III. Condensation phenomenon

Yaoyu Zhang

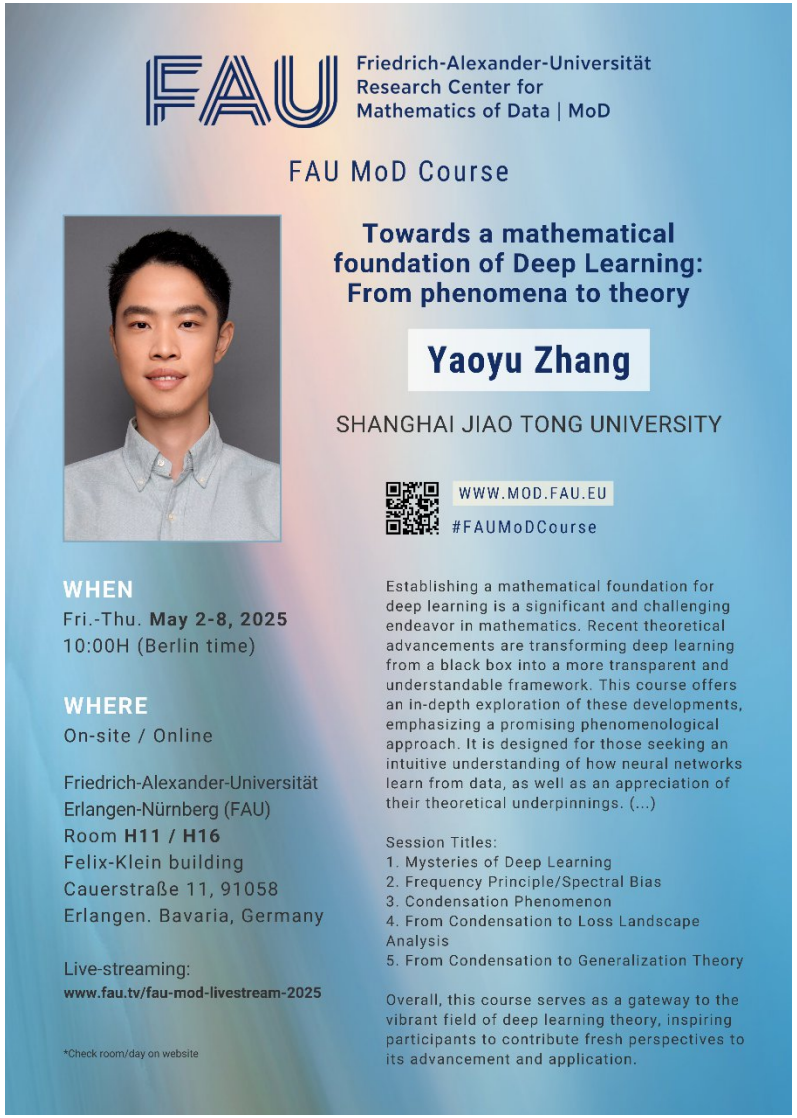
Institute of Natural Sciences & School of Mathematical Sciences

Shanghai Jiao Tong University

FAU MoD Course

饮 水 思 源 • 爱 国 荣 校

Deep learning is no longer a black-box



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

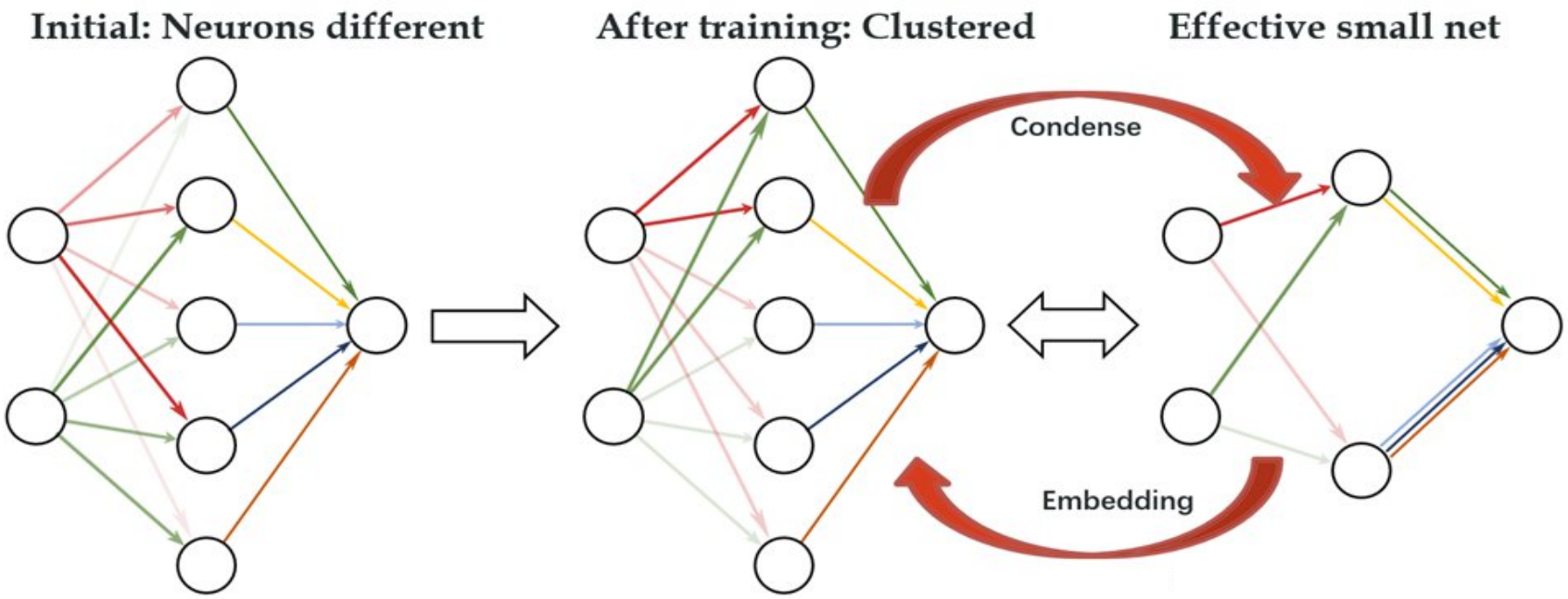
Date

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

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

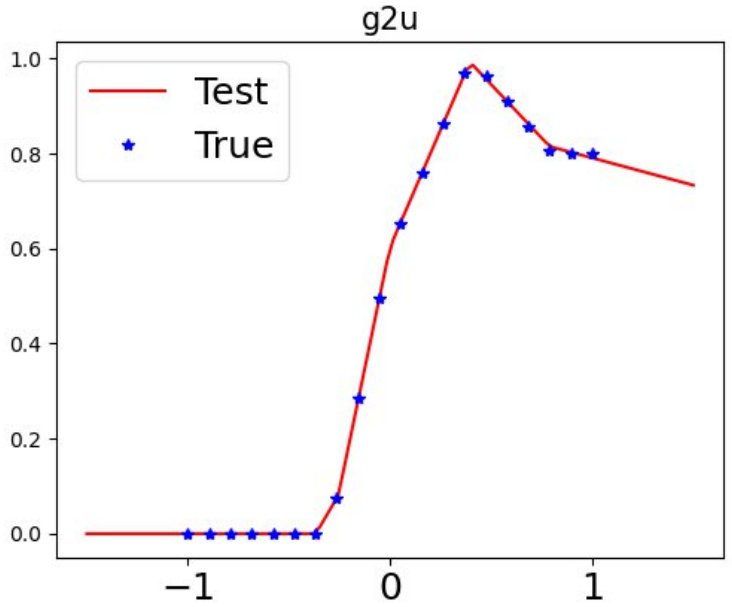
Effect: equiv to small net

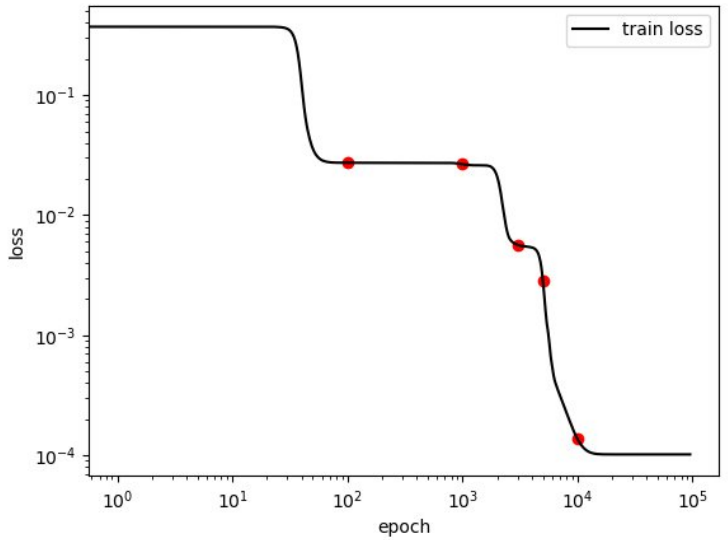
1d example: condensation with small initialization

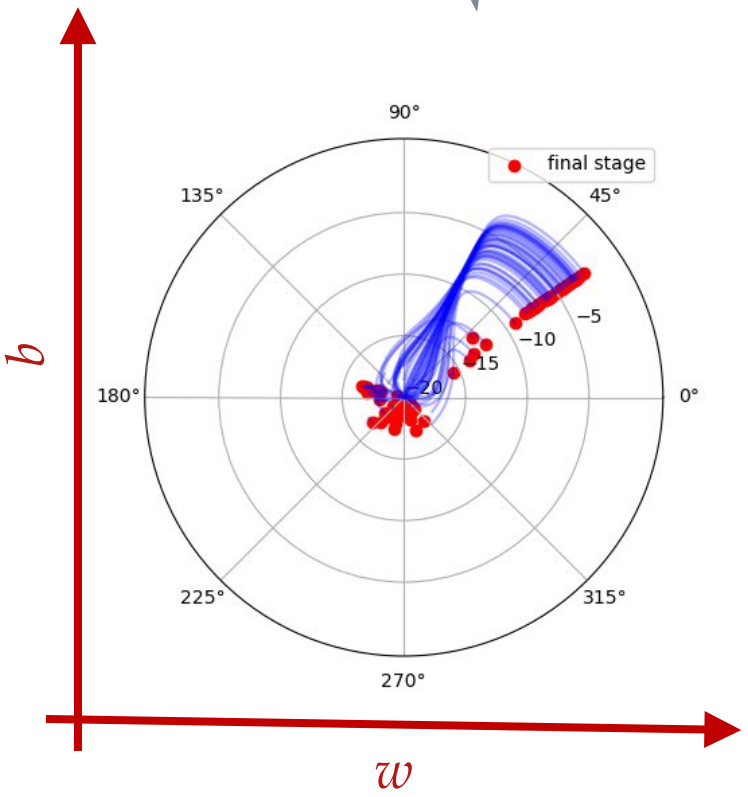
$$f_\theta(x) = \sum_{j=1}^{m} a_j \,\mathrm{relu}\,(w_j x + b_j)$$

