

# III. Condensation phenomenon

**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



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](http://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](http://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.

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

### Date

Fri. – Thu. May 2 – 8, 2025

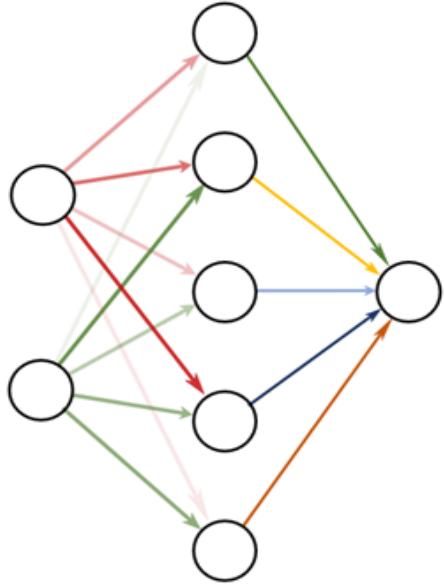
### Session Titles

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



# Illustration of Condensation

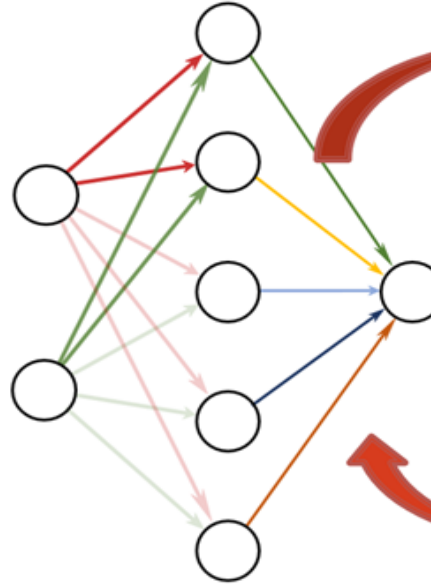
Initial: Neurons different



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

Initial: random

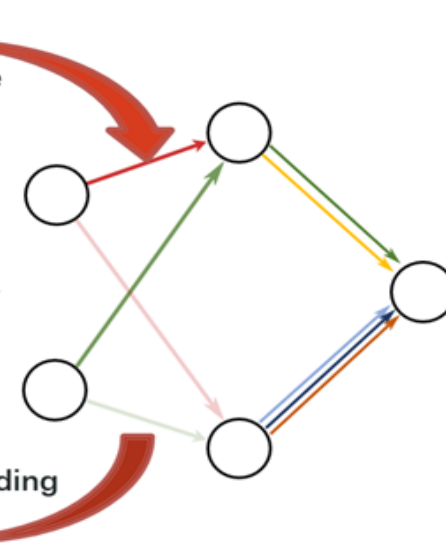
After training: Clustered



$$\begin{aligned} \mathbf{w}_1 &= \mathbf{w}_2, \\ \mathbf{w}_3 &= \mathbf{w}_4 = \mathbf{w}_5 \end{aligned}$$

Training: condense

Effective small net



$$\begin{aligned} f(x) &= \\ &(a_1 + a_2) \sigma(\mathbf{w}_1^T \mathbf{x}) + \\ &(a_3 + a_4 + a_5) \sigma(\mathbf{w}_3^T \mathbf{x}) \end{aligned}$$

Effect: equiv to small net



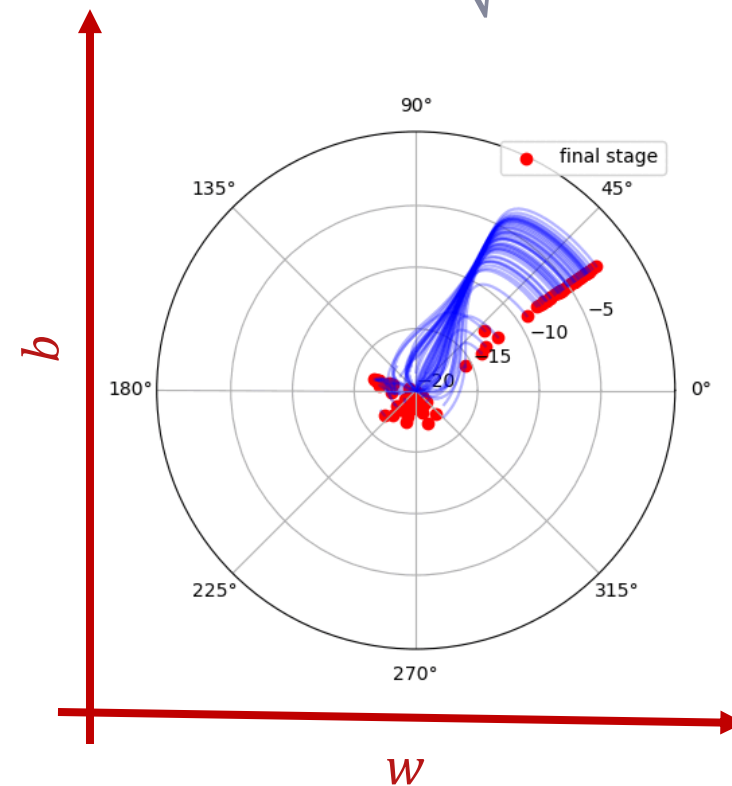
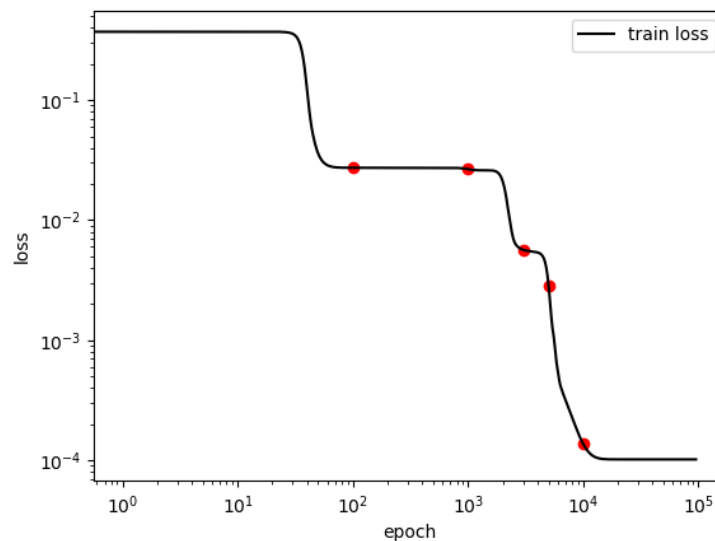
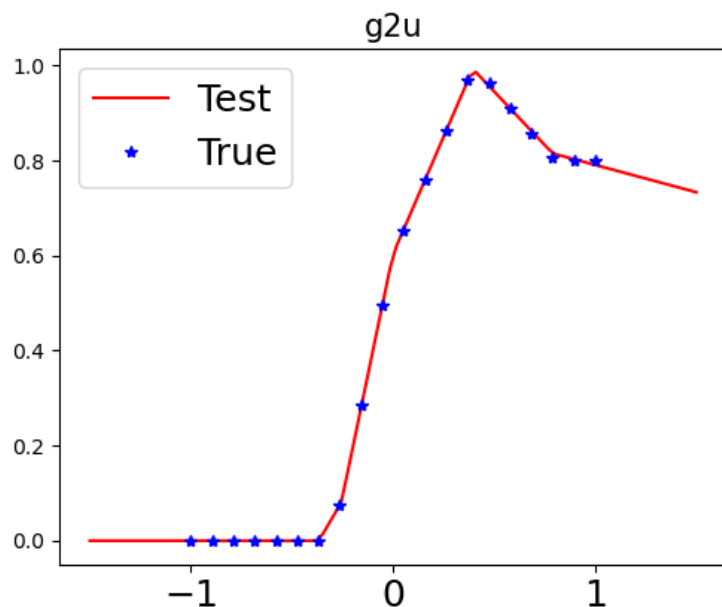
# 1d example: condensation with small initialization



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

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

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

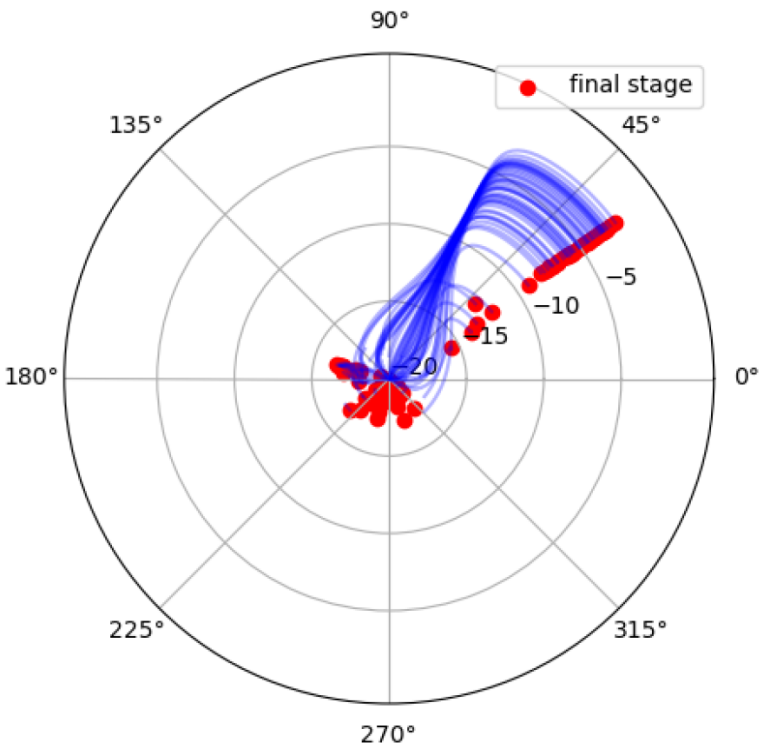


**Small initialization:**  $a_j(0), w_j(0), b_j(0) \sim N(0, \sigma^2)$  with small  $\sigma$

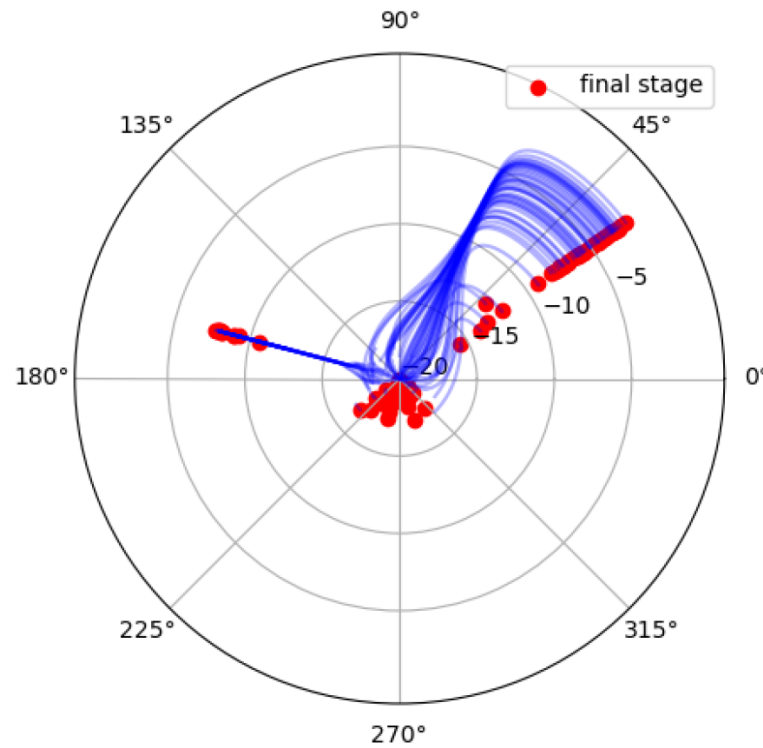




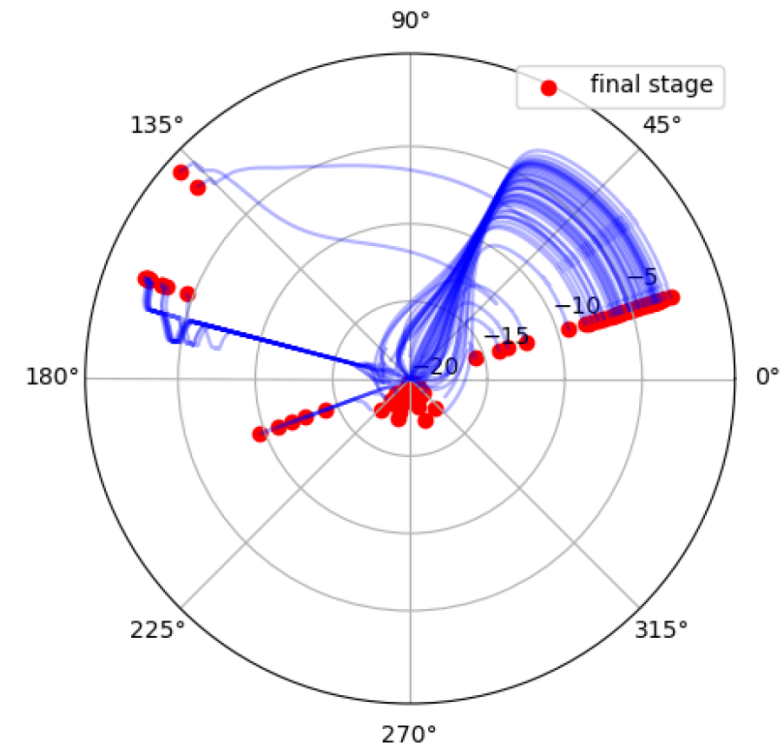
# Evolution trajectory: change significantly



