IV.From condensation to loss landscape analysis

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

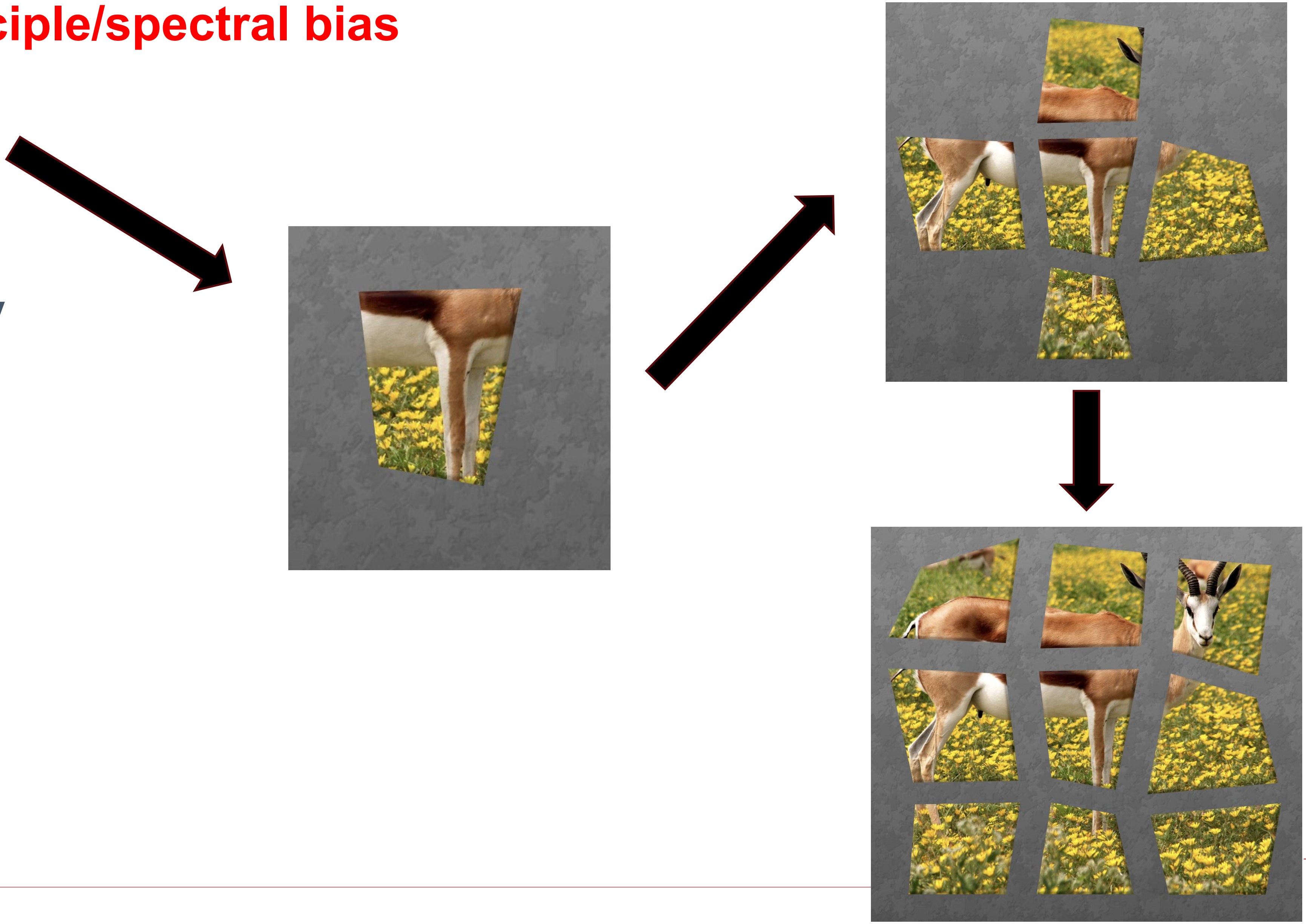
Fri. – Thu. May 2 – 8, 2025

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

Phenomenon as a key family of facts to uncover

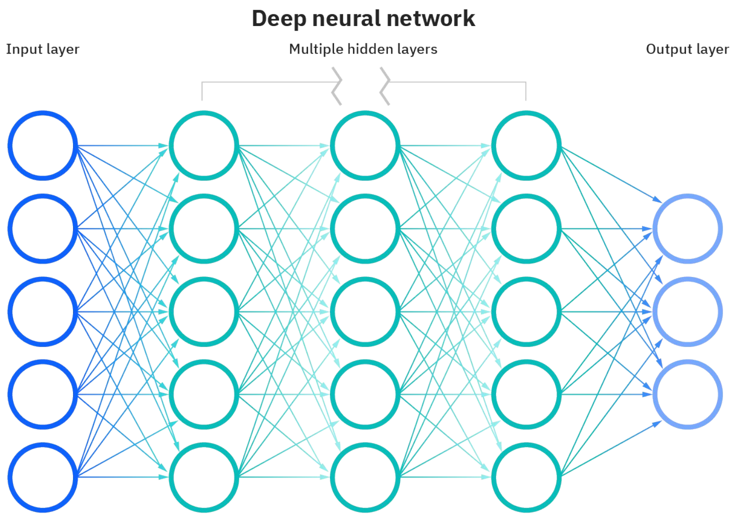
- Frequency principle/spectral bias
- Condensation
- Double descent
- Edge of stability
- Lottery ticket
- Neural collapse
- Grokking
- ……

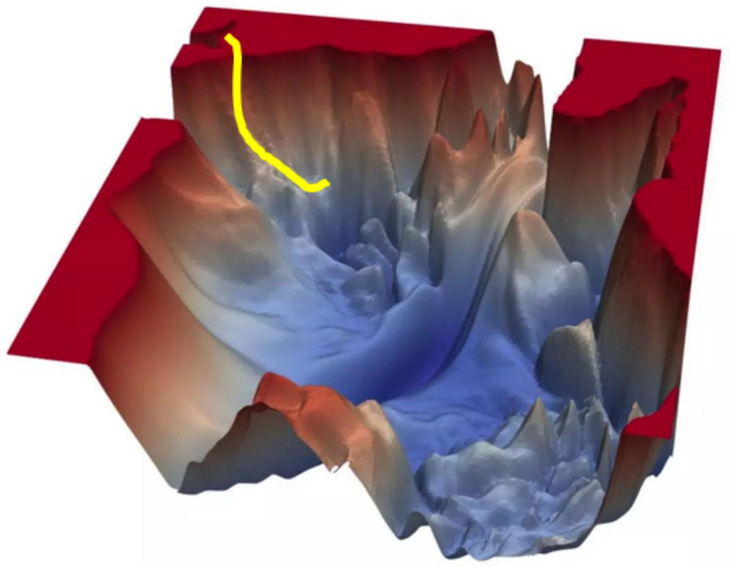


Deep learning loss landscape

$$R_S(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{f}(\boldsymbol{x}_i, \theta), \boldsymbol{y}_i)$$

Model: $\boldsymbol{f}(\boldsymbol{x}_i, \theta)$
Data: $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{n}$
Loss: $\ell(\cdot, \cdot)$



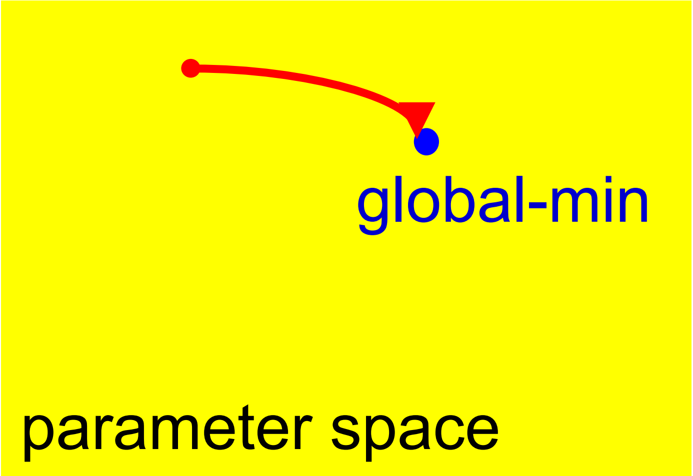


Role of loss landscape (conventional ML)

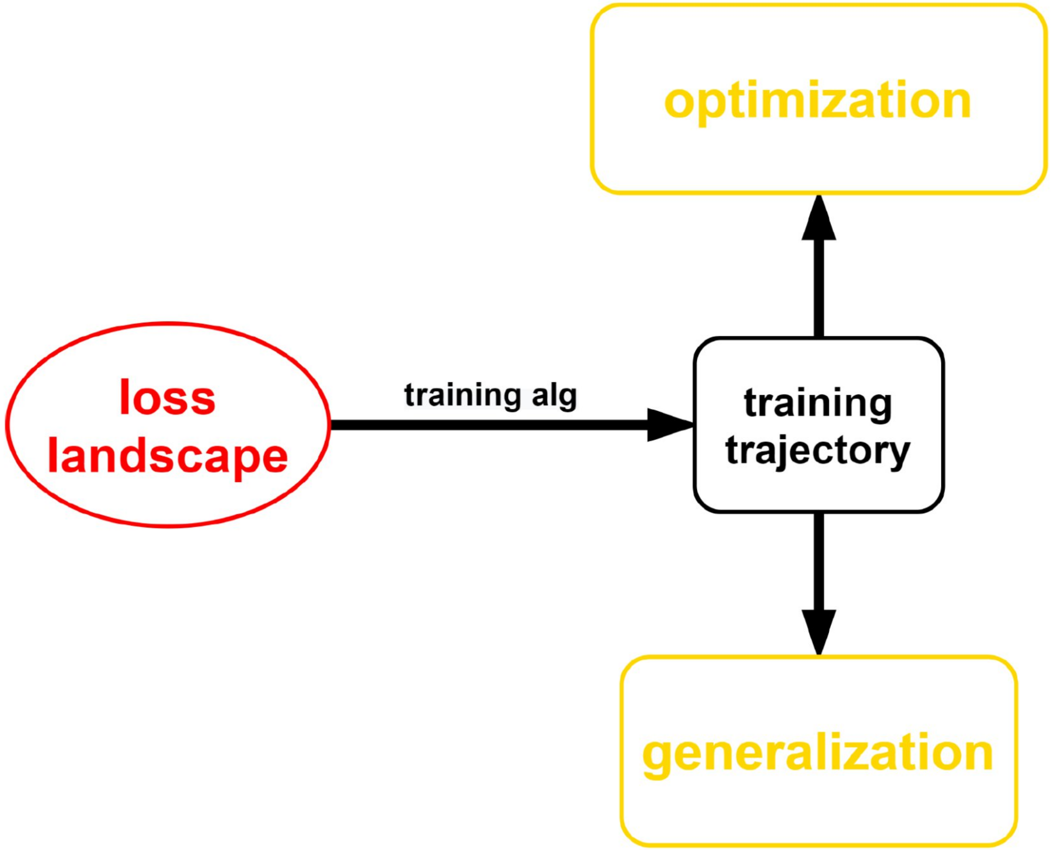
loss landscape $\xrightarrow{\text{training alg}}$ optimization

solution /minima $\longrightarrow$ generalization

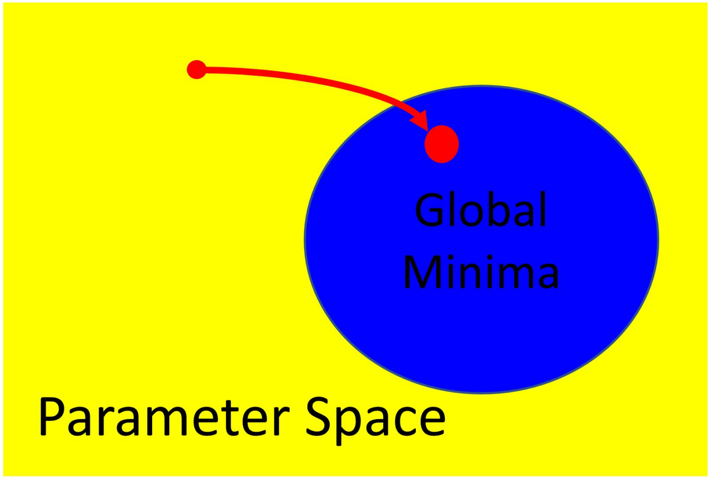
$$\#\text{param} \leq \#\text{data}$$



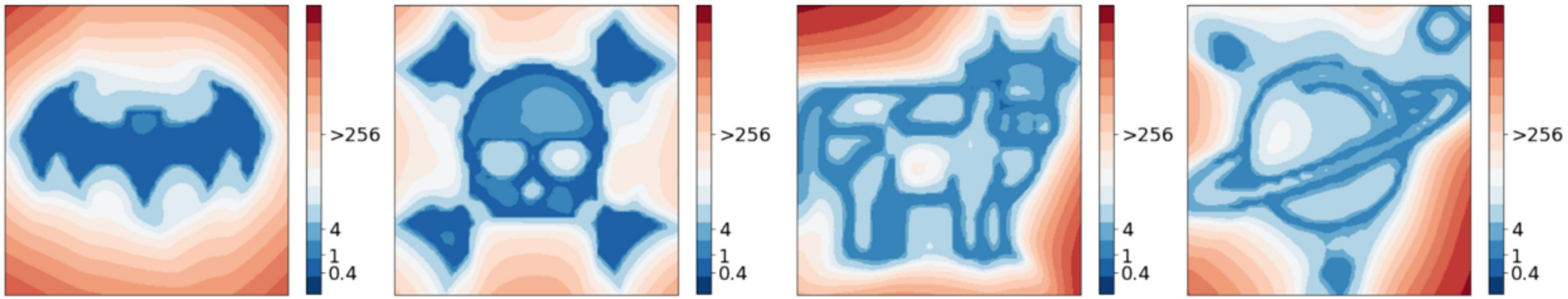
Role of loss landscape (deep learning)



#param > #data



DNN loss landscape is complex



(a) Loss surface on FashionMNIST dataset

(b) Loss surface on CIFAR10 dataset

I. Skorokhodov, M. Burtsev, 2019

Which picture captures NN loss landscape?

