Empirical Phase Diagram for Three-layer Neural Networks with infinite Width

Hanxu Zhou[1], Qixuan Zhou[1], Zhenyuan Jin[1], Tao Luo[1,2], Yaoyu Zhang[1,3], Zhi-Qin John Xu[1,*]

(*corresponding author)

[1]School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC and Qing Yuan Research Institute, Shanghai Jiao Tong University

[2]CMA-Shanghai

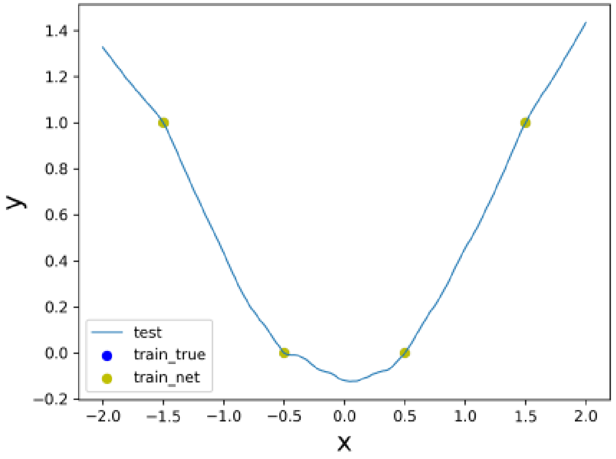
[3]Shanghai Center for Brain Science and Brain-Inspired Technology

2022.10.11

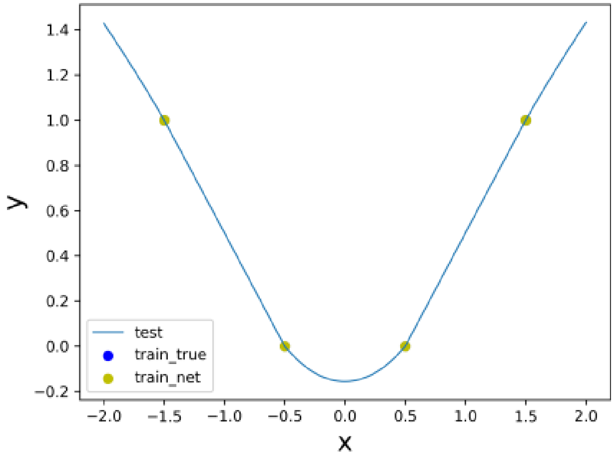
饮 水 思 源 • 爱 国 荣 校

Motivation

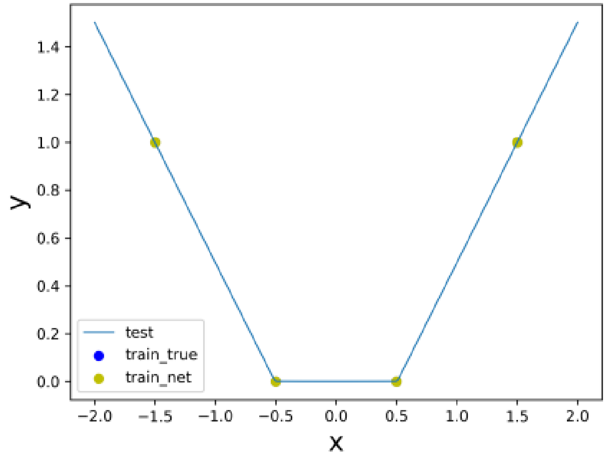
The output of different initialization methods has differentiated properties



NTK



Xavier

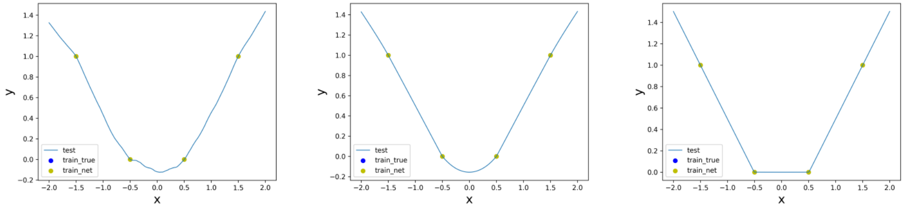


Condensed

- Learning four data points by three-layer ReLU NNs with different initialization methods.

Motivation

The output of different initialization methods has differentiated properties

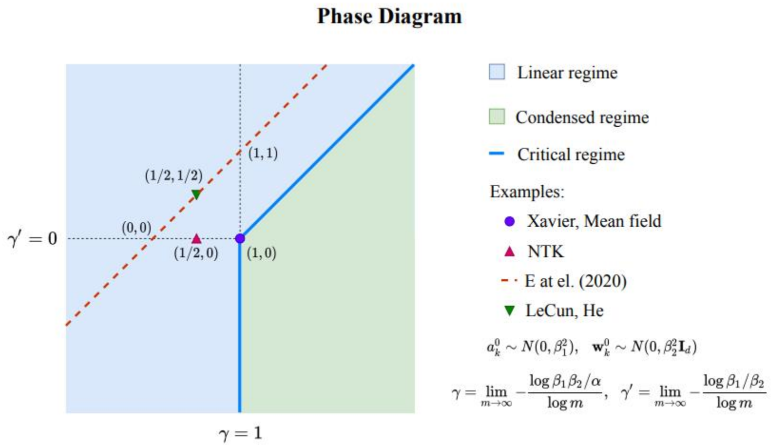


NTK   *Xavier*   *Condensed*

- Learning four data points by three-layer ReLU NNs with different initialization methods.

Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit

Tao Luo#, Zhi-Qin John Xu#, Zheng Ma, Yaoyu Zhang*

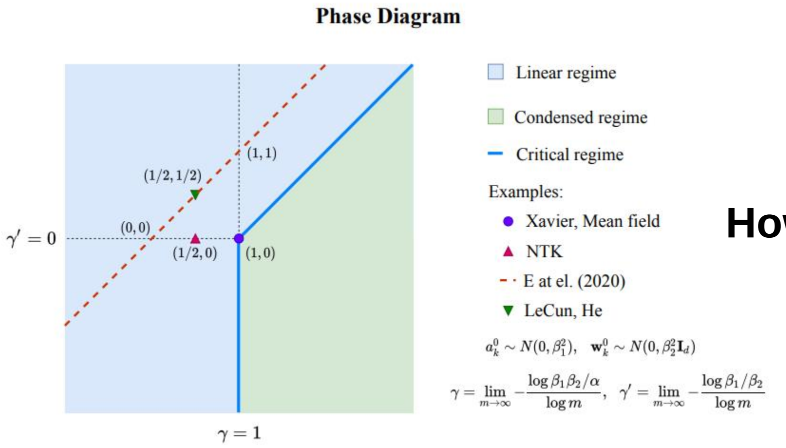


Motivation

The output of different initialization methods has differentiated properties

Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit

Tao Luo[#], Zhi-Qin John Xu[#], Zheng Ma, Yaoyu Zhang*



How about the more general case?

Difficulty:

- Multi-layer structure
- Non-linearity
- Distinct characteristics

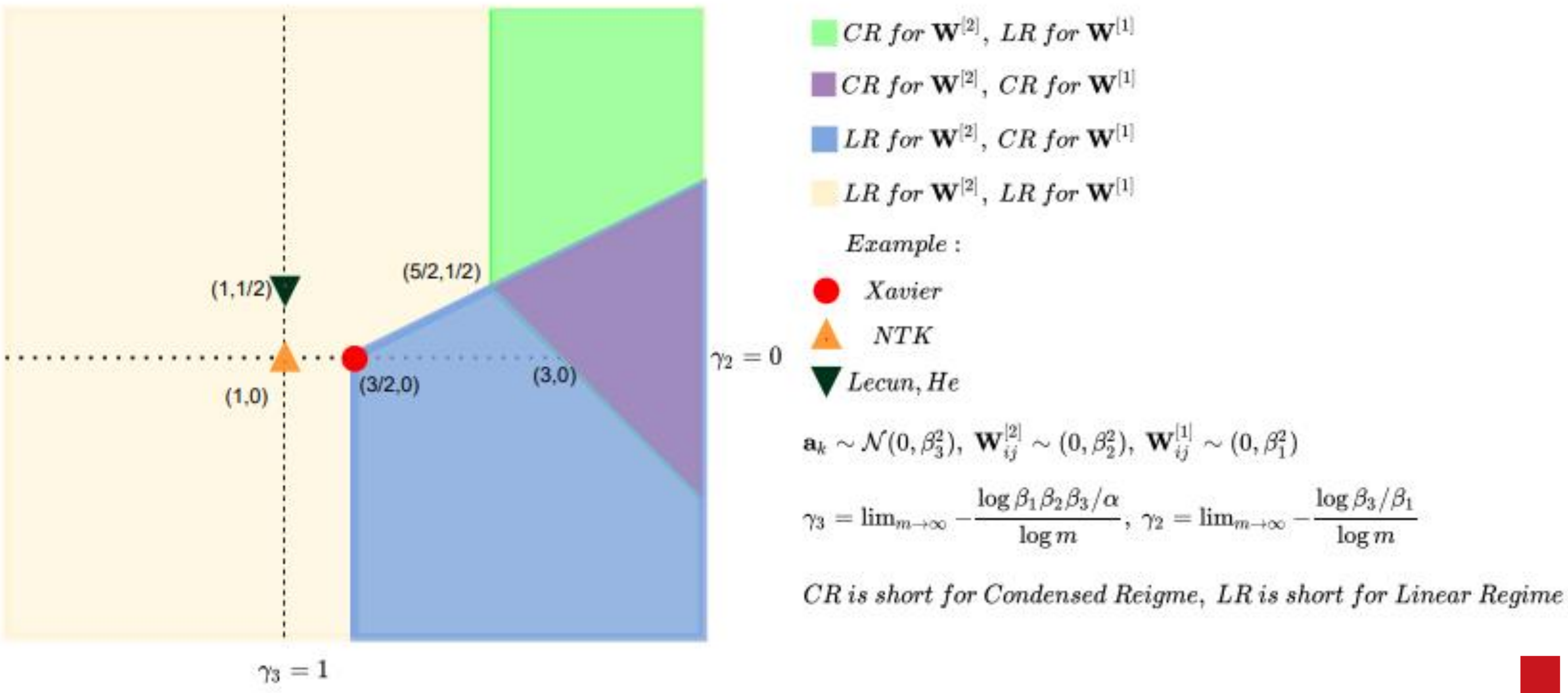
Curiosity:

- Different frow two-layer
- Distinct dynamics in one NN

Overview

This study: **make a step towards drawing a phase diagram for three-layer ReLU NNs with infinite width**

- Figure out **key quantities** and **divide** the dynamics into:
  - a linear regime
  - a condensed regime
  - a critical regime.
- Identify the **condensation** as the strong non-linear signature behavior
- Suggest a complicated **dynamical regimes** consisting of three possible regimes, together with their mixture.



Preliminary

A three-layer NN with m hidden neurons for each layer is,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\alpha}\boldsymbol{a}^T\sigma\left(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})\right)$$

where, $\boldsymbol{x} = [\boldsymbol{x}^T, 1]^T$, $\boldsymbol{W}^{[1]} = [\boldsymbol{W}^{[1]}, b_k^{[1]}]^T$, $\overline{\boldsymbol{x}} = \sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})$, $\overline{\boldsymbol{x}} = [\overline{\boldsymbol{x}}^T, 1]^T$, $\boldsymbol{W}^{[2]} = [\boldsymbol{W}^{[2]}, b_k^{[2]}]^T$, and $\boldsymbol{a}_k^0 \sim \mathcal{N}(0, \beta_3^2)$, $\boldsymbol{W}_{kk'}^{[2],0} \sim \mathcal{N}(0, \beta_2^2)$, $\boldsymbol{W}_{kk'}^{[1],0} \sim \mathcal{N}(0, \beta_1^2)$,

The empirical risk is,

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

SJTU

Preliminary

A three-layer NN with m hidden neurons for each layer is,

$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\alpha}\boldsymbol{a}^T\sigma\left(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})\right)$, $\boldsymbol{a}_k^0 \sim \mathcal{N}(0,\beta_3^2)$, $\boldsymbol{W}_{kk'}^{[2],0} \sim \mathcal{N}(0,\beta_2^2)$, $\boldsymbol{W}_{kk'}^{[1],0} \sim \mathcal{N}(0,\beta_1^2)$,

where, $\boldsymbol{x} = [\boldsymbol{x}^T, 1]^T$, $\boldsymbol{W}^{[1]} = [\boldsymbol{W}^{[1]}, b_k^{[1]}]^T$, $\overline{\boldsymbol{x}} = \sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})$, $\overline{\boldsymbol{x}} = [\overline{\boldsymbol{x}}^T, 1]^T$, $\boldsymbol{W}^{[2]} = [\boldsymbol{W}^{[2]}, b_k^{[2]}]^T$.

The gradient flow of $\boldsymbol{\theta} = \mathrm{vec}\{\boldsymbol{a}, \boldsymbol{W}^{[2]}, \boldsymbol{W}^{[1]}\}$,

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\sigma(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i))e_i,$$

$$\frac{\mathrm{d}\boldsymbol{W}^{[2]}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\boldsymbol{a}\odot\sigma'(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i))\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)^{\mathbf{T}}e_i,$$

$$\frac{\mathrm{d}\boldsymbol{W}^{[1]}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\boldsymbol{W}^{[2]^{\mathbf{T}}}(\boldsymbol{a}\odot\sigma'(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)))\odot\sigma'(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)\boldsymbol{x}_i^{\mathbf{T}}e_i,$$

where $e_i = \left(\frac{1}{\alpha}\boldsymbol{a}^T\sigma\left(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x})\right) - y_i\right)$, the operation $\odot$ is the Hadamard product.

Rescaling and the normalized model

The gradient flow of $\boldsymbol{\theta} = \mathrm{vec}\{\boldsymbol{a}, \boldsymbol{W}^{[2]}, \boldsymbol{W}^{[1]}\}$,

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\sigma(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i))e_i,$$

$$\frac{\mathrm{d}\boldsymbol{W}^{[2]}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\boldsymbol{a}\odot\sigma'(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i))\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)^{\mathbf{T}}e_i,$$

$$\frac{\mathrm{d}\boldsymbol{W}^{[1]}}{\mathrm{d}t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\boldsymbol{W}^{[2]\mathbf{T}}(\boldsymbol{a}\odot\sigma'(\boldsymbol{W}^{[2]}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)))\odot\sigma'(\boldsymbol{W}^{[1]}\boldsymbol{x}_i)\boldsymbol{x}_i^{\mathbf{T}}e_i,$$

$\sigma(a\boldsymbol{u}) = a\sigma(\boldsymbol{u}), \sigma'(a\boldsymbol{u}) = \sigma'(\boldsymbol{u})$

