Mean-field theory and dynamical isometry of deep neural networks

Feng Wang

AIRD, Coretronic Co.

Mar 25, 2019

References list & slides can be find at

https://github.com/fwcore/mean-field-theory-deep-learning

Outlines

- **Introduction**
- **Mean-field theory framework and its predictions**
  - Initialization strategies
    - MLP
    - CNN
  - Architectures
    - ResNet
    - Dropout
    - batch normalization
- **Details of the theory**
  - Assumptions
  - Possible pitfalls

Introduction

Initialization
- He's & Xavier's

Activation functions

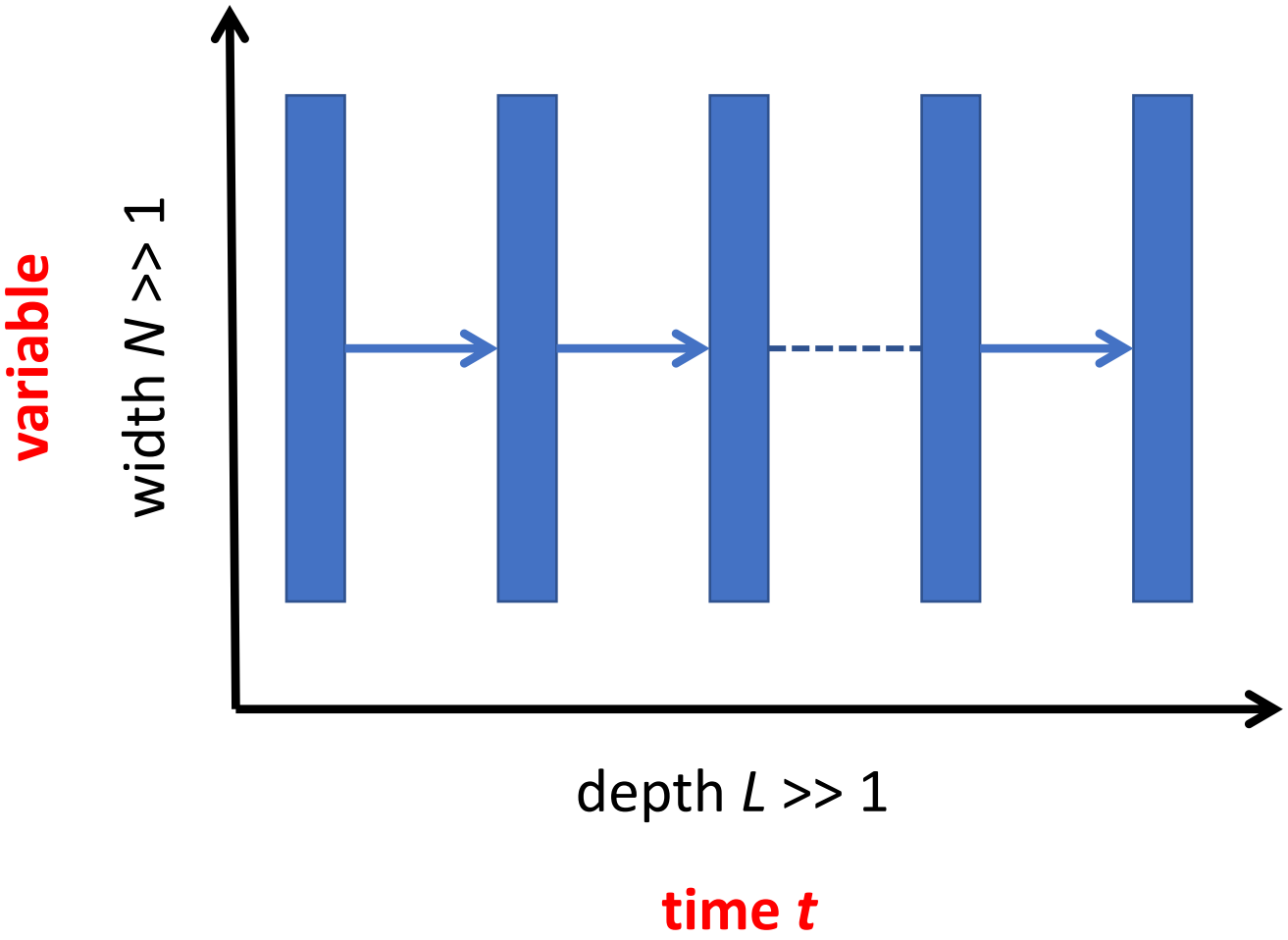
Dropout

BatchNorm

LSTM for RNN

To enhance the information propagation

Is there a theory for deep learning?