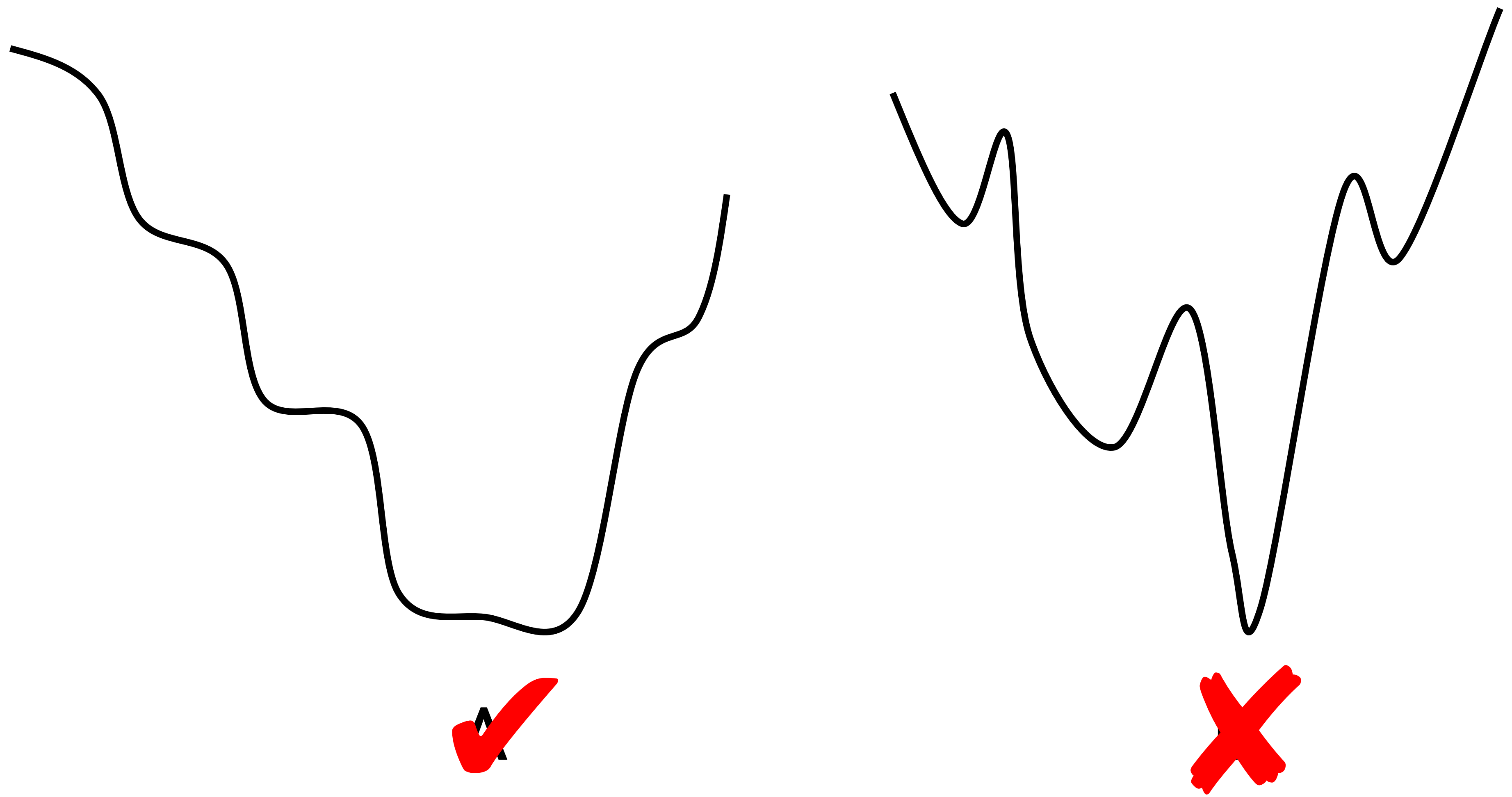
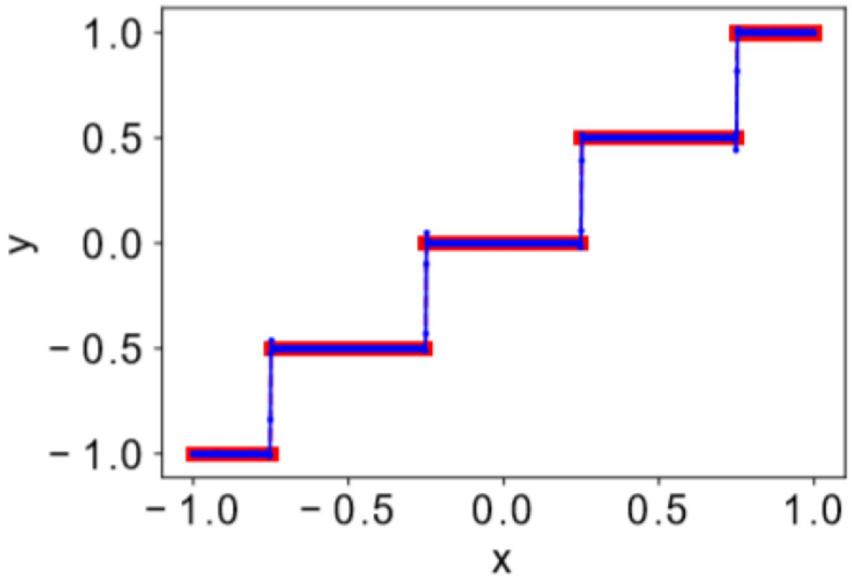
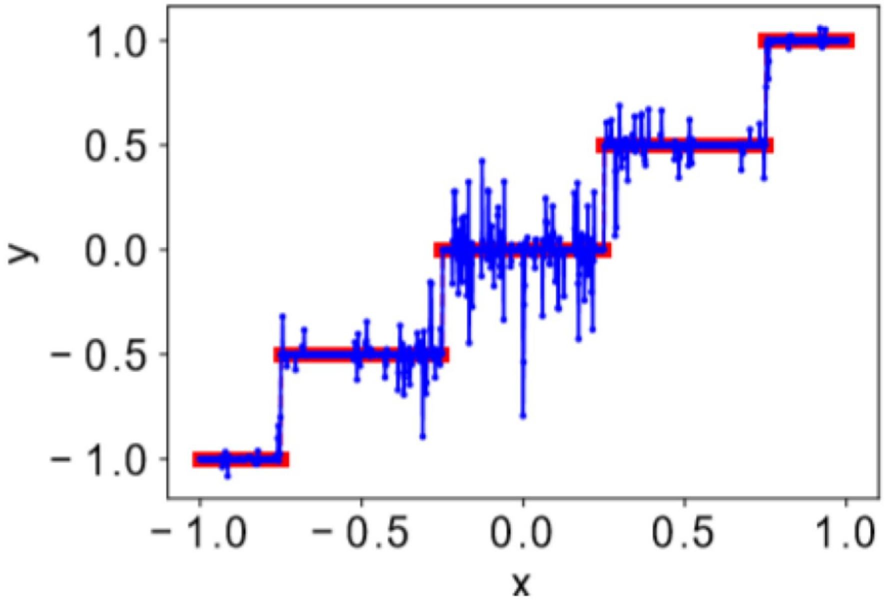


Illustration of different global-minima





Loss landscape structure underlying condensation

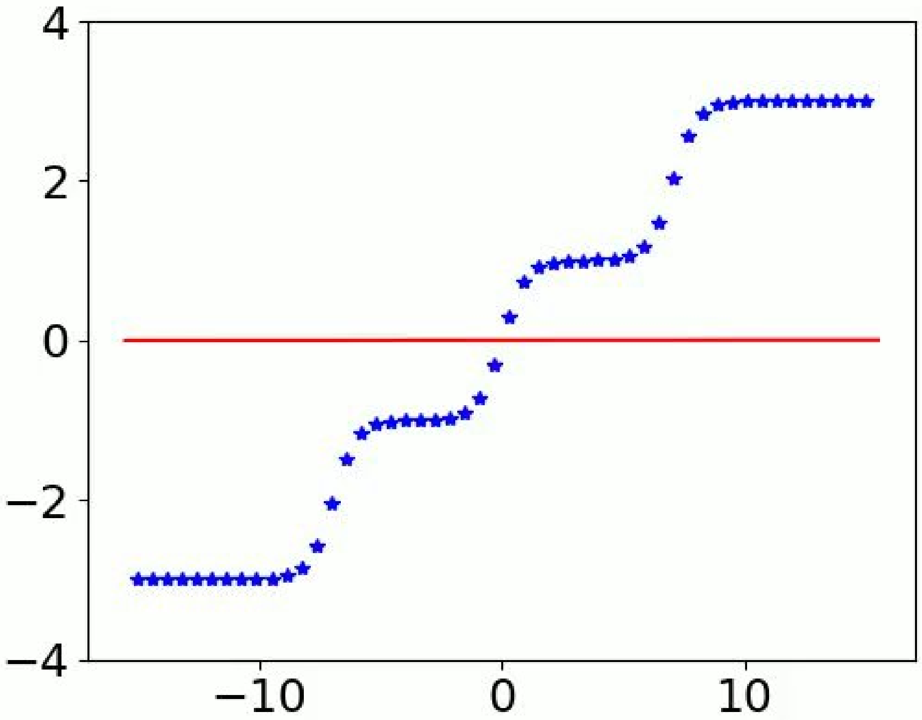
1.Yaoyu Zhang, Zhongwang Zhang, Tao Luo, Zhi-Qin John Xu, “Embedding Principle of Loss Landscape of Deep Neural Networks,” NeurIPS 2021 spotlight.

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.

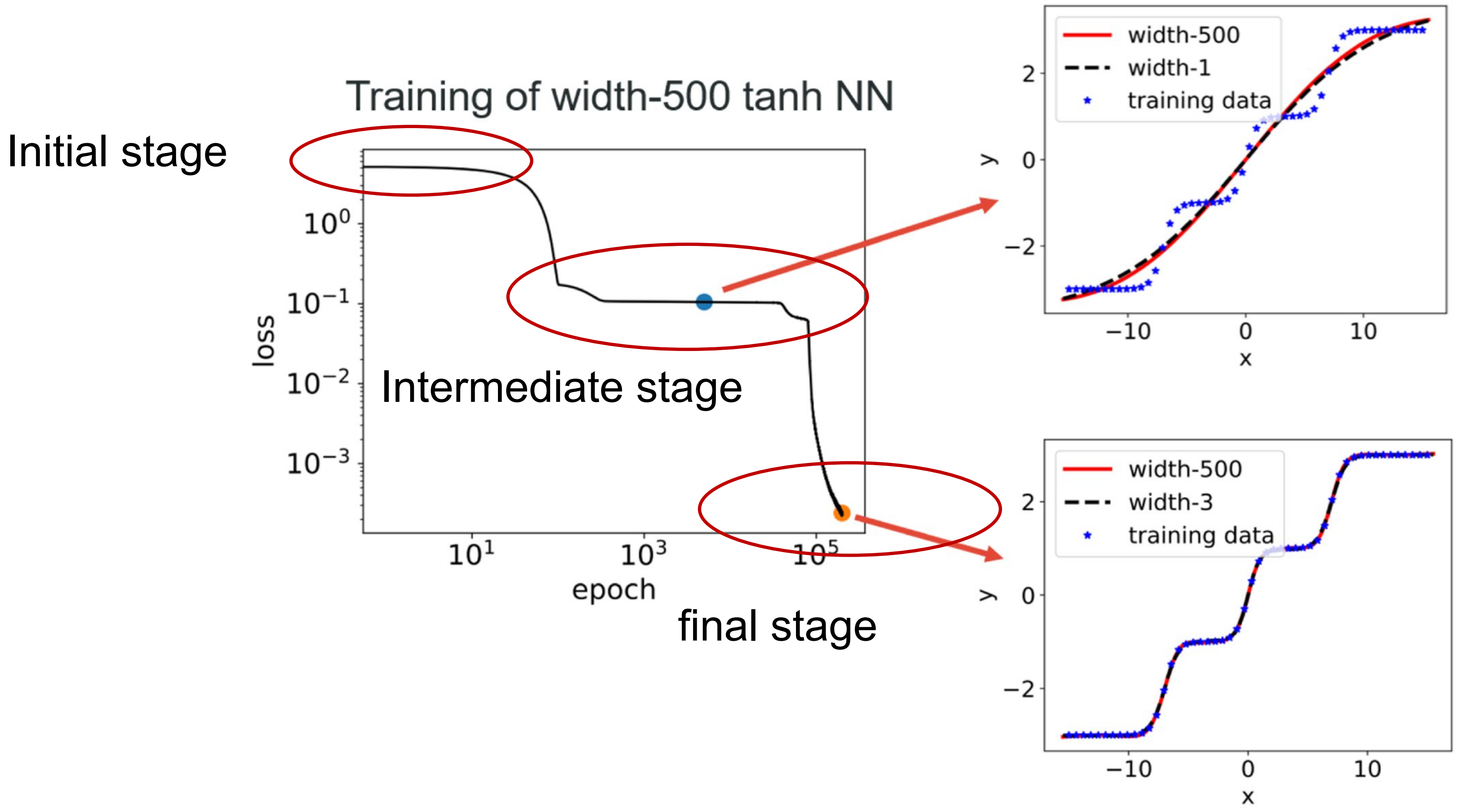
4.Tao Luo, Leyang Zhang, Yaoyu Zhang, “Structure and Gradient Dynamics Near Global Minima of Two-layer Neural Networks,” arXiv:2309.00508 (2023).