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

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

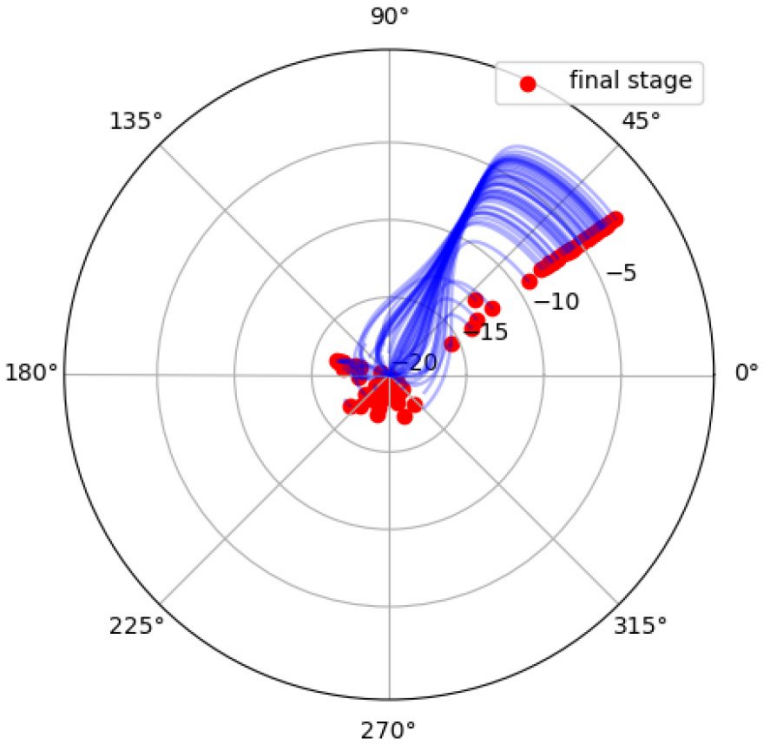




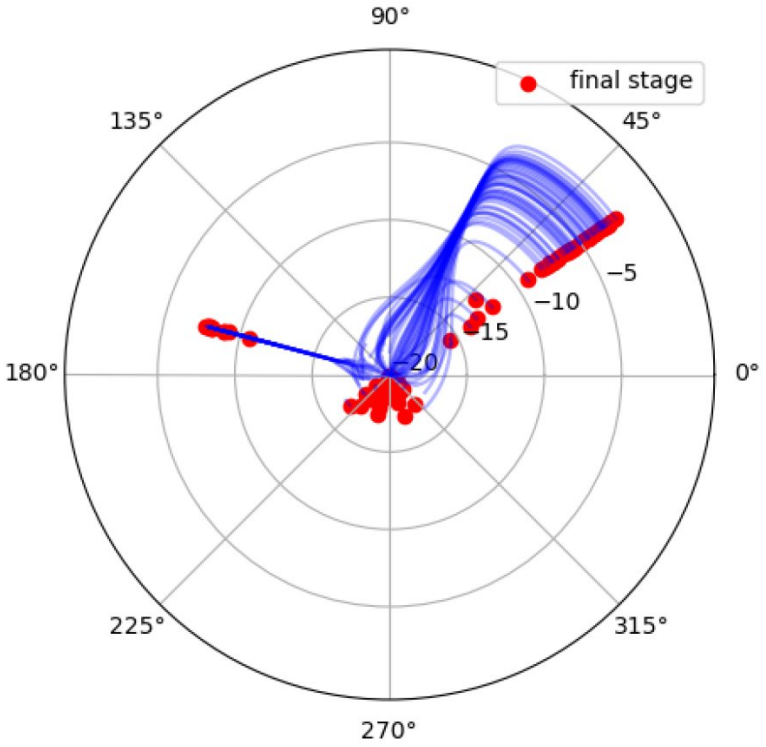


Small initialization: $a_j(0), w_j(0), b_j(0) \sim N(0, \sigma^2)$ with small σ

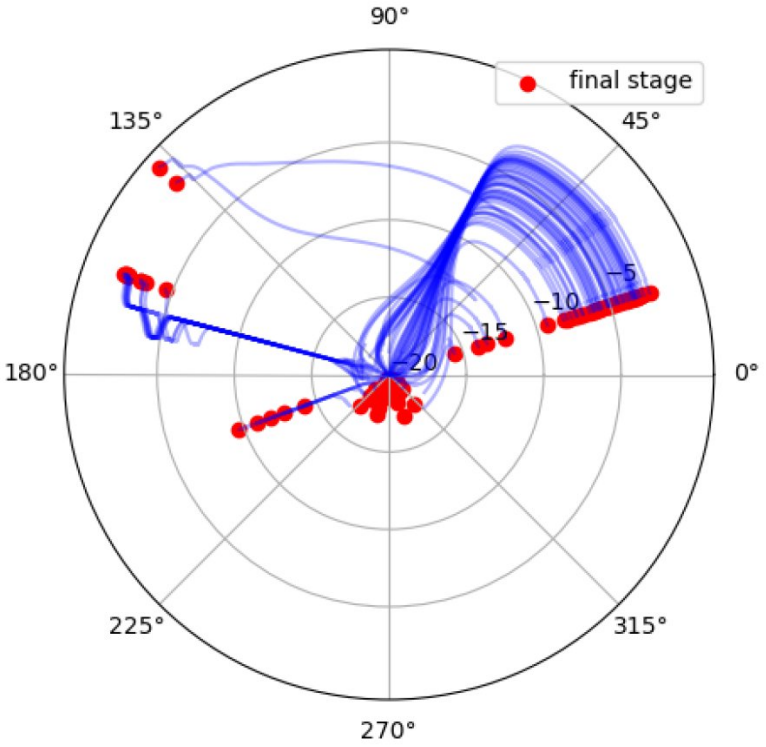
Evolution trajectory: change significantly



(a) epoch=100

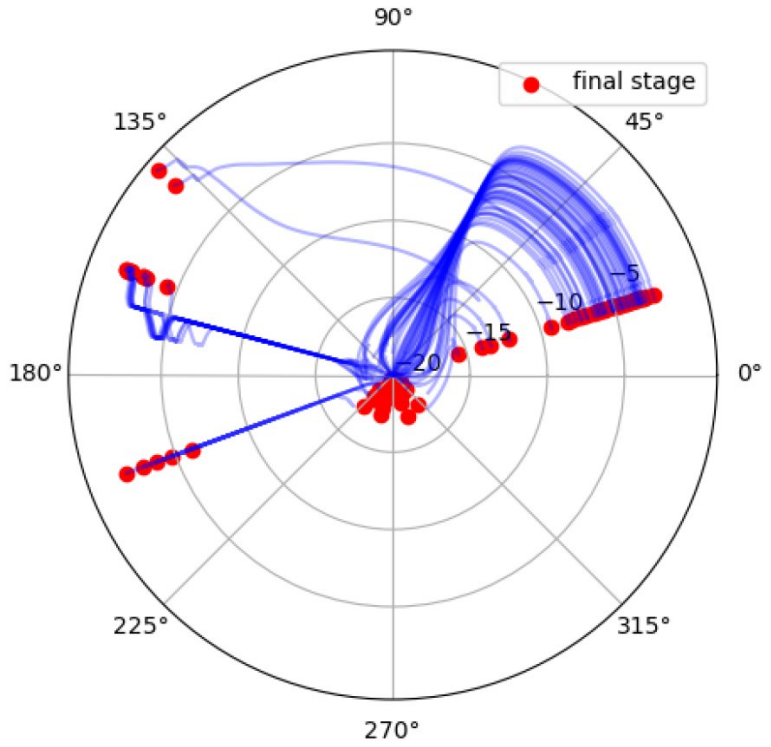


(b) epoch=1000

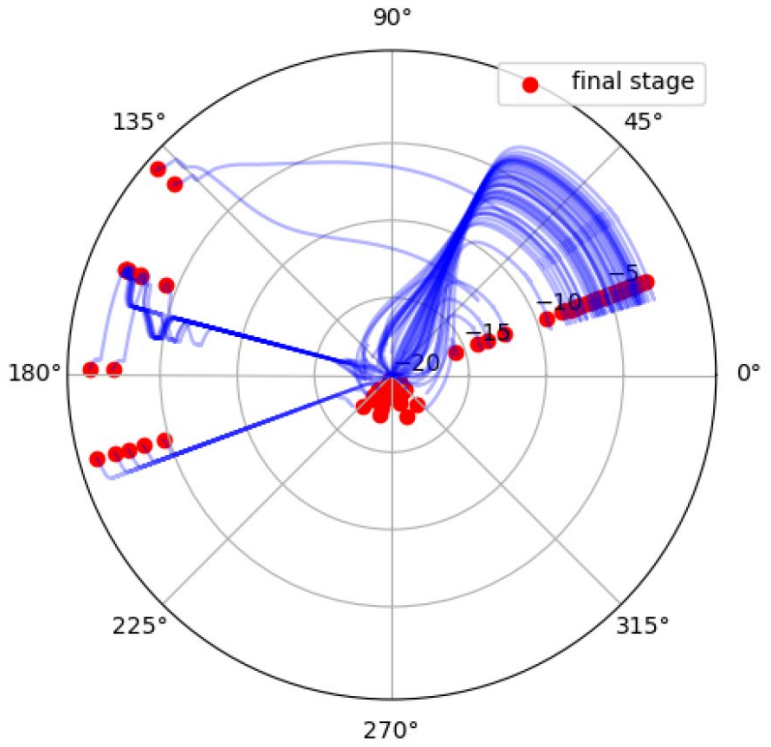


(c) epoch=3000

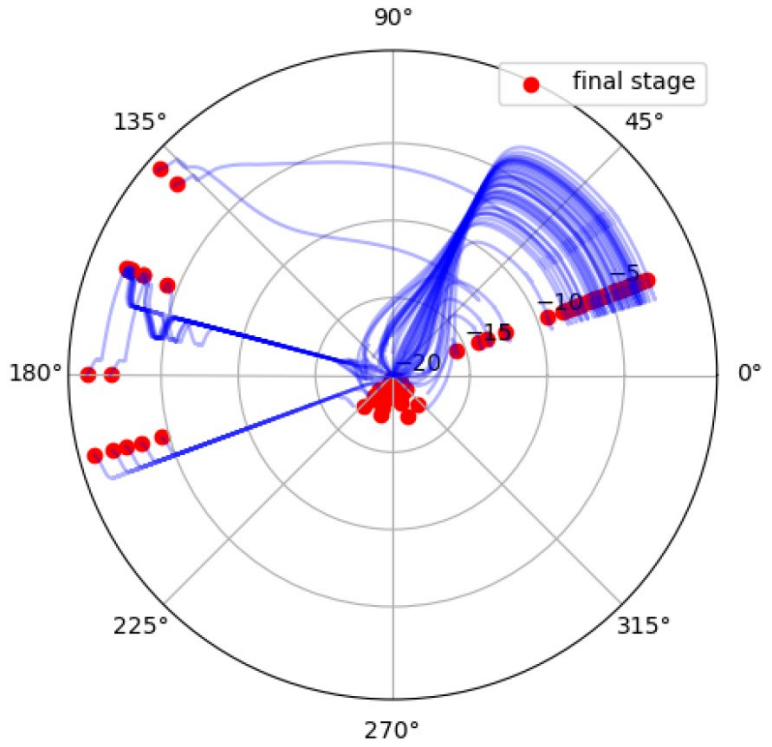
Evolution trajectory: change significantly



(d) epoch=5000

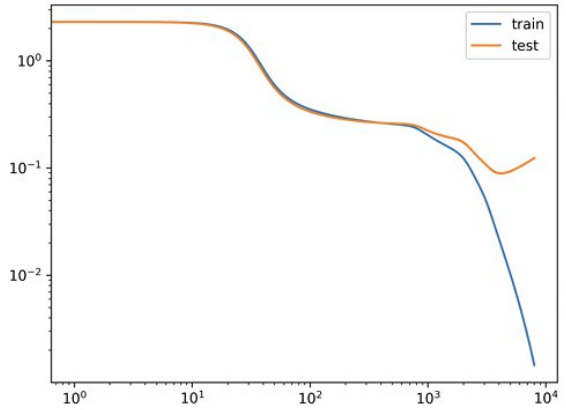


(e) epoch=10000

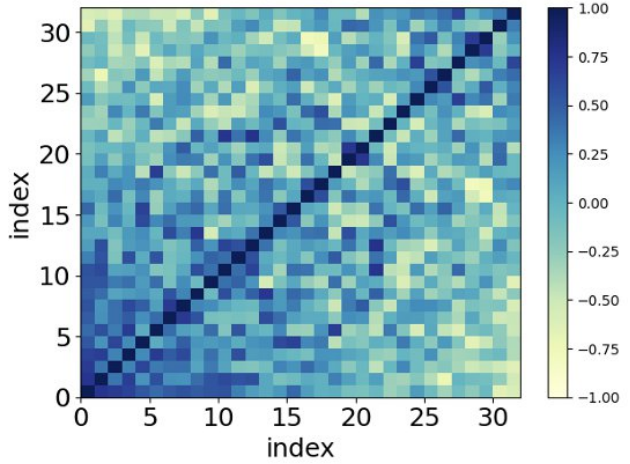


(f) epoch=100000

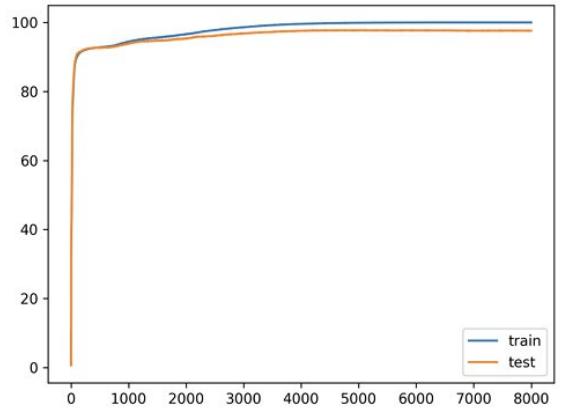
Condensation in CNN on MNIST



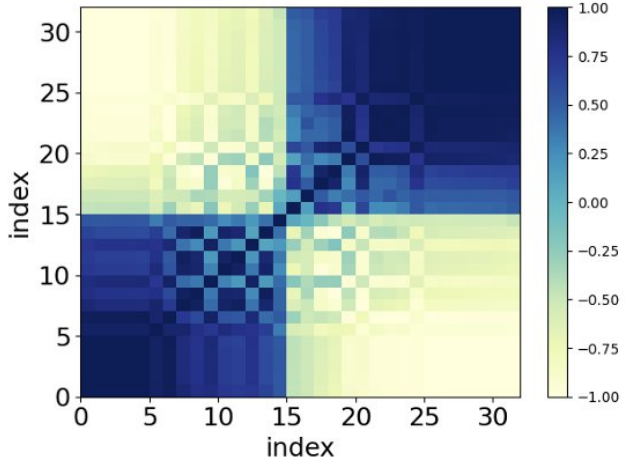
(a) Loss



(b) initial weight



(d) Accuracy



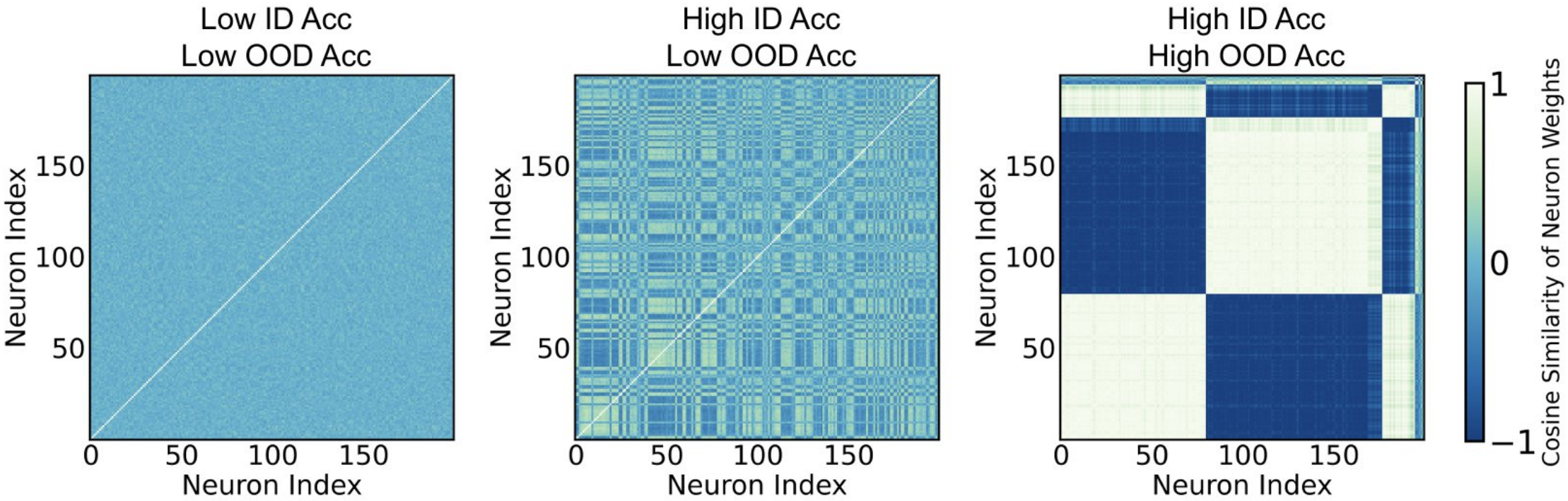
(e) final weight

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.