deep uniform neural networks

=

dynamical system

$$z_i^l = \sum_j W_{ij}^l y_j^l + b_i^l$$

$$y_i^{l+1} = \phi(z_i^l)$$

***But* …**

***z, y* are random variables and hard to track**

Mean-field theory for deep learning

$$z_i^l = \sum_j W_{ij}^l y_j^l + b_i^l$$

sum of random numbers

random numbers

Mean-field treatment

- **Replace *z* as Gaussian**
- **Match *z*'s mean and variance**

Law of large numbers

- mean(z) = 0
- Hence we only need to track var(z)

width N >> 1

Deep Information Propagation | arXiv:1611.01232

Mean-field flow of var(*z*)

$$\mathbb{E}[z^l_{i;a} z^l_{j;a}] = q^l_{aa} \delta_{ij}$$

$$q^l_{aa} = \sigma_w^2 \int \mathcal{D}z \phi^2 \left(\sqrt{q^{l-1}_{aa}} z \right) + \sigma_b^2$$

i, j: neural index
l: layer index
a, b: sample index

covariance

$$\mathbb{E}[z^l_{i;a} z^l_{j;b}] = q^l_{ab} \delta_{ij}$$

$$q^l_{ab} = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1) \phi(u_2) + \sigma_b^2$$

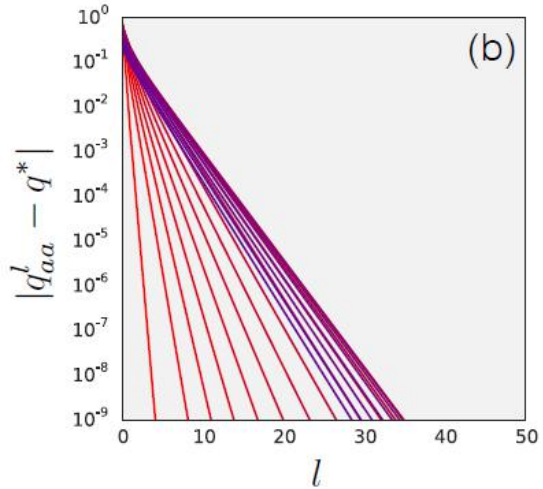
$$u_1 = \sqrt{q^{l-1}_{aa}} z_1$$

$$u_2 = \sqrt{q^{l-1}_{bb}} \left(c^{l-1}_{ab} z_1 + \sqrt{1 - (c^{l-1}_{ab})^2} z_2 \right)$$

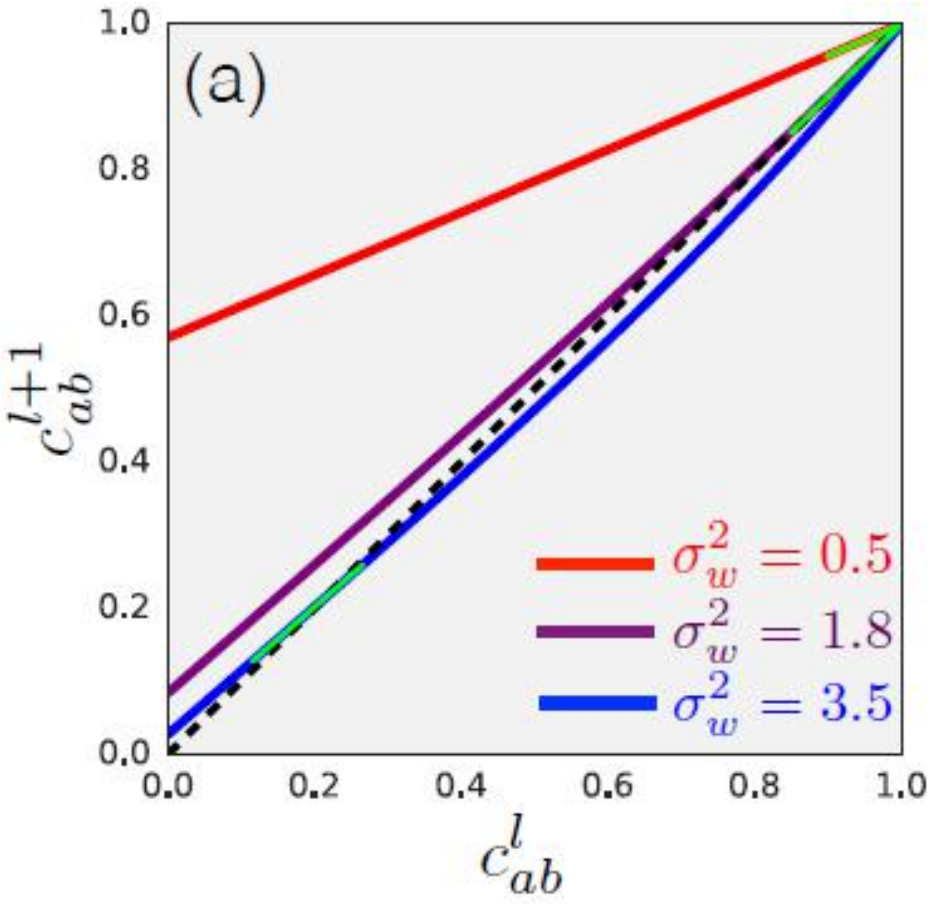
$$c^l_{ab} = q^l_{ab} / \sqrt{q^l_{aa} q^l_{bb}}$$