(a) epoch=100



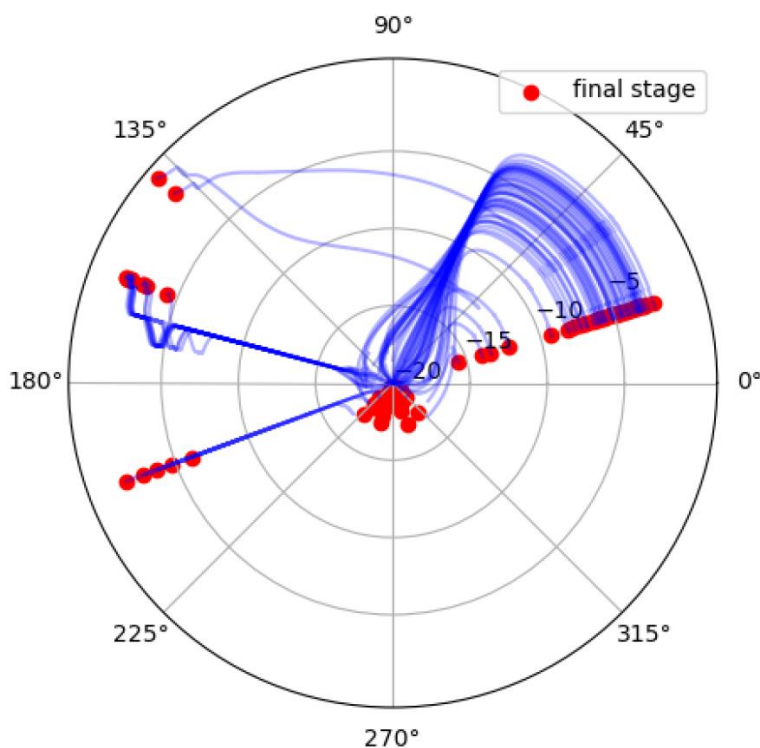
(b) epoch=1000



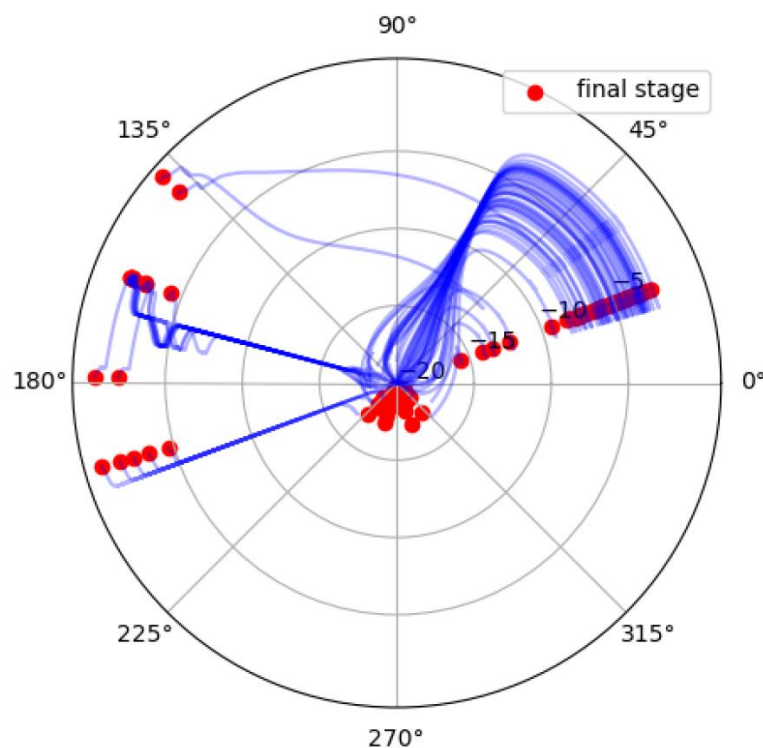
(c) epoch=3000



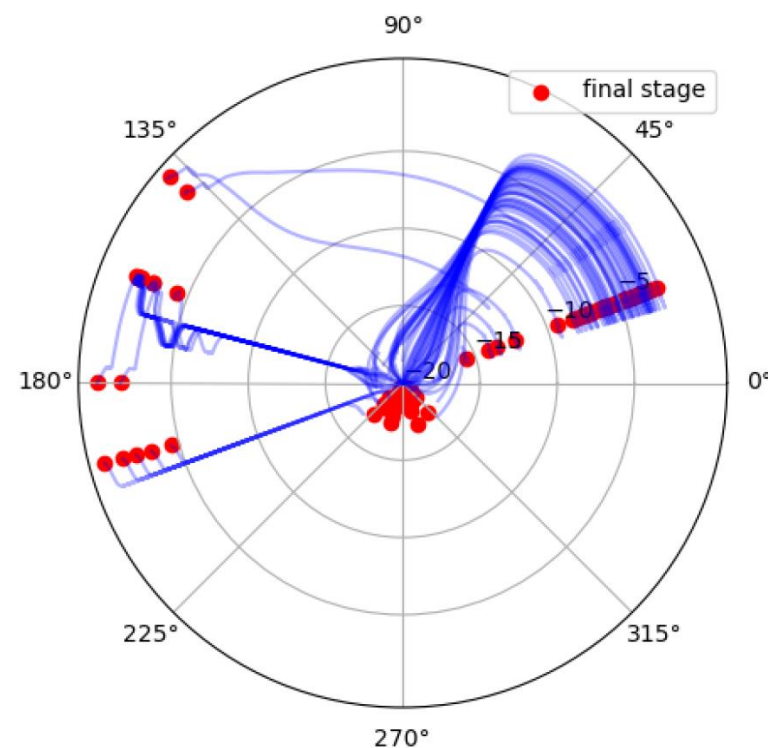
# Evolution trajectory: change significantly



(d) epoch=5000



(e) epoch=10000

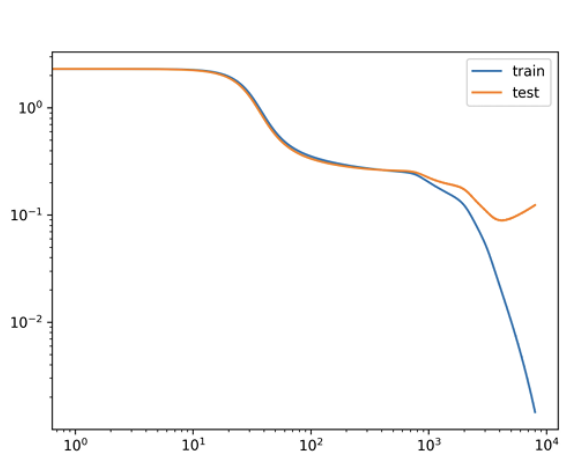


(f) epoch=100000

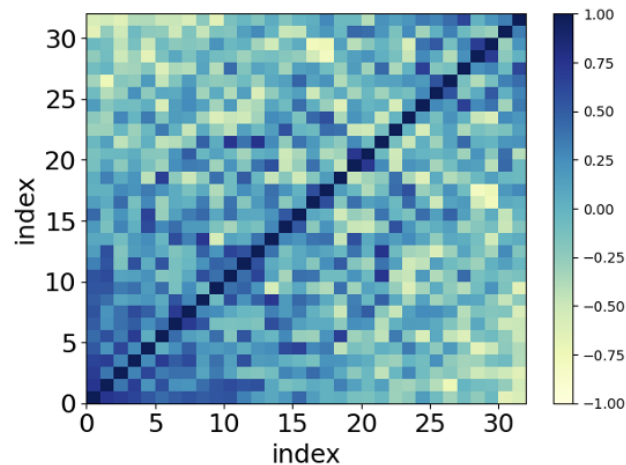




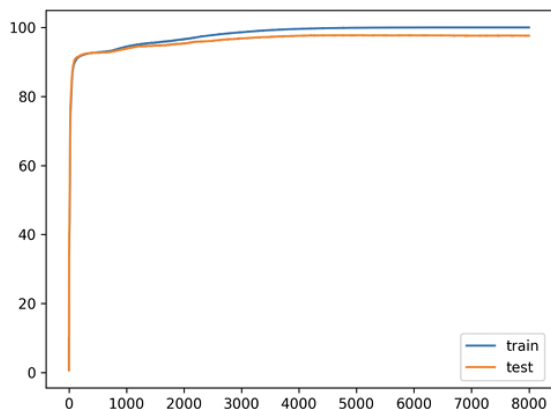
# Condensation in CNN on MNIST



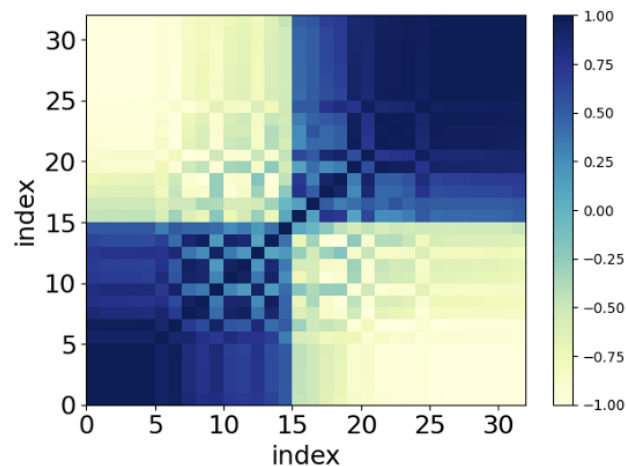
(a) Loss



(b) initial weight



(d) Accuracy



(e) final weight

**Cosine similarity:**

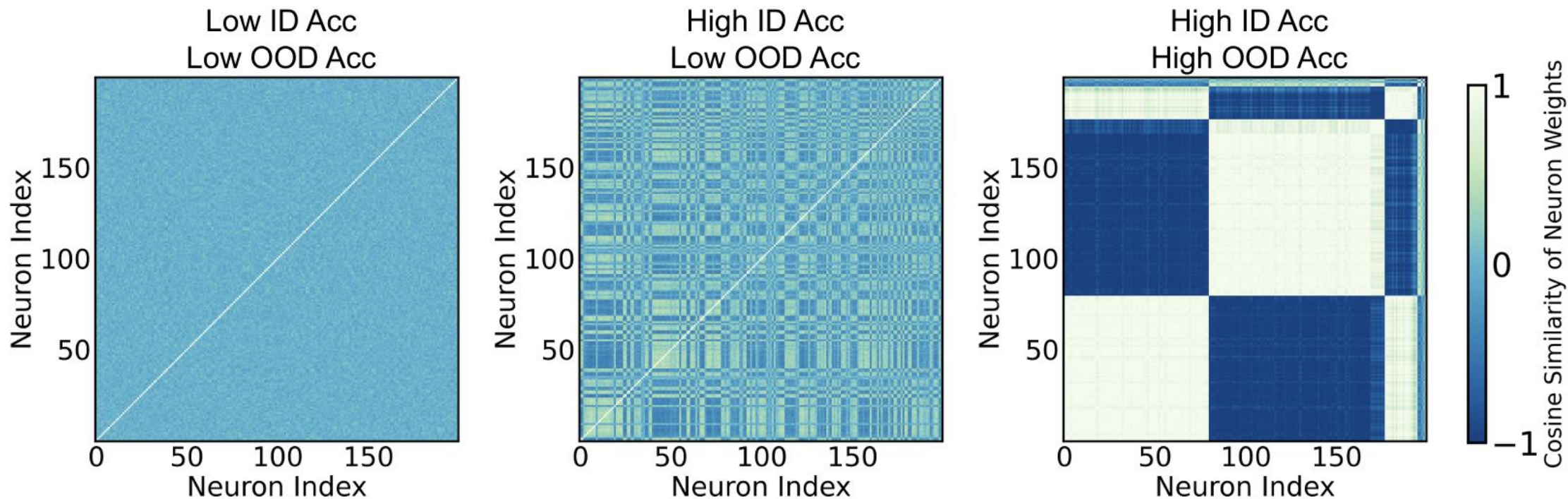
$$D(\mathbf{u}_1, \mathbf{u}_2) = \frac{\mathbf{u}_1^T \mathbf{u}_2}{(\mathbf{u}_1^T \mathbf{u}_1)^{1/2} (\mathbf{u}_2^T \mathbf{u}_2)^{1/2}}.$$

100% training and 97.62% test accuracy





# Condensation in transformer



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



# 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.

$$x = [x^T, 1]^T$$

$$w_k = [w_k^T, b_k]^T$$

- Data:  $\{x_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}_{i=1}^n$
- Two layer ReLU network

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

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

- Loss

$$R_S(\theta) = \frac{1}{2n} \sum_{i=1}^n (f_{\theta}^{\alpha}(x_i) - y_i)^2.$$

- Gradient flow dynamics

$$\frac{d\theta}{dt} = -\nabla_{\theta} R_S(\theta).$$

## Overparameterized setup:

$$M = m(d + 1) \gg n,$$

## Properties:

1. Global minima is  $M - n$  dimensional (proved by Yaim Cooper 2018)
2. Often non-overfitting
3. Evolution of  $\theta(t)$  and  $\theta(\infty)$  depend on  $\alpha, \beta_1, \beta_2$

## Goal:

Identify dynamical regimes of training over  $\alpha, \beta_1, \beta_2$  at infinite-width limit.



