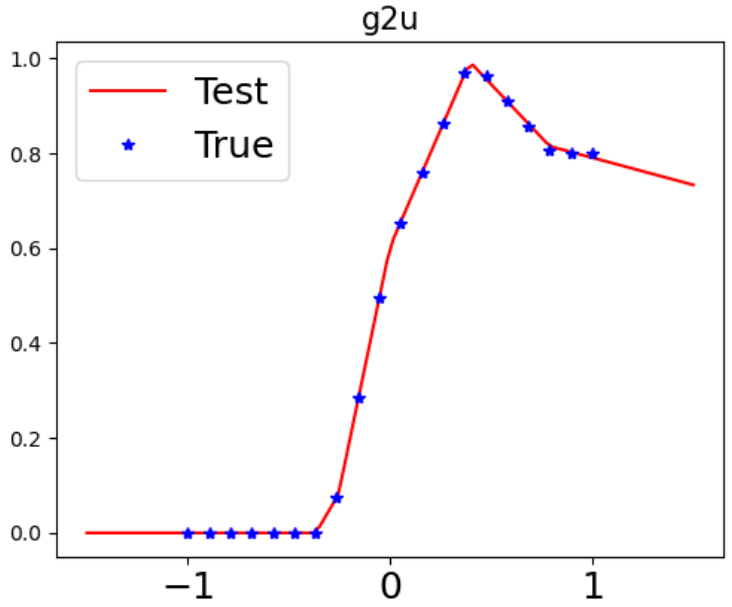
3. Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

Generalization consequence of condensation

Large initialization
(no condensation)

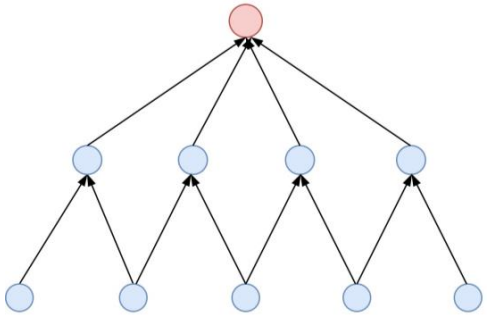


Small initialization
(Strong condensation)

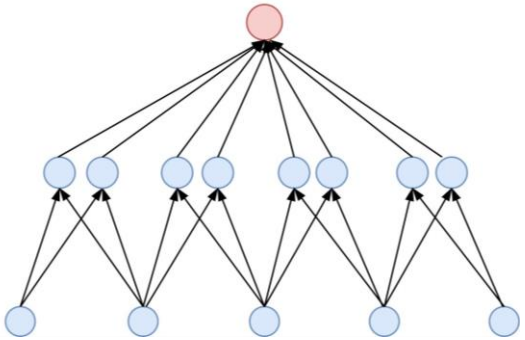


Condensation improves sample efficiency

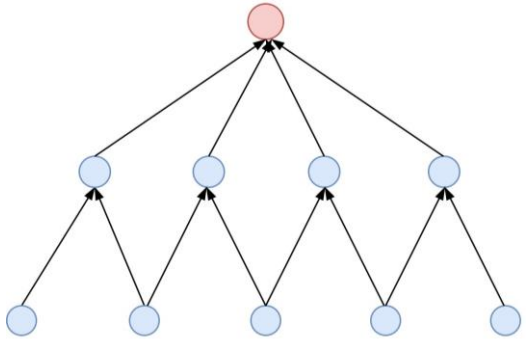
How many samples are required to recover f^* by $\mathrm{NN}_{\mathrm{wide}}$?