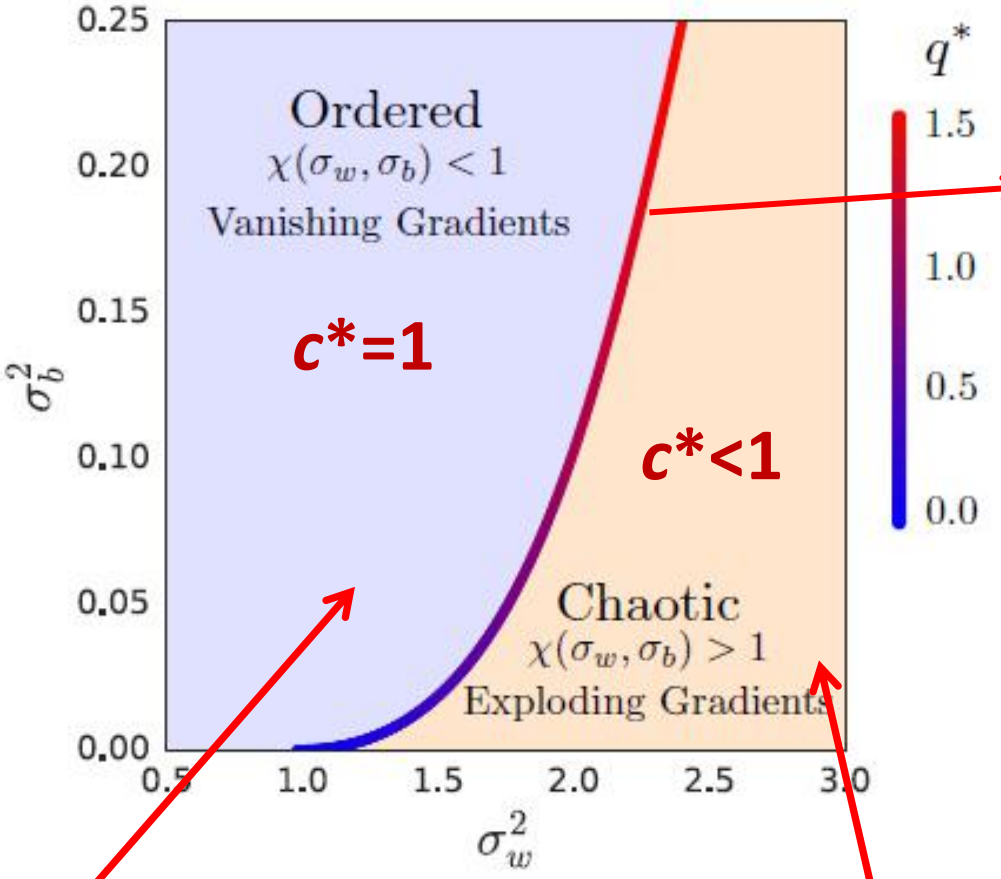
$q^* = \lim_{l \to \infty} q^l_{aa}$ is a smooth function.