# Initialization methods with their scaling parameters

Name (related works)	$\alpha$	$\beta_1$	$\beta_2$	$\kappa$ $(\frac{\beta_1 \beta_2}{\alpha})$	$\kappa'$ $(\frac{\beta_1}{\beta_2})$	$\gamma$ $(\lim_{m \rightarrow \infty} \frac{\log 1/\kappa}{\log m})$	$\gamma'$ $(\lim_{m \rightarrow \infty} \frac{\log 1/\kappa'}{\log m})$
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)	$m$	1	1	$\frac{1}{m}$	1	1	0
(Sirignano and Spiliopoulos, 2020)							
(Rotskoff and Vanden-Eijnden, 2018)							
E et al. (E et al., 2020)	1	$\beta$	1	$\beta$	$\beta$	$\lim_{m \rightarrow \infty} \frac{\log 1/\beta}{\log m}$	$\lim_{m \rightarrow \infty} \frac{\log 1/\beta}{\log m}$





# Normalization and scaling parameters

- Two layer ReLU network

$$f_{\theta}^{\alpha}(\mathbf{x}) = \frac{1}{\alpha} \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^{\top} \mathbf{x}) \quad a_k^0 \sim N(0, \beta_1^2), \quad \mathbf{w}_k^0 \sim N(0, \beta_2^2 \mathbf{I}_d) \quad \begin{aligned} \mathbf{x} &= [\mathbf{x}^T, 1]^T \\ \mathbf{w}_k &= [\mathbf{w}_k^T, b_k]^T \end{aligned}$$

- Normalized gradient flow

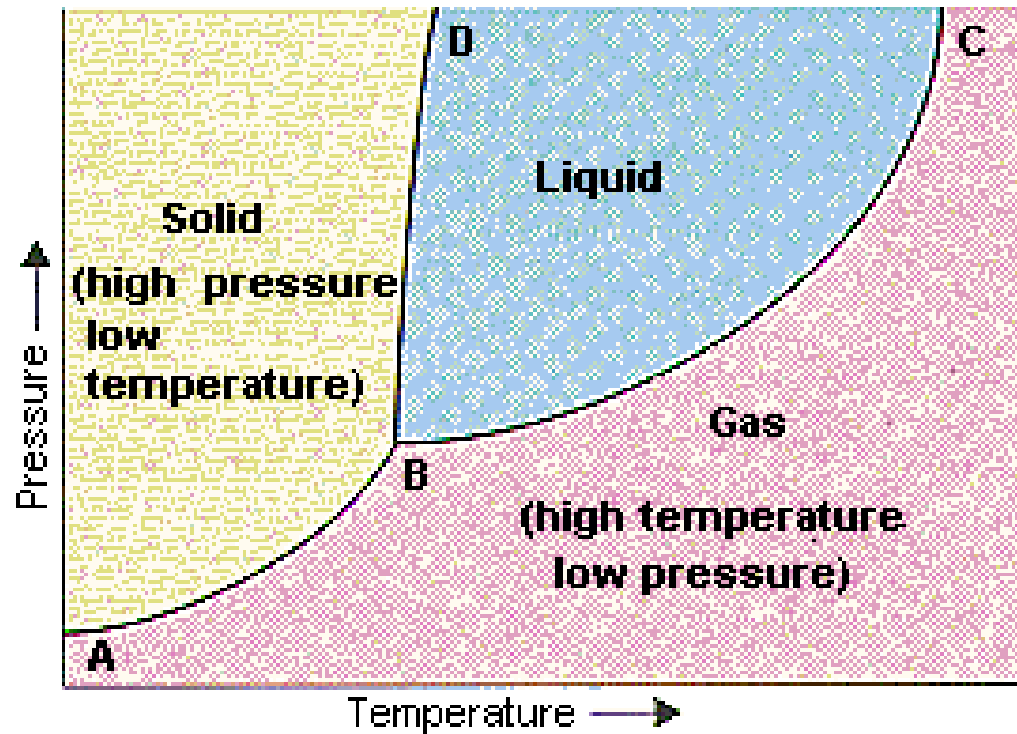
$$\begin{aligned} \bar{a}_k &= \beta_1^{-1} a_k, \quad \bar{\mathbf{w}}_k = \beta_2^{-1} \mathbf{w}_k, \quad \bar{t} = \frac{1}{\beta_1 \beta_2} t, \\ \frac{d\bar{a}_k}{d\bar{t}} &= -\frac{1}{\kappa'} \frac{1}{n} \sum_{i=1}^n \kappa \sigma(\bar{\mathbf{w}}_k^{\top} \mathbf{x}_i) \left( \kappa \sum_{k'=1}^m \bar{a}_{k'} \sigma(\bar{\mathbf{w}}_{k'}^{\top} \mathbf{x}_i) - y_i \right), \\ \frac{d\bar{\mathbf{w}}_k}{d\bar{t}} &= -\kappa' \frac{1}{n} \sum_{i=1}^n \kappa \bar{a}_k \sigma'(\bar{\mathbf{w}}_k^{\top} \mathbf{x}_i) \mathbf{x}_i \left( \kappa \sum_{k'=1}^m \bar{a}_{k'} \sigma(\bar{\mathbf{w}}_{k'}^{\top} \mathbf{x}_i) - y_i \right). \end{aligned}$$

- Scaling parameters and infinite-width limit

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



# Phase diagram



- **Phase diagram for matter**  
distinctive states of matter  $\leftrightarrow$  environment  
(phase transition happens at infinite size limit)  
solid, liquid, gas  $\leftrightarrow$  pressure, temperature

- **Phase diagram for two-layer ReLU NN**  
training dynamics  $\leftrightarrow$  initialization ( $m \rightarrow \infty$ )  
?  $\leftrightarrow$  ?

**Identification of coordinates of phase diagram (in analogy to pressure, temperature)**

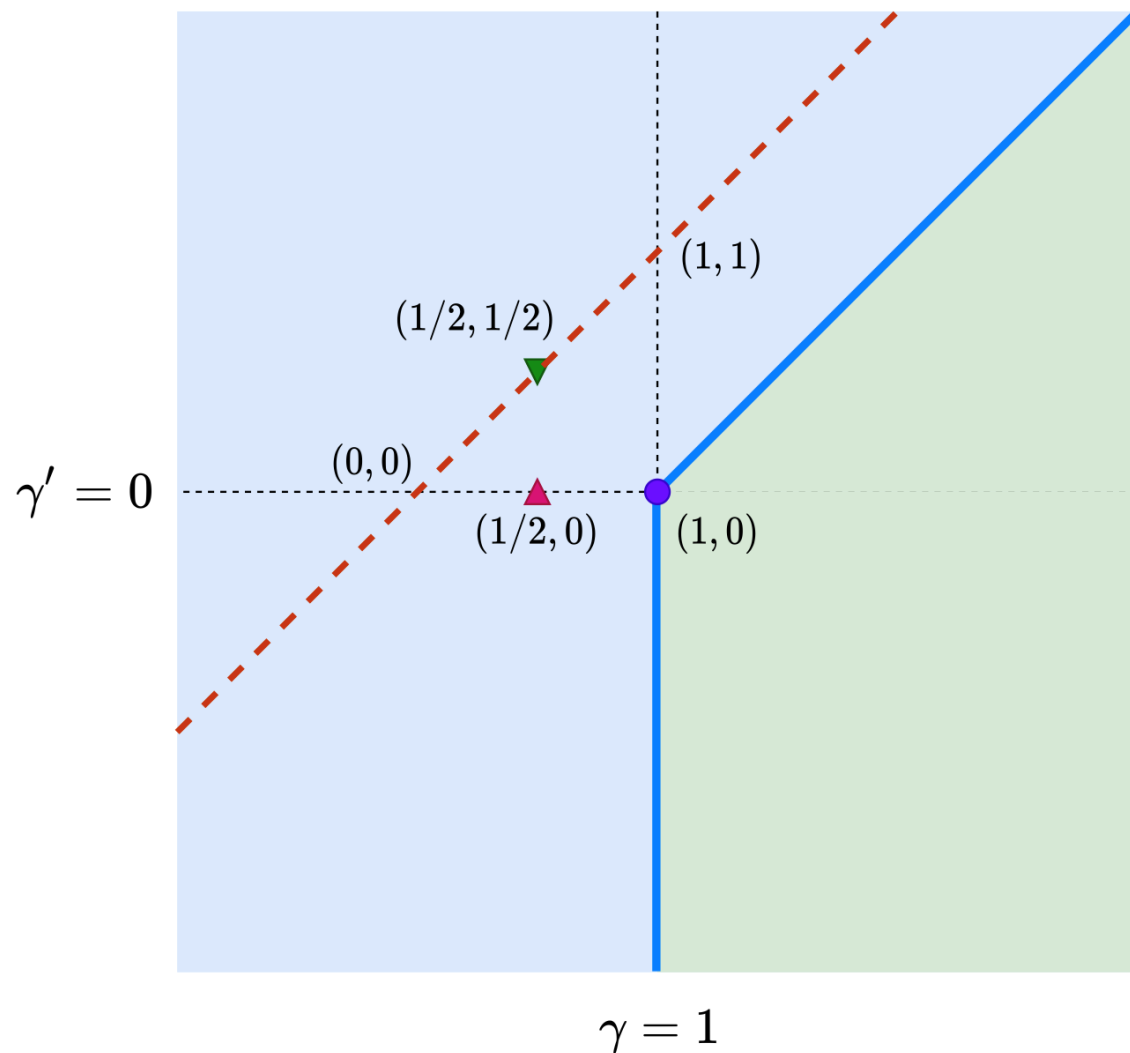
1. Effectively independent
2. Dynamical similarity
3. Differentiation capability

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





# Phase Diagram



Linear regime

Condensed regime

Critical regime

Examples:

● Xavier, Mean field

▲ NTK

- · - E at el. (2020)

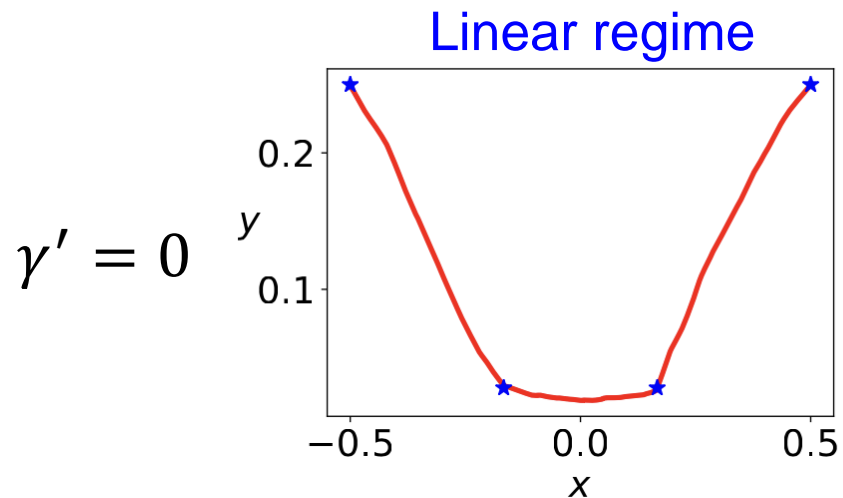
▼ LeCun, He

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

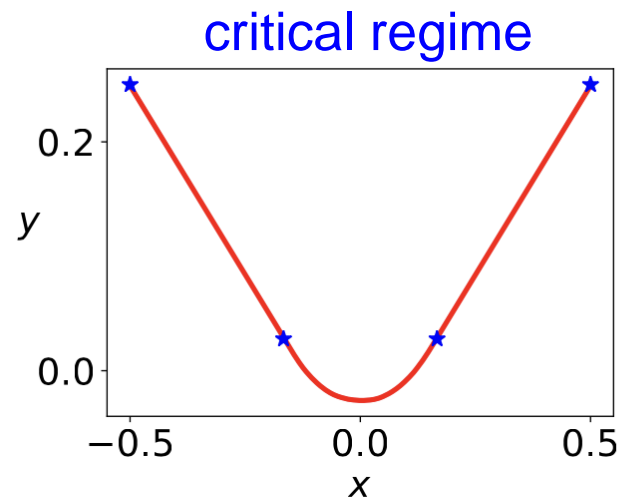
$$\gamma = \lim_{m \rightarrow \infty} -\frac{\log \beta_1 \beta_2 / \alpha}{\log m}, \quad \gamma' = \lim_{m \rightarrow \infty} -\frac{\log \beta_1 / \beta_2}{\log m}$$



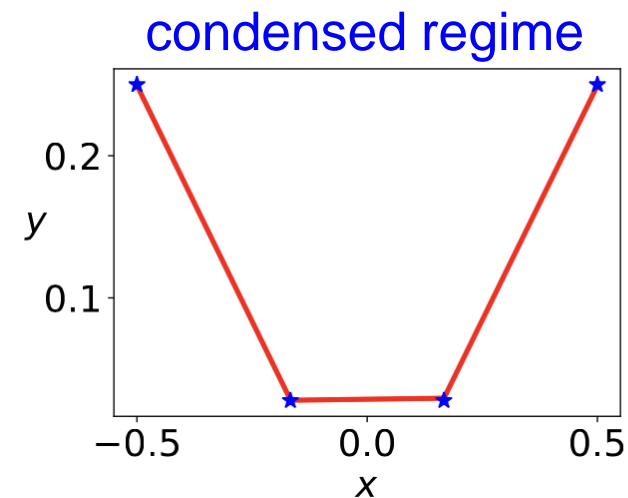
# Typical cases across the phase diagram