The normalized gradient flow of $\boldsymbol{\theta}$,

$$\frac{\mathrm{d}\bar{\boldsymbol{a}}}{\mathrm{d}\bar{t}} = -\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\sigma(\bar{\boldsymbol{W}}^{[2]}\sigma(\bar{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i))\right)e_i,$$

$$\frac{\mathrm{d}\bar{\boldsymbol{W}}^{[2]}}{\mathrm{d}\bar{t}} = -\kappa_1^2\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\bar{\boldsymbol{a}}\odot\sigma'(\bar{\boldsymbol{W}}^{[2]}\sigma(\bar{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i))\sigma(\bar{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)^{\mathbf{T}}\right)e_i,$$

$$\frac{\mathrm{d}\bar{\boldsymbol{W}}^{[1]}}{\mathrm{d}\bar{t}} = -\kappa_2^2\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\bar{\boldsymbol{W}}^{[2]\mathbf{T}}(\bar{\boldsymbol{a}}\odot\sigma'(\bar{\boldsymbol{W}}^{[2]}\sigma(\bar{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)))\odot\sigma'(\bar{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)\boldsymbol{x}_i^{\mathbf{T}}\right)e_i,$$

where $\bar{\boldsymbol{a}} = \frac{1}{\beta_3}\boldsymbol{a}$, $\bar{\boldsymbol{W}}^{[2]} = \frac{1}{\beta_2}\boldsymbol{W}^{[2]}$, $\bar{\boldsymbol{W}}^{[1]} = \frac{1}{\beta_1}\boldsymbol{W}^{[1]}$, $\kappa_1 = \frac{\beta_3}{\beta_2}$, $\kappa_2 = \frac{\beta_3}{\beta_1}$, $\kappa_3 = \frac{\beta_1\beta_2\beta_3}{\alpha}$, $t = \left(\alpha\prod_{i=1}^{3}\kappa_i\right)^{-\frac{2}{3}}\bar{t}$.

Rescaling and the normalized model

The normalized gradient flow of θ,

$$\frac{\mathrm{d}\bar{\boldsymbol{a}}}{\mathrm{d}\bar{t}} = -\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\sigma(\overline{\boldsymbol{W}}^{[2]}\sigma(\overline{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i))\right)e_i,$$

$$\frac{\mathrm{d}\overline{\boldsymbol{W}}^{[2]}}{\mathrm{d}\bar{t}} = -\kappa_1^2\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\bar{\boldsymbol{a}}\odot\sigma'(\overline{\boldsymbol{W}}^{[2]}\sigma(\overline{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i))\sigma(\overline{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)^{\mathbf{T}}\right)e_i,$$

$$\frac{\mathrm{d}\overline{\boldsymbol{W}}^{[1]}}{\mathrm{d}\bar{t}} = -\kappa_2^2\left(\frac{1}{n}\sum_{i=1}^{n}\kappa_3\overline{\boldsymbol{W}}^{[2]^{\mathbf{T}}}(\bar{\boldsymbol{a}}\odot\sigma'(\overline{\boldsymbol{W}}^{[2]}\sigma(\overline{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)))\odot\sigma'(\overline{\boldsymbol{W}}^{[1]}\boldsymbol{x}_i)\boldsymbol{x}_i^{\mathbf{T}}\right)e_i,$$

The scaling parameters and infinite-width limit,

$$\kappa_1 = \frac{\beta_3}{\beta_2},\ \kappa_2 = \frac{\beta_3}{\beta_1},\ \kappa_3 = \frac{\beta_1\beta_2\beta_3}{\alpha},\ \bar{t} = \left(\alpha\prod_{i=1}^{3}\kappa_i\right)^{-\frac{2}{3}}t,$$

Assumption 3.1: $m_1 = m_2 = m$
Assumption 3.2: $\beta_2 = B\beta_3$

$$\gamma_1 = \lim_{m\to\infty} -\frac{\log\kappa_1}{\log m} = 0,\ \gamma_2 = \lim_{m\to\infty} -\frac{\log\kappa_2}{\log m},\gamma_3 = \lim_{m\to\infty} -\frac{\log\kappa_3}{\log m}$$

Rescaling and the normalized model

The scaling parameters and infinite-width limit,

$$\kappa_1 = \frac{\beta_3}{\beta_2},\ \kappa_2 = \frac{\beta_3}{\beta_1},\ \kappa_3 = \frac{\beta_1\beta_2\beta_3}{\alpha},\ \bar{t} = \left(\alpha \prod_{i=1}^{3} \kappa_i\right)^{-\frac{2}{3}} t,\ \gamma_2 = \lim_{m\to\infty} -\frac{\log \kappa_2}{\log m},\gamma_3 = \lim_{m\to\infty} -\frac{\log \kappa_3}{\log m}$$