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

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

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

- Loss

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

- Gradient flow dynamics

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}).$$

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

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

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 α, β_1, β_2

Goal:

Identify dynamical regimes of training over α, β_1, β_2 at infinite-width limit.

Initialization methods with their scaling parameters

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	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$	$\lim_{m\to\infty}\frac{\log 1/\beta}{\log m}$

Normalization and scaling parameters

- Two layer ReLU network

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

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

- Normalized gradient flow

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

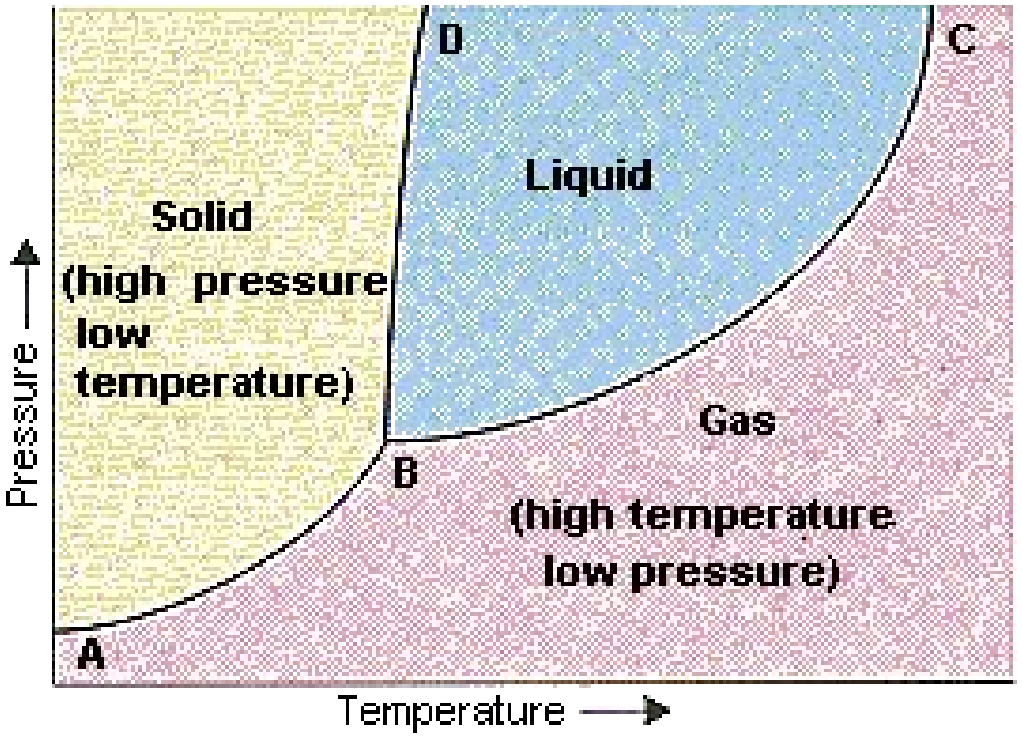
$$\frac{\mathrm{d}\bar{a}_k}{\mathrm{d}\bar{t}} = -\frac{1}{\kappa'}\frac{1}{n}\sum_{i=1}^{n} \kappa \sigma(\bar{\boldsymbol{w}}_k^{\intercal}\boldsymbol{x}_i)\left(\kappa \sum_{k'=1}^{m} \bar{a}_{k'}\sigma(\bar{\boldsymbol{w}}_{k'}^{\intercal}\boldsymbol{x}_i) - y_i\right),$$

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

- Scaling parameters and infinite-width limit

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

Phase diagram



- **Phase diagram for matter**

distinctive states of matter <-> environment (phase transition happens at infinite size limit)

solid, liquid, gas <-> pressure, temperature

- **Phase diagram for two-layer ReLU NN**

training dynamics <-> initialization ($m \to \infty$)

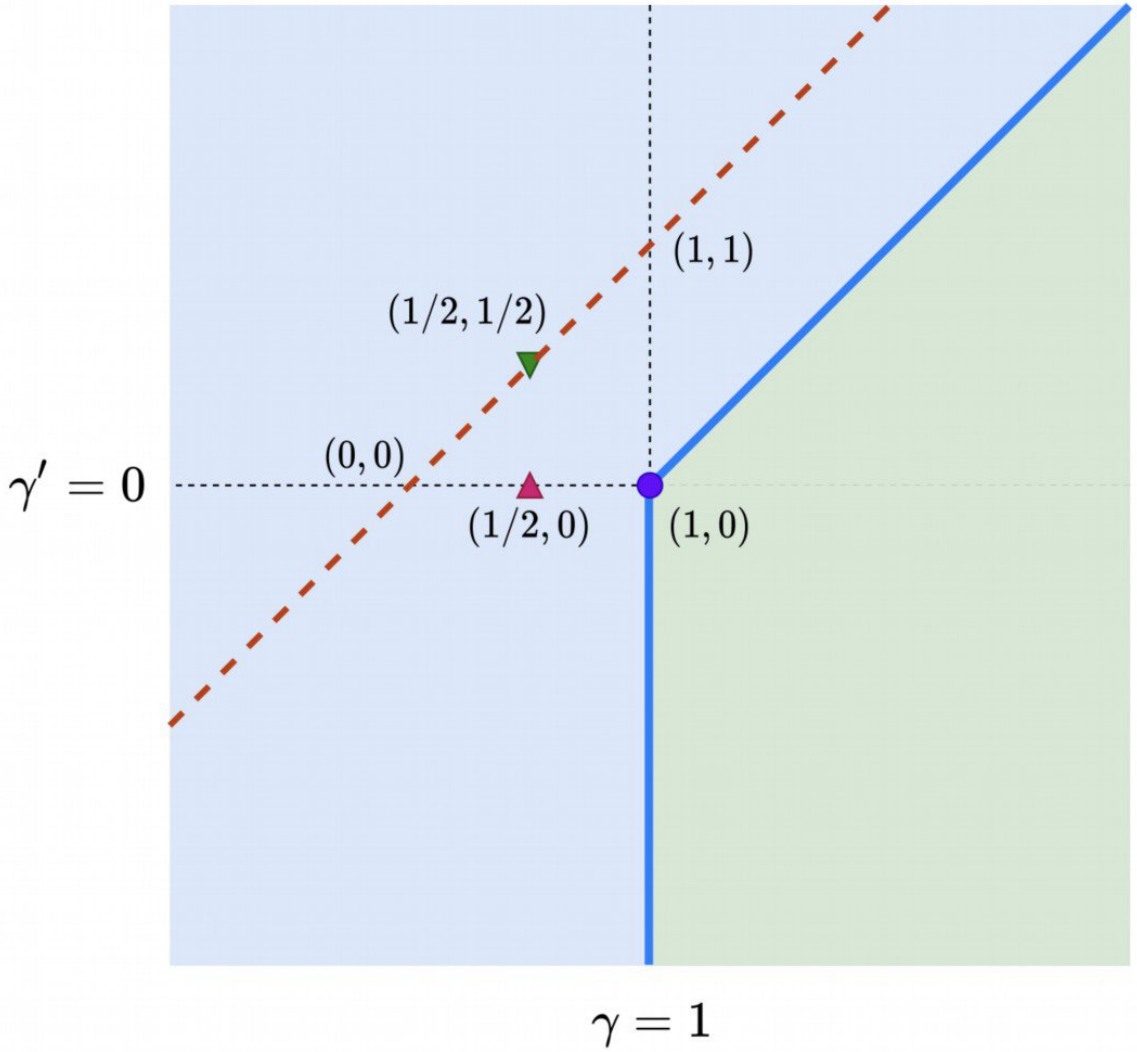
? <->?

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

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

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

Phase Diagram



Linear regime

Condensed regime

Critical regime

Examples:

- Xavier, Mean field
- NTK
- E at el. (2020)
- LeCun, He

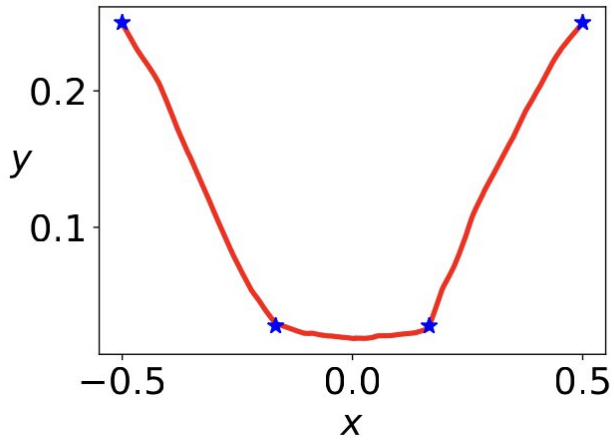
$$a_k^0 \sim N(0, \beta_1^2), \quad \mathbf{w}_k^0 \sim N(0, \beta_2^2 \mathbf{I}_d)$$

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

Typical cases across the phase diagram

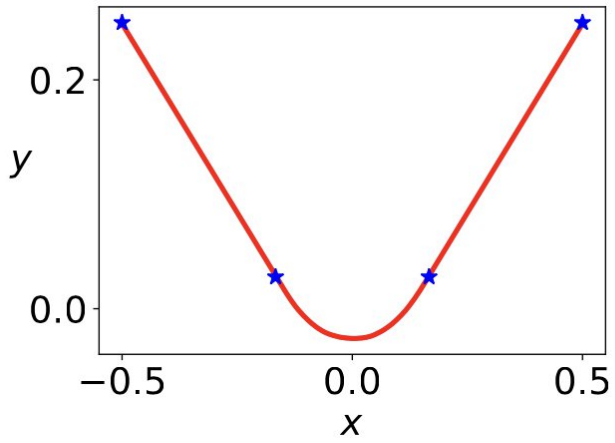
$\gamma' = 0$

Linear regime



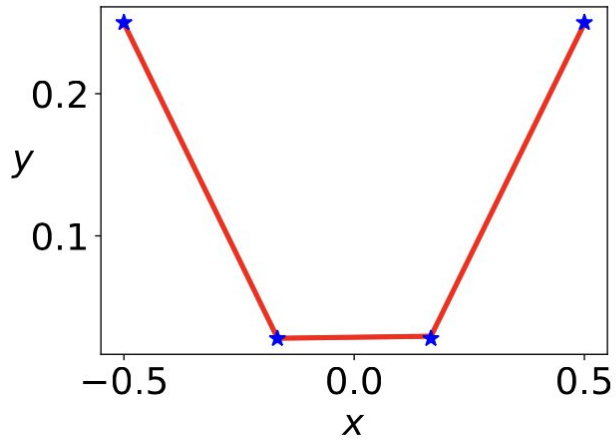
(a) $\gamma = 0.5$

critical regime



(b) $\gamma = 1$

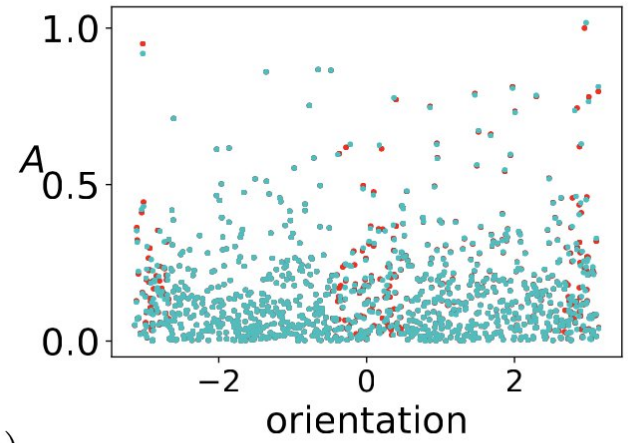
condensed regime



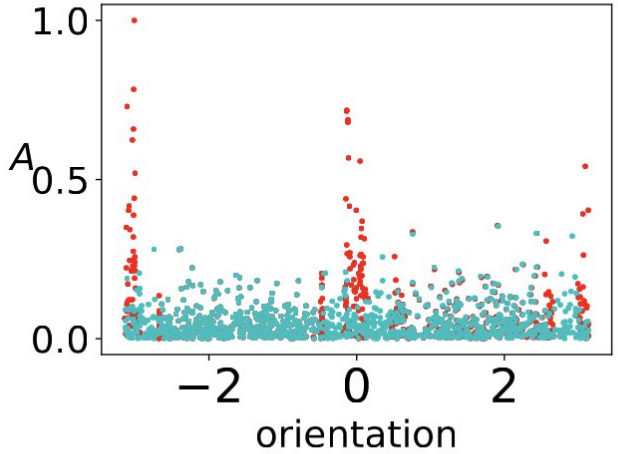
(c) $\gamma = 1.75$

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

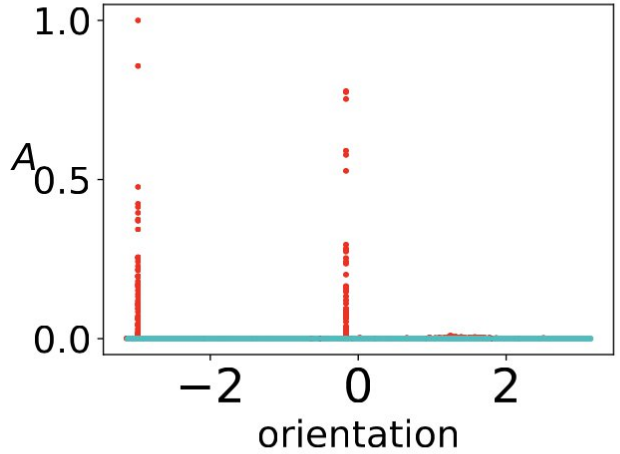
$A = |a| \|\boldsymbol{w}\|_2$



(d) $\gamma = 0.5$



(e) $\gamma = 1$



(f) $\gamma = 1.75$

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

Regime identification

- Linear regime (with ASI)

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

- Relative distance

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

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

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

As $m \to \infty$,

- Linear regime: $\sup_{t \in [0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to 0$
- Condensed regime: $\sup_{t \in [0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to +\infty$
- Critical regime: $\sup_{t \in [0,+\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) \to O(1)$