Typical training behavior with strong condensation



Width-500 tanh-NN (~1500 parameters)

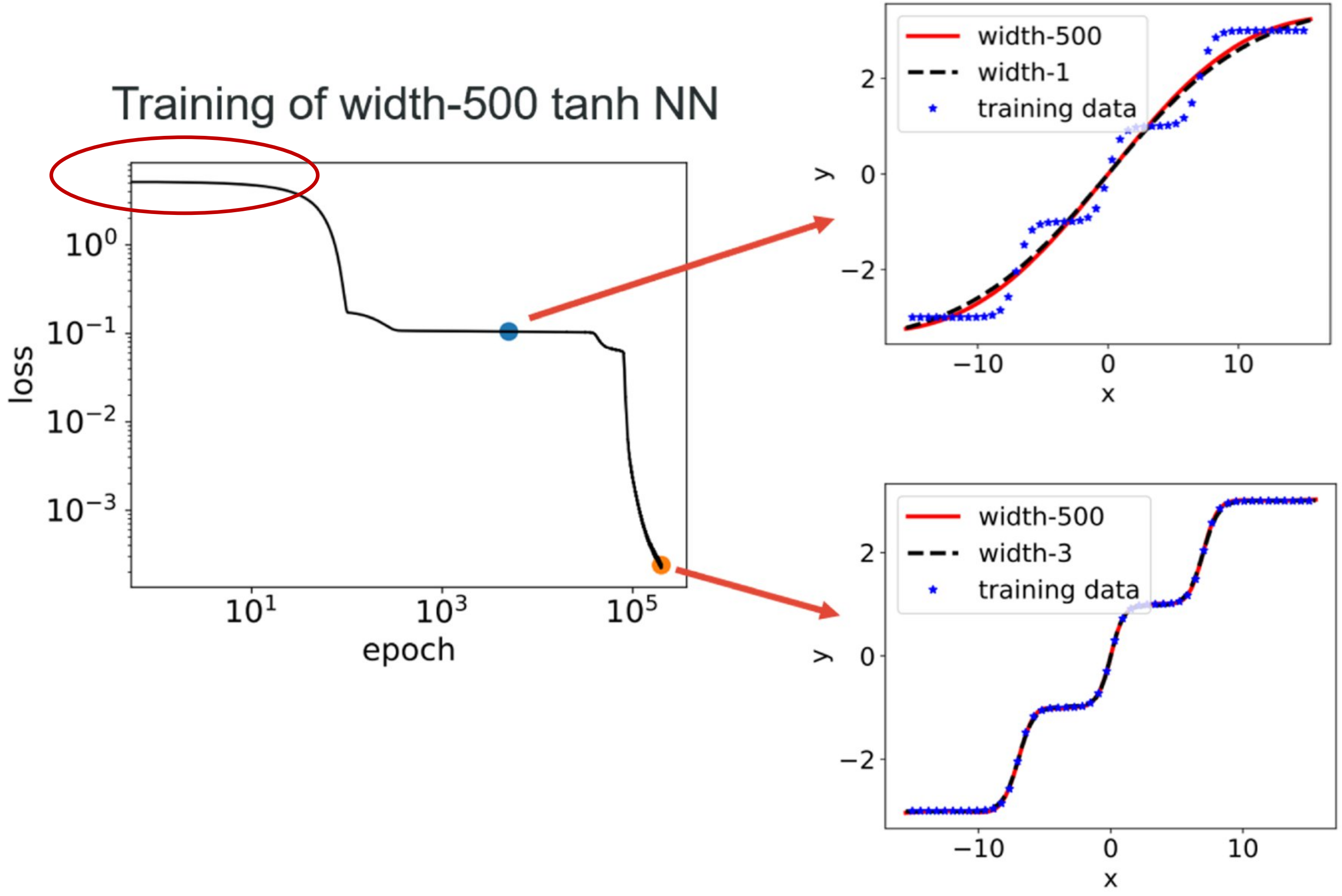
Trajectory of training loss



Initial condensation

Initial condensation

Initial stage



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

Loss landscape around 0 and Initial condensation

$$\dot{w}_j = \sum_{i=1}^{m} (y_i - f_\theta(x_i))\, a_j \sigma'\left(w_j^{\mathrm{T}} x_i\right) x_i$$

When $\theta \approx 0$, then $f_\theta(\cdot) \approx 0(\cdot)$:

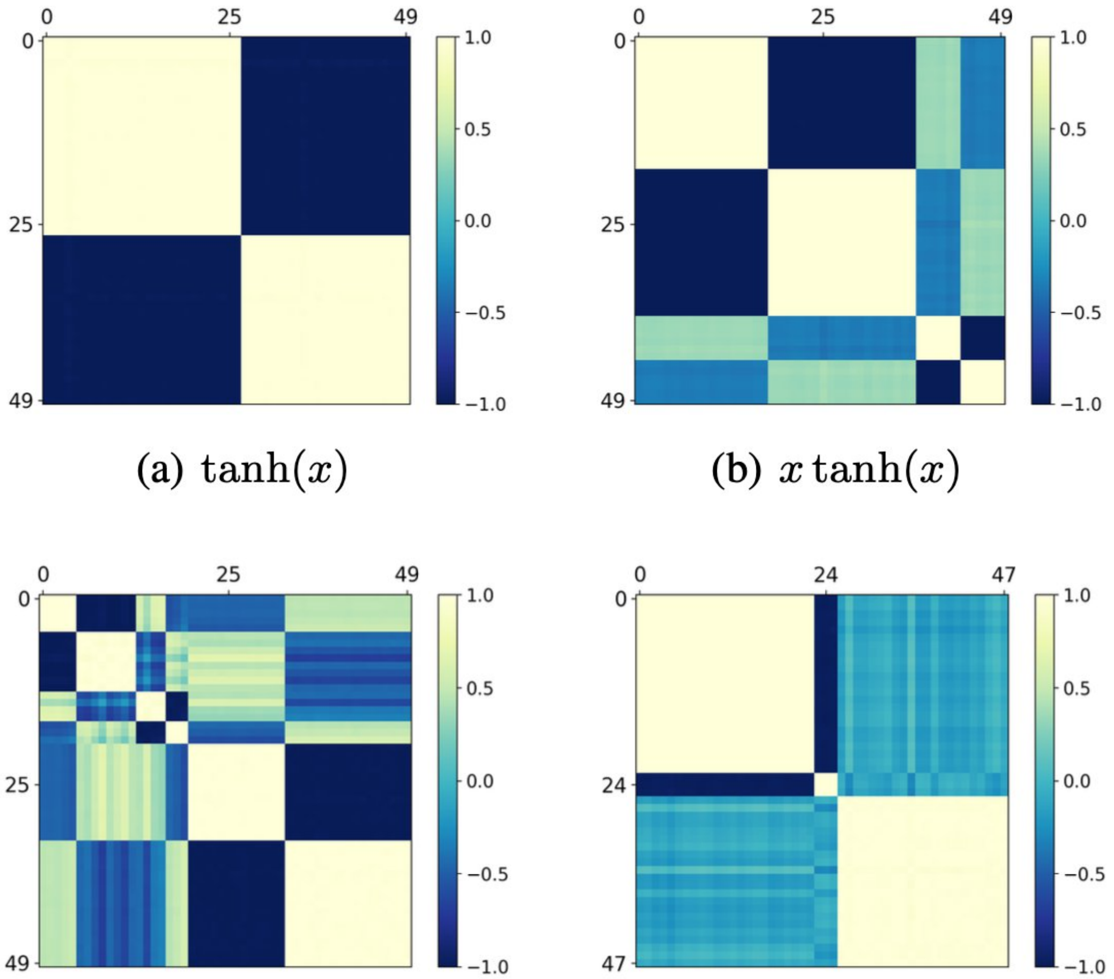
$$\dot{w}_j \approx a_j \sum_{i=1}^{m} y_i \sigma'\left(w_j^{\mathrm{T}} x_i\right) x_i$$

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

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

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

ii. **2 limiting directions: $\pm\sum_{i=1}^{m} y_i x_i$.**



(a) $\tanh(x)$ (b) $x\tanh(x)$

(c) $x^2\tanh(x)$ (d) $\mathrm{ReLU}(x)$

Intermediate condensation and embedding principle

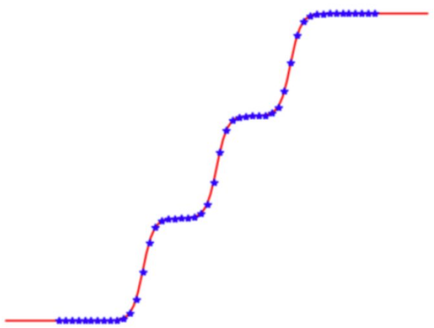
Exercise

Target: $f^*(x) = \tanh(x-7) + \tanh(x) + \tanh(x+7)$

Data: $S = \{(x_i, f^*(x_i))\}_{i=1}^{50}$

Model : $f_\theta(x) = \sum_{j=1}^{m} a_j \tanh(w_j x + b_j)$ $\quad (\theta = [a_j, w_j, b_j]_{j=1}^{m})$

Loss: $R_S(\theta) = \frac{1}{50}\sum_{i=1}^{50} |f_\theta(x_i) - f^*(x_i)|^2$



m=3,

1. Dimension of $R_S(\theta)$?
2. Existance of zero loss global minima?
3. Is $R_S(\theta)$ convex?
4. Are there non-global critical points? How many?
5. How many critical functions $\mathscr{F}^c := \{f_\theta(\cdot) | \nabla R_S(\theta) = 0\}$?
   What are they?

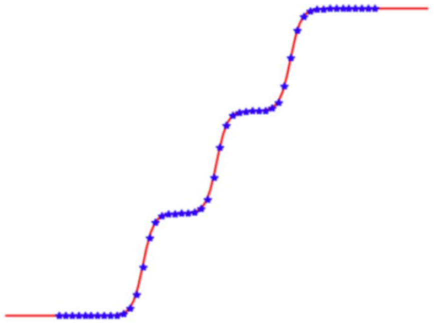
Answer

Target: $f^*(x) = \tanh(x-7) + \tanh(x) + \tanh(x+7)$

Data: $S = \{(x_i, f^*(x_i))\}_{i=1}^{50}$

Model : $f_\theta(x) = \sum_{j=1}^{m} a_j \tanh(w_j x + b_j)$ $(\theta = [a_j, w_j, b_j]_{j=1}^{m})$