However, ***limit c is differen***t



Mean-field flow of *c*: a phase transition





Edge of chaos

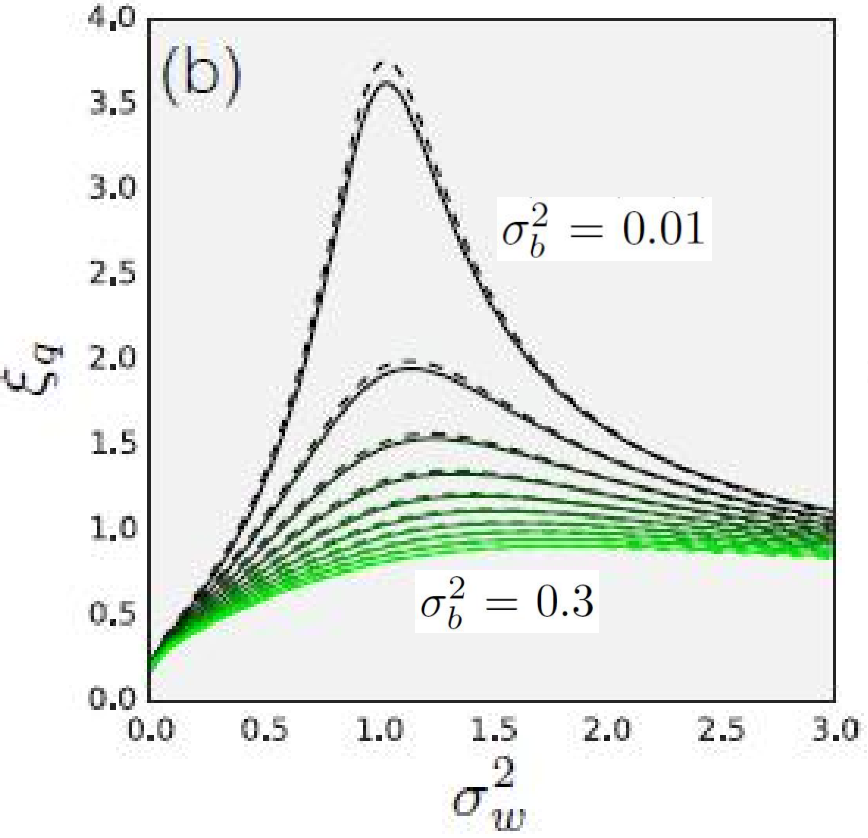
- Information collapse.
- Gradients vanish.

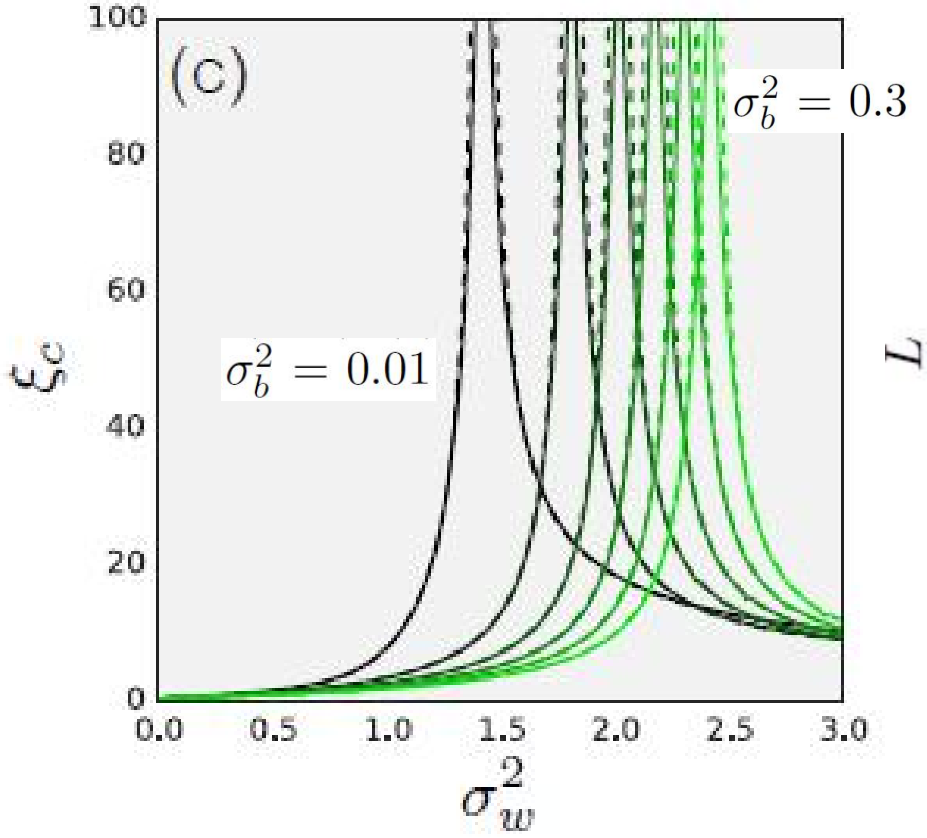
- Information decorrelates.
- Gradients explode.