$f^* =$ 

optimistic sample size



condense



parameter count

12

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

$\mathrm{NN}_{\mathrm{narr}}$（equiv）

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

Quantification of condensation--model rank

Model:

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

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

lower model rank signifies stronger condensation

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**($w_1 \neq \pm w_2 \neq 0, a_1 \neq 0, a_2 \neq 0$):

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

Optimistic sample size estimate

Model:

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

Model rank:

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

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

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

$F^{-1}(f^*)$: Target set

Intuitive procedure:

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

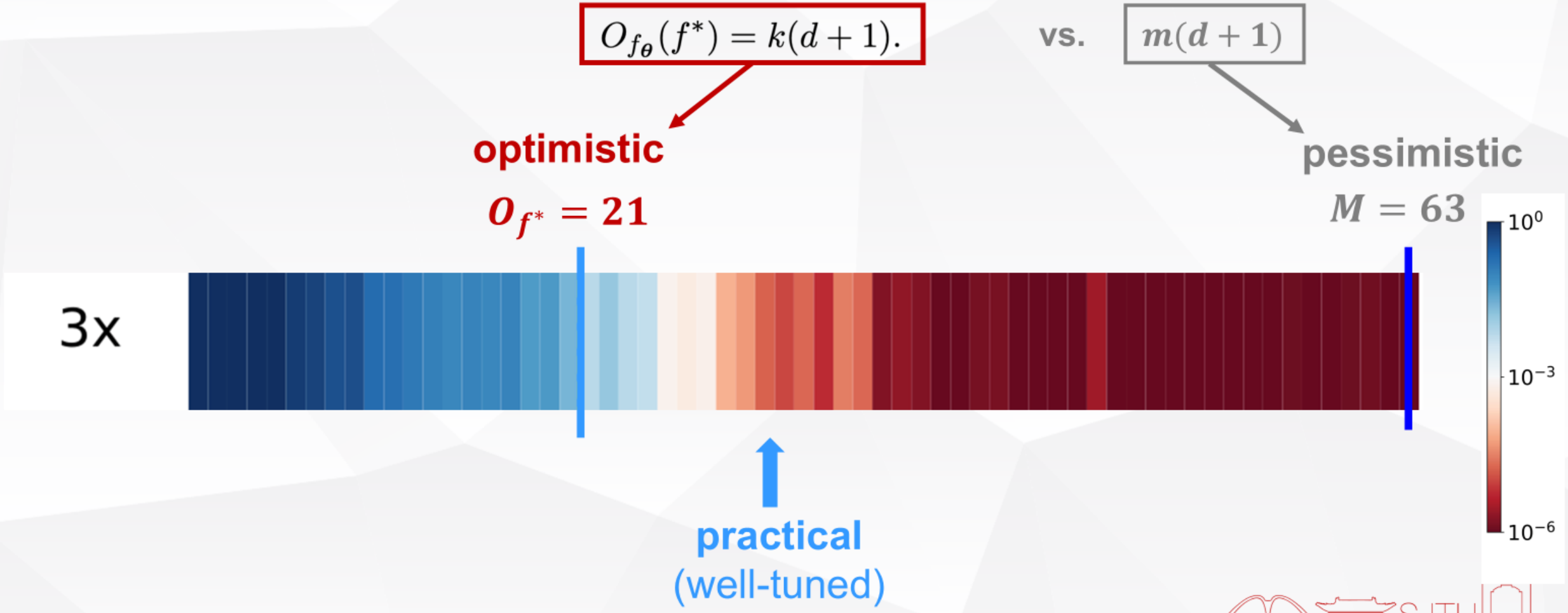
Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai,
Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. arXiv:2406.18035, (2024).