Loss: $R_S(\theta) = \frac{1}{50}\sum_{i=1}^{50} |f_\theta(x_i) - f^*(x_i)|^2$

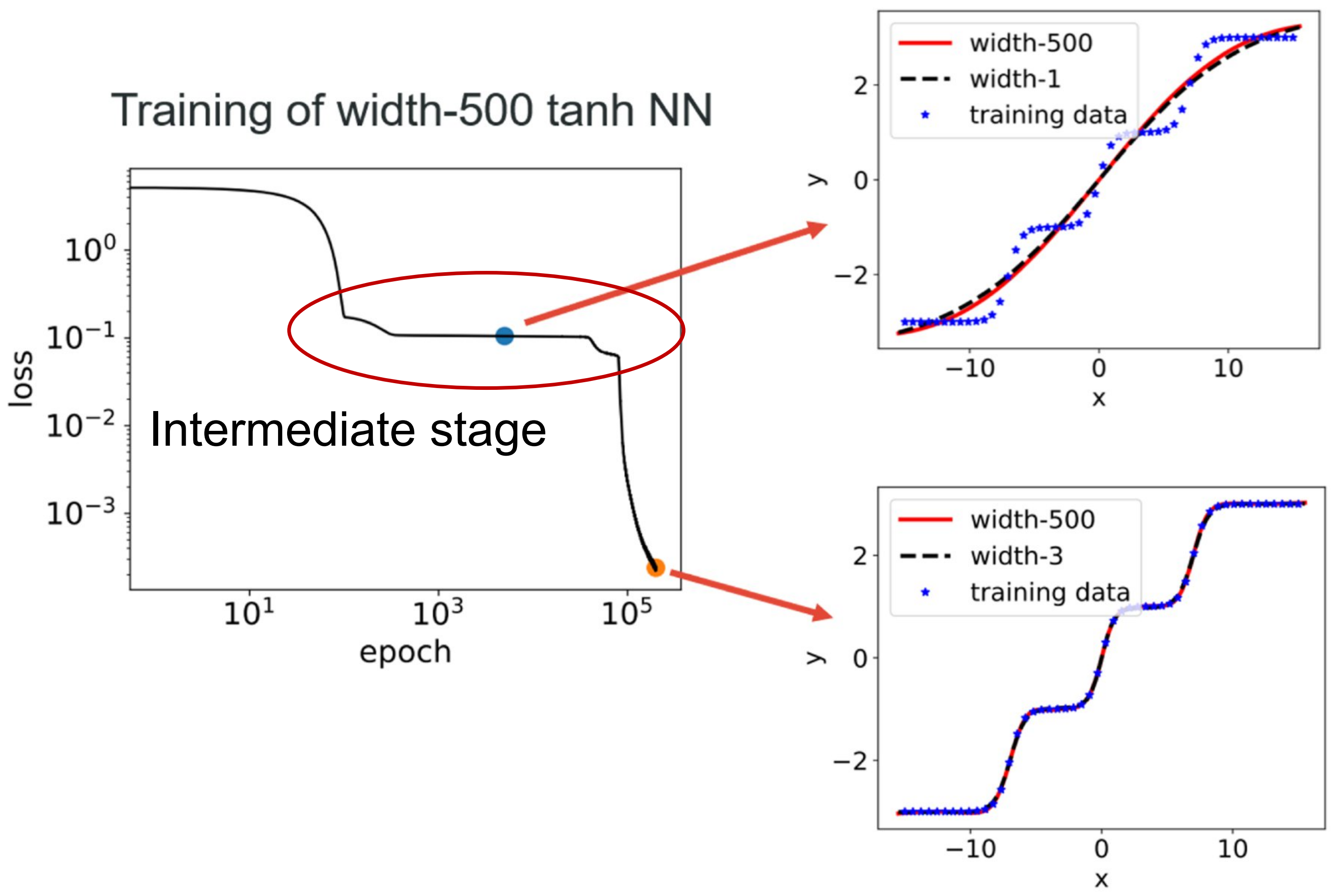


m=3,

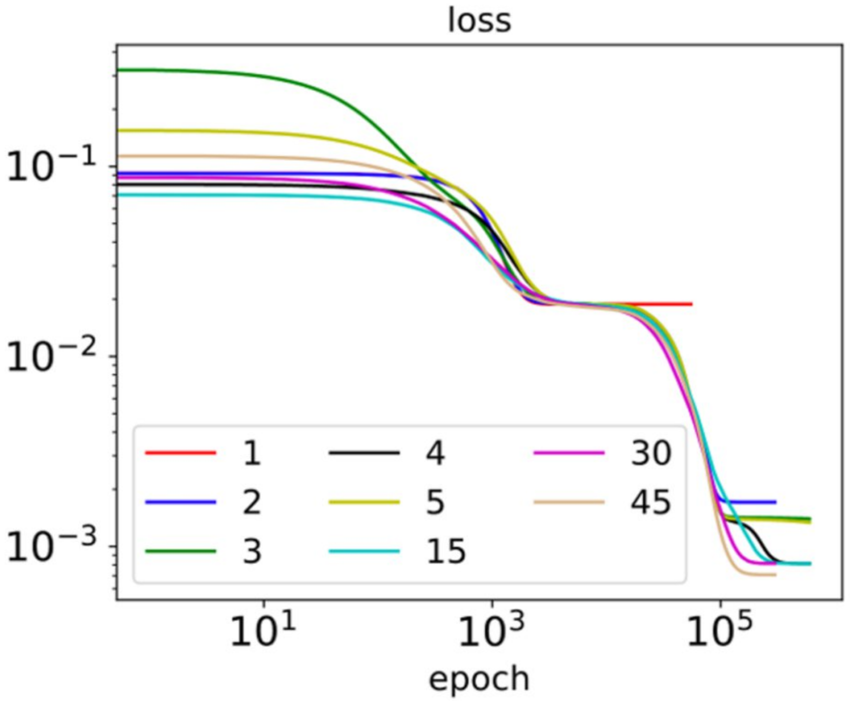
1. Dimension of $R_S(\theta)$? 9
2. Existance of zero loss global minima? exist
3. Is $R_S(\theta)$ convex? no
4. Are there non-global critical points? How many? infinite
5. How many critical functions $\mathscr{F}^c := \{f_\theta(\cdot) | \nabla R_S(\theta) = 0\}$? What are they? ≥ 5

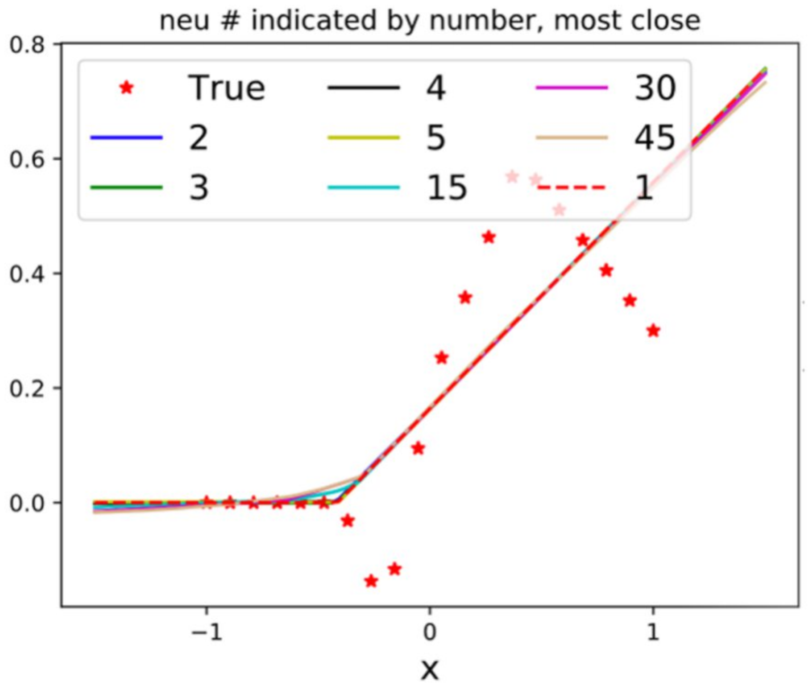
SJTU

Intermediate condensation



Training similarity of NNs with different widths





Embedding Principle in width

Embedding Principle

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

Theorem

Critical functions of narrow NNs are critical functions of any wider NNs, i.e.,

$$\mathscr{F}^{\mathbf{c}}_{\mathbf{narr}} \subset \mathscr{F}^{\mathbf{c}}_{\mathbf{wide}},$$

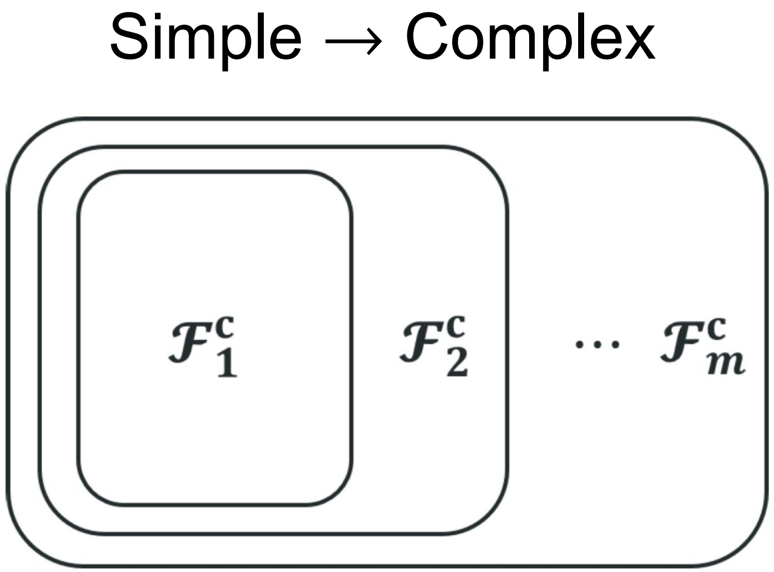
where $\mathscr{F}^{\mathrm{c}} := \{f_\theta(\cdot)|\nabla R_S(\theta) = 0\}$.

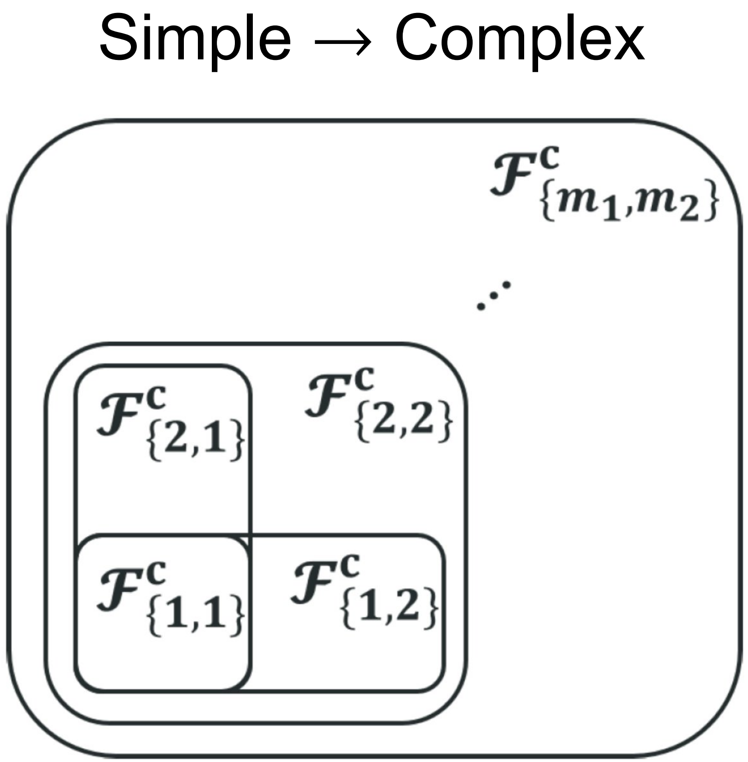
Implication

"simple" critical points always exist!

[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.* Journal of Machine Learning, 2022.

hierarchical structure of DNN loss landscape





Exercise

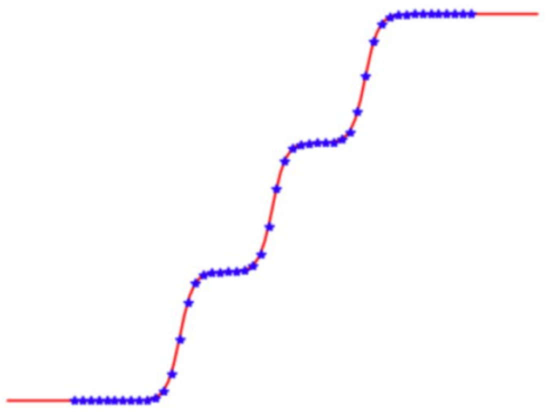
Objective: $f^*(x) = \tanh(x-7) + \tanh(x) + \tanh(x+7)$

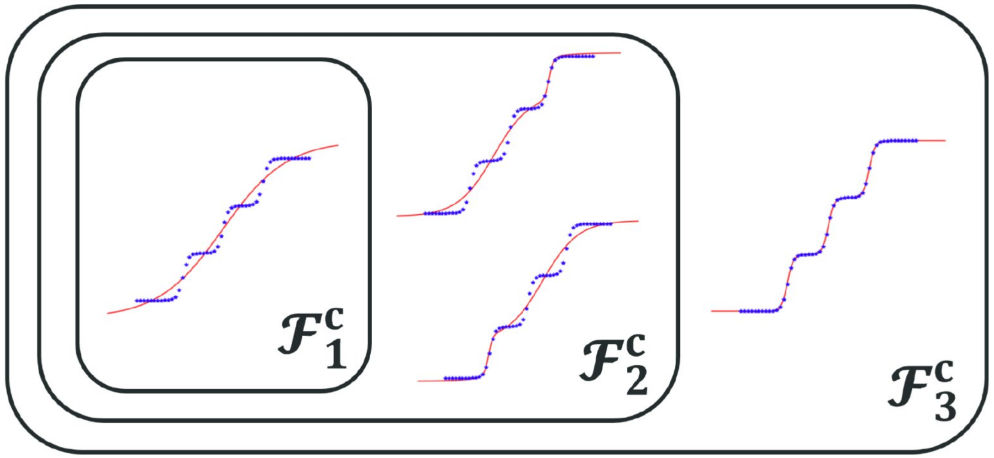
Data : $S = \{(x_i, f^*(x_i))\}_{i=1}^{50}$

Model : $f_\theta(x) = \sum_{j=1}^{3} a_j \tanh(w_j x + b_j)$ $\quad (\theta = [a_j, w_j, b_j]_{j=1}^{3})$

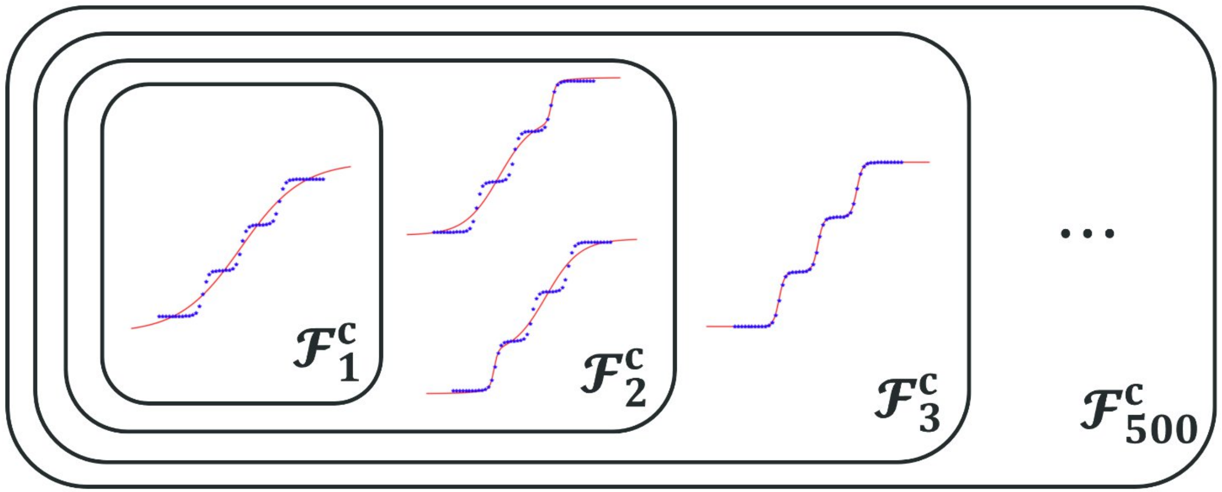
Loss landscape : $R_S(\theta) = \frac{1}{50}\sum_{i=1}^{50} |f_\theta(x_i) - f^*(x_i)|^2$

Critical functions $\mathscr{F}^c := \{f_\theta(\cdot) \mid \nabla R_S(\theta) = 0\}$?