(a)  $\gamma = 0.5$



(b)  $\gamma = 1$

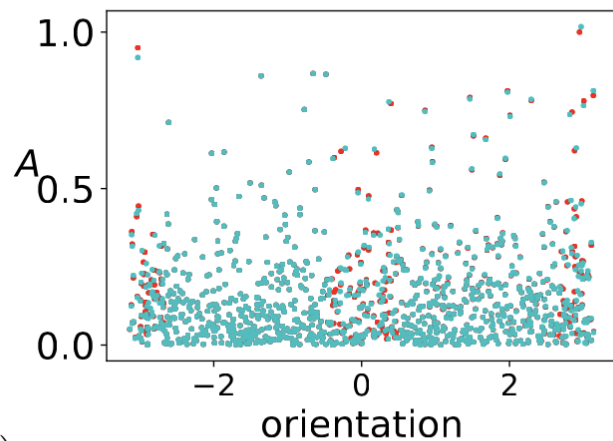


(c)  $\gamma = 1.75$

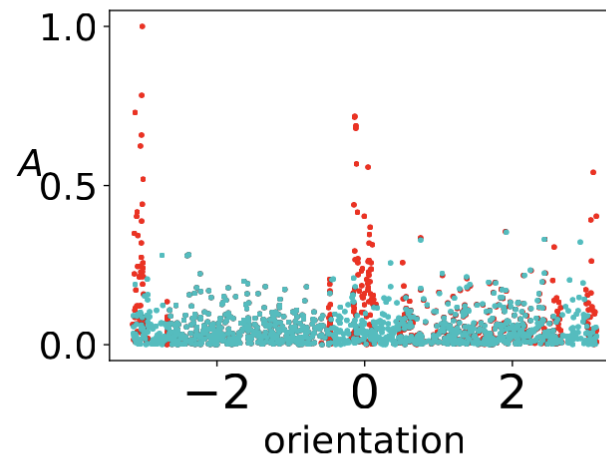
$$\{(A_k, \hat{\mathbf{w}}_k)\}_{k=1}^m$$

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

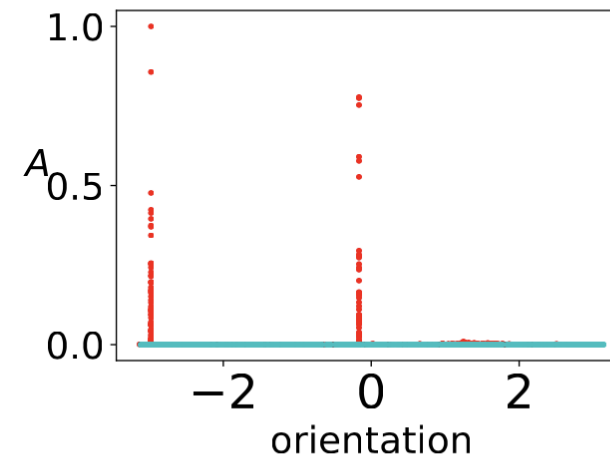
$$f_{\theta}^{\alpha}(\mathbf{x}) = \frac{1}{\alpha} \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^{\top} \mathbf{x})$$



(d)  $\gamma = 0.5$



(e)  $\gamma = 1$



(f)  $\gamma = 1.75$



- Linear regime (with ASI)

$$f_{\boldsymbol{\theta}}^{\text{lin}} = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(0)} \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)).$$

- Relative distance

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

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

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

As  $m \rightarrow \infty$ ,

- Linear regime:

$$\sup_{t \in [0, +\infty)} \text{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \rightarrow 0$$

- Condensed regime:

$$\sup_{t \in [0, +\infty)} \text{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \rightarrow +\infty$$

- Critical regime:

$$\sup_{t \in [0, +\infty)} \text{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \rightarrow O(1).$$





# Scaling analysis

- Two layer ReLU network at infinite-width limit

$$f_{\theta}^{\alpha}(\mathbf{x}) = \frac{1}{\alpha} \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^{\top} \mathbf{x}) \quad a_k^0 \sim N(0, \beta_1^2), \quad \mathbf{w}_k^0 \sim N(0, \beta_2^2 \mathbf{I}_d) \quad \begin{aligned} \mathbf{x} &= [x^T, 1]^T \\ \mathbf{w}_k &= [w_k^T, b_k]^T \end{aligned}$$

$$\kappa := \frac{\beta_1 \beta_2}{\alpha}, \quad \kappa' := \frac{\beta_1}{\beta_2},$$

- “capability” of NN:  $C = m\beta_1\beta_2/\alpha = m\kappa \gtrsim O(1)$

$$\kappa \gtrsim 1/m$$

- output-layer dominant:  $C = m\beta_2\mathbb{E}(|a|)/\alpha \ll m\beta_2^2/\alpha = m\kappa/\kappa'$ .