Scaling analysis

- Two layer ReLU network at infinite-width limit

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

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

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

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

$$\kappa \gtrsim 1/m$$

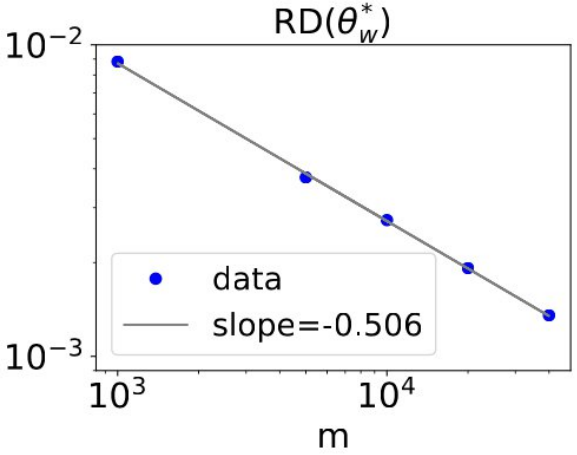
- output-layer dominant: $C = m\beta_2 \mathbb{E}(|a|)/\alpha \ll m\beta_2^2/\alpha = m\kappa/\kappa'.$

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

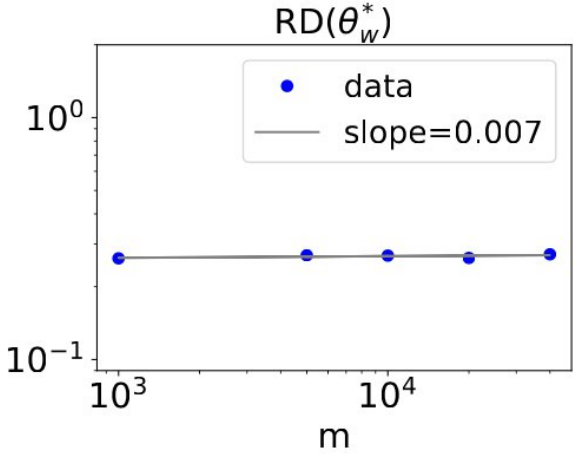
Regime identification through experiments

$\gamma' = 0$

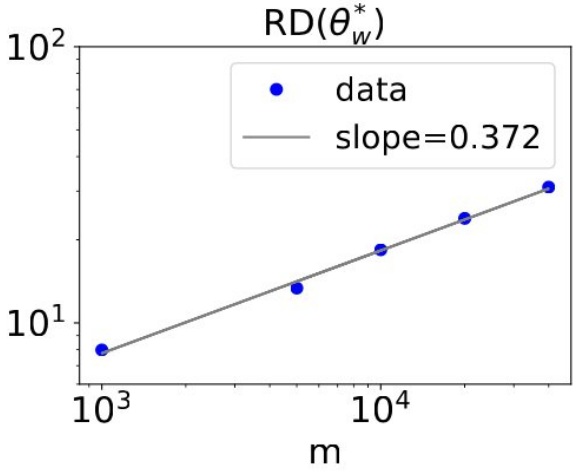
$$S_{\boldsymbol{w}} = \lim_{m \to \infty} \frac{\log \mathrm{RD}(\boldsymbol{\theta}^*_{\boldsymbol{w}})}{\log m}$$