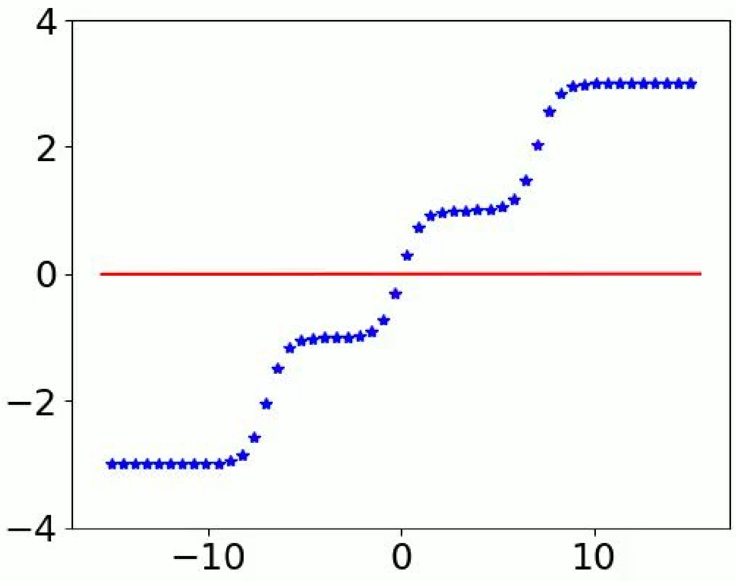




Example: analysis of hierarchical structure



500 tanh neuron



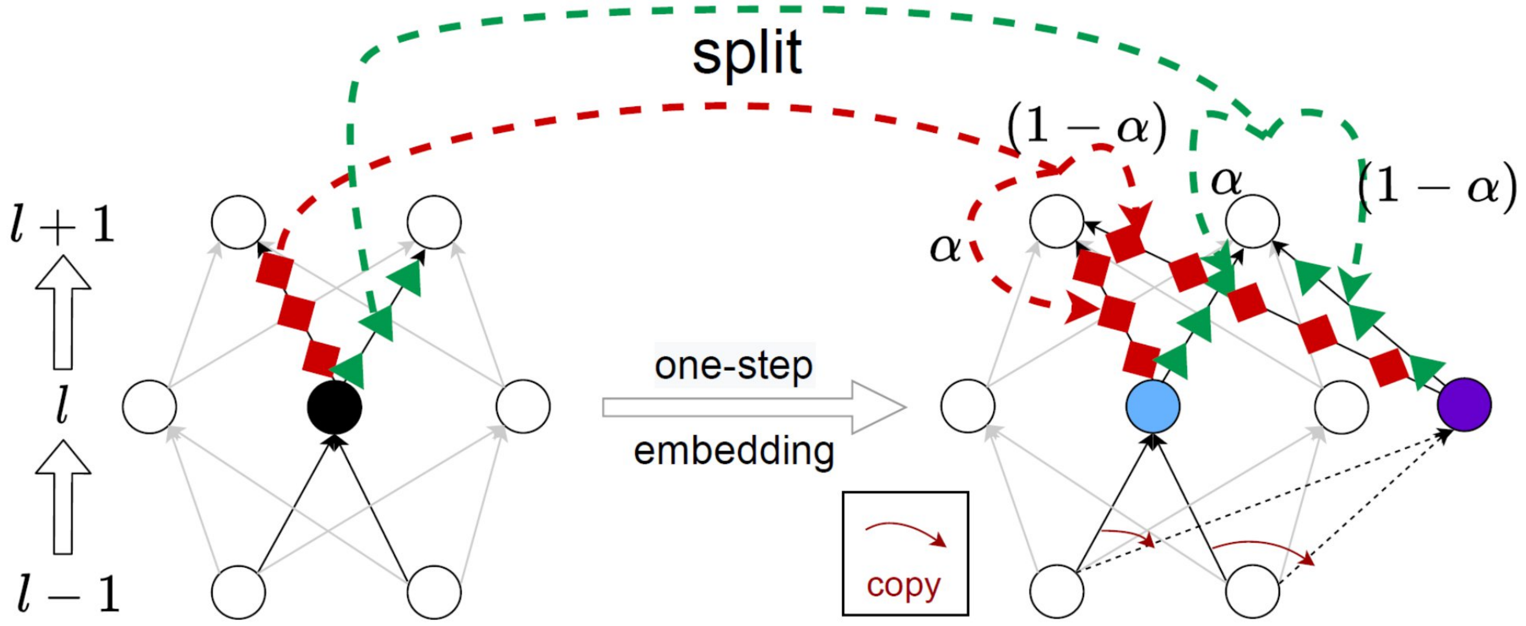
Key to the proof of Embedding Principle

Key: discovering critical embedding

Discover embedding $\mathcal{T} : \mathbb{R}^{M_{\rm narr}} \to \mathbb{R}^{M_{\rm wide}}$ such that for $\boldsymbol{\theta}_{\rm wide} = \mathcal{T}(\boldsymbol{\theta}_{\rm narr})$
(i) **output preserving:** $\boldsymbol{f}_{\theta_{\rm narr}} = \boldsymbol{f}_{\theta_{\rm wide}}$;
(ii) **criticality preserving:** if $\boldsymbol{\theta}_{\rm narr}$ is a critical point, then $\boldsymbol{\theta}_{\rm wide}$ is also a critical point.

critical embedding exists $\Rightarrow$ Embedding Principle

One-step splitting embedding



One-step embedding $\mathcal{T}_{l,s}^{\alpha}$.

Splitting embedding is critical embedding

Proposition (output and representation preserving)

For any point θ_{narr} of a DNN, a point θ_{wide} of a wider DNN obtained from θ_{narr} by **one-step embedding** satisfies

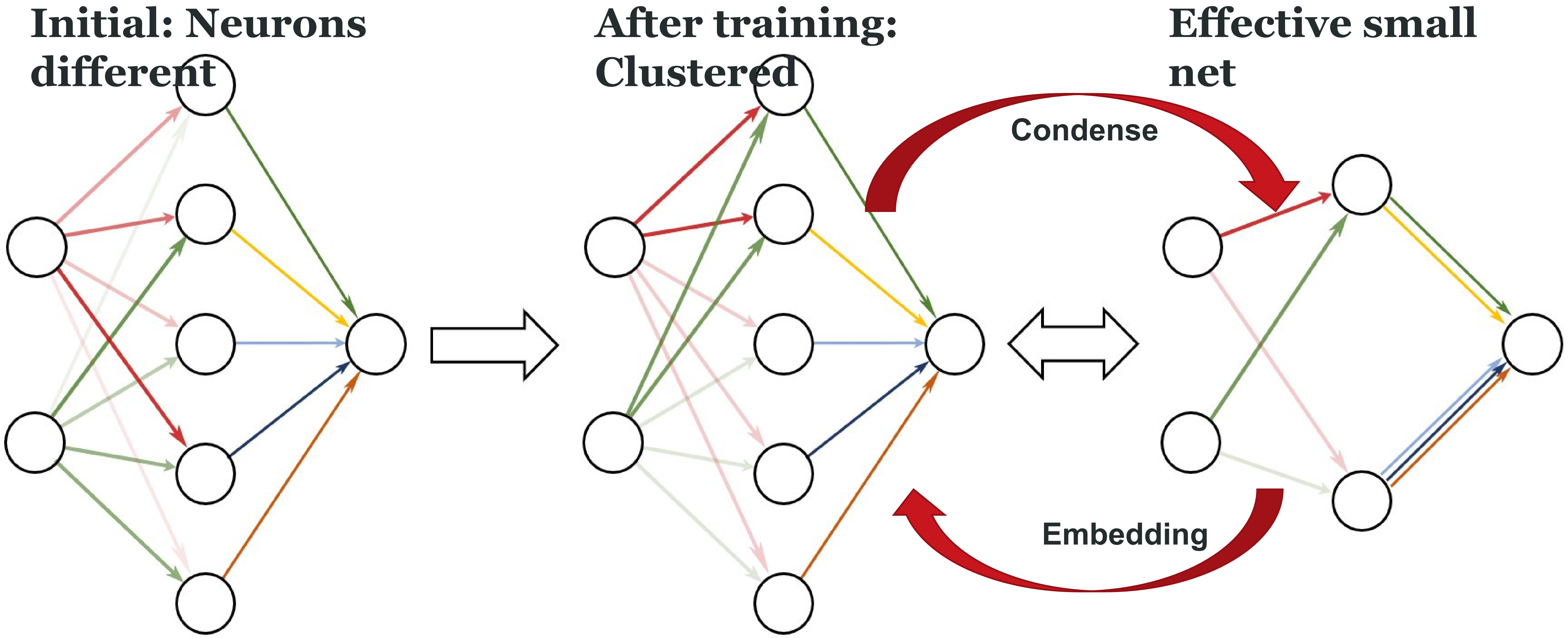
$$\boldsymbol{f}_{\theta_{\text{narr}}}(\boldsymbol{x}) = \boldsymbol{f}_{\theta_{\text{wide}}}(\boldsymbol{x}) \text{ for any } \boldsymbol{x}.$$

Theorem (criticality preserving)

*For any critical point θ_{narr} of a DNN, a point θ_{wide} of a wider DNN obtained from θ_{narr} by **one-step embedding** is a critical point.*

Remark: Obviously, **multi-step embedding**, i.e., composition of **one-step embedding**, is also critical embedding.

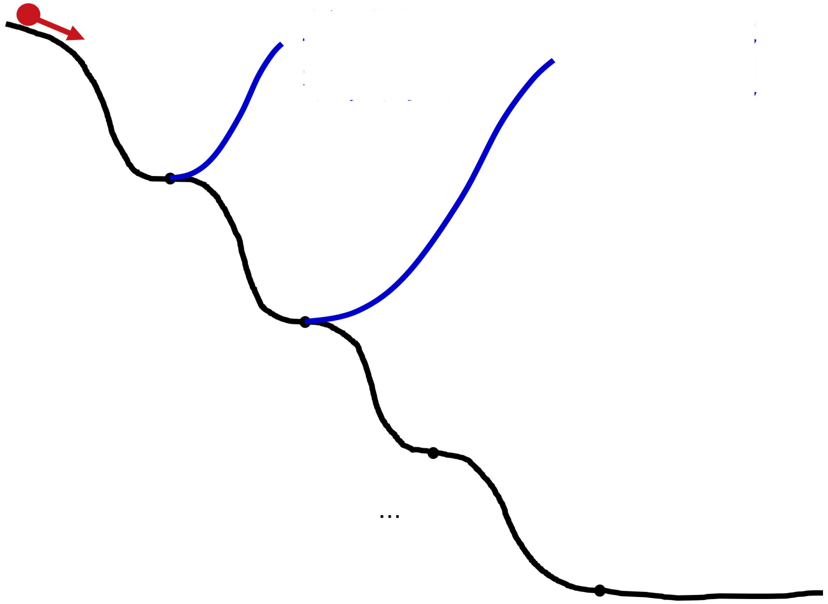
Embedding as “inverse” of condensation

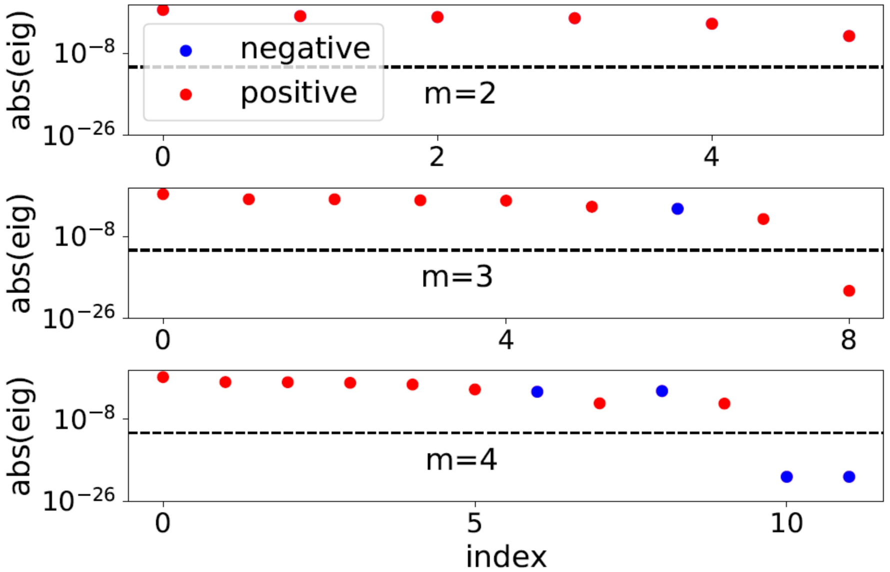


Implication of Embedding Principle

Implication—optimization

local-min of narrow NN→ saddles of wide NN





(a) synthetic data

Optimization benefit of wide NNs

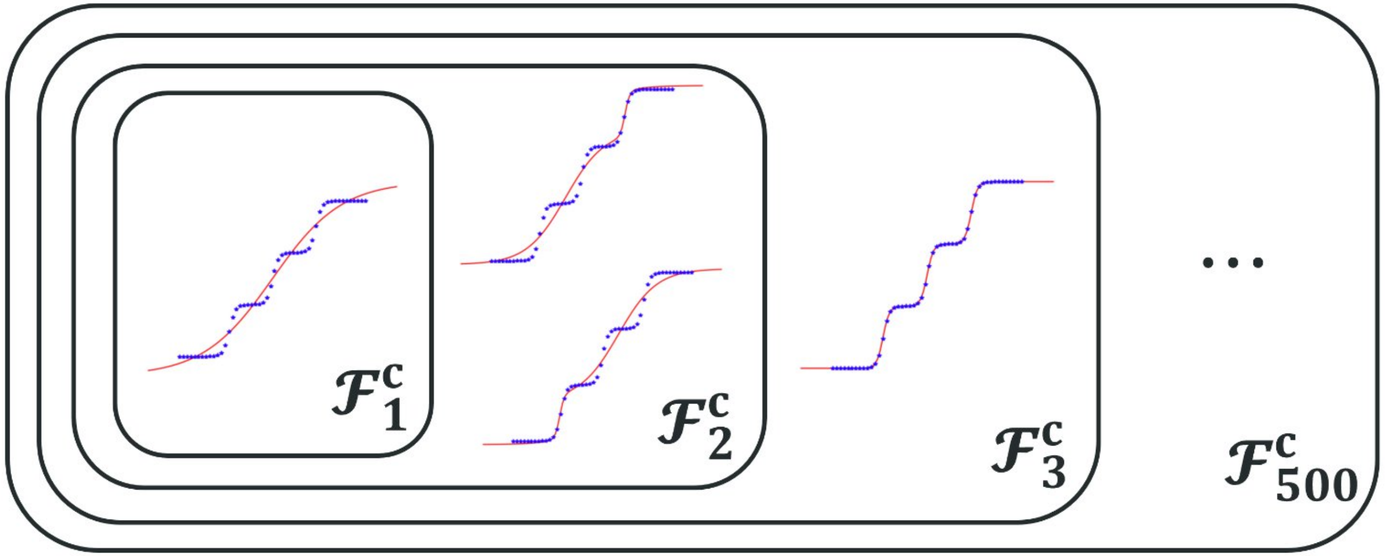
Transition to strict saddle points is Irreversible.