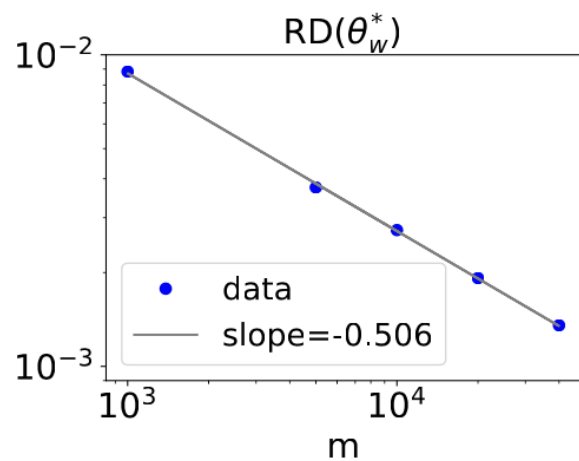
$$1/\kappa' \gg 1/m\kappa$$



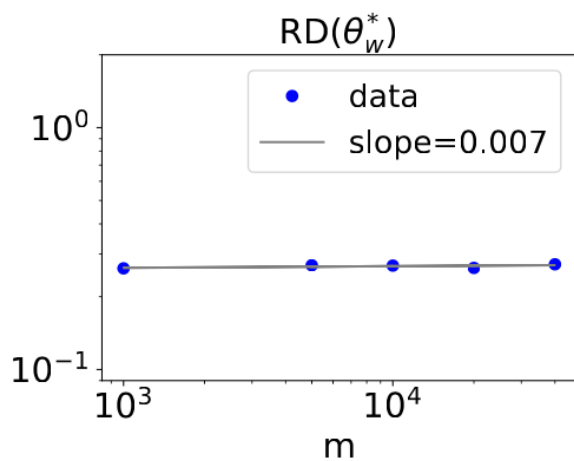


# Regime identification through experiments

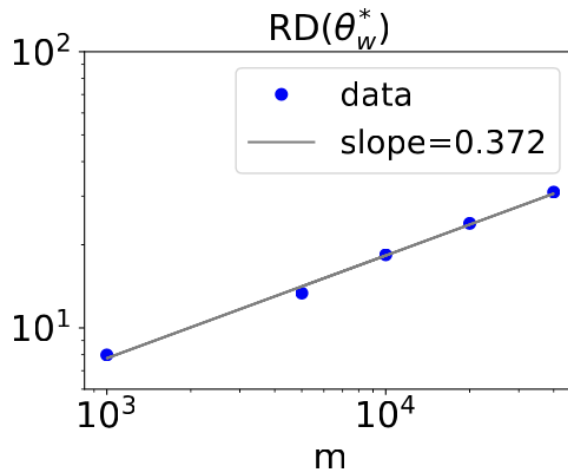
$$\gamma' = 0$$



(a)  $\gamma = 0.5$

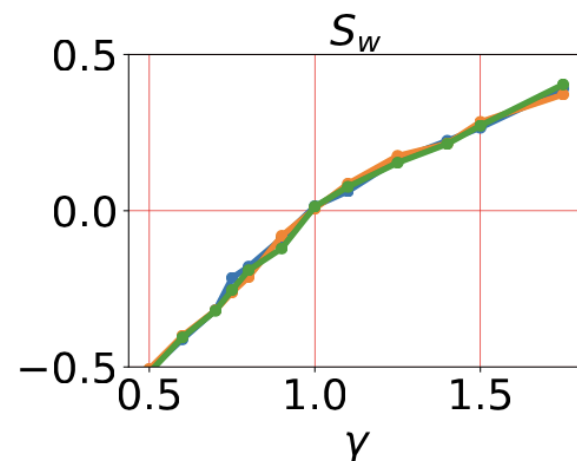


(b)  $\gamma = 1$

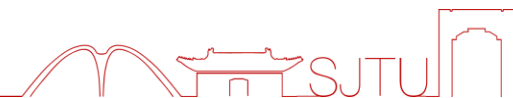


(c)  $\gamma = 1.75$

$$S_w = \lim_{m \rightarrow \infty} \frac{\log RD(\theta_w^*)}{\log m}$$



(d)  $S_w$  vs.  $\gamma$

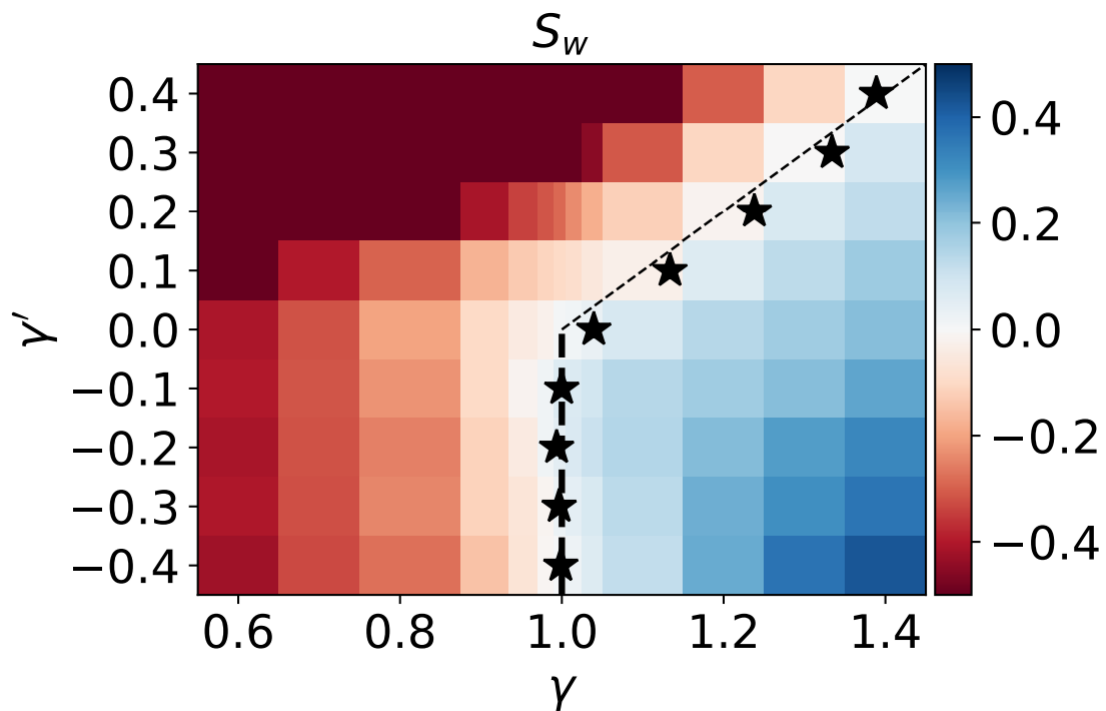




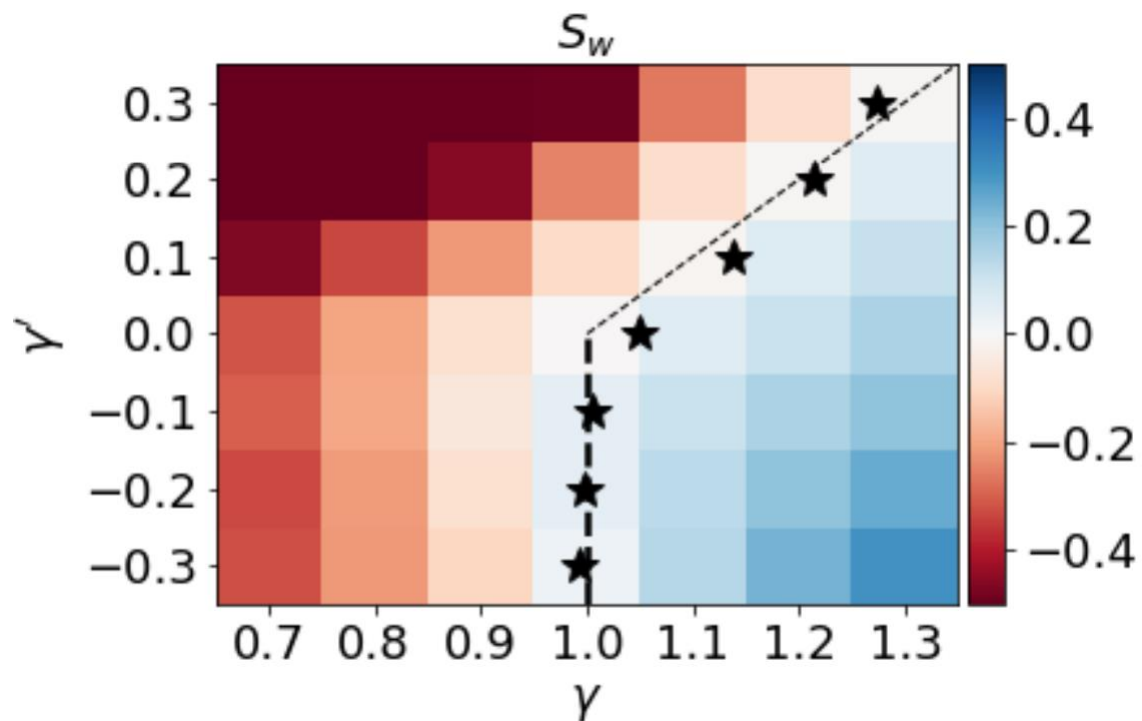


# Regime identification through experiments

Synthetic data



MNIST data



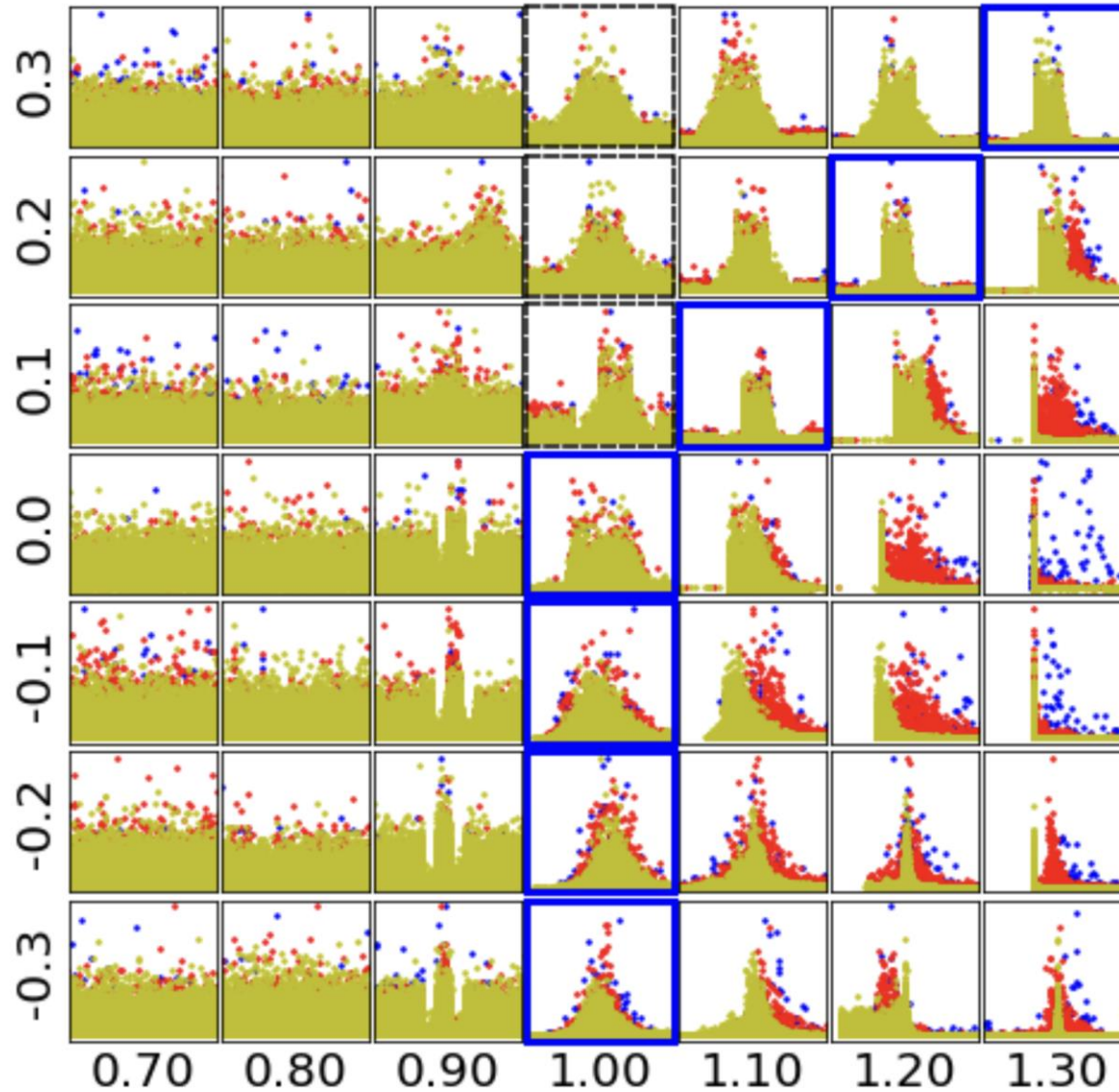


# Feature distribution at the condensed regime-synthetic



$$\{(A_k, \hat{\mathbf{w}}_k)\}_{k=1}^m$$
$$A = |a| \|\mathbf{w}\|_2$$

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



Blue:  $m = 10^3$   
red:  $m = 10^4$   
Yellow:  $m = 10^6$



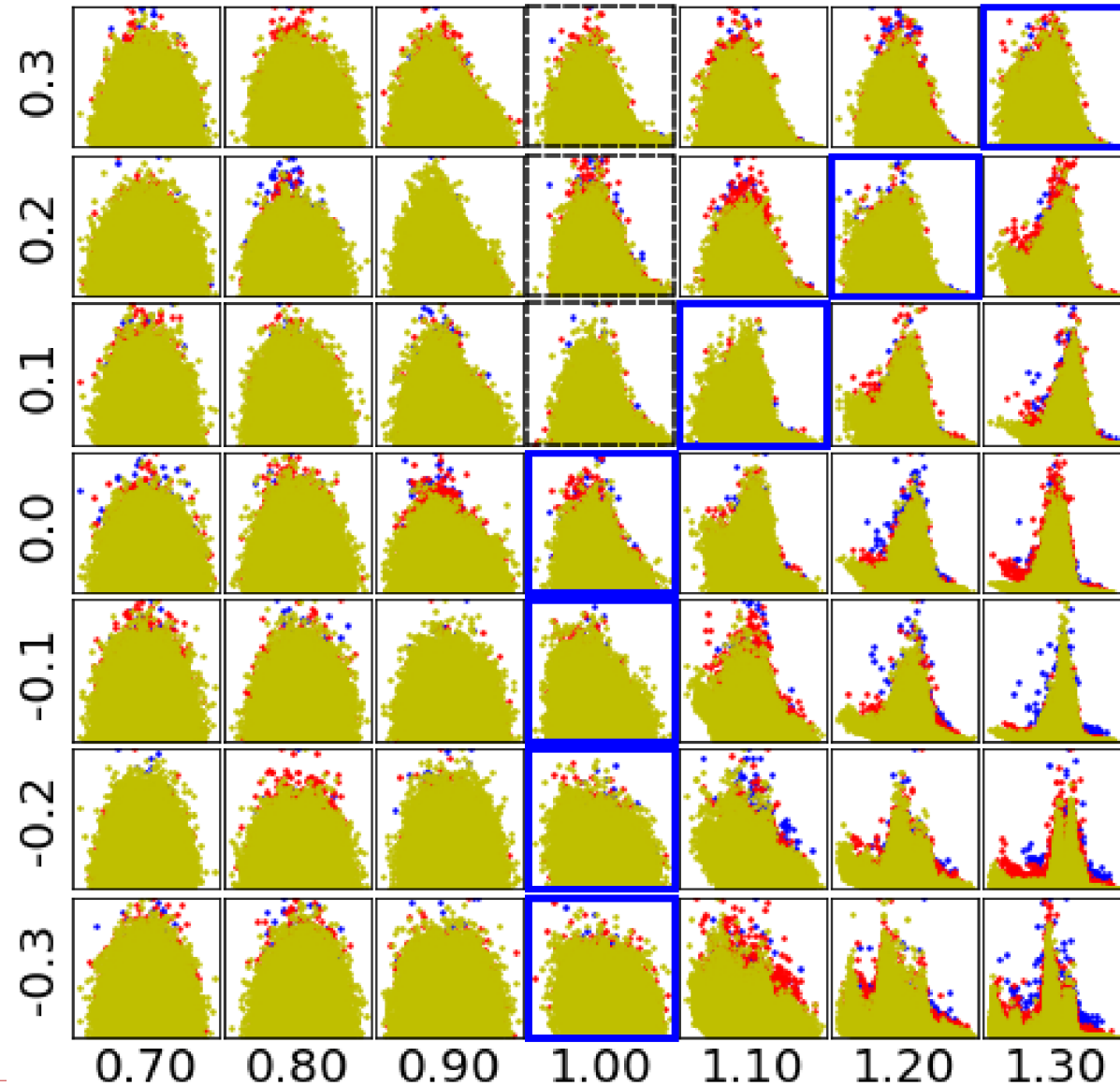


# Feature distribution at the condensed regime-MNIST



$$\{(A_k, \hat{\mathbf{w}}_k)\}_{k=1}^m$$
$$A = |a| \|\mathbf{w}\|_2$$

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



Blue:  $m = 10^3$   
red:  $m = 10^4$   
Yellow:  $m = 10^6$





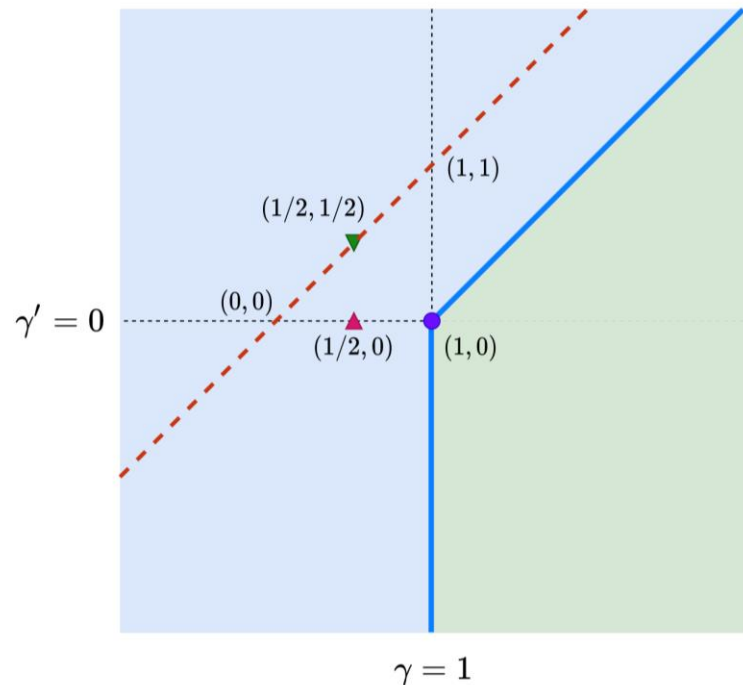
# Regime separation -- theorems

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

$$\lim_{m \rightarrow +\infty} \sup_{t \in [0, +\infty)} \text{RD}(\theta_w(t)) = 0. \quad (20)$$

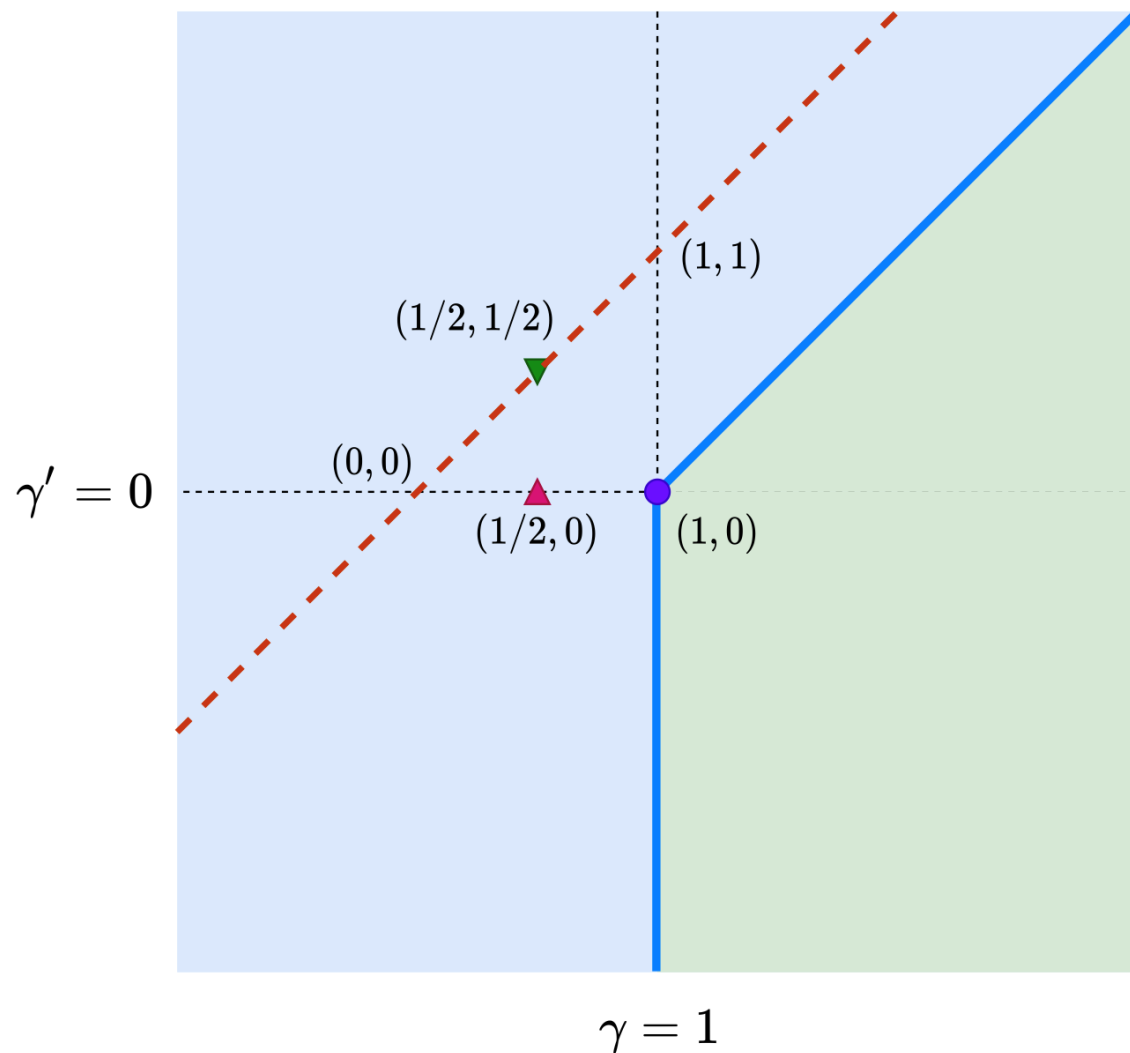
**Theorem 2\*.** (Informal statement of Theorem 8) If  $\gamma > 1$  and  $\gamma' < \gamma - 1$ , then with a high probability over the choice of  $\theta^0$ , we have

$$\lim_{m \rightarrow +\infty} \sup_{t \in [0, +\infty)} \text{RD}(\theta_w(t)) = +\infty. \quad (21)$$





# Phase Diagram



Linear regime

Condensed regime

Critical regime

Examples:

● Xavier, Mean field

▲ NTK

- · - E at el. (2020)

▼ LeCun, He

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

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

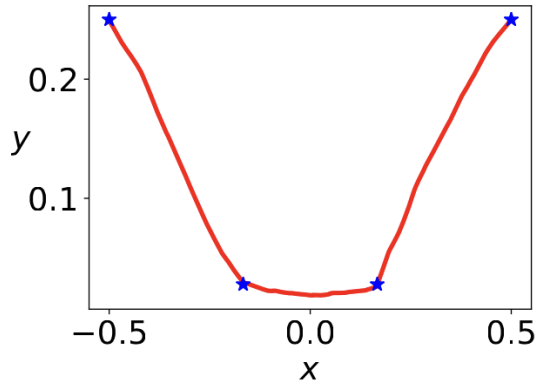




# Typical cases across the phase diagram

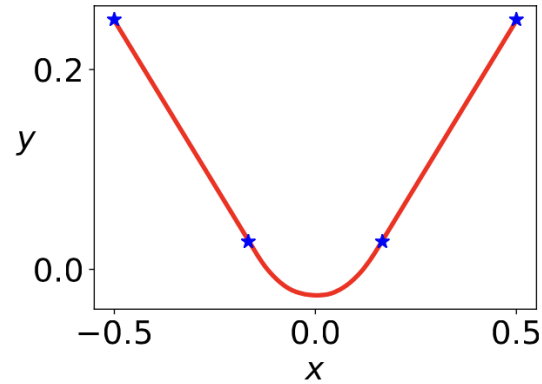
$$\gamma' = 0$$

Linear regime



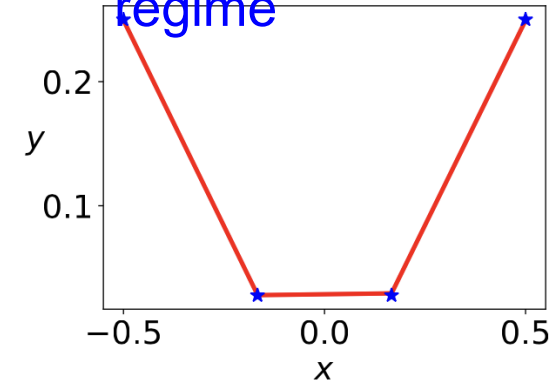
(a)  $\gamma = 0.5$

critical regime

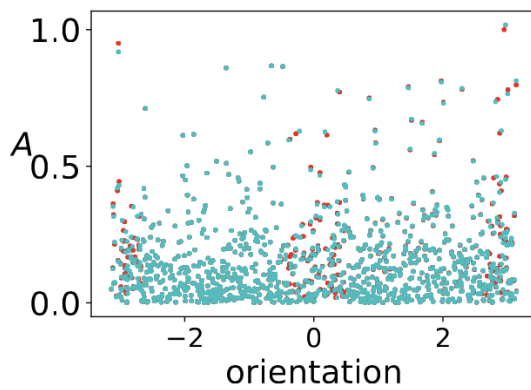


(b)  $\gamma = 1$

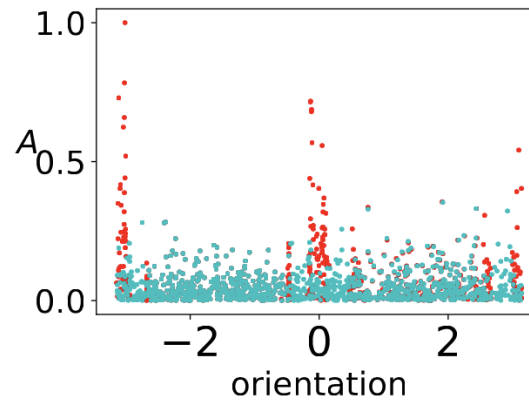
condensed regime



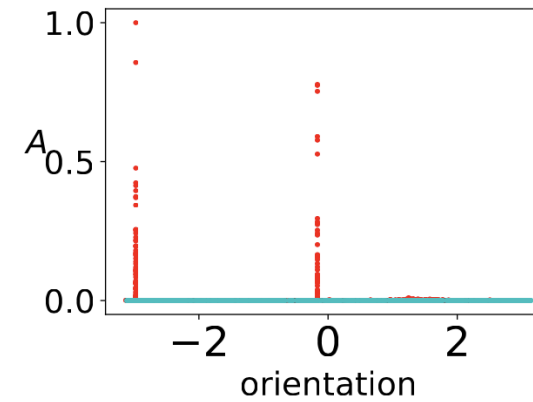
(c)  $\gamma = 1.75$



(d)  $\gamma = 0.5$



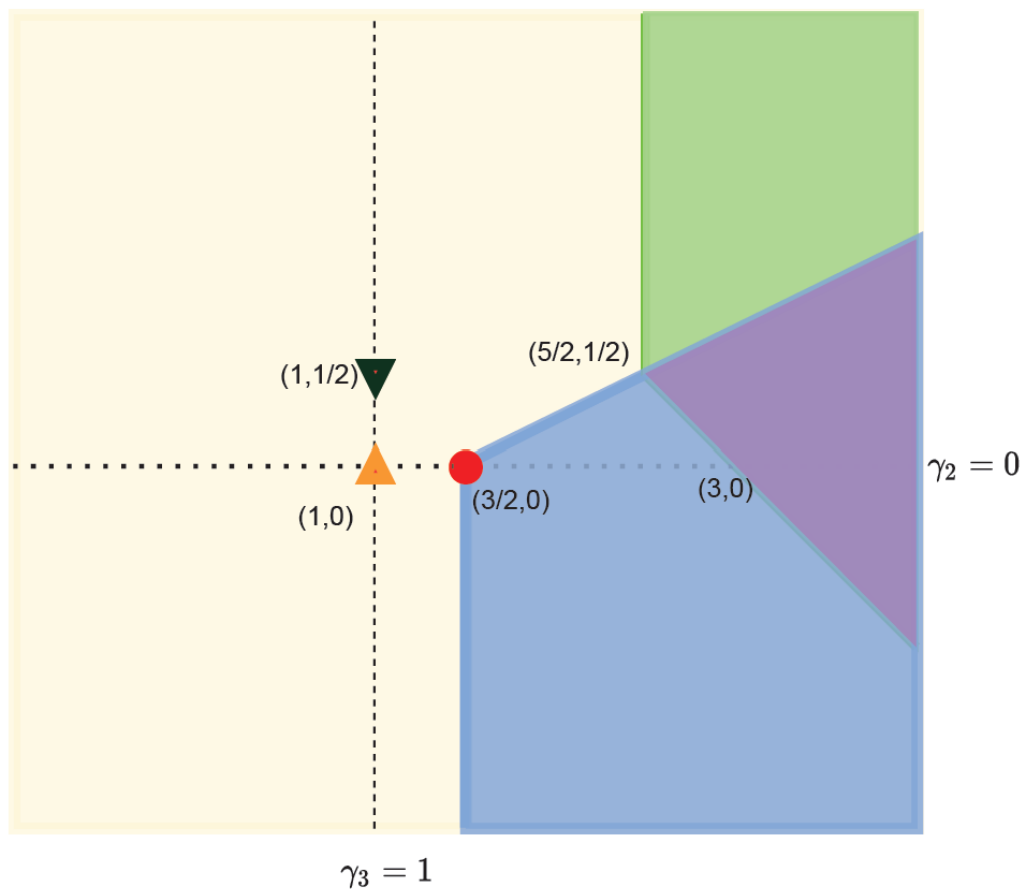
(e)  $\gamma = 1$



(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*

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

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

*CR is short for Condensed Reigme, LR is short for Linear Regime*



# Condensation facilitates reasoning



# Composite Anchor function

Single anchors

14 seen inferential composite anchors

Input data examples

Target

1	:	+5
2	:	+1
3	:	-2
4	:	-8

1	1	:	+10
---	---	---	-----

1	2	:	+6
---	---	---	----

...

4	4	:	-16
---	---	---	-----

Composition

Padding

1 seen non-inferential composite anchors

3	4	:	-6
---	---	---	----

1 unseen composite anchor

4	3	:	?
---	---	---	---

55 46 32 52 **28** **1** **1** 34 33

**38**

20 95 **43** **3** **1** 44 34 76 32

**46**

...

...

28 53 44 78 32 **62** **3** **4** 44

**56**

77 43 23 63 89 33 **52** **4** **3**

**?**

seen inferential anchor

seen non-inferential anchor

unseen anchor

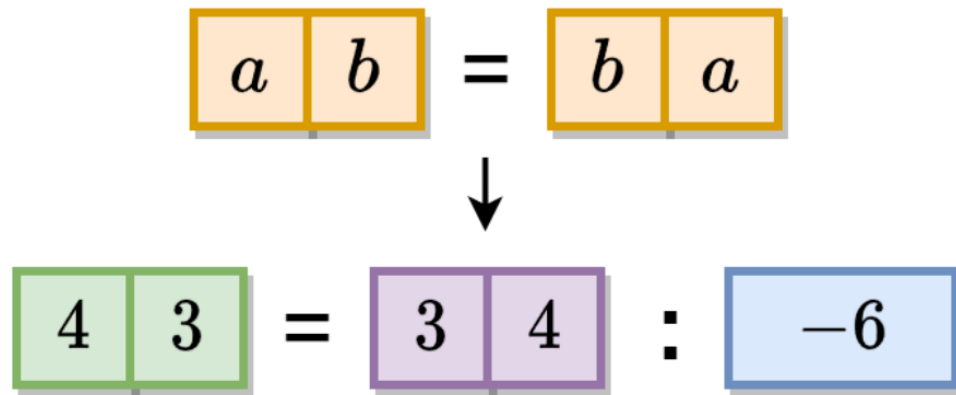
key item  
(item before anchor)