Some common initialization methods

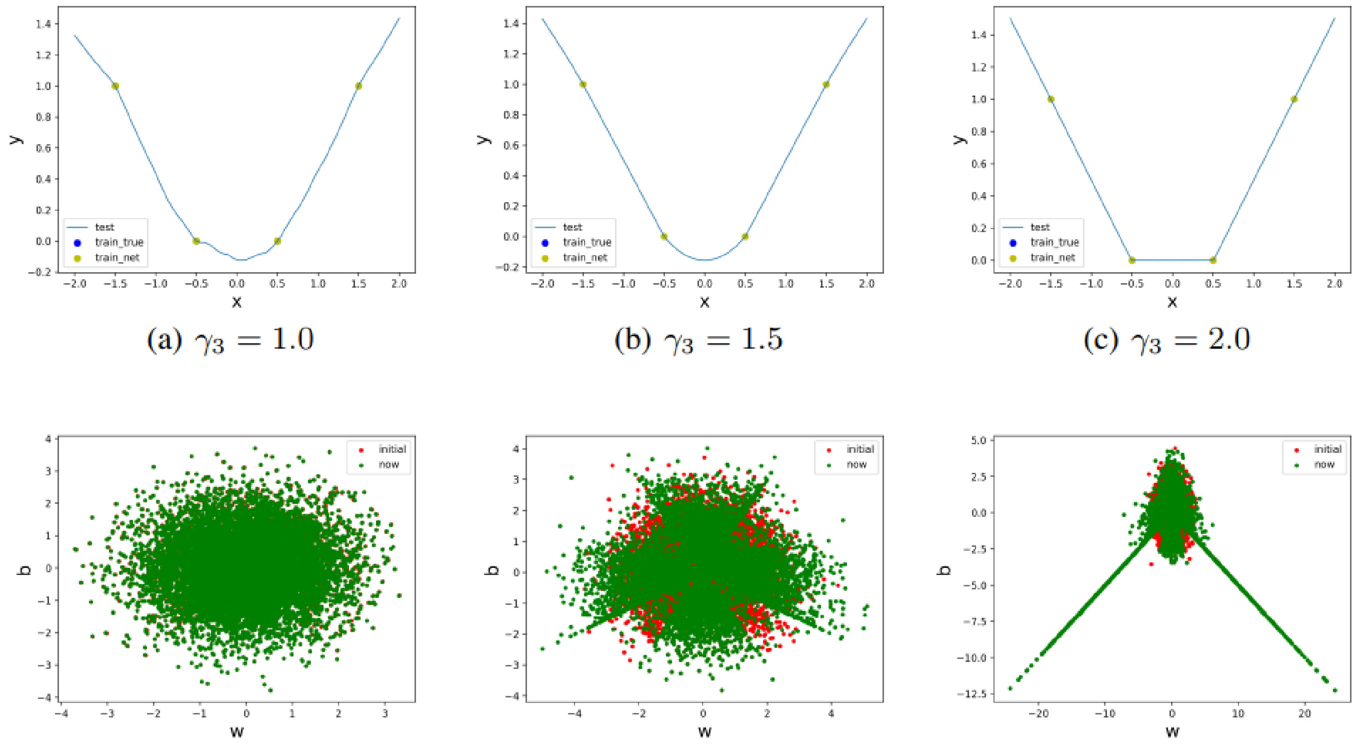
Table 1: Common initialization methods with their scaling parameters

Name	α	$\boldsymbol{a}$	$\boldsymbol{W}^{[2]}$	$\boldsymbol{W}^{[1]}$	κ_2	κ_3	γ_2	γ_3
NTK Jacot et al. (2018)	$\sqrt{m_1 m_2}$	1	1	1	1	$\sqrt{\frac{1}{m_1 m_2}}$	0	1
Lecun LeCun et al. (2012)	1	$\sqrt{\frac{1}{m_2}}$	$\sqrt{\frac{1}{m_1}}$	$\sqrt{\frac{1}{d}}$	$\sqrt{\frac{d}{m_2}}$	$\sqrt{\frac{1}{m_1 m_2 d}}$	$\frac{1}{2}$	1
He He et al. (2015)	1	$\sqrt{\frac{2}{m_2}}$	$\sqrt{\frac{2}{m_1}}$	$\sqrt{\frac{2}{d}}$	$\sqrt{\frac{d}{m_2}}$	$\sqrt{\frac{8}{m_1 m_2 d}}$	$\frac{1}{2}$	1
Xavier Glorot and Bengio (2010)	1	$\sqrt{\frac{2}{m_2+1}}$	$\sqrt{\frac{2}{m_1+m_2}}$	$\sqrt{\frac{2}{d+m_1}}$	$\sqrt{\frac{d+m_1}{m_2+1}}$	$\sqrt{\frac{8/(m_1+m_2)}{(m_2+1)(d+m_1)}}$	0	$\frac{3}{2}$

Empirical phase diagram

Empirical phase diagram

Intuitive experiments of synthetic data



(a) $\gamma_3 = 1.0$ (b) $\gamma_3 = 1.5$ (c) $\gamma_3 = 2.0$

(d) $\gamma_3 = 1.0$ (e) $\gamma_3 = 1.5$ (f) $\gamma_3 = 2.0$

Learning four data points by three-layer ReLU NNs with $m = 10000$ and $\gamma_2 = 0$. The scatter plots in the second row are $\left\{W_K^{[1]}\right\}_{k=1}^{m} = \left\{(w_k^{[1]}, b_k^{[1]})\right\}_{k=1}^{m}$, where red plots represent initial position and green plots represent final position.