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*

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

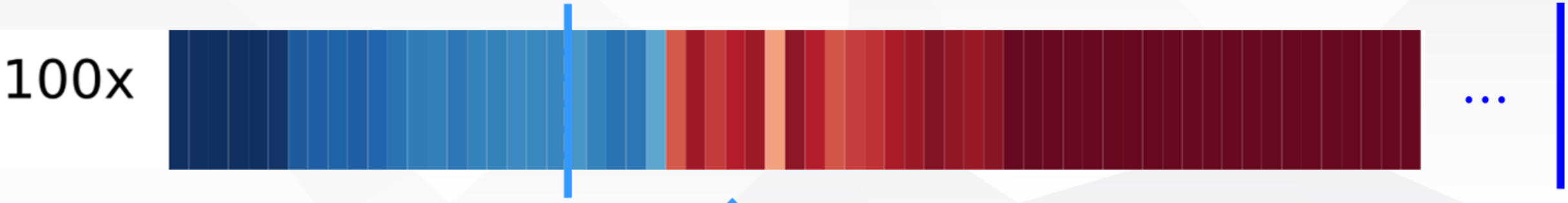
vs. $m(d+1)$

optimistic

$O_{f^*} = 21$

pessimistic

$M = 2100$



practical

(well-tuned)

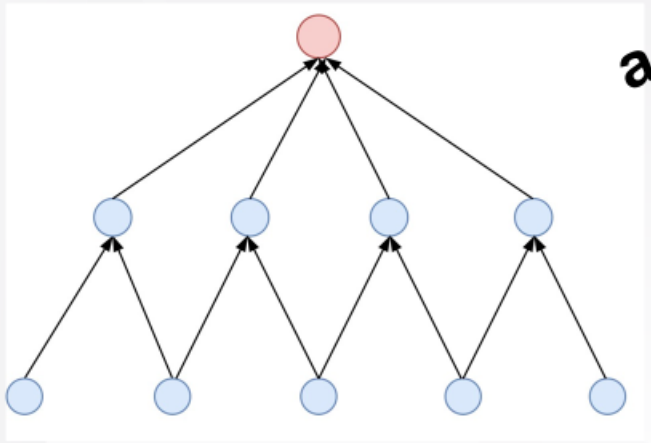
Impact of width—Deep NNs

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

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

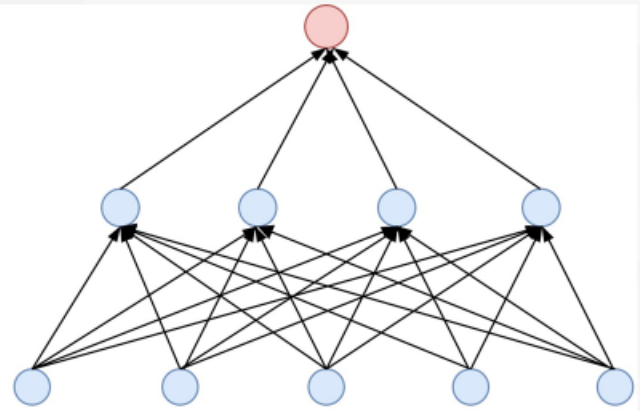
Specialty of DNN models via optimistic estimate