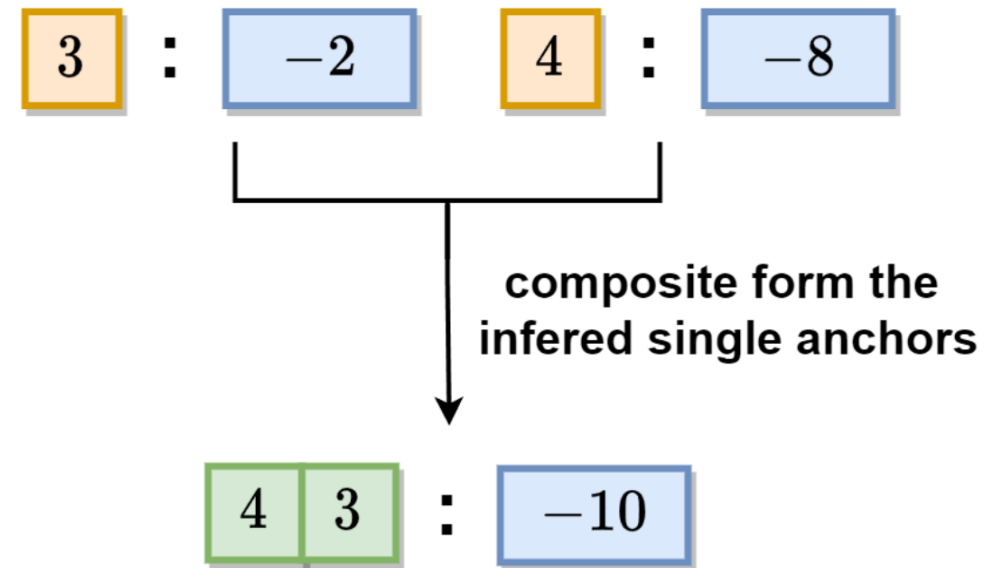


# Symmetric or inferred solutions?

Mechanism 1: learn symmetric structure



Mechanism 2: infer single anchor mappings

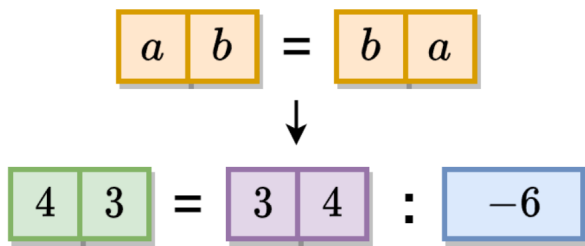




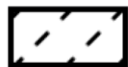
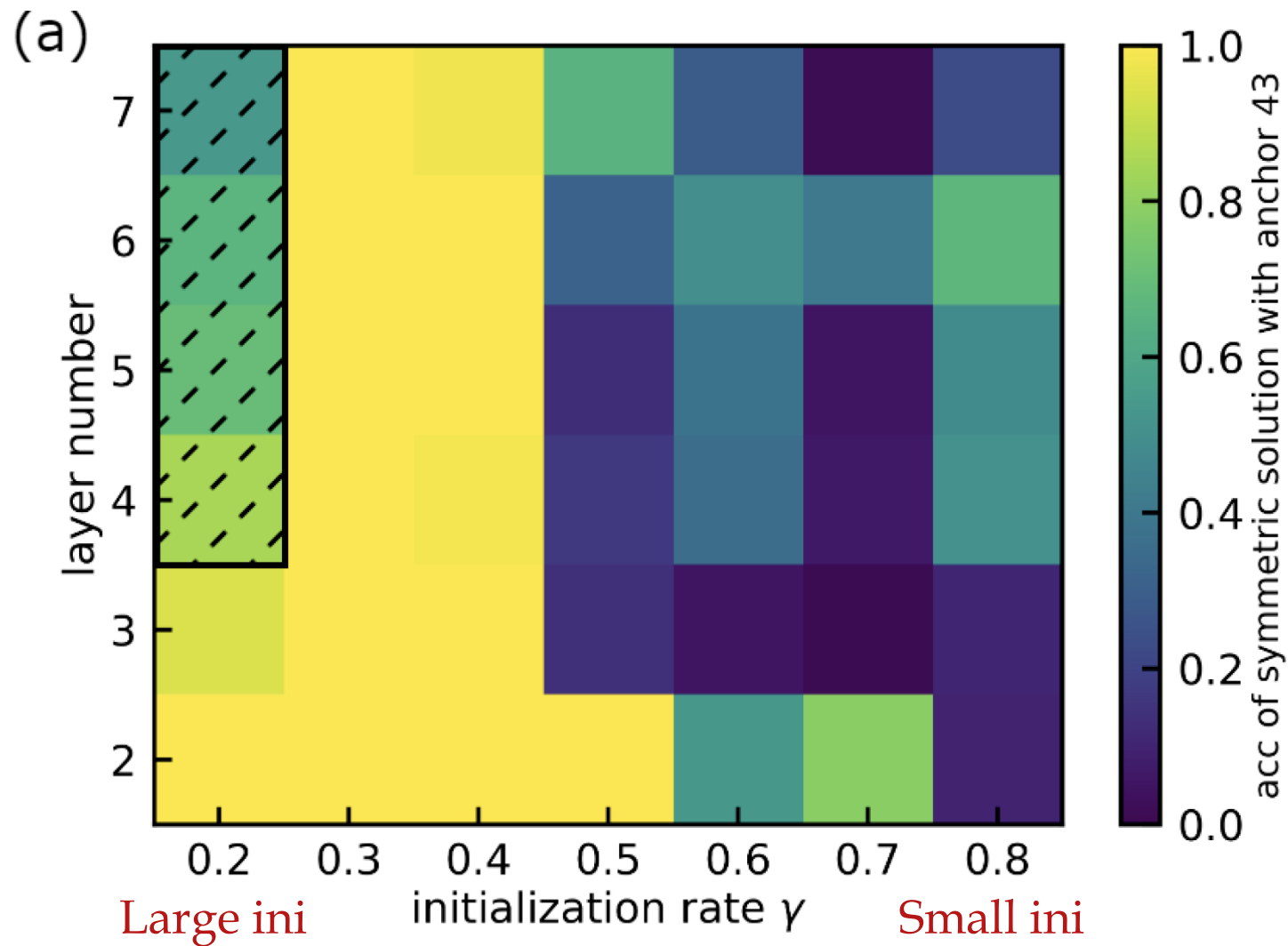


# Phase diagram of symmetric solution

Mechanism 1: learn symmetric structure



$$\text{Initialization} \sim N\left(0, \frac{1}{d_{in}^\gamma}\right)$$



bad generalization on seen anchors (test accuracy < 90%)





# Phase diagram of inferential solution

Mechanism 2: infer single anchor mappings

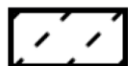
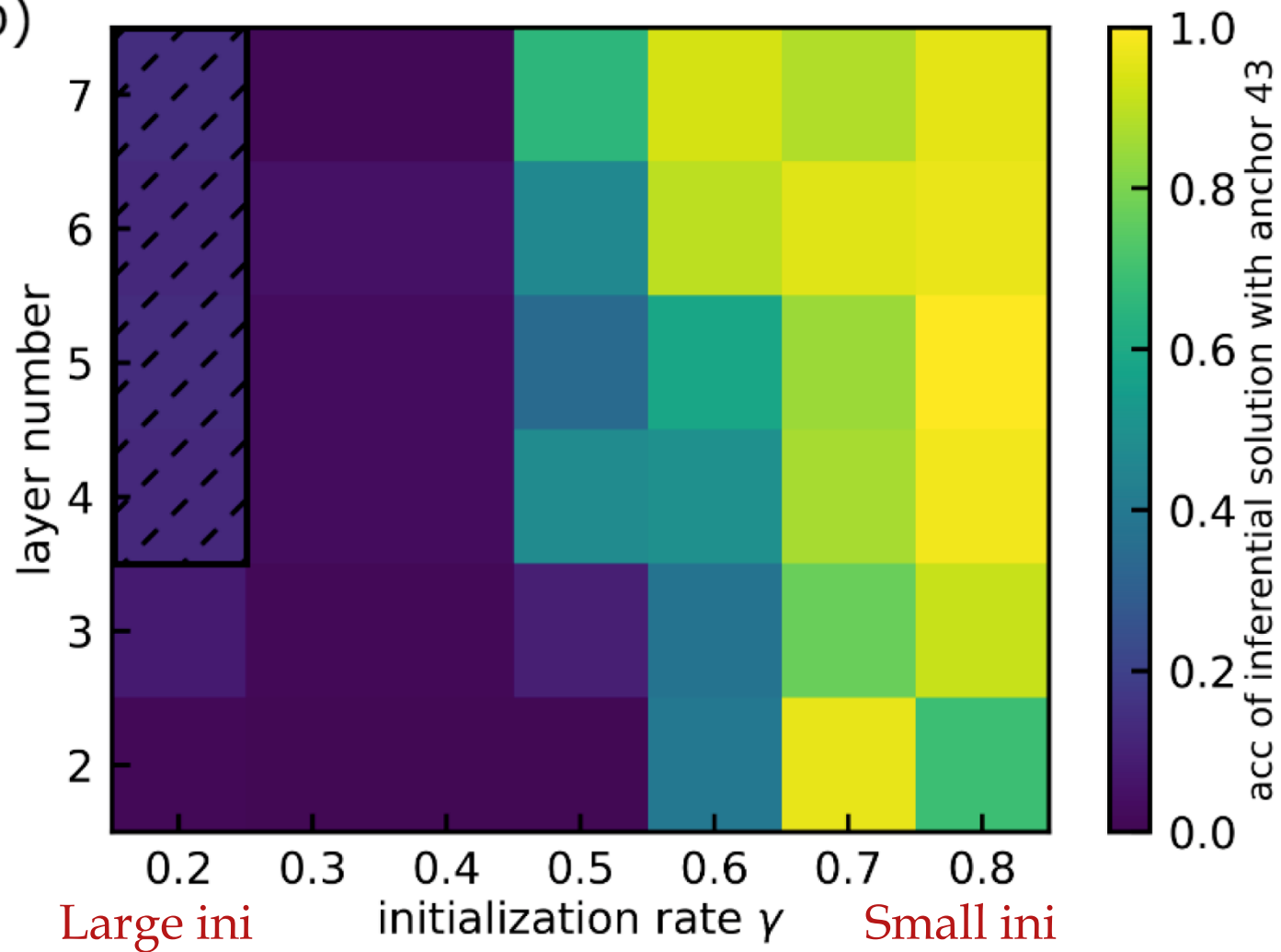


composite form the  
inferred single anchors



$$\text{Initialization} \sim N\left(0, \frac{1}{d_{in}^\gamma}\right)$$

(b)



bad generalization on seen anchors (test accuracy < 90%)





# Condensation of $W^{Q(1)}$ by column

$$Q^{(l)} = X^{(l)} W^{Q(l)}$$

$X^{(1)}$



$W^{Q(1)}$

output:  $Q^{(1)}$

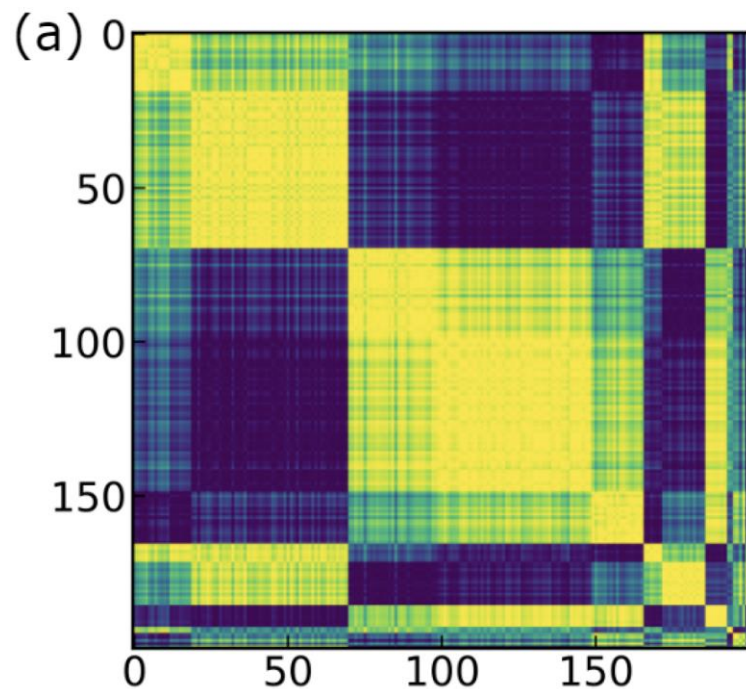


Cosine Similarity b/w

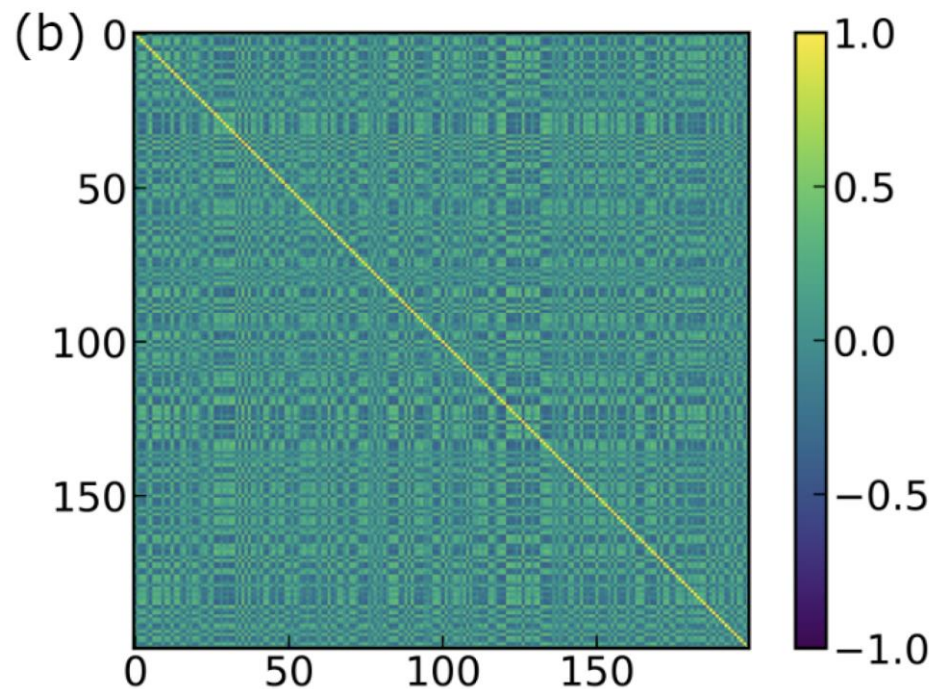
Columns of  $W^{Q(1)}$

Input weight of Q neurons

Small ini: clear condensation



Large ini: no condensation





# Complexity of solutions

Mechanism 1: learn symmetric structure

$$\begin{array}{c} \boxed{a} \boxed{b} = \boxed{b} \boxed{a} \\ \downarrow \\ \boxed{4} \boxed{3} = \boxed{3} \boxed{4} : \boxed{-6} \end{array}$$

Not only one pair

But ten pairs in training

11;12,21; 13, 31; 14, 41; 23, 32;...

Mechanism 2: infer single anchor mappings

$$\begin{array}{c} \boxed{3} : \boxed{-2} \quad \boxed{4} : \boxed{-8} \\ \downarrow \\ \text{composite form the} \\ \text{inferred single anchors} \\ \boxed{4} \boxed{3} : \boxed{-10} \end{array}$$

Need to learn four functions

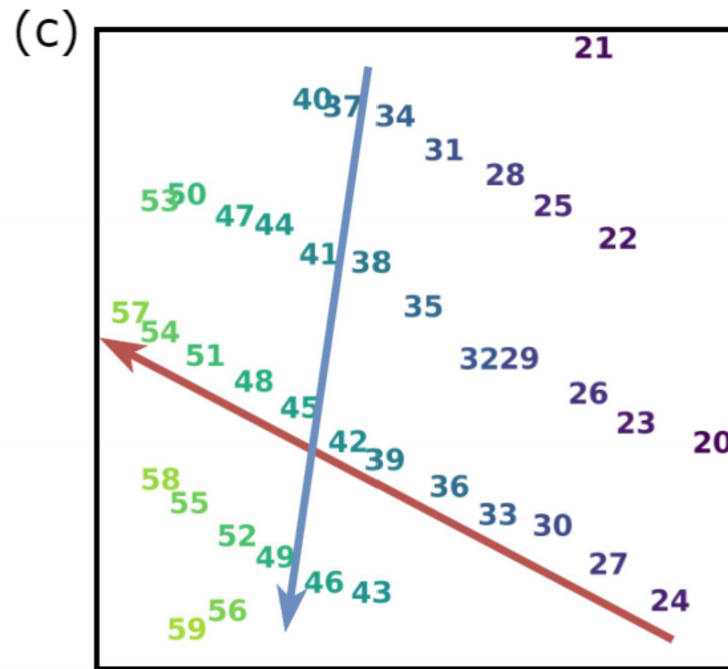




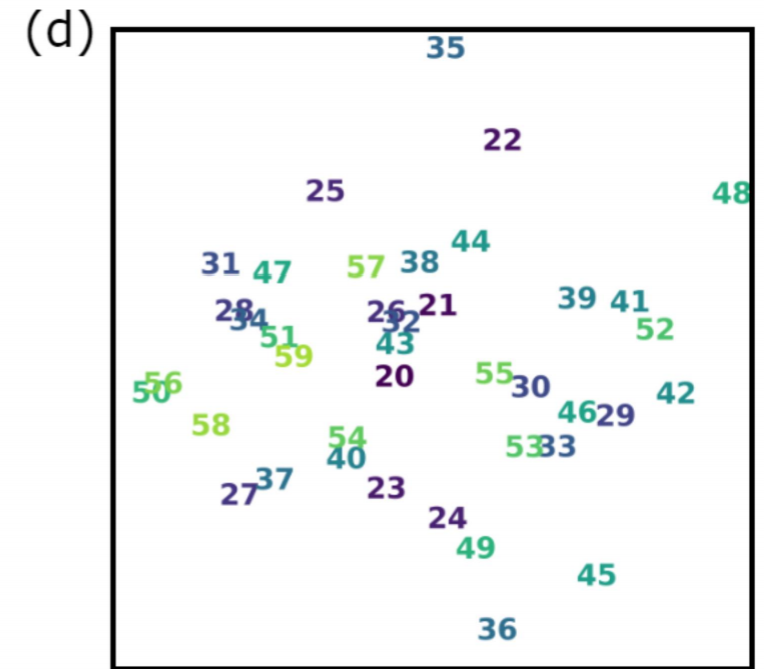
# Structure of embedding space

$$f_a(x) = \begin{cases} x + 5, & \text{if } a = 1 \\ x + 1, & \text{if } a = 2 \\ x - 2, & \text{if } a = 3 \\ x - 8, & \text{if } a = 4 \end{cases}$$