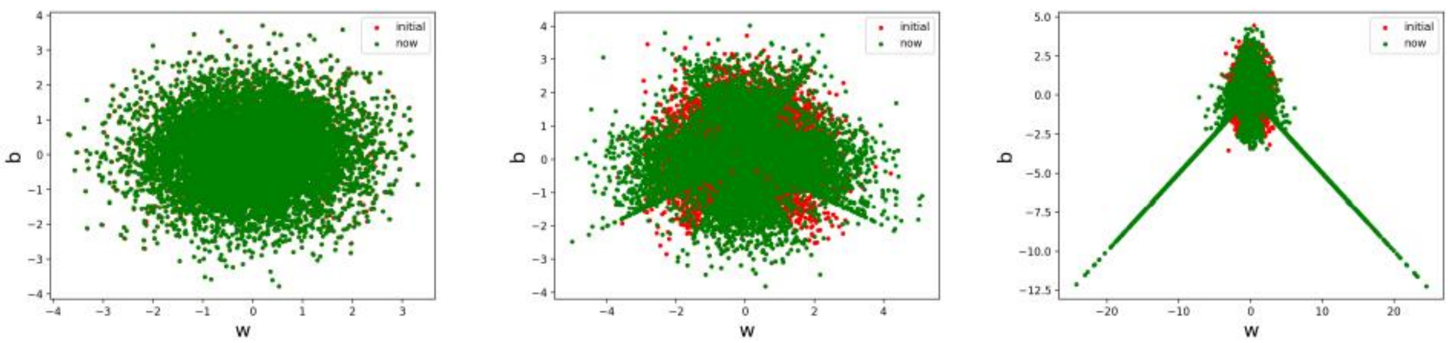
Empirical phase diagram

Regime identification and separation

- Relative distance,

$$\mathrm{RD}(\boldsymbol{W}^{[1]}) = \frac{\|\boldsymbol{\theta}^*_{\boldsymbol{W}_1} - \boldsymbol{\theta}_{\boldsymbol{W}_1}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}_1}(0)\|_2},\ \mathrm{RD}(\boldsymbol{W}^{[2]}) = \frac{\|\boldsymbol{\theta}^*_{\boldsymbol{W}_2} - \boldsymbol{\theta}_{\boldsymbol{W}_2}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}_2}(0)\|_2},$$

- We empirically consider that as $m \to \infty$,
  - Linear regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to 0, i = 1,2$
  - Condensed regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to +\infty\ i = 1,2$
  - Critical regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to O(1), i = 1,2$



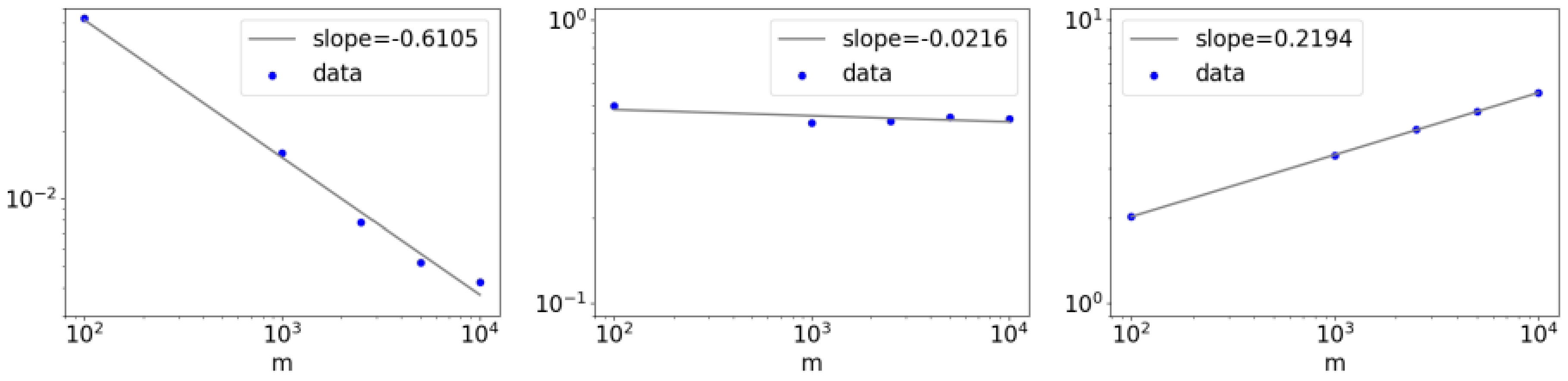
(d) $\gamma_3 = 1.0$ (e) $\gamma_3 = 1.5$ (f) $\gamma_3 = 2.0$

Empirical phase diagram

Regime identification and separation

- Relative distance, $\mathrm{RD}(\boldsymbol{W}^{[1]}) = \frac{\|\boldsymbol{\theta}^*_{\boldsymbol{W}_1} - \boldsymbol{\theta}_{\boldsymbol{W}_1}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}_1}(0)\|_2}$, $\mathrm{RD}(\boldsymbol{W}^{[2]}) = \frac{\|\boldsymbol{\theta}^*_{\boldsymbol{W}_2} - \boldsymbol{\theta}_{\boldsymbol{W}_2}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}_2}(0)\|_2}$,
- We empirically found that as $m \to \infty$,
  - Linear regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to 0, i = 1,2$
  - Condensed regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to +\infty\ i = 1,2$
  - Critical regime: $\sup_{t\in[0,+\infty)} \mathrm{RD}\left(\boldsymbol{W}^{[i]}(t)\right) \to O(1), i = 1,2$

$\mathrm{RD}(\boldsymbol{W}^{[1]})$ v.s. m. Still learn four data points by three-layer ReLU NNs with different γ_3's and $\gamma_2 = 0$.