12

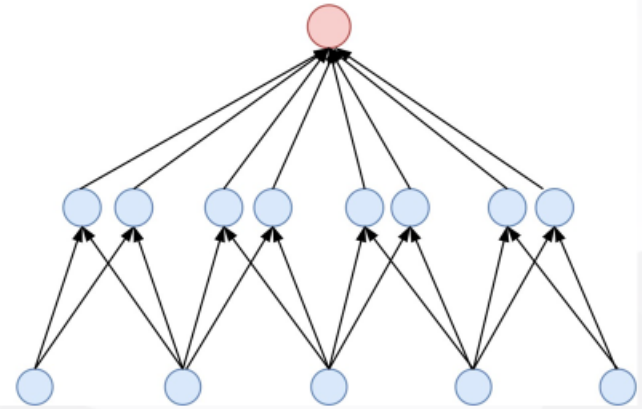
add connection

inefficient



24

add width



12

efficient

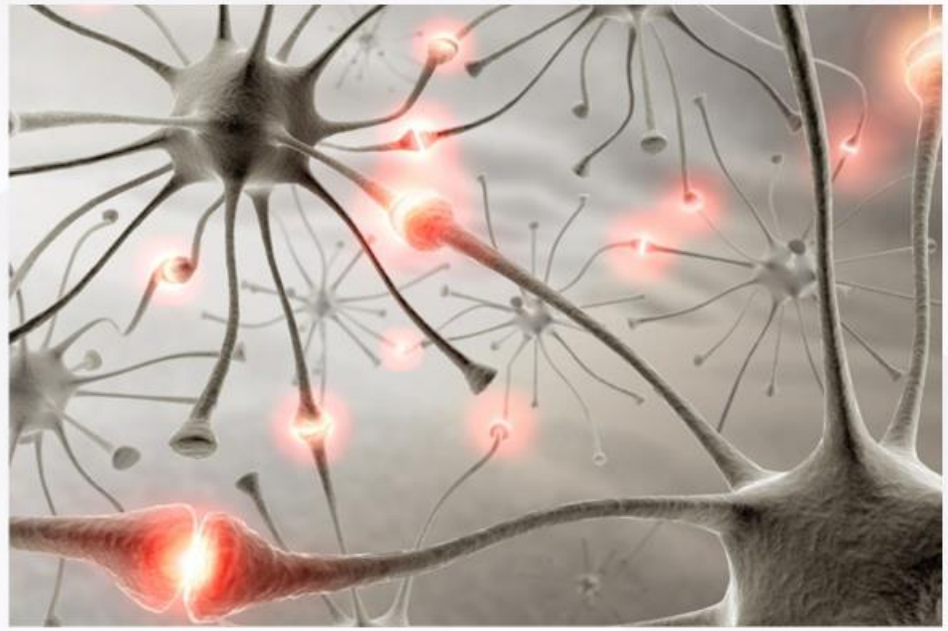
Sample inefficient:
Adding (unnecessary) connections worsens generalization

Sample efficient:
Increasing width doesn't harm generalization

Inspiration

Principle of model scaling

- Freely increase width
- Refrain from adding connection

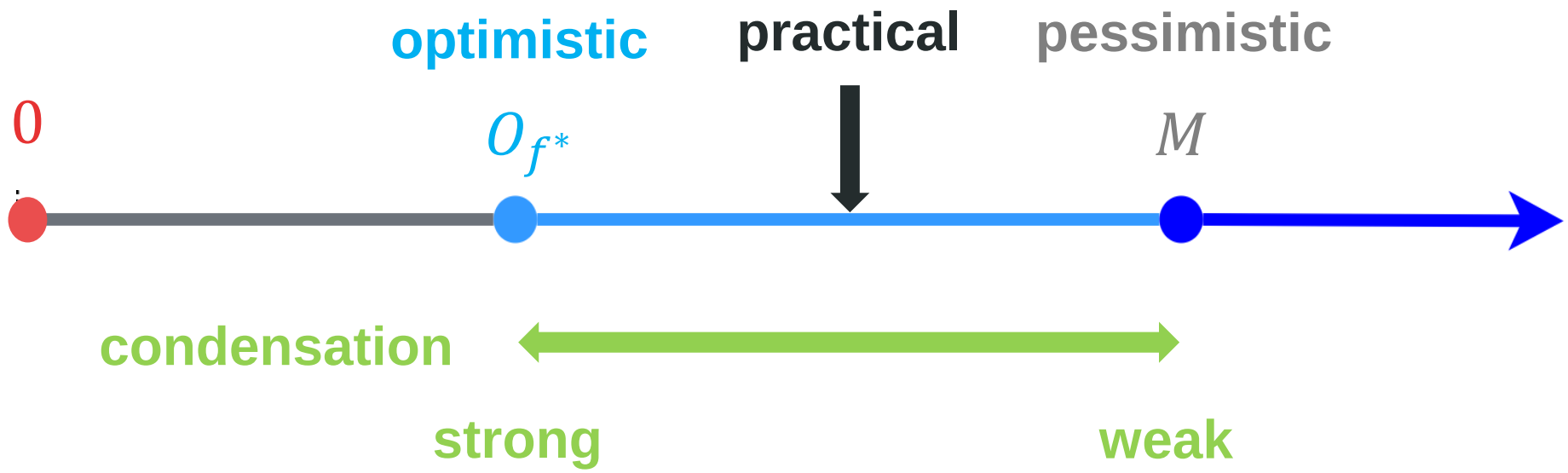


vs. Scaling of brain

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

Condensation improves sample efficiency

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



Ways to facilitate condensation:

smaller initialization, dropout, larger weight decay, large learning rate

The origin of condensation

Permutation symmetry → sample efficiency preserving

Permutation symmetry: 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

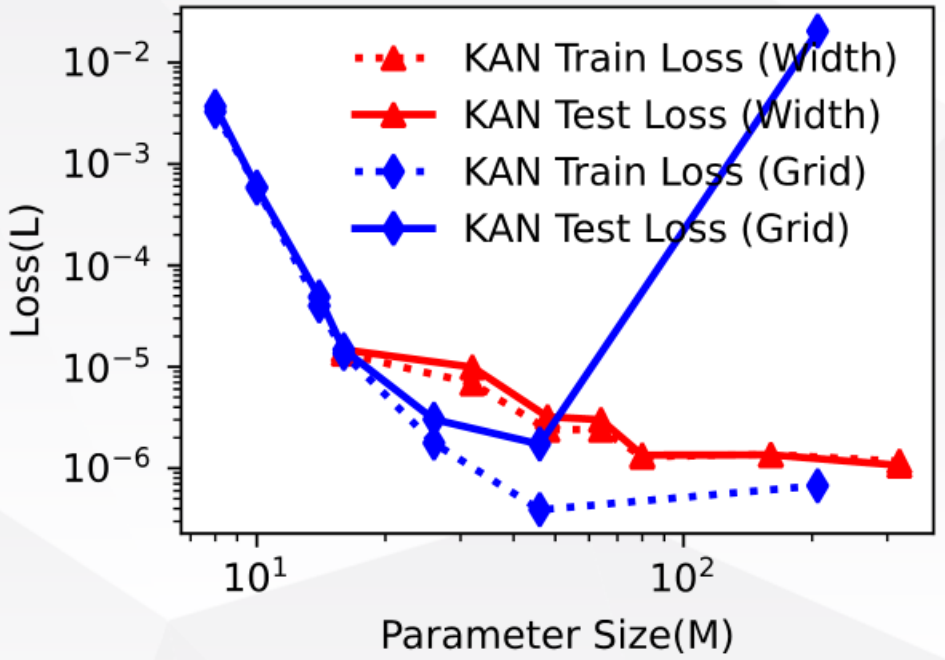
Increase m preserves sample efficiency

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

grid: no symmetry

Increase G reduces sample efficiency

Experiment



Our series works on condensation

A1. Regime of condensation—phase diagram series

1.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).

2.Hanxu Zhou, Qixuan Zhou, Zhenyuan Jin, Tao Luo, Yaoyu Zhang, Zhi-Qin John Xu, "Empirical Phase Diagram for Three-layer Neural Networks with Infinite Width," NeurIPS 2022.

A2. Loss landscape structure—embedding principle series

1.Yaoyu Zhang, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, "Embedding Principle of Loss Landscape of Deep Neural Networks," NeurIPS 2021 spotlight.

2.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.

3.Hanxu Zhou, Qixuan Zhou, Tao Luo, Yaoyu Zhang, Zhi-Qin John Xu, "Towards Understanding the Condensation of Neural Networks at Initial Training," NeurIPS 2022.

4.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.