Small init: clear structure



Large init: no structure

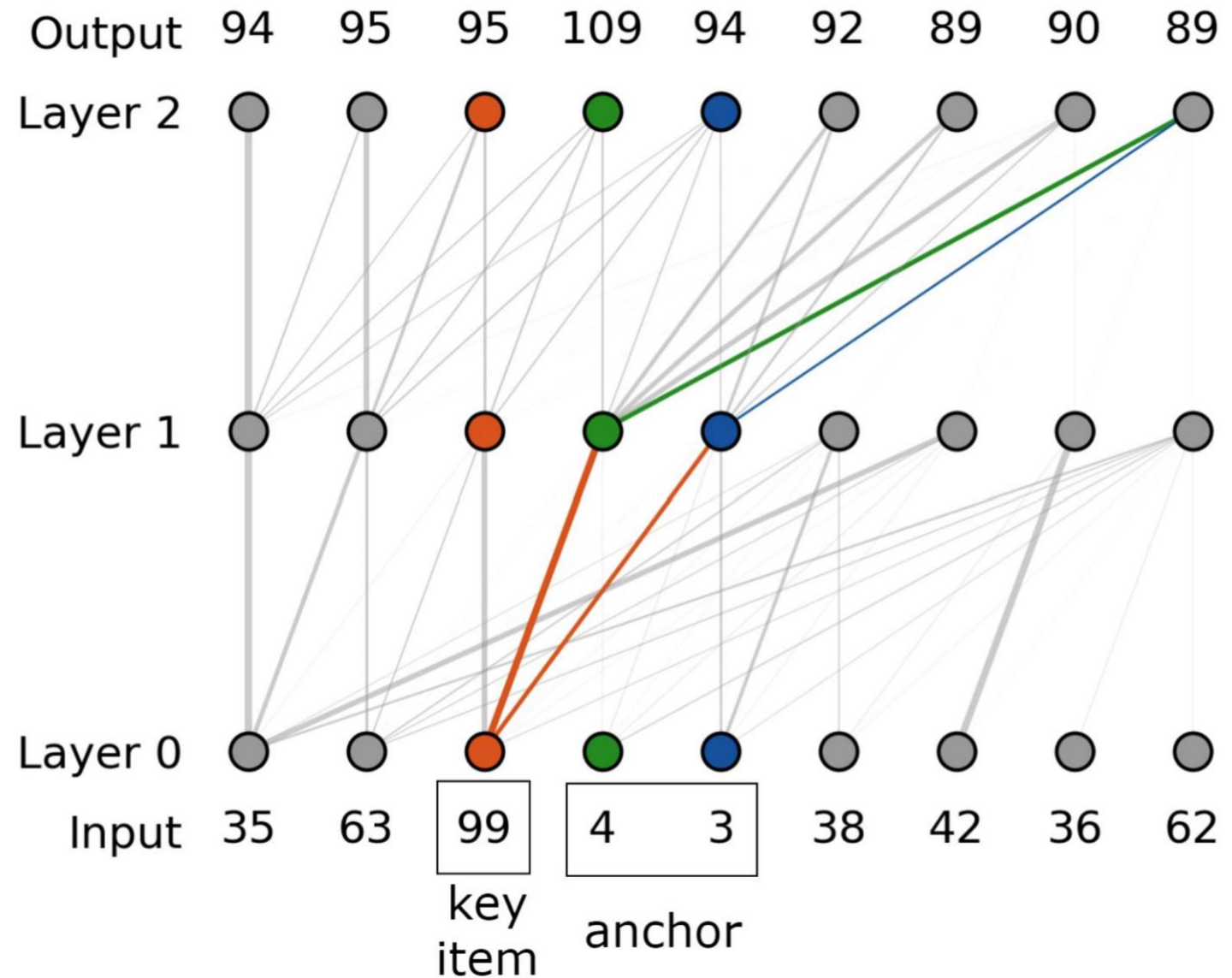
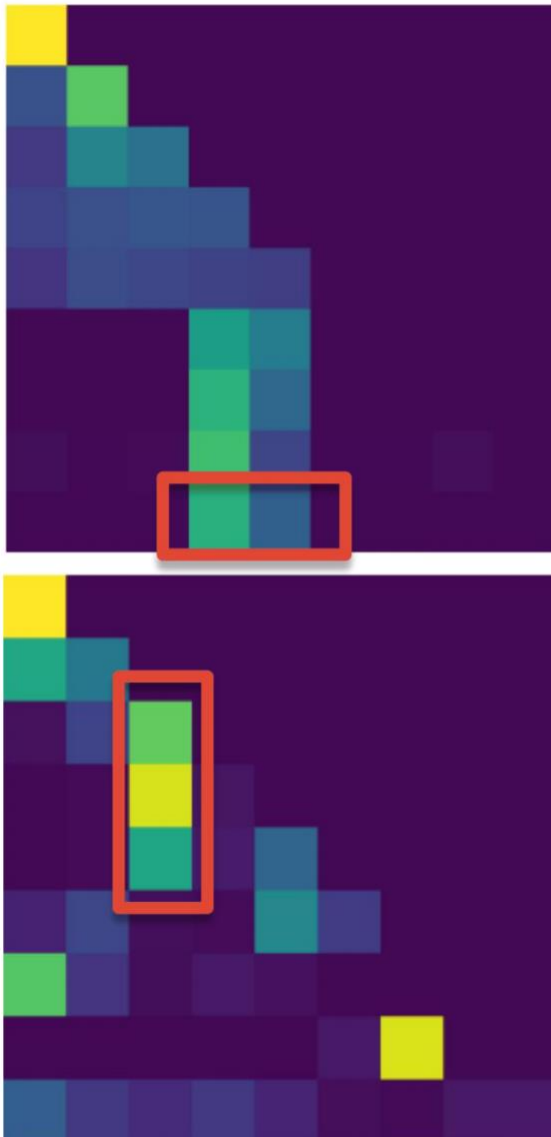








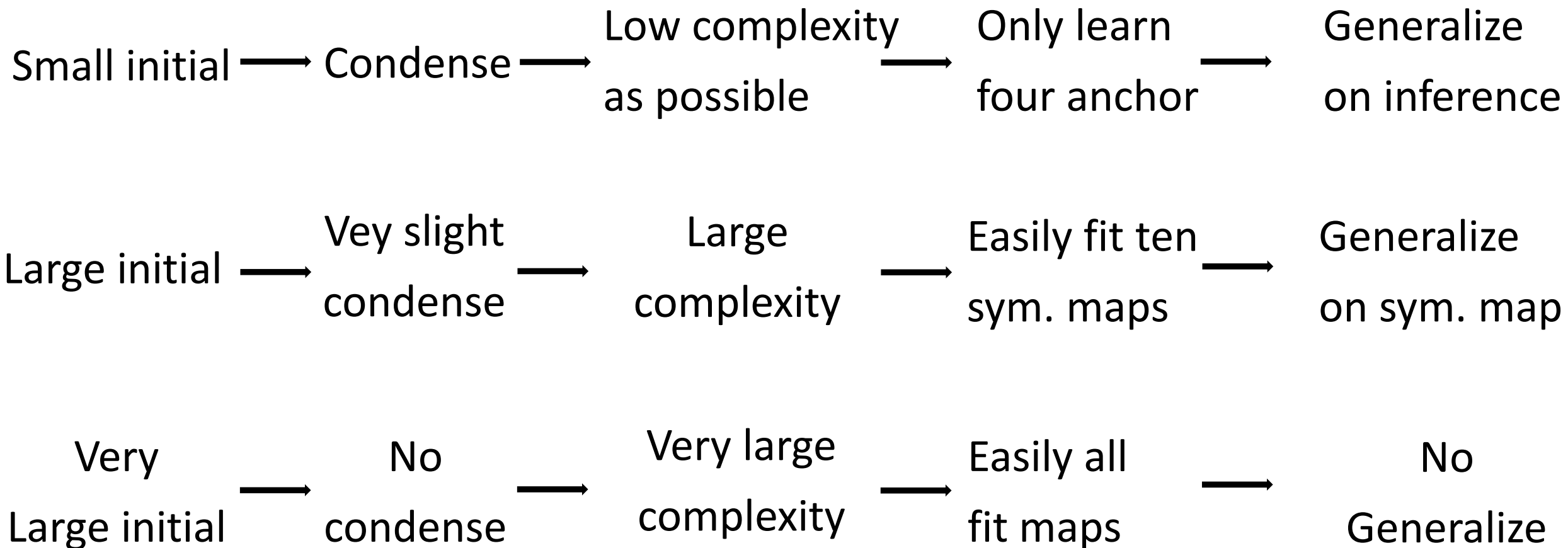
# Inferential solution





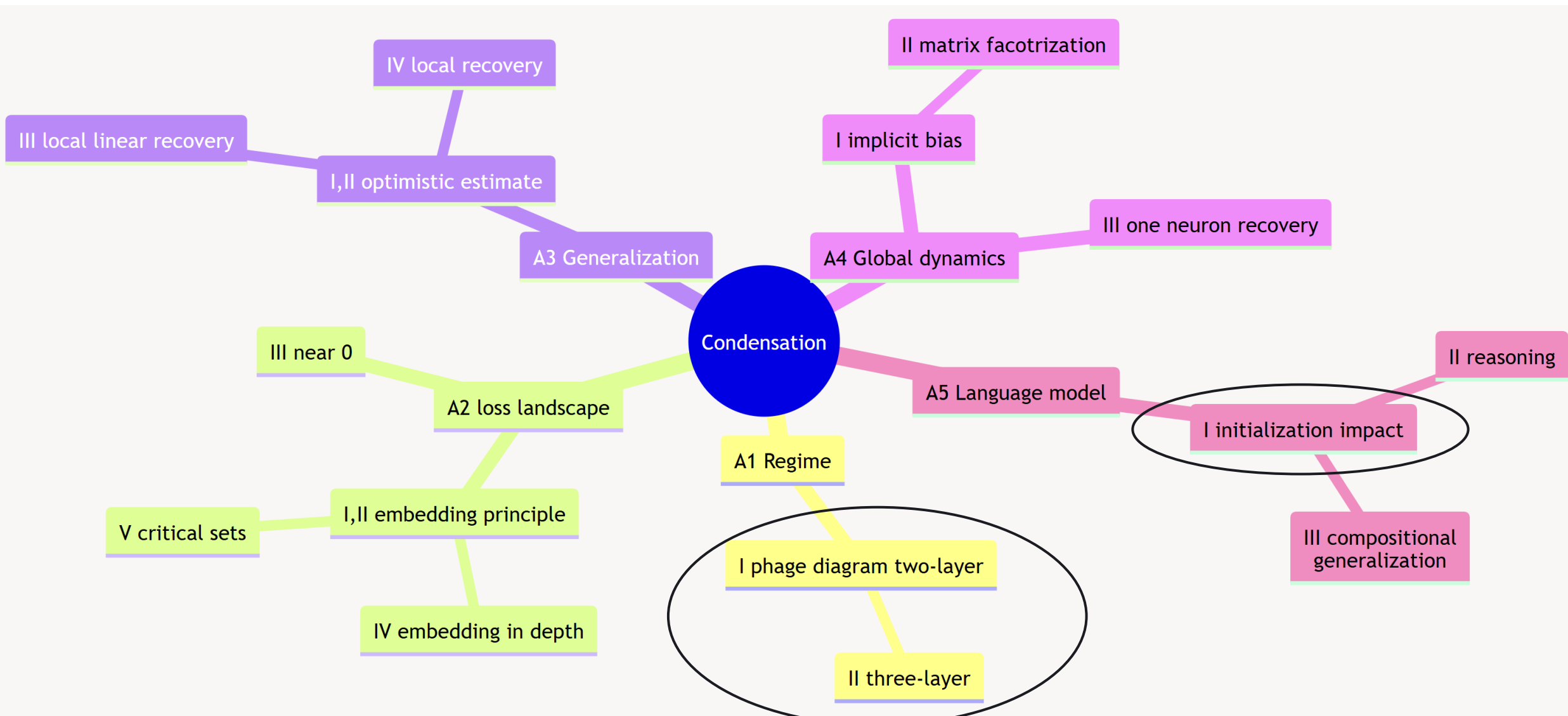


# Mechanisms underlying initialization effect





# Condensation





- ① How can condensation be facilitated in a neural network?
- ① Is it valid to compare the performance of wide and narrow networks when the initialization variance is fixed?
- ① What initialization strategy can be used for a three-layer network to induce condensation in the first hidden layer but not in the second?
- ① Where condensation can happen within a transformer?



---

# Thanks!

---

饮水思源 爱国荣校