Theorem

Given an $\mathrm{NN}(\{m_l\}_{l=0}^L)$ *and any of its parameters* $\boldsymbol{\theta} \in \mathbb{R}^M$*, for any critical embedding* $\mathcal{T} : \mathbb{R}^M \to \mathbb{R}^{M'}$ *to any wider* $\mathrm{NN}(\{m'_l\}_{l=0}^L)$*, the number of positive, zero, negative eigenvalues of* $\boldsymbol{H}_S(\mathcal{T}(\boldsymbol{\theta}))$ *is no less than the counterparts of* $\boldsymbol{H}_S(\boldsymbol{\theta})$*.*

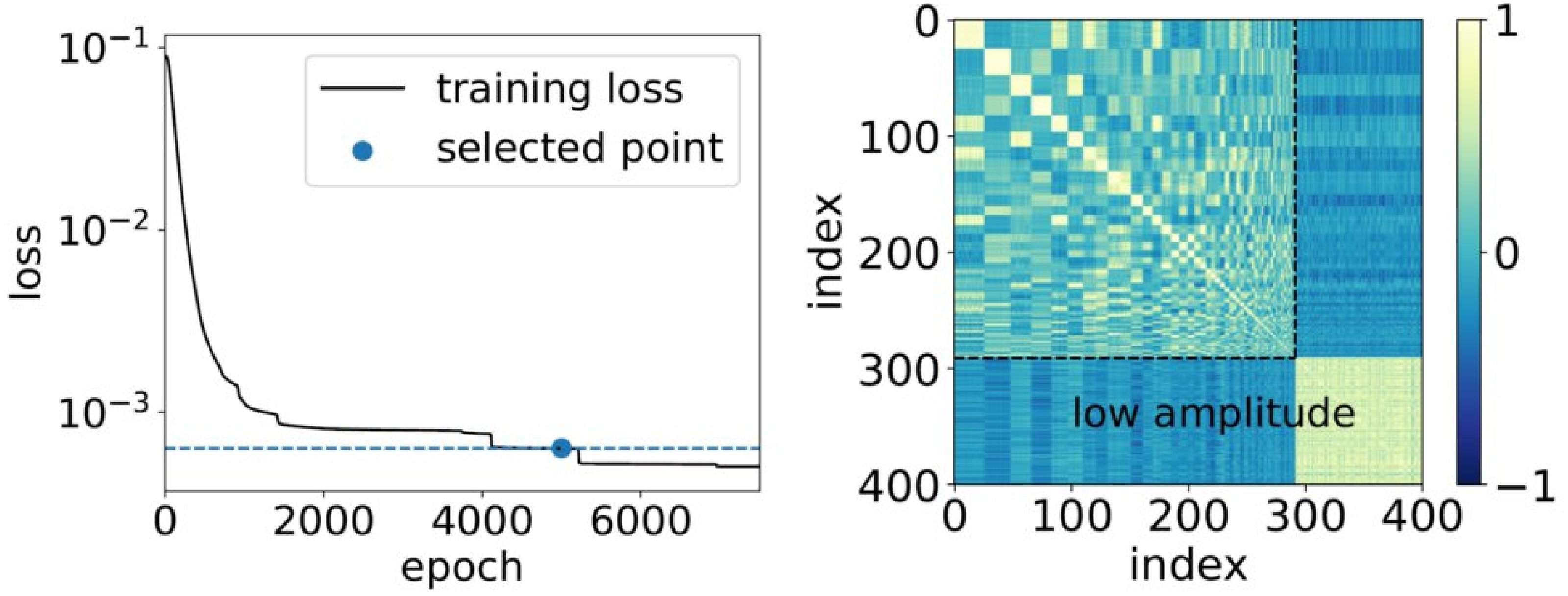
Observation

Nonlinear training of DNNs tends to learn simple critical functions.



Implication—pruning

"simple" critical points has huge pruning potential

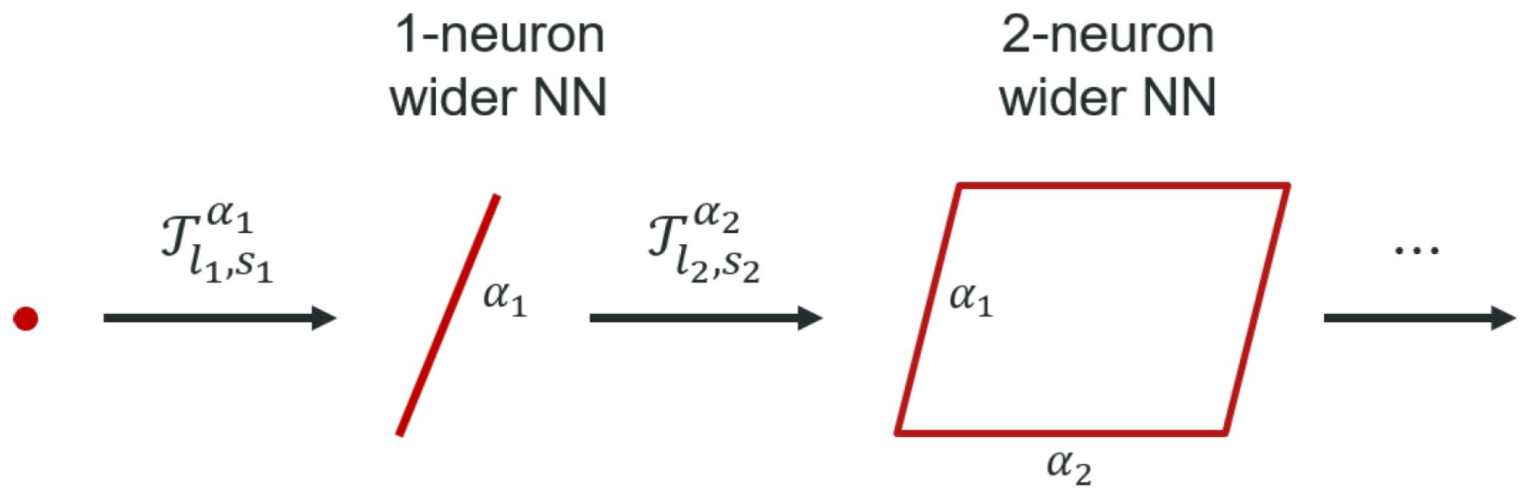


(a) initial loss (b) orientation similarity

Dimension of critical submanifolds

Theorem (informal)

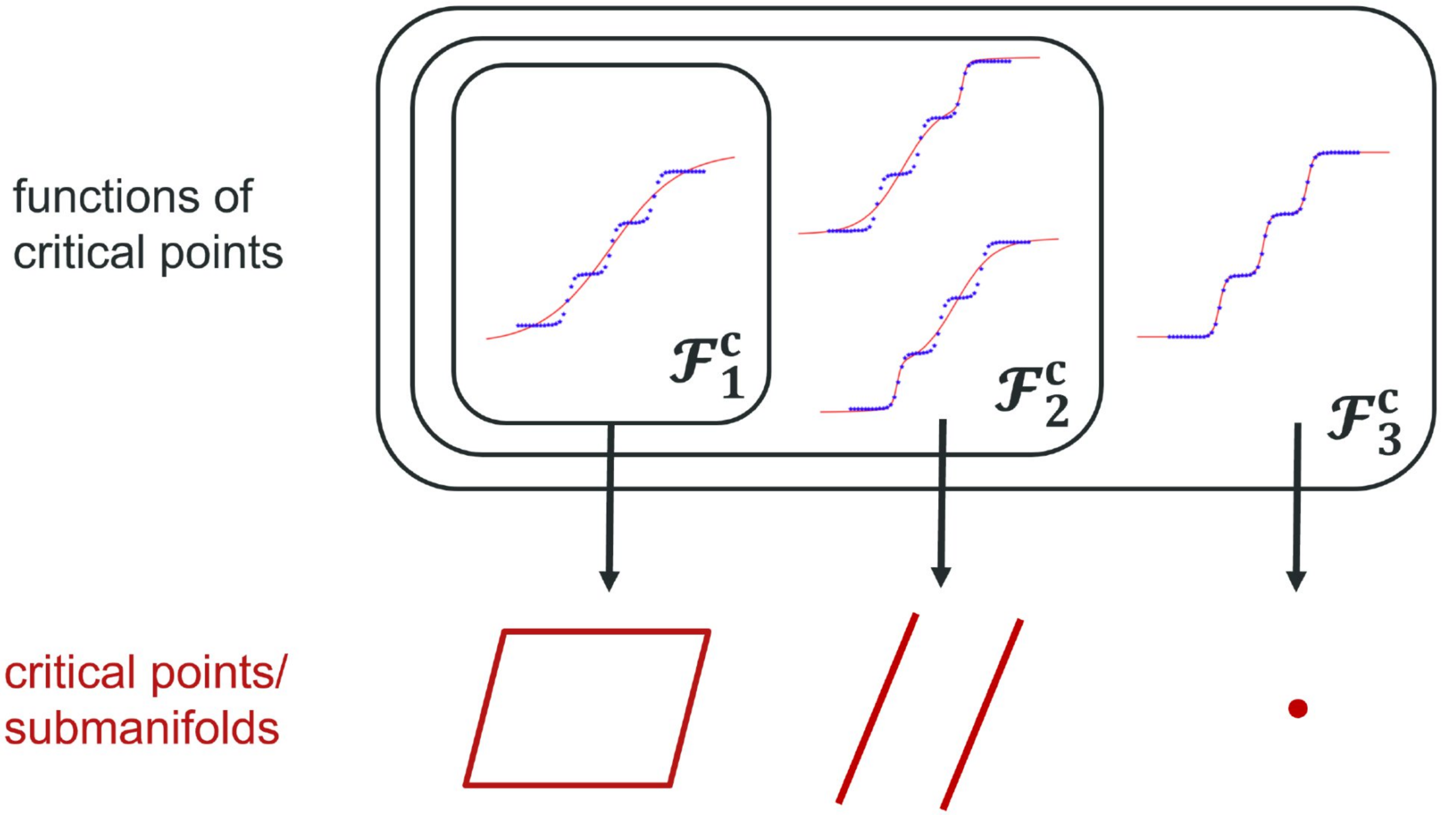
(Under mild assumption) Any critical point θ^c of a DNN can be embedded to K-dimensional critical affine subspaces of a K-neuron wider DNN.



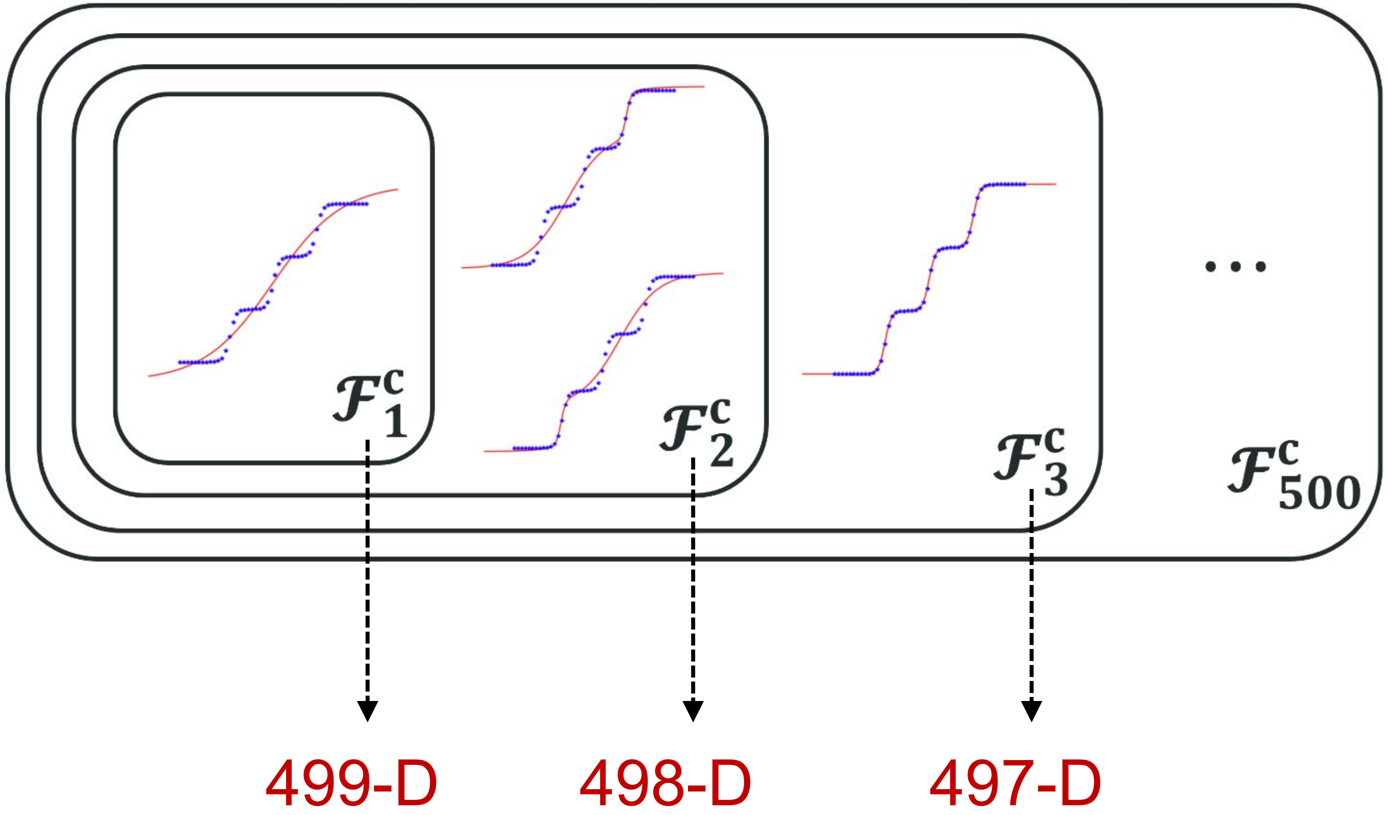
"Simple" critical functions possess high dimensional critical submanifolds.