A3. Generalization advantage—optimistic estimate series

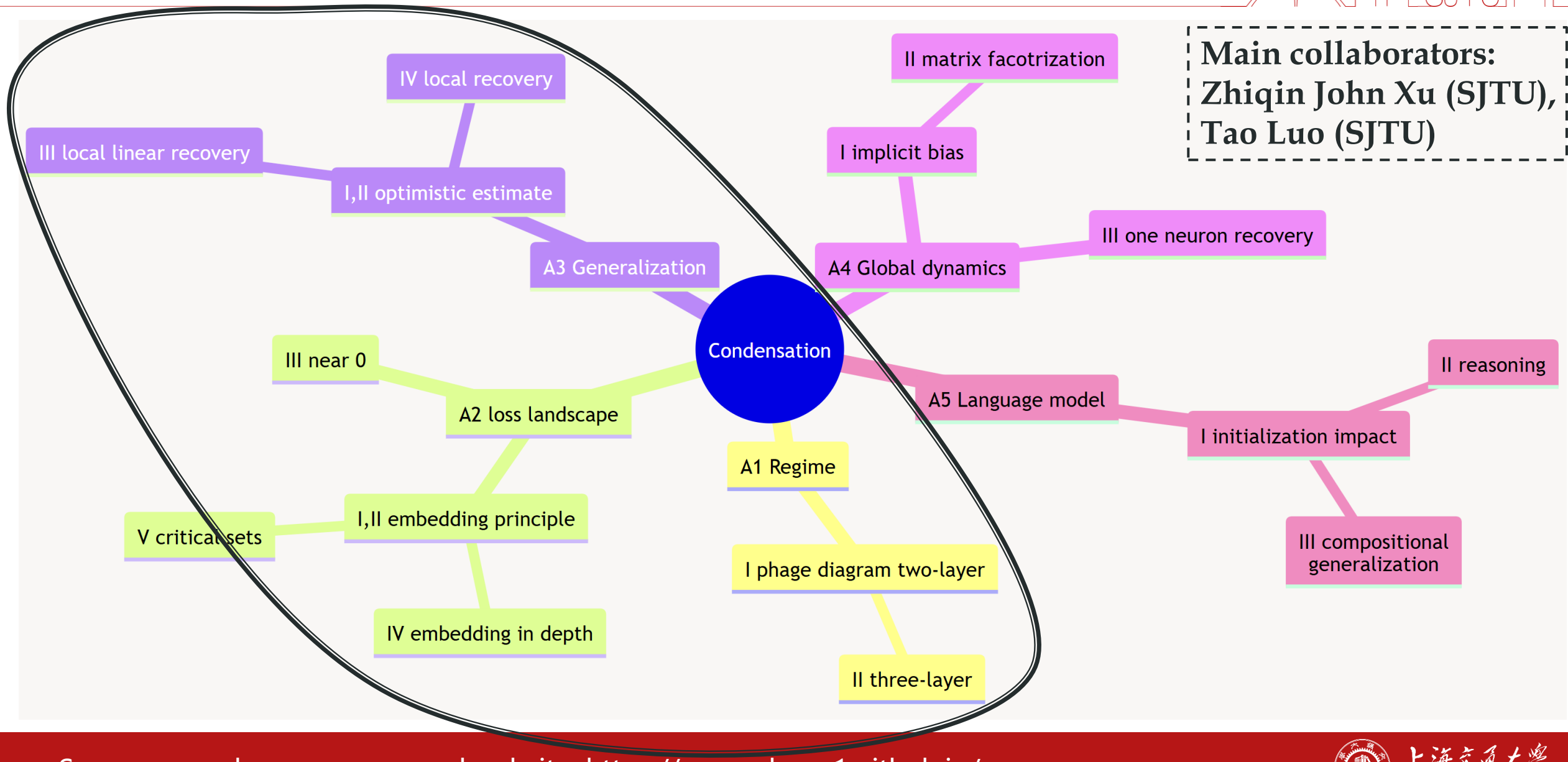
1.Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, "Linear Stability Hypothesis and Rank Stratification for Nonlinear Models," arXiv:2211.11623 (2022).

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, Zhiwei Bai, "Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization," arXiv:2406.18035 (2024).

4.Tao Luo, Leyang Zhang, Yaoyu Zhang, "Structure and Gradient Dynamics Near Global Minima of Two-layer Neural Networks," arXiv:2309.00508 (2023).

Overview of our works on condensation



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

Thanks!

饮 水 思 源　爱 国 荣 校