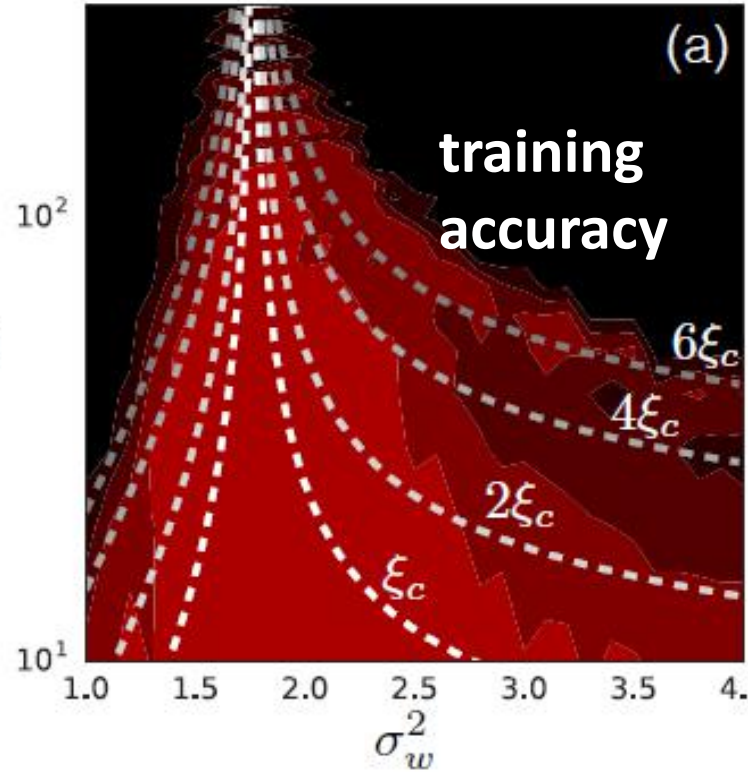
Deep Information Propagation | arXiv:1611.01232

Perturbation around c^* = asymptotic behavior near the edge of chaos

$$\epsilon^{l+1} = \epsilon^l \left[\sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1^*) \phi'(u_2^*) \right] + \mathcal{O}((\epsilon^l)^2) \sim e^{-l/\xi_c}$$