Loss landscape analysis of width-3 tanh-NN

$$R_S(\theta) = \frac{1}{50}\sum_{i=1}^{50}\left(f_\theta(x_i) - y_i\right)^2, \quad f_\theta(x) = \sum_{j=1}^{3} a_j \tanh\left(w_j x + b_j\right)$$



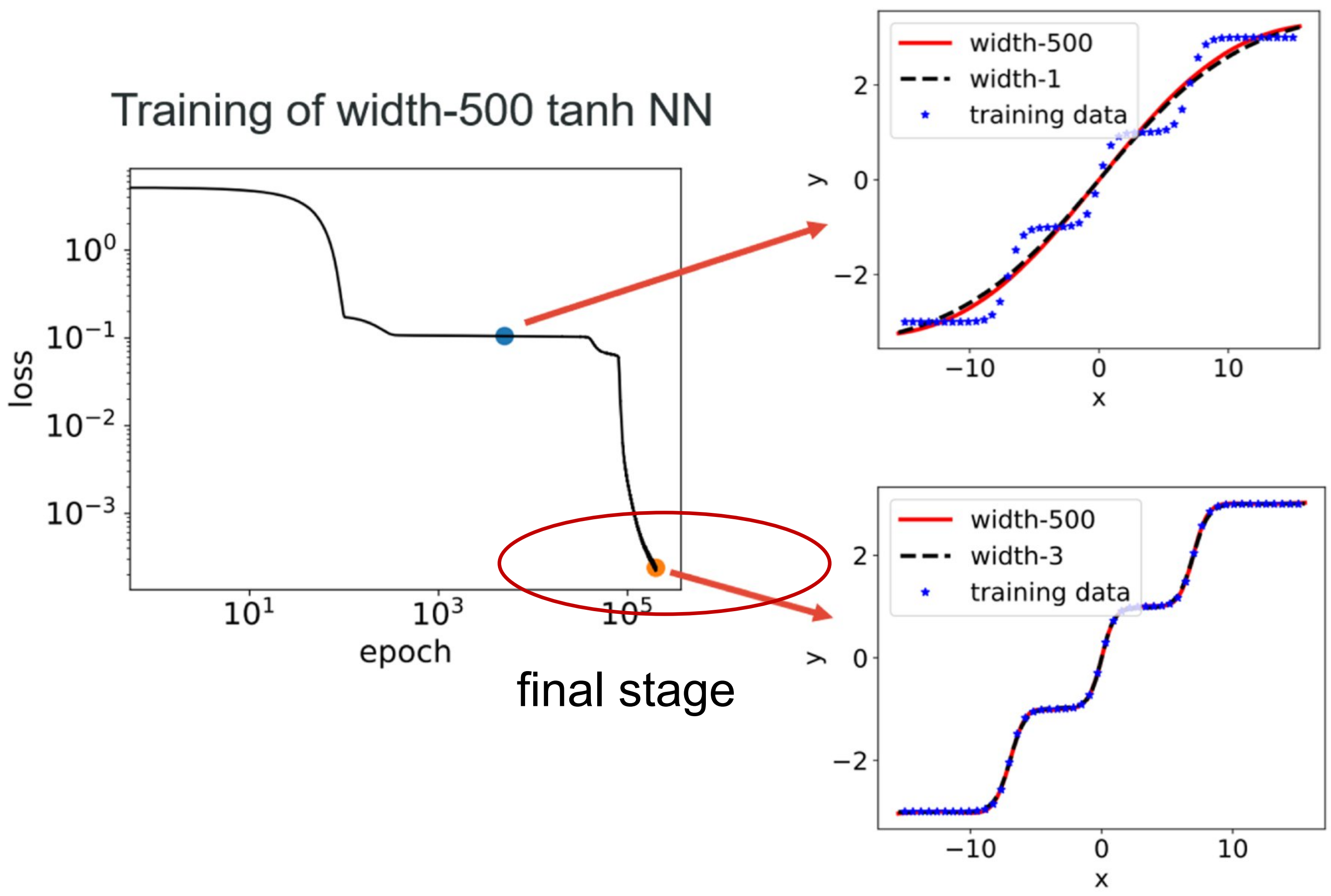
Hierarchical structure for two-layer NN



Final condensation

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

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: $f^* = \bar{a}\sigma\left(\overline{w}^{\mathrm{T}} x\right)$
- Target Set $Q^* = F^{-1}(f^*)$ generally consists of three "branches" (sets)
  - (a) $Q_1 = \left\{(a_k, w_k)_{k=1}^2 : w_1 = w_2 = \overline{w}, a_1 + a_2 = \bar{a}\right\}$.
  - (b) $Q_2 = \left\{(a_k, w_k)_{k=1}^2 : w_1 = \overline{w}, a_1 = \bar{a}, a_2 = 0\right\}$.
  - (c) $Q_3 = \left\{(a_k, w_k)_{k=1}^2 : w_2 = \overline{w}, a_2 = \bar{a}, a_1 = 0\right\}$.

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

$L_S^{-1}(0)$

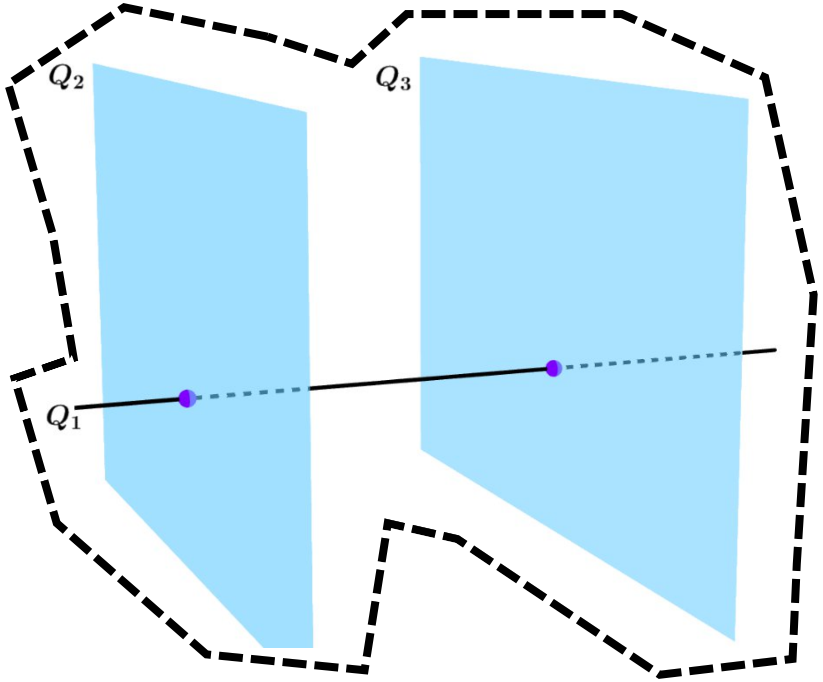
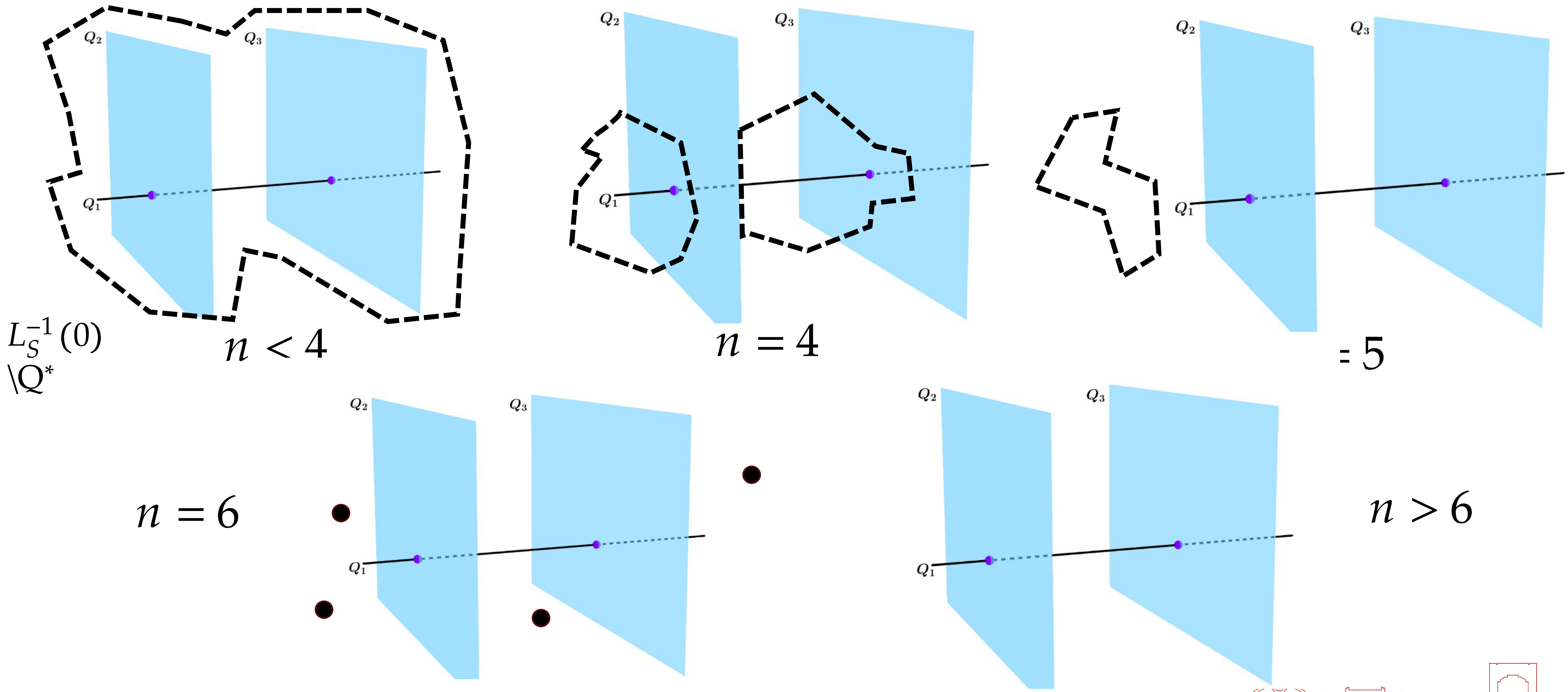


Illustration of Q^1, Q^2, Q^3

Geometry of global minima for final condensation

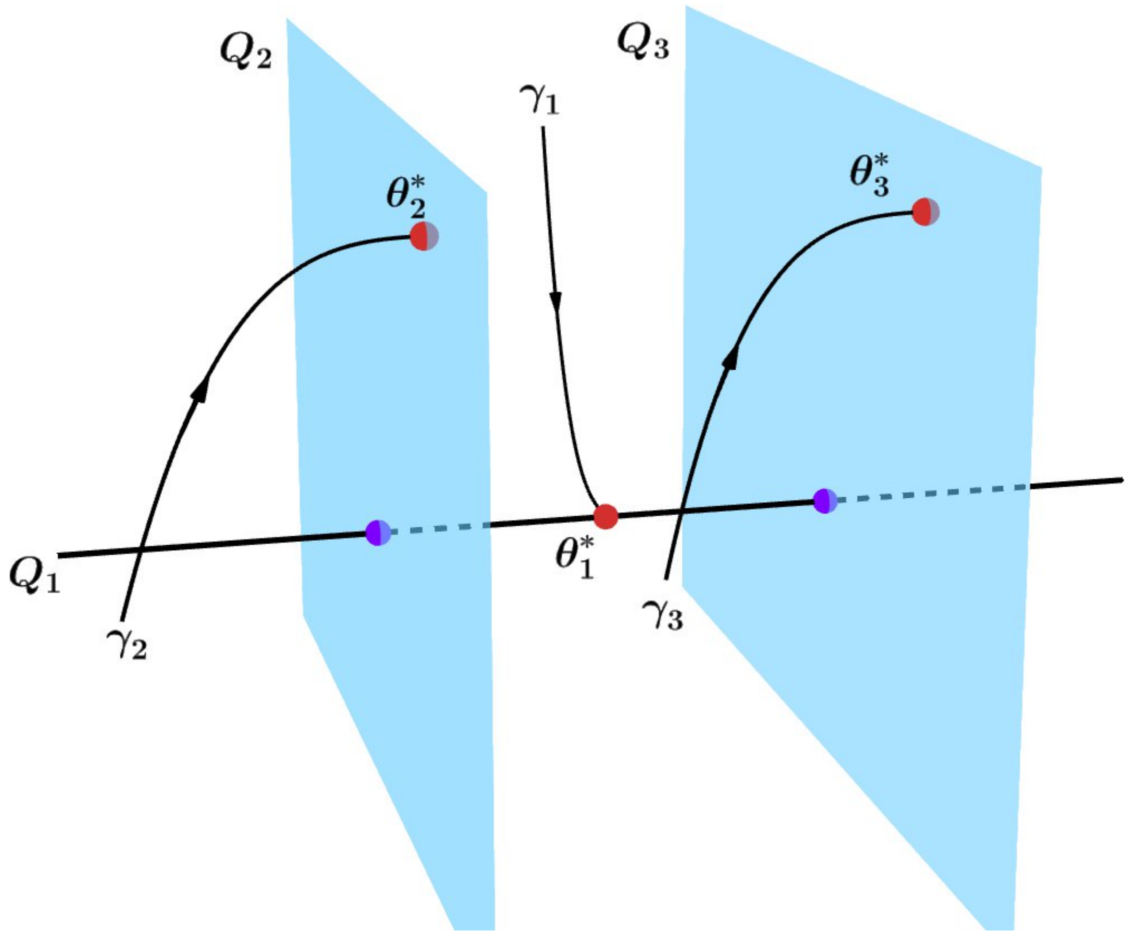


Typical convergence rate for final condensation

Gradient flows near Q^{*}:

γ_1 : sublinear rate;

γ_2,γ_3: linear rate.