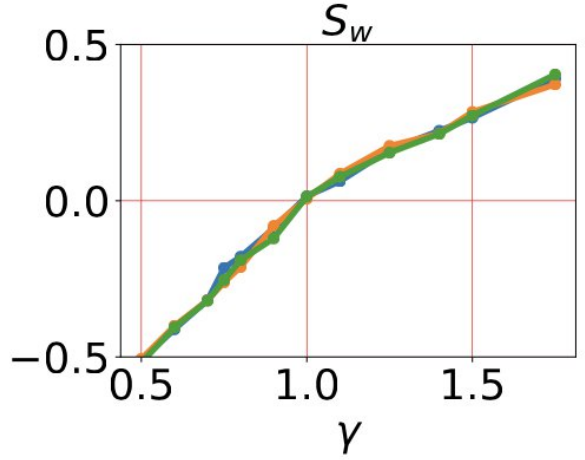
(a) $\gamma = 0.5$



(b) $\gamma = 1$



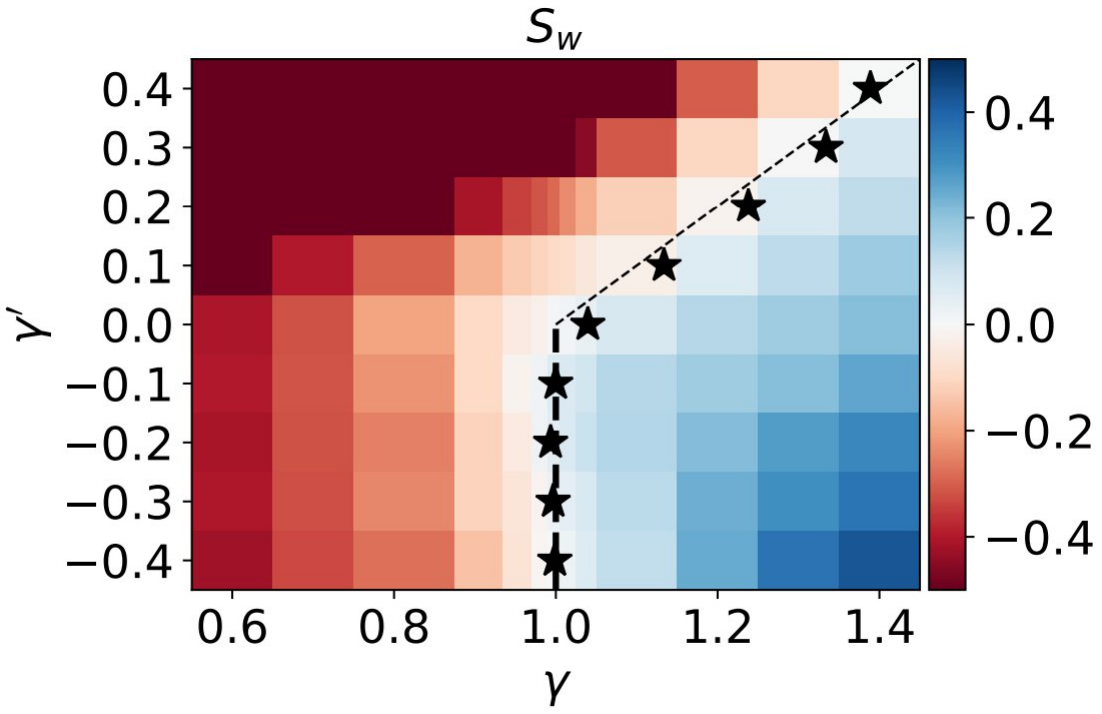
(c) $\gamma = 1.75$



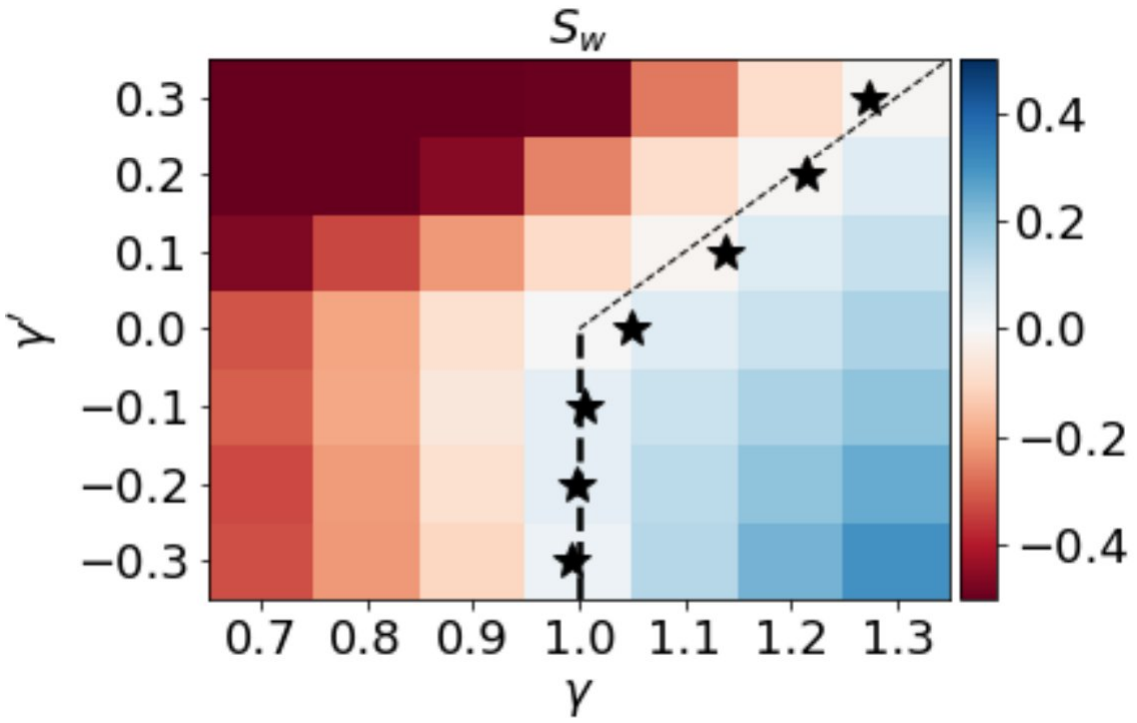
(d) S_w vs. γ

Regime identification through experiments

Synthetic data



MNIST data

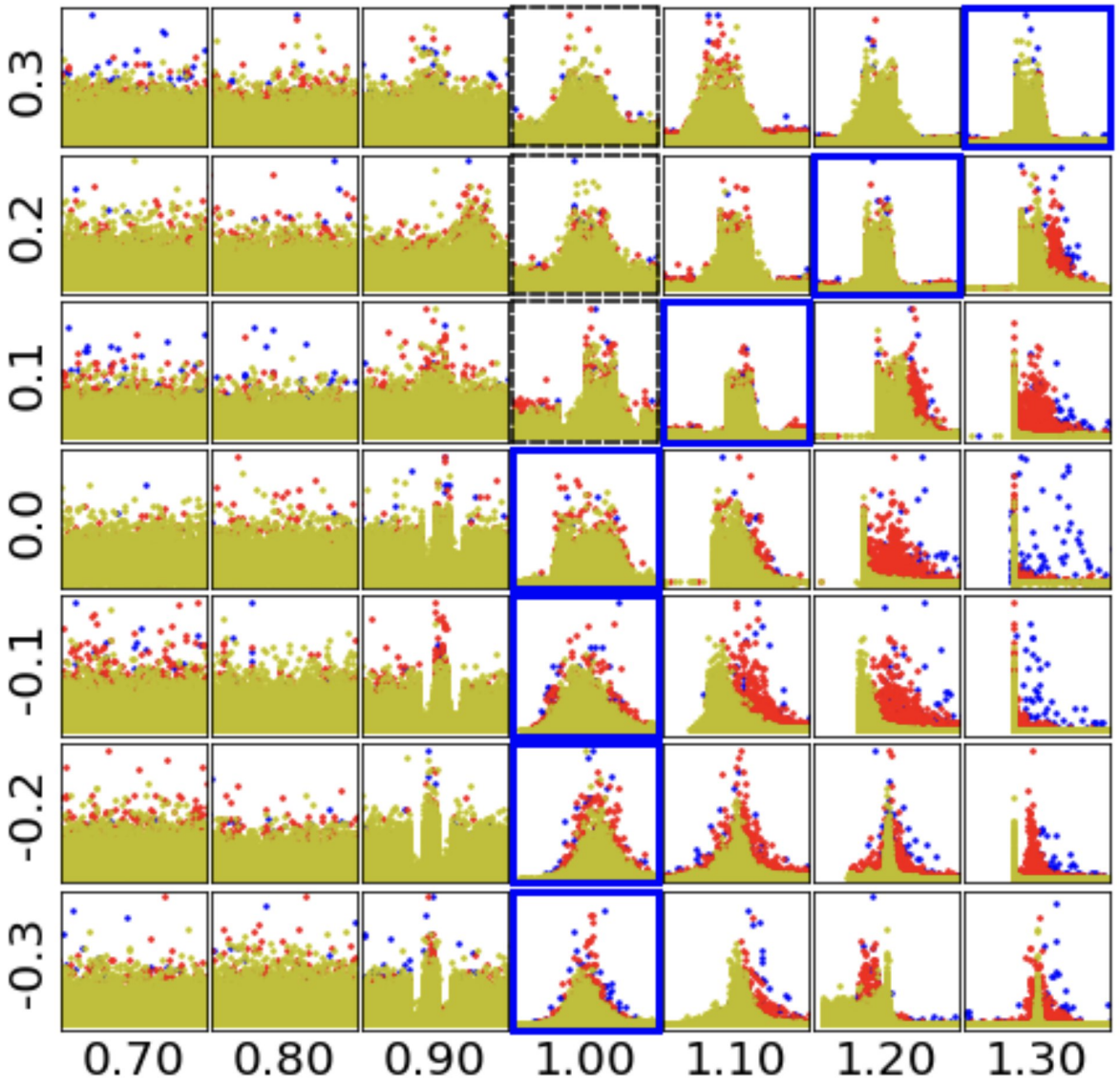


Feature distribution at the condensed regime-synthetic

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

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

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



Blue: $m = 10^3$

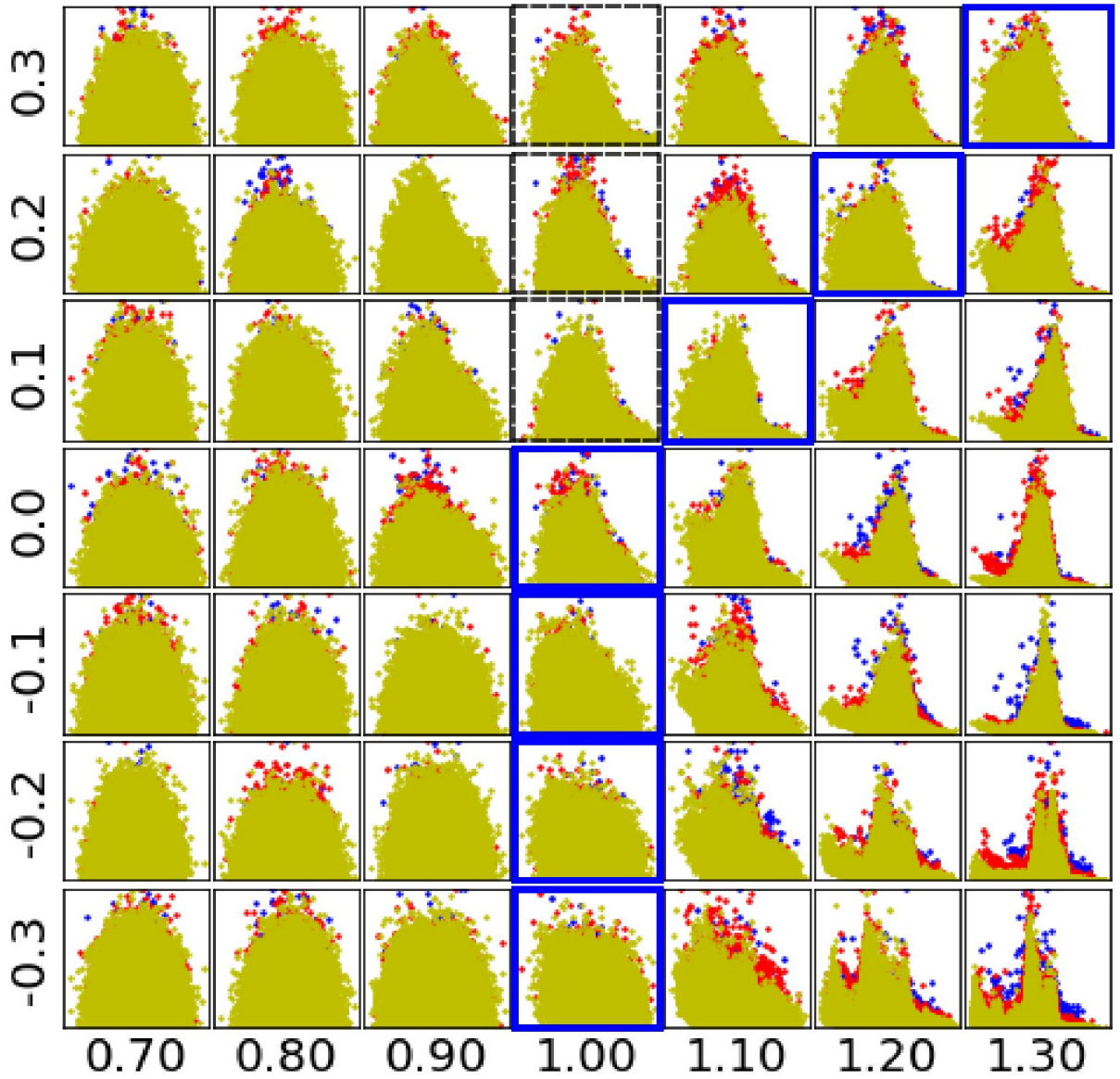
red: $m = 10^4$

Yellow: $m = 10^6$

Feature distribution at the condensed regime-MNIST

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

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



Blue: $m = 10^3$

red: $m = 10^4$

Yellow: $m = 10^6$

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

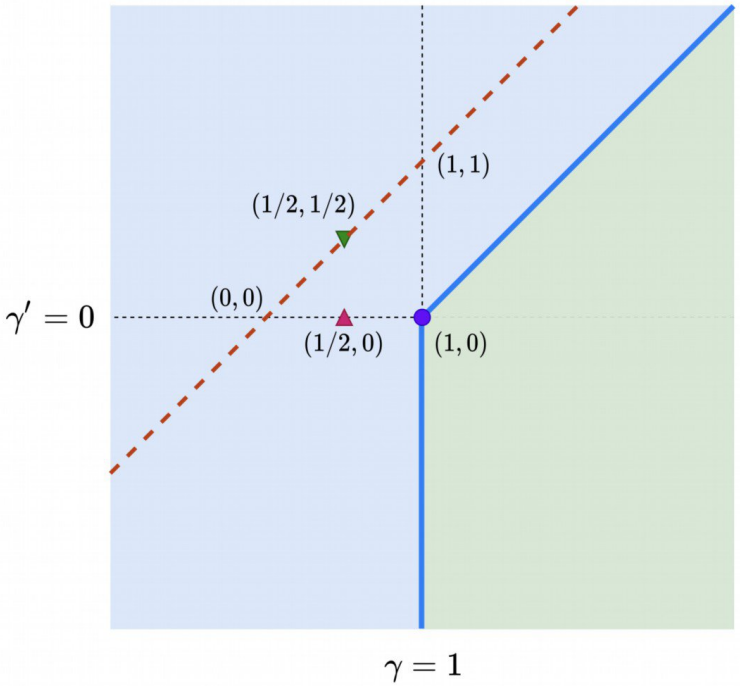
Regime separation -- theorems

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*

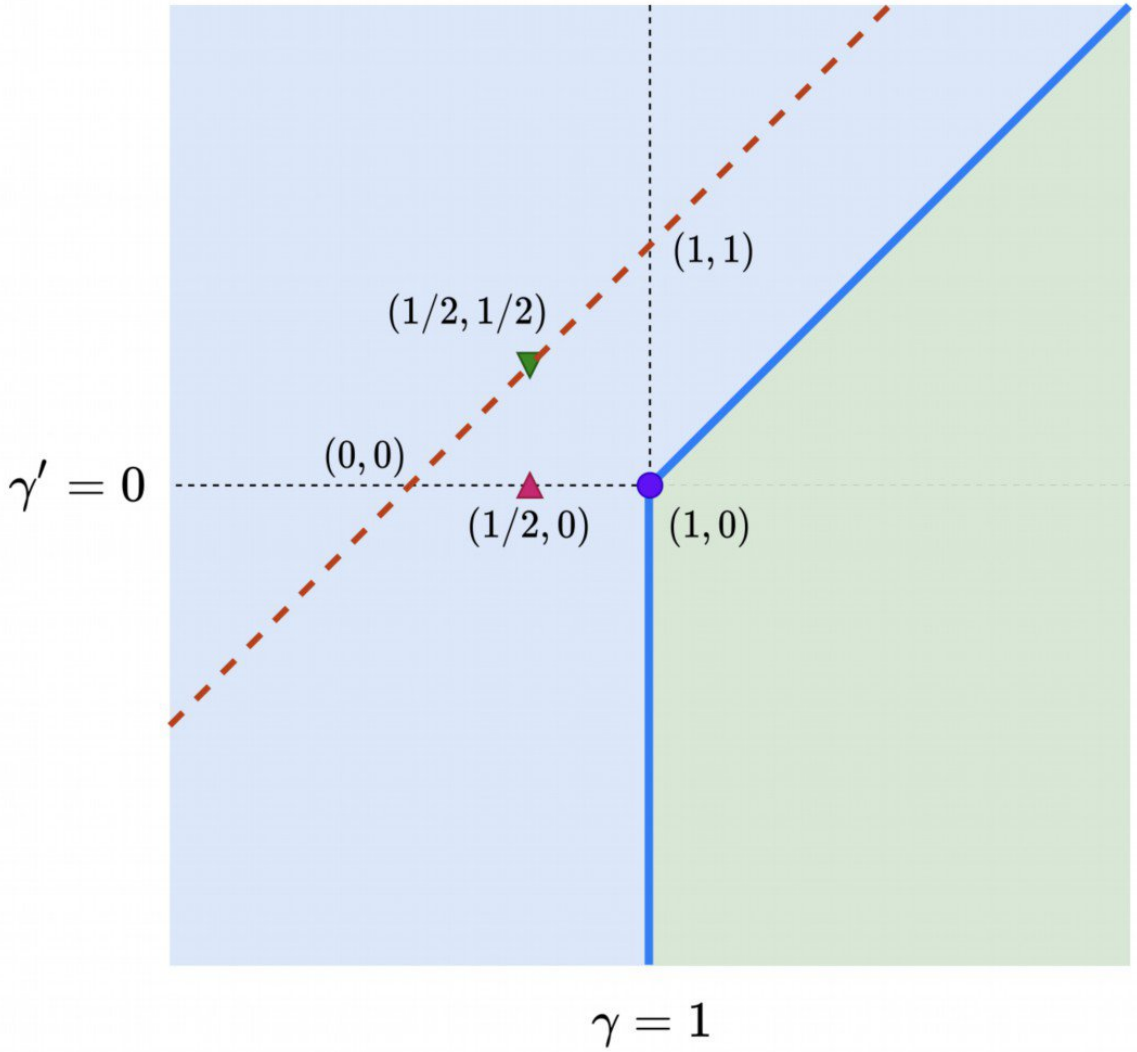
$$\lim_{m \to +\infty} \sup_{t \in [0, +\infty)} \mathrm{RD}(\boldsymbol{\theta}_{\boldsymbol{w}}(t)) = 0. \tag{20}$$

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*

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



Phase Diagram



Linear regime

Condensed regime

Critical regime

Examples:

- Xavier, Mean field
- NTK
- E at el. (2020)
- LeCun, He

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

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

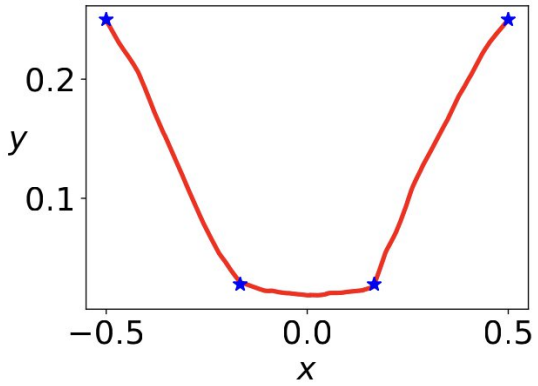
Typical cases across the phase diagram

$\gamma' = 0$

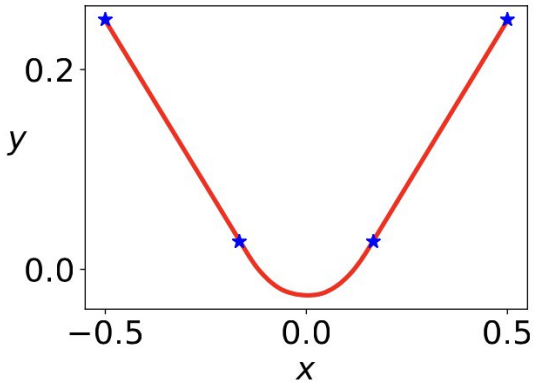
Linear regime

critical regime

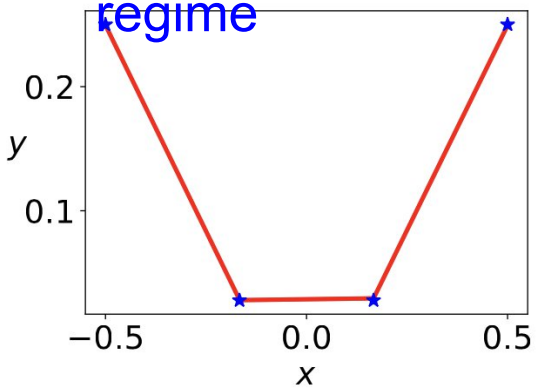
condensed regime



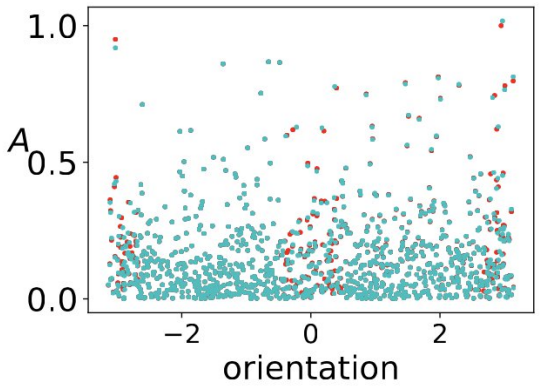
(a) $\gamma = 0.5$



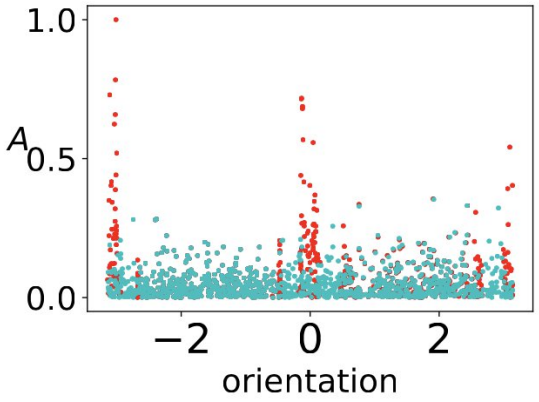
(b) $\gamma = 1$



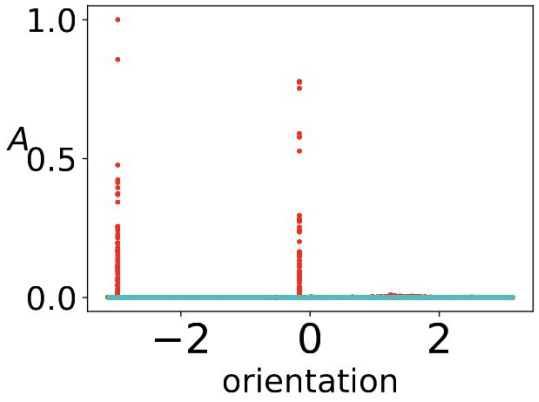
(c) $\gamma = 1.75$



(d) $\gamma = 0.5$

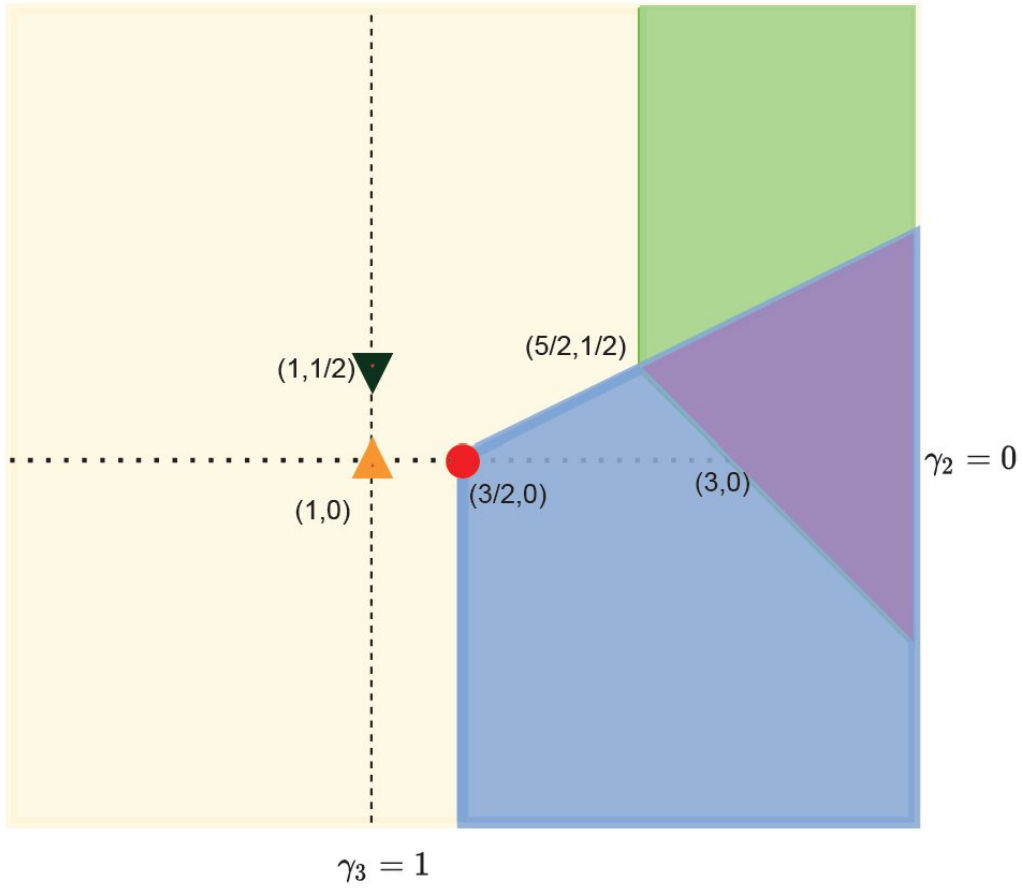


(e) $\gamma = 1$



(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

$Lecun, He$

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

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

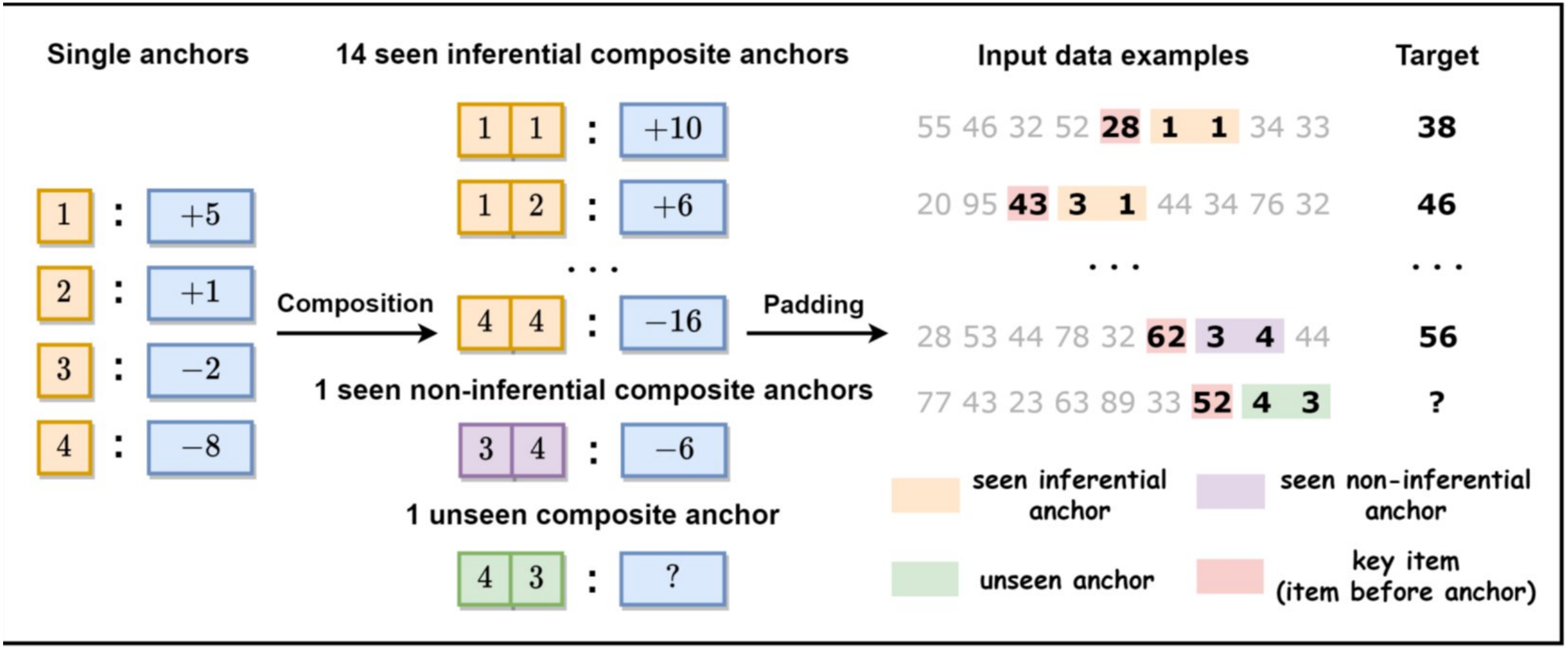
CR is short for Condensed Reigme, LR is short for Linear Regime

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

Condensation facilitates reasoning

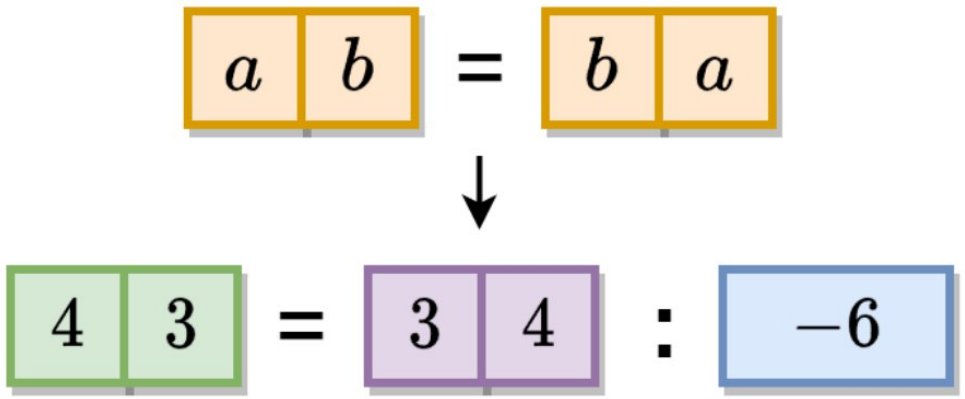
Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu, “Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing,” NeurIPS 2024.

Composite Anchor function

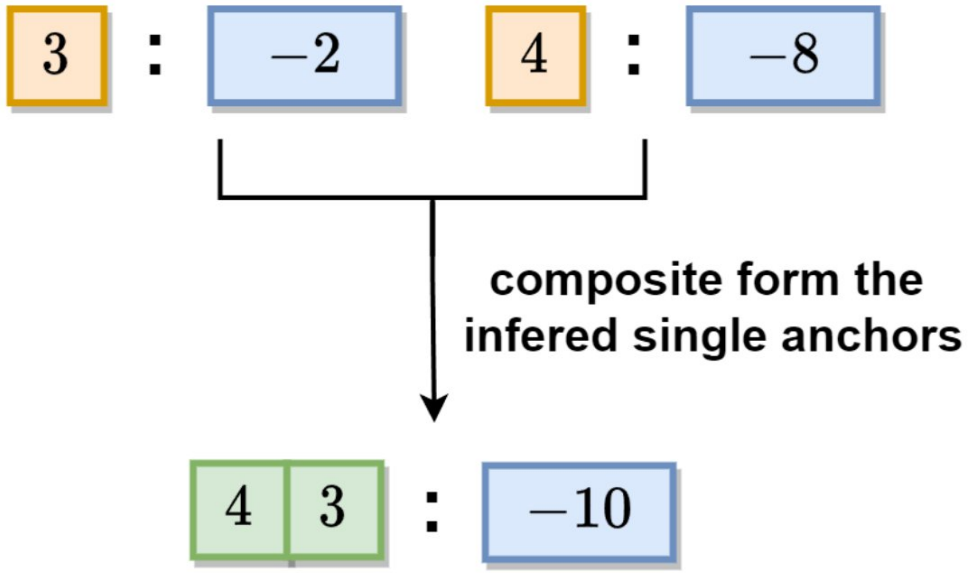


Symmetric or inferred solutions?

Mechanism 1: learn symmetric structure



Mechanism 2: infer single anchor mappings



Phase diagram of symmetric solution

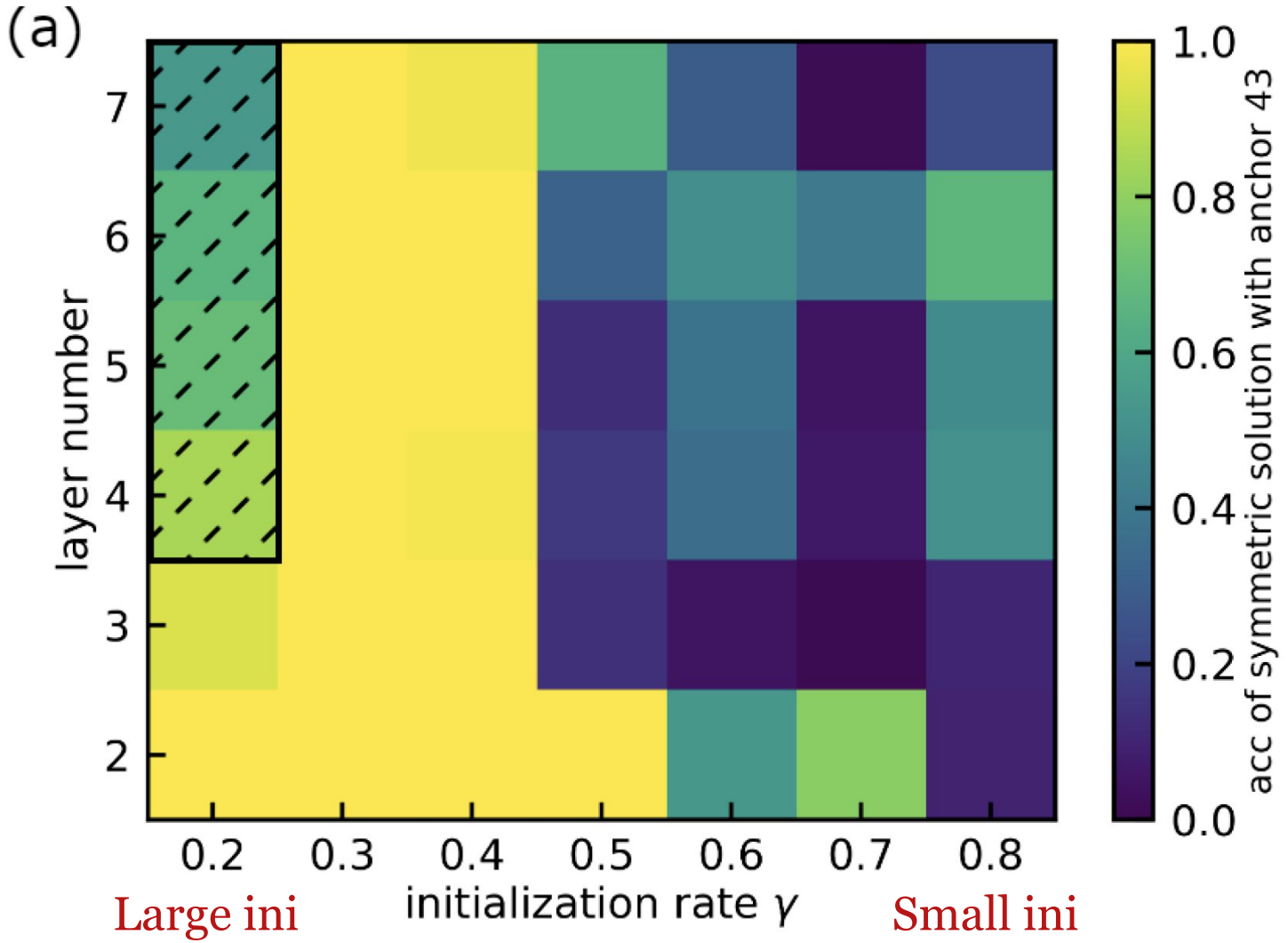
Mechanism 1: learn symmetric structure

$$\boxed{a}\boxed{b} = \boxed{b}\boxed{a}$$

$$\downarrow$$

$$\boxed{4}\boxed{3} = \boxed{3}\boxed{4} : \boxed{-6}$$

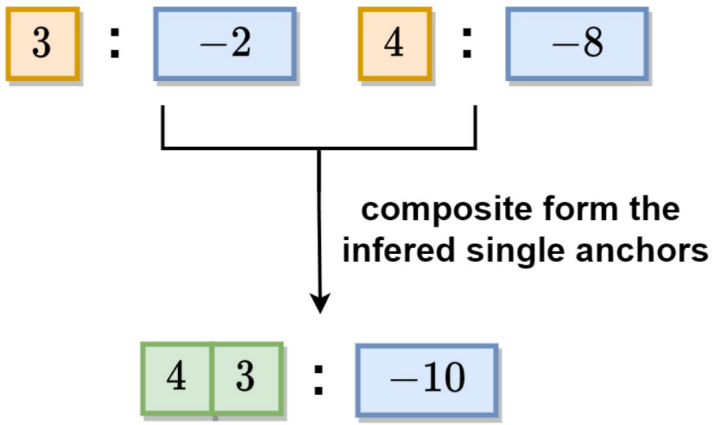
$$\text{Initialization} \sim N\left(0, \frac{1}{d_{in}^{\gamma}}\right)$$