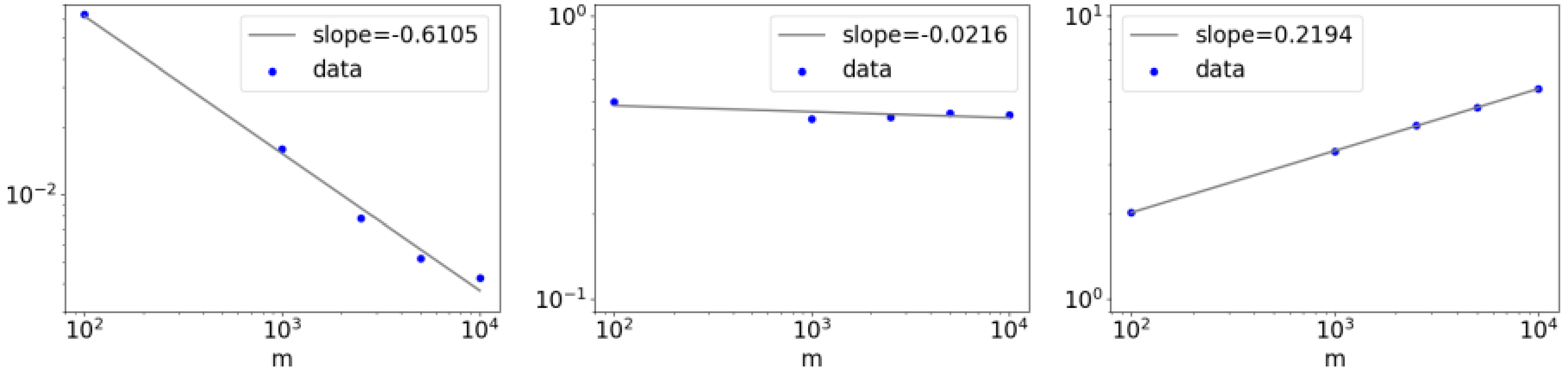


(a) $\gamma_3 = 0.9$ (b) $\gamma_3 = 1.5$ (c) $\gamma_3 = 2.1$

Empirical phase diagram

Regime identification and separation

$\mathrm{RD}\left(\boldsymbol{W}^{[1]}\right)$ v.s. m. Still learn four data points by three-layer ReLU NNs with different γ_3's and $\gamma_2 = 0$.



(a) $\gamma_3 = 0.9$ (b) $\gamma_3 = 1.5$ (c) $\gamma_3 = 2.1$

We quantify the growth of $\mathrm{RD}\left(\boldsymbol{W}^{[i]}\right),\ i = 1,2$, as $m \to \infty$, by defining,

$$S_{\boldsymbol{W}^{[i]}} = \lim_{m\to\infty} \frac{\log \mathrm{RD}(\boldsymbol{W}^{[i]})}{\log m}$$

- Linear regime: $S_{\boldsymbol{W}^{[i]}} < 0$
- Condensed regime: $S_{\boldsymbol{W}^{[i]}} > 0$
- Critical regime: $S_{\boldsymbol{W}^{[i]}} = 0$

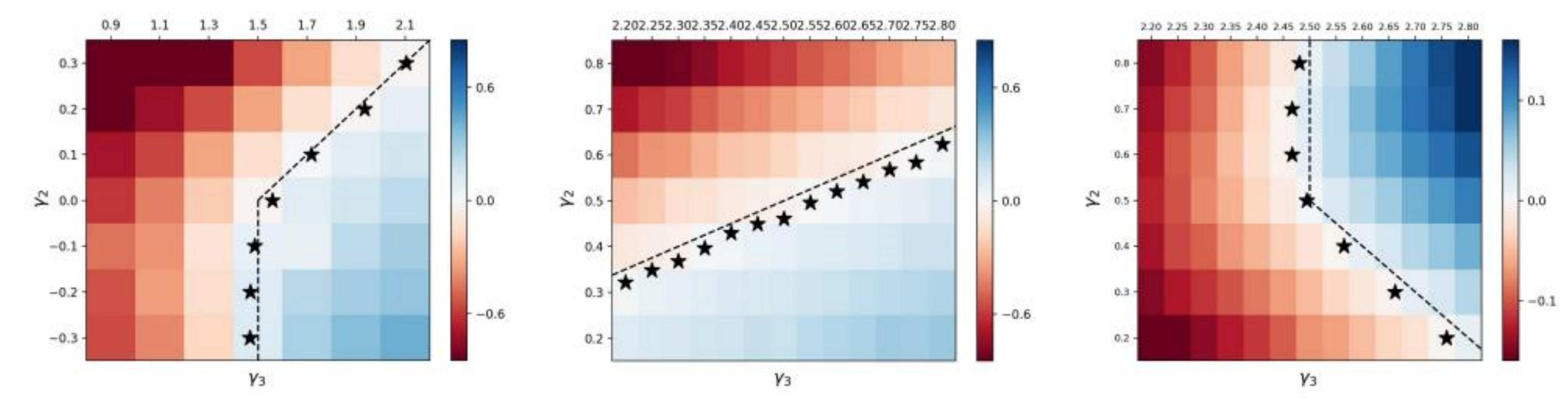
Regime identification and separation

We quantify the growth of $\mathrm{RD}\left(\boldsymbol{W}^{[i]}\right),\ i = 1,2$, as $m \to \infty$, by defining,