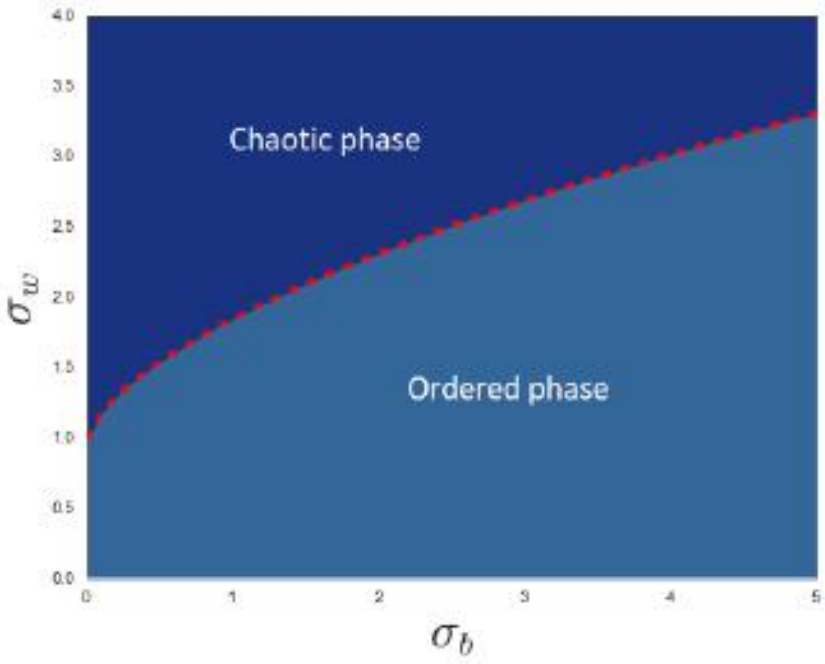




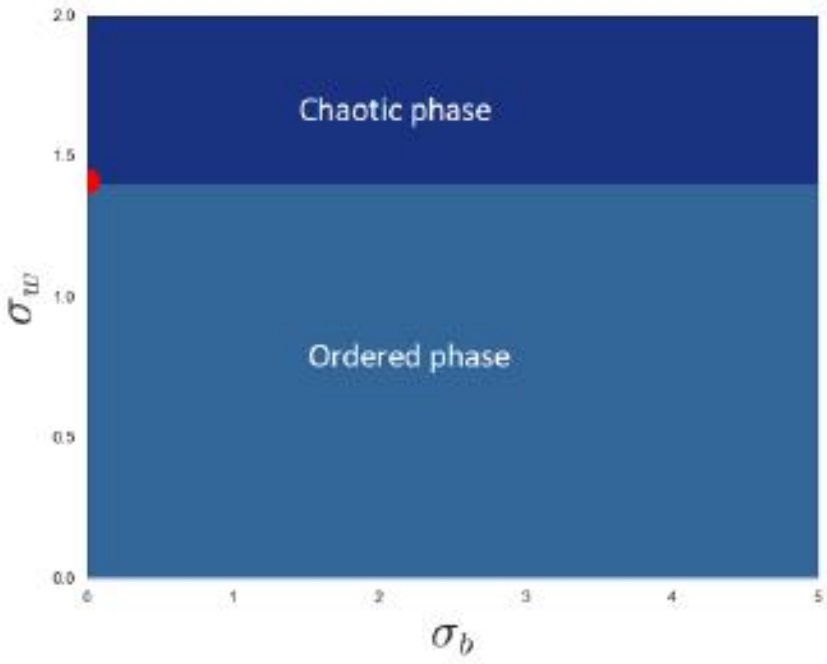


Deep Information Propagation | arXiv:1611.01232

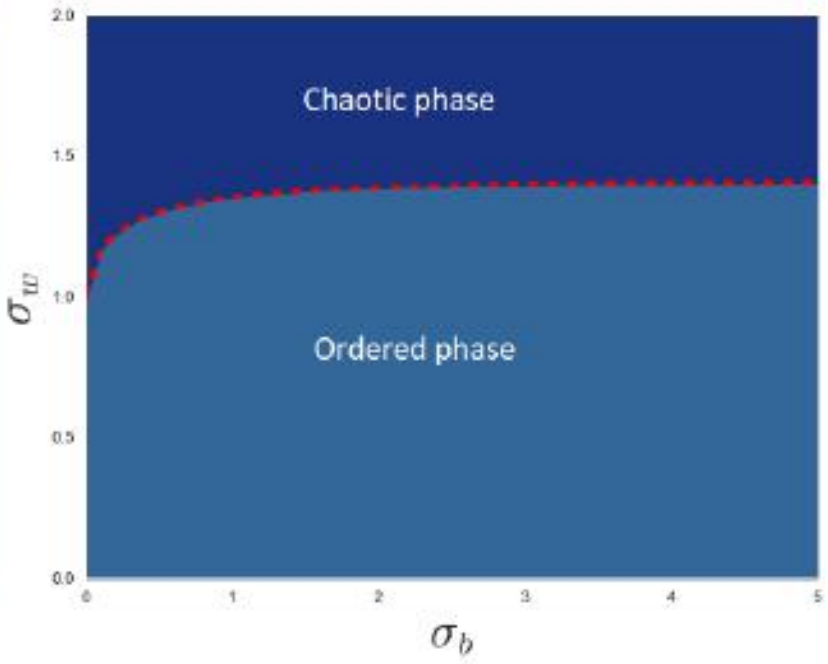
Activation functions



tanh



ReLU



ELU

Gaussian initialization is enough?