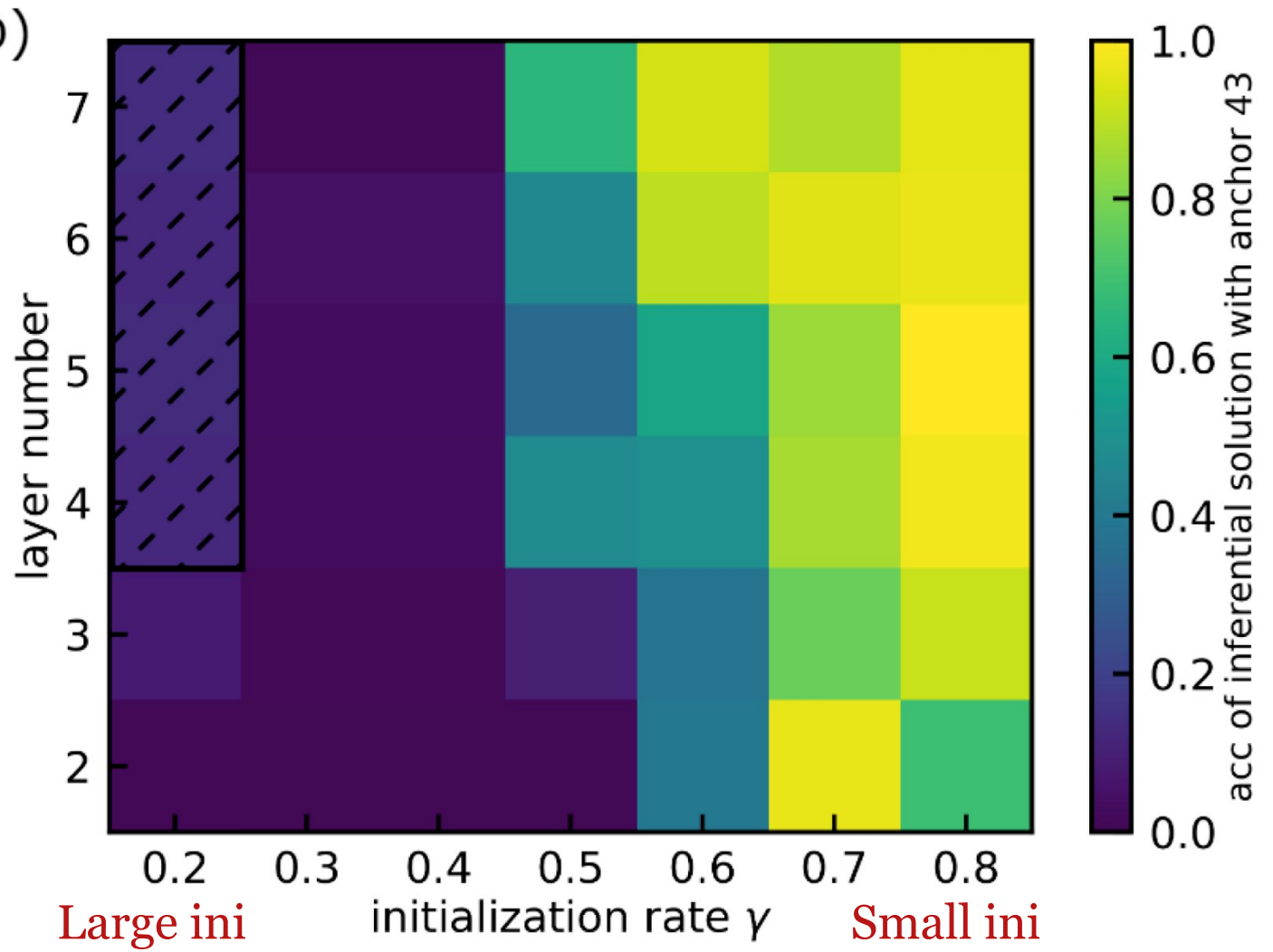
bad generalization on seen anchors (test accuracy < 90%)

Phase diagram of inferential solution

Mechanism 2: infer single anchor mappings



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



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: $Q^{(1)}$**

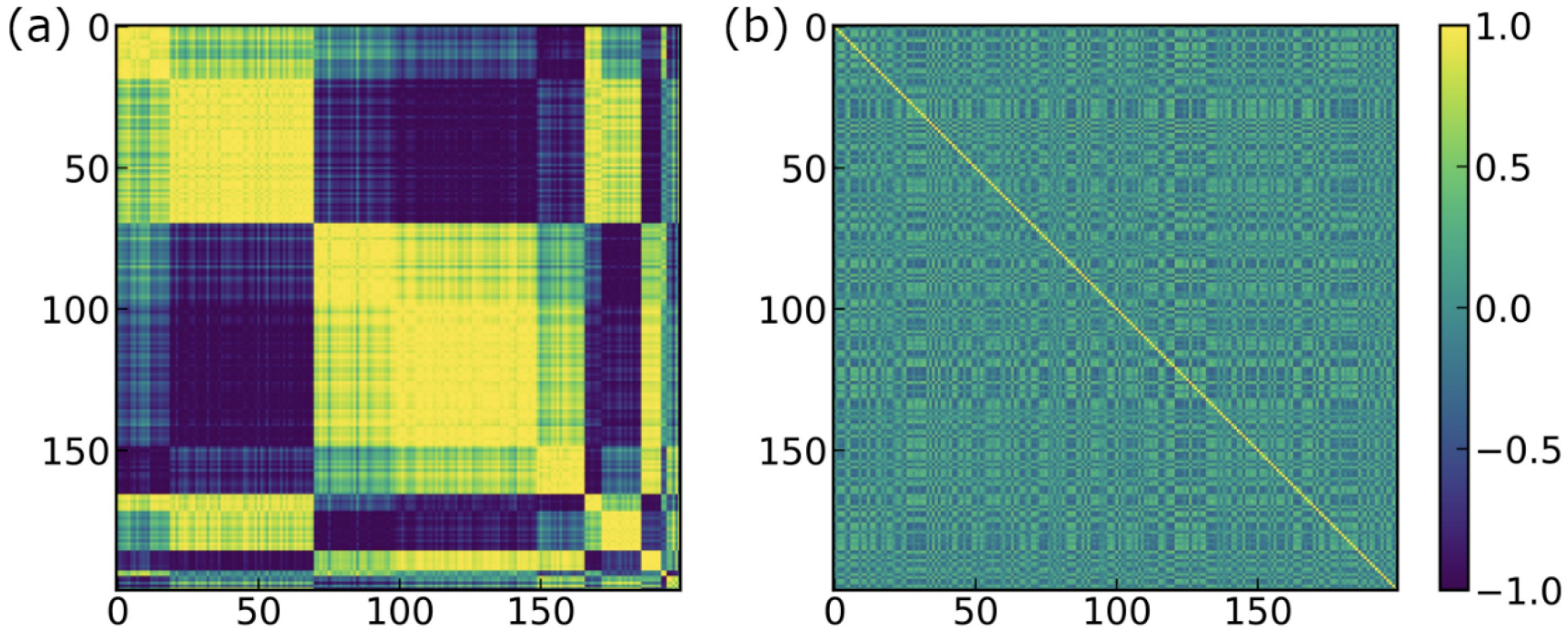
Cosine Similarity b/w

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

Input weight of Q neurons

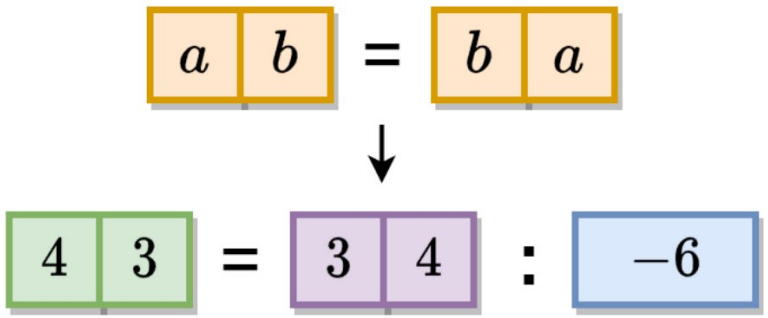
Small ini: clear condensation

Large ini: no condensation



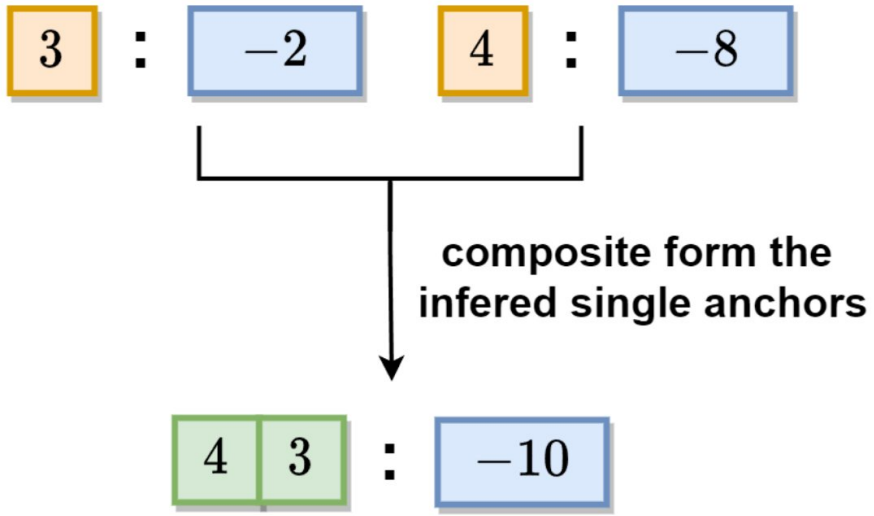
Complexity of solutions

Mechanism 1: learn symmetric structure



Not only one pair
But ten pairs in training
11;12,21; 13, 31; 14, 41; 23, 32;…

Mechanism 2: infer single anchor mappings



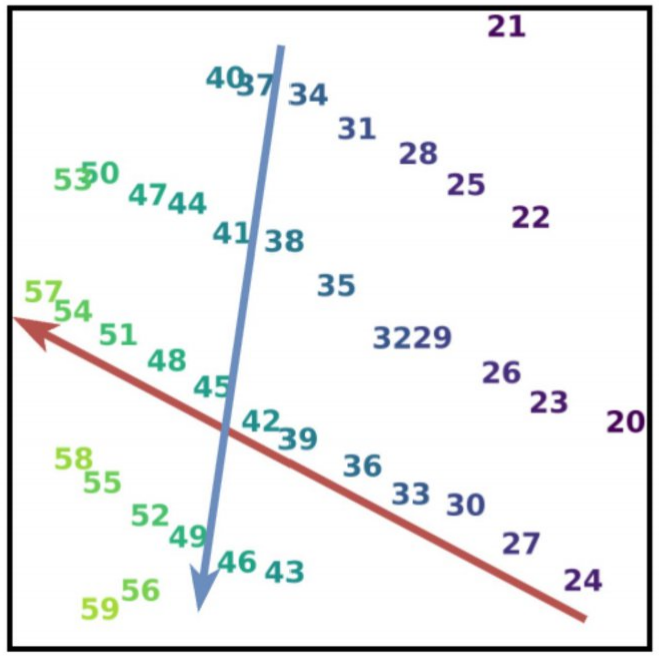
Need to learn four functions

Structure of embedding space

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

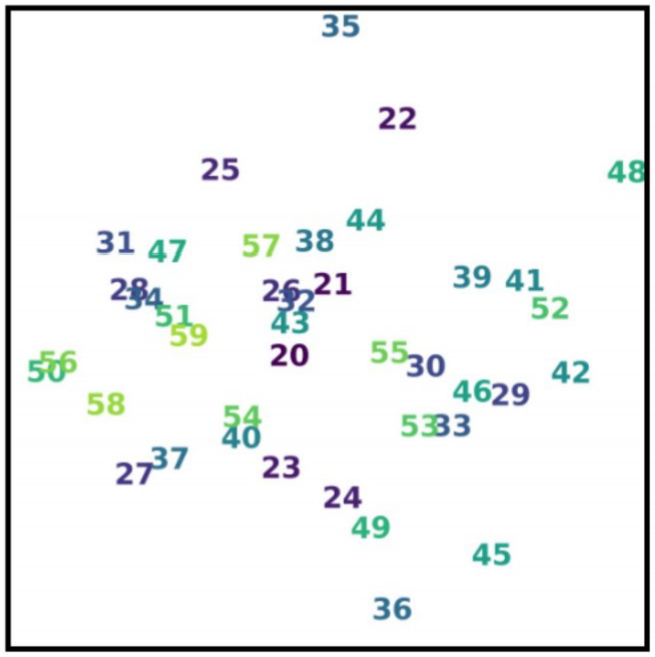
Small init: clear structure

(c)