$$S_{\boldsymbol{W}^{[i]}} = \lim_{m\to\infty} \frac{\log \mathrm{RD}(\boldsymbol{W}^{[i]})}{\log m}$$

- Linear regime: $S_{\boldsymbol{W}^{[i]}} < 0$
- Condensed regime: $S_{\boldsymbol{W}^{[i]}} > 0$
- Critical regime: $S_{\boldsymbol{W}^{[i]}} = 0$

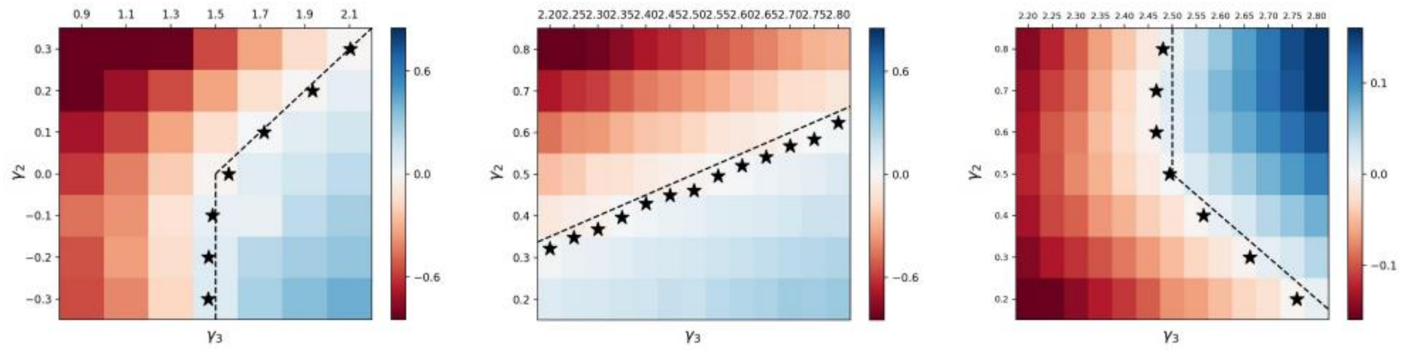
For synthetic data,



(a) S_{W_1} (b) S_{W_1} (c) S_{W_2}

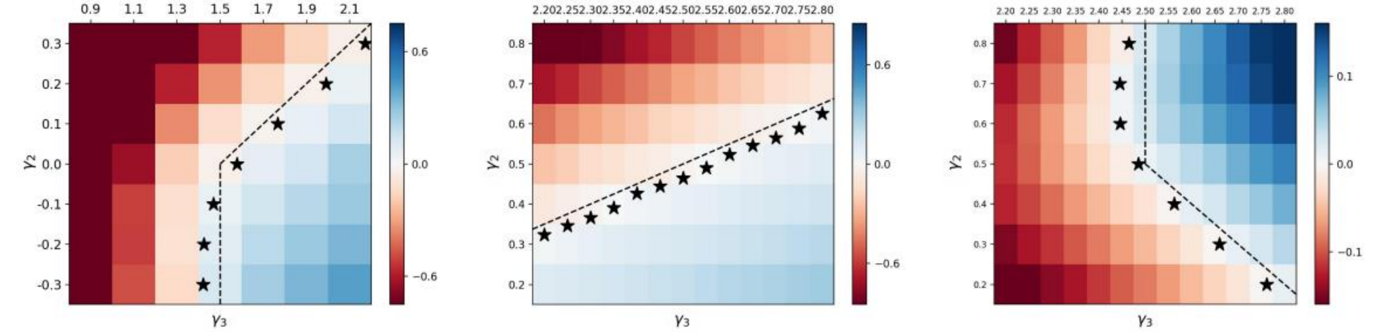
Regime identification and separation

For synthetic data,



(a) $S_{\mathbf{w}_1}$ (b) $S_{\mathbf{w}_1}$ (c) $S_{\mathbf{w}_2}$

For mnist data,



(a) $S_{\mathbf{w}_1}$ of mnist (b) $S_{\mathbf{w}_1}$ of mnist (c) $S_{\mathbf{w}_2}$ of mnist

Summary

- Characterize the linear, critical, and condensed regimes
- Identify the condensation as non-linear
- Figure out the relation between the training dynamics and initialization
- Draw the phase diagram
- Reveal different training dynamics within a neural network

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

Thank You

饮 水 思 源　爱 国 荣 校