ξ_c controls information propagation depth, and determines training accuracy,

How about ***training speed*** & ***generalization***?

Edge of chaos
guarantees slow changes of correlation

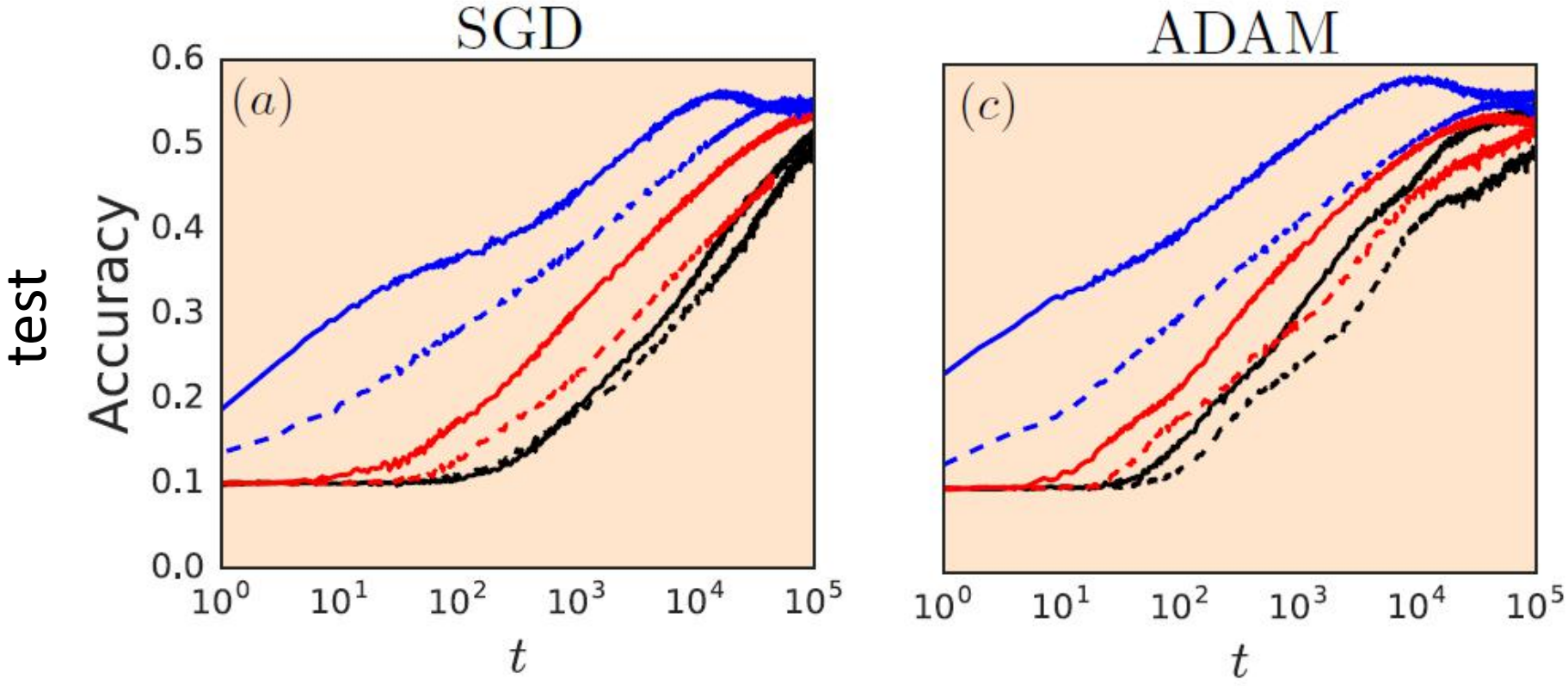
Dynamical isometry
guarantees correlation preservation

$$\mathbf{J} = \frac{\partial \mathbf{x}^L}{\partial \mathbf{h}^0} = \prod_{l=1}^{L} \mathbf{D}^l \mathbf{W}^l \qquad D_{ij}^l = \phi'(h_i^l)\,\delta_{ij}$$

All singular values have norm 1

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice | arXiv:1711.04735

Dynamical isometry initialization speeds up training