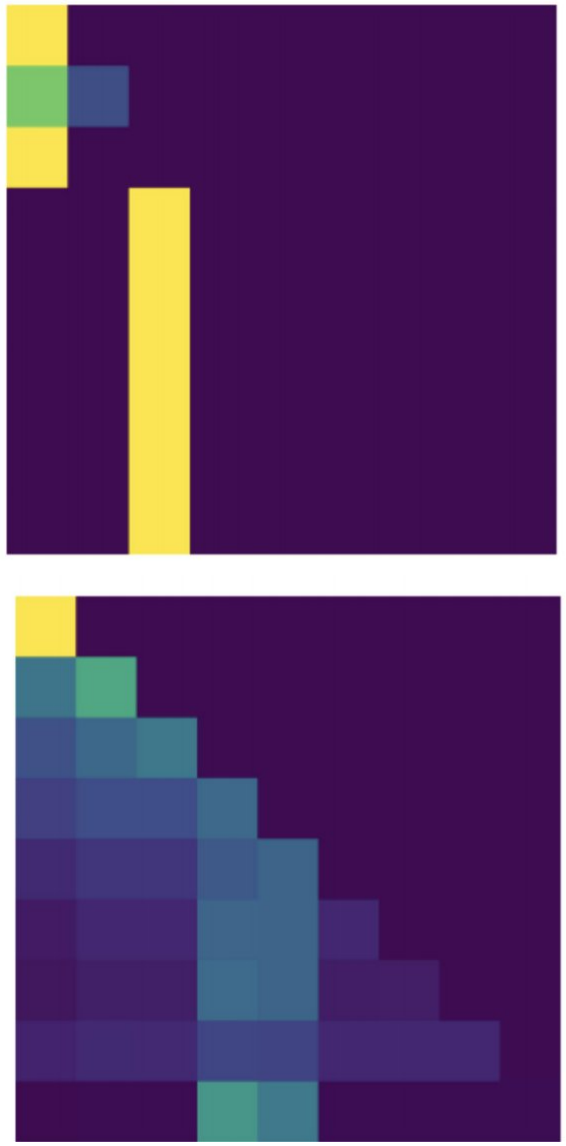


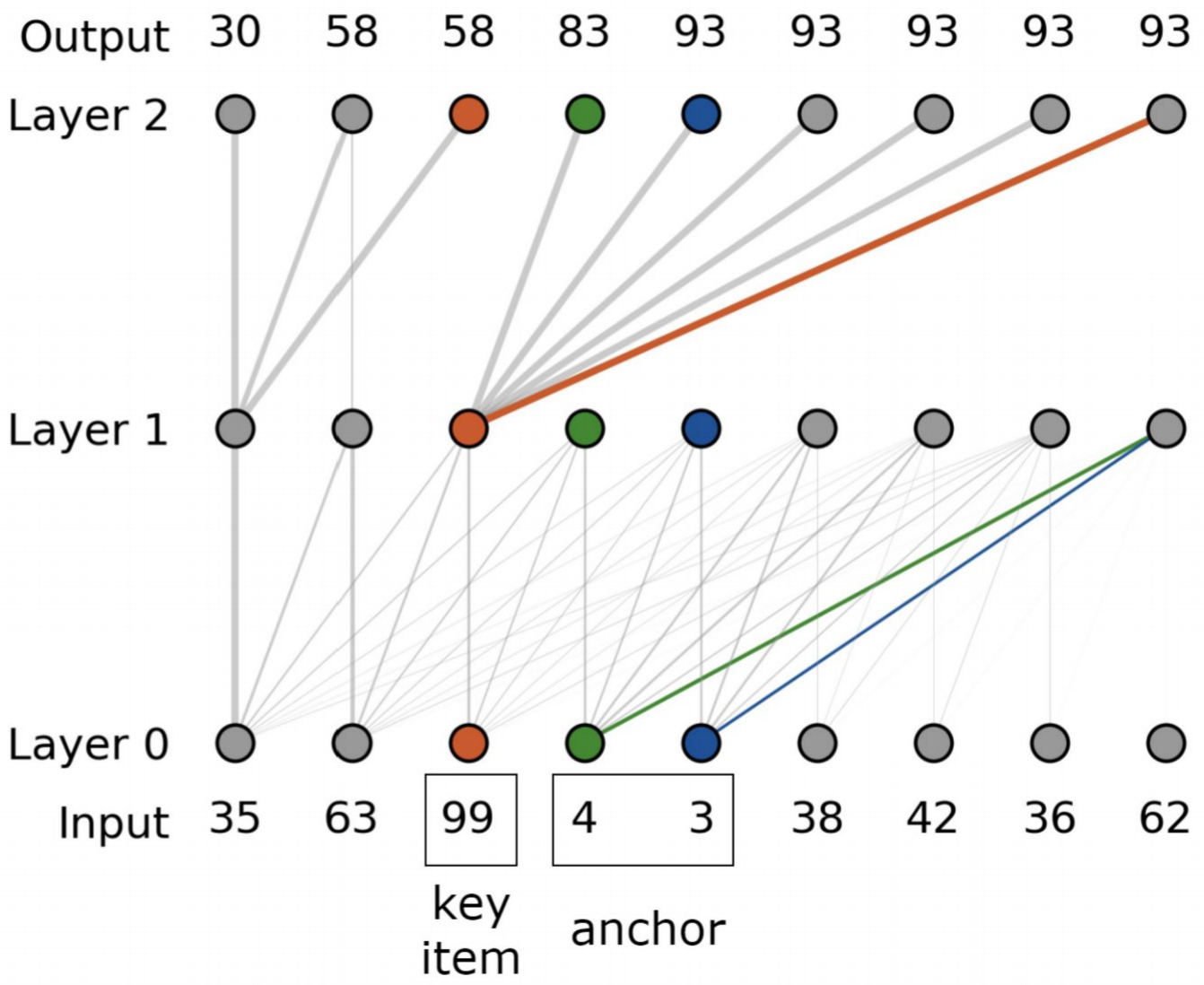
Large init: no structure

(d)

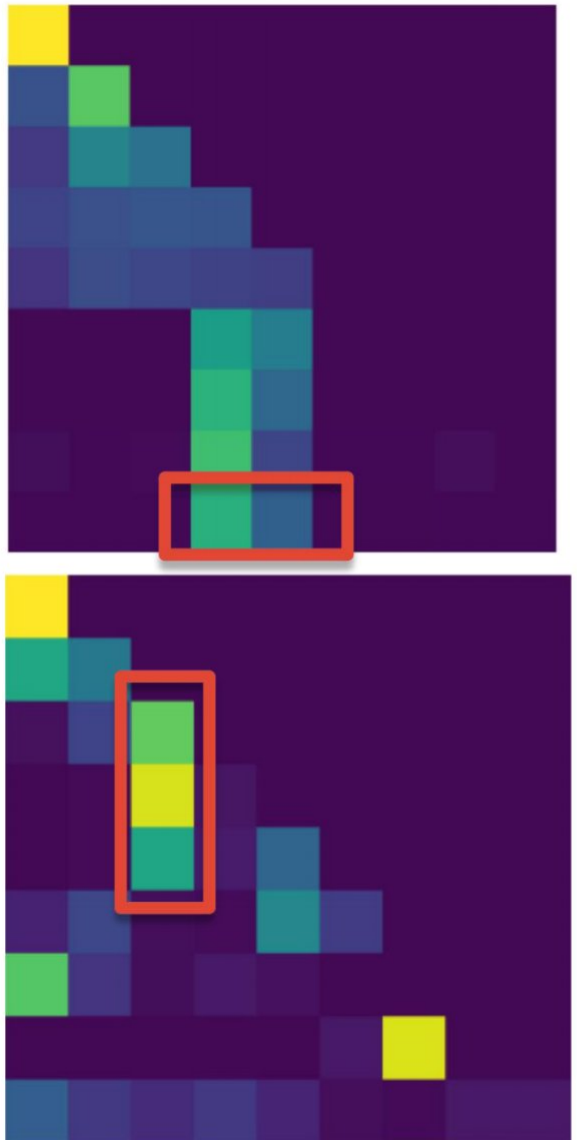


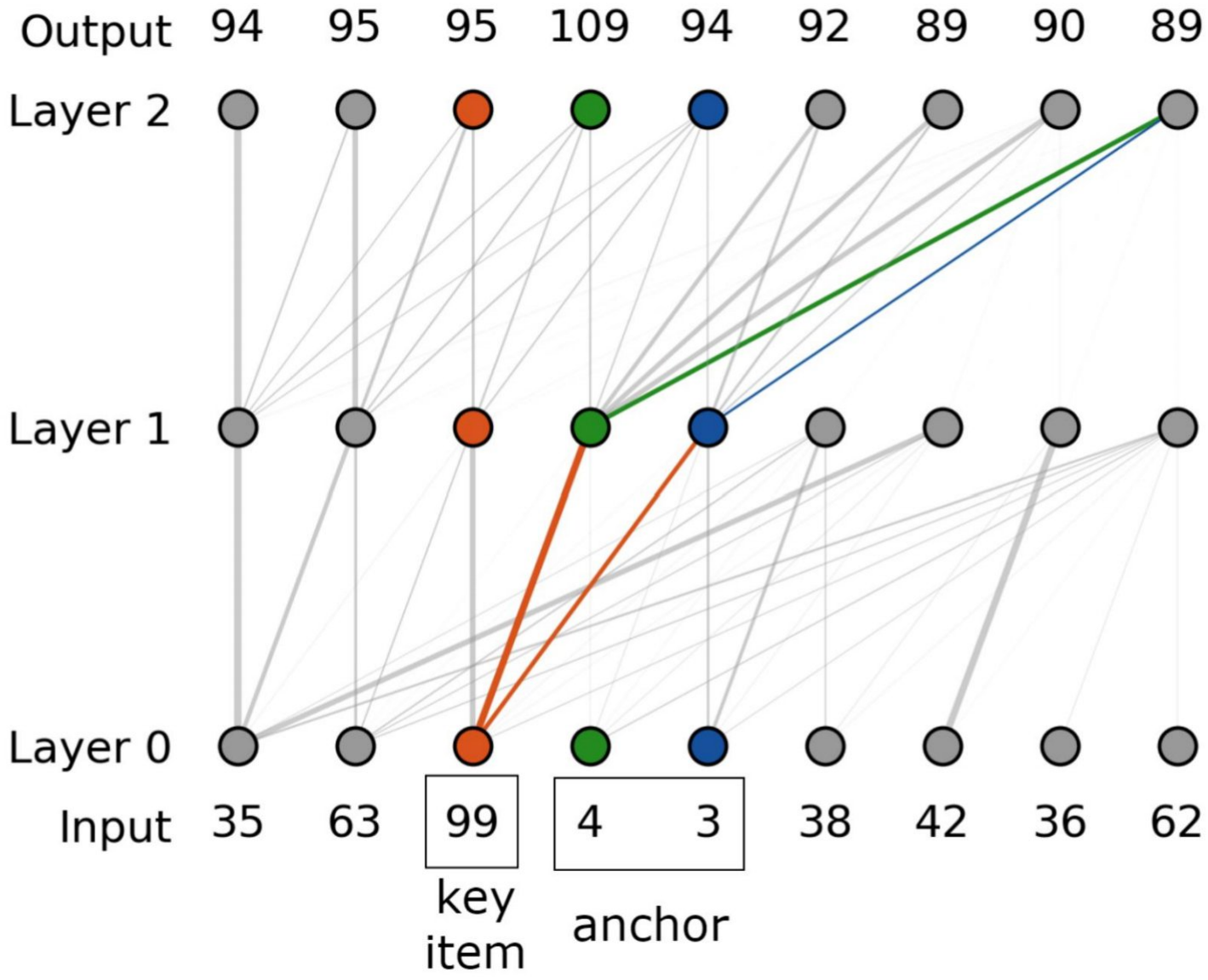
Symmetric solution





Inferential solution

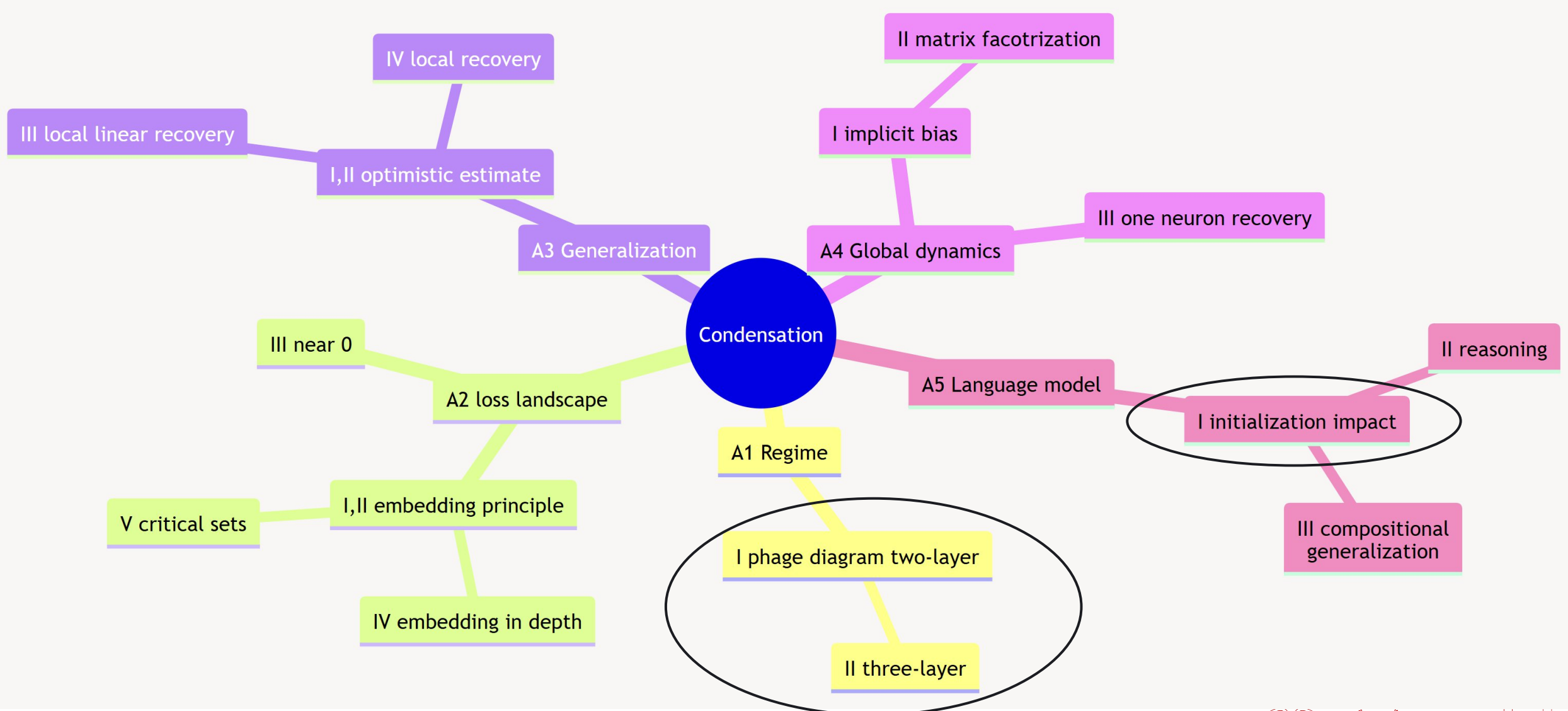




Mechanisms underlying initialization effect

Small initial	Condense	Low complexity as possible	Only learn four anchor	Generalize on inference
Large initial	Vey slight condense	Large complexity	Easily fit ten sym. maps	Generalize on sym. map
Very Large initial	No condense	Very large complexity	Easily all fit maps	No Generalize

Condensation



Problems

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

Thanks!

饮 水 思 源　　爱 国 荣 校