Stability of target branches underlies final condensation

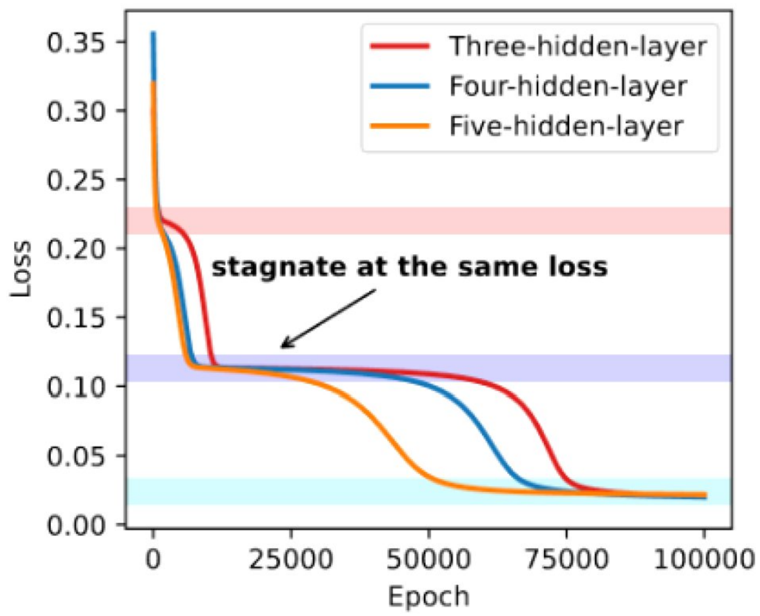
Theorem 5.4 (recovery stability). *Given $m_0 \le r \le m$, partition P and permutation π and separating inputs $\{x_i\}_{i=1}^n$. Then no point in $Q^r_{P,\pi}$ is recovery stable when $n \le r + (r-l)d$ (l is the deficient number of P), and almost all points in $Q^r_{P,\pi}$ are recovery stable when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

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

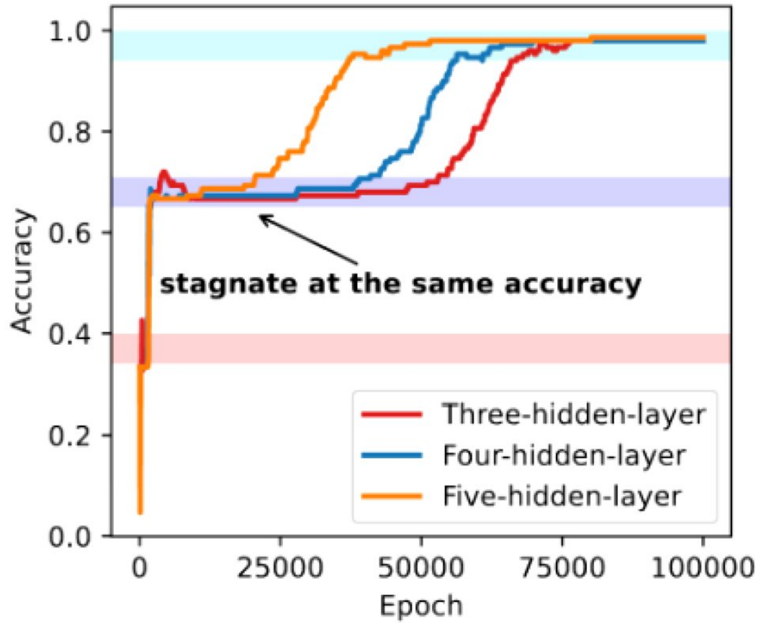
Embedding principle in depth

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.

Phenomenon: training similarity in depth



(a) Loss (Iris)



(b) Accuracy (Iris)

Embedding Principle in depth

Embedding principle in depth

the loss landscape of any network “contains” all critical points of all shallower networks.

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.

Key to the Proof

Goal: discover an embedding operator

Discover a mapping $\mathcal{T}: \mathbb{R}^{M_{\mathrm{shal}}} \to \mathbb{R}^{M_{\mathrm{deep}}}$ such that for any $\boldsymbol{\theta}_{\mathrm{deep}} \in \mathcal{T}(\boldsymbol{\theta}_{\mathrm{shal}})$, we have

(1) **output preserving:**

$$f_{\boldsymbol{\theta}_{\mathrm{deep}}}(\boldsymbol{x}) = f_{\boldsymbol{\theta}_{\mathrm{shal}}}(\boldsymbol{x}), \forall \boldsymbol{x} \in S_{\boldsymbol{x}}.$$

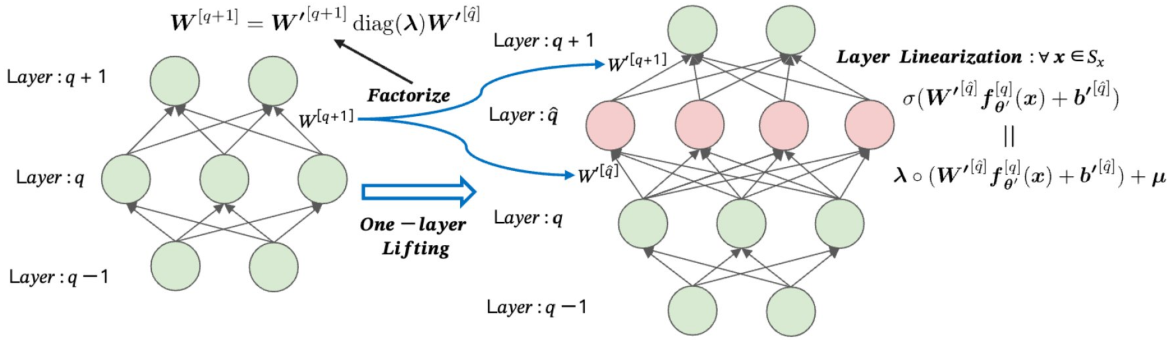
(2) **criticality preserving:**

if $\boldsymbol{\theta}_{\mathrm{shal}}$ is a critical point, then $\boldsymbol{\theta}_{\mathrm{deep}}$ is also a critical point.

$$\nabla R_S(\boldsymbol{\theta}_{\mathrm{shal}}) = \mathbf{0} \implies \nabla R_S(\boldsymbol{\theta}_{\mathrm{deep}}) = \mathbf{0}.$$

One-layer lifting

Assumption 1. Activation function σ has at least non-constant **linear piece**, e.g., ReLU, Leaky ReLU, ELU, etc.



1. Layer linearization condition

$$\sigma\left(W'^{[\hat{q}]} f^{[q]}_{\theta'}(x) + b'^{[\hat{q}]}\right) = \lambda \circ \left(W'^{[\hat{q}]} f^{[q]}_{\theta'}(x) + b'^{[\hat{q}]}\right) + \mu, \forall x \in S_x$$

2. Output preserving condition

$$\begin{cases} W'^{[q+1]}\,\mathrm{diag}(\lambda)W'^{[\hat{q}]} = W^{[q+1]}, \\ W'^{[q+1]}\,\mathrm{diag}(\lambda)b'^{[\hat{q}]} + W'^{[q+1]}\mu + b'^{[q+1]} = b^{[q+1]}. \end{cases}$$

Embedding Principle in depth

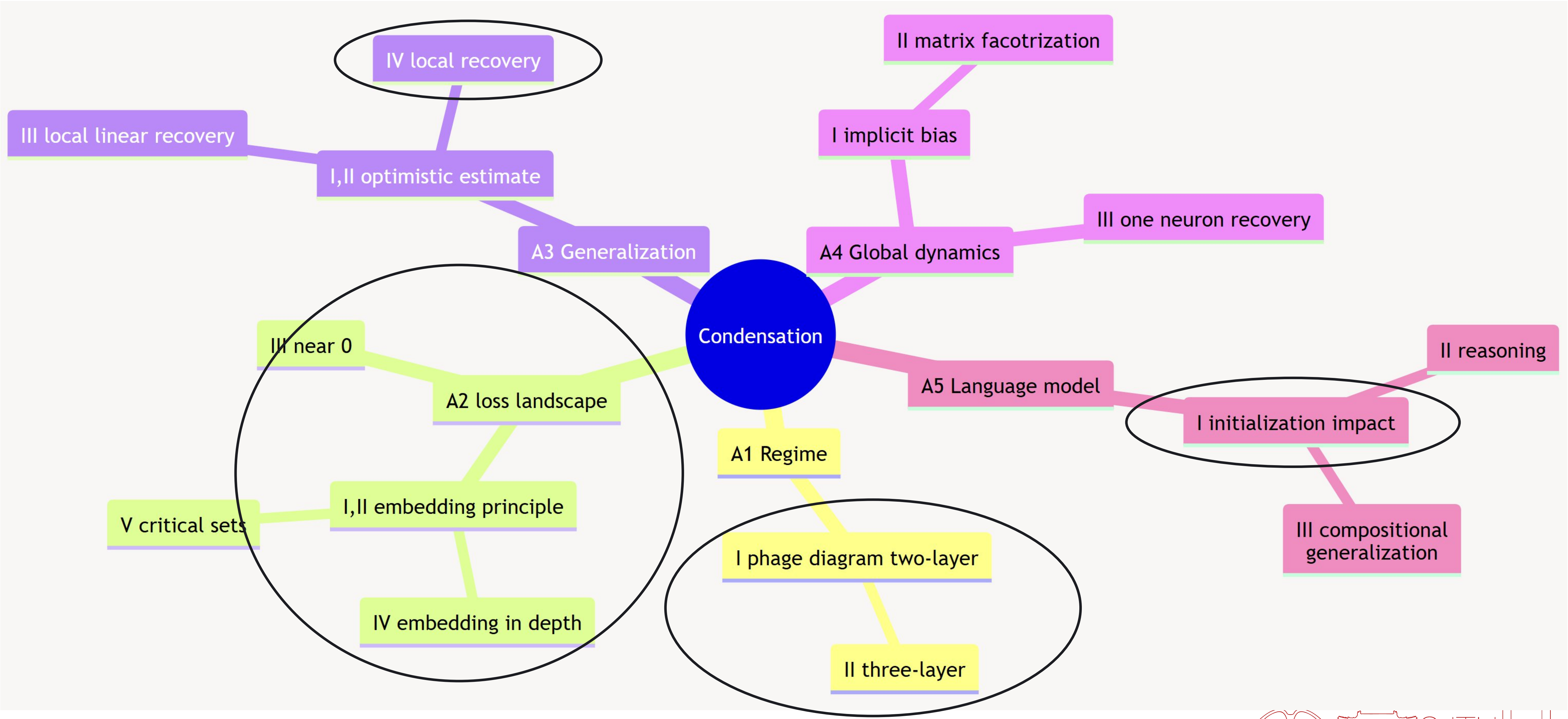
Embedding principle in depth [Bai et al.]

Theorem 1 (embedding principle in depth). Given any $\mathrm{NN}\left(\{m_l\}_{l=0}^{L}\right)$ and data S, for any $\boldsymbol{\theta}_{\mathrm{c}}'$ of any shallower $\mathrm{NN}'\left(\{m_l'\}_{l=0}^{L'}\right)$ satisfying $\nabla_{\boldsymbol{\theta}} R_S\left(\boldsymbol{\theta}_{\mathrm{c}}'\right)=\mathbf{0}$, 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) **Output Preserving:** $\boldsymbol{f}_{\boldsymbol{\theta}_{\mathrm{c}}}(\boldsymbol{x})=\boldsymbol{f}_{\boldsymbol{\theta}_{\mathrm{c}}'}(\boldsymbol{x})$ for $\boldsymbol{x}\in S_{\boldsymbol{x}}$;

(ii) **Criticality Preserving:** $\nabla_{\boldsymbol{\theta}} R_S\left(\boldsymbol{\theta}_{\mathrm{c}}\right)=\mathbf{0}$.

Condensation



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

Problems

- Why does small initialization lead to (initial) condensation in neural networks?
- How does the empirical risk landscape transform as the width of a neural network increases?
- What are the reasons why wider neural networks are often easier to optimize than narrower ones?
- Is it possible for neural networks to achieve zero generalization error for a target function under overparameterization?

Thanks！

饮 水 思 源　爱 国 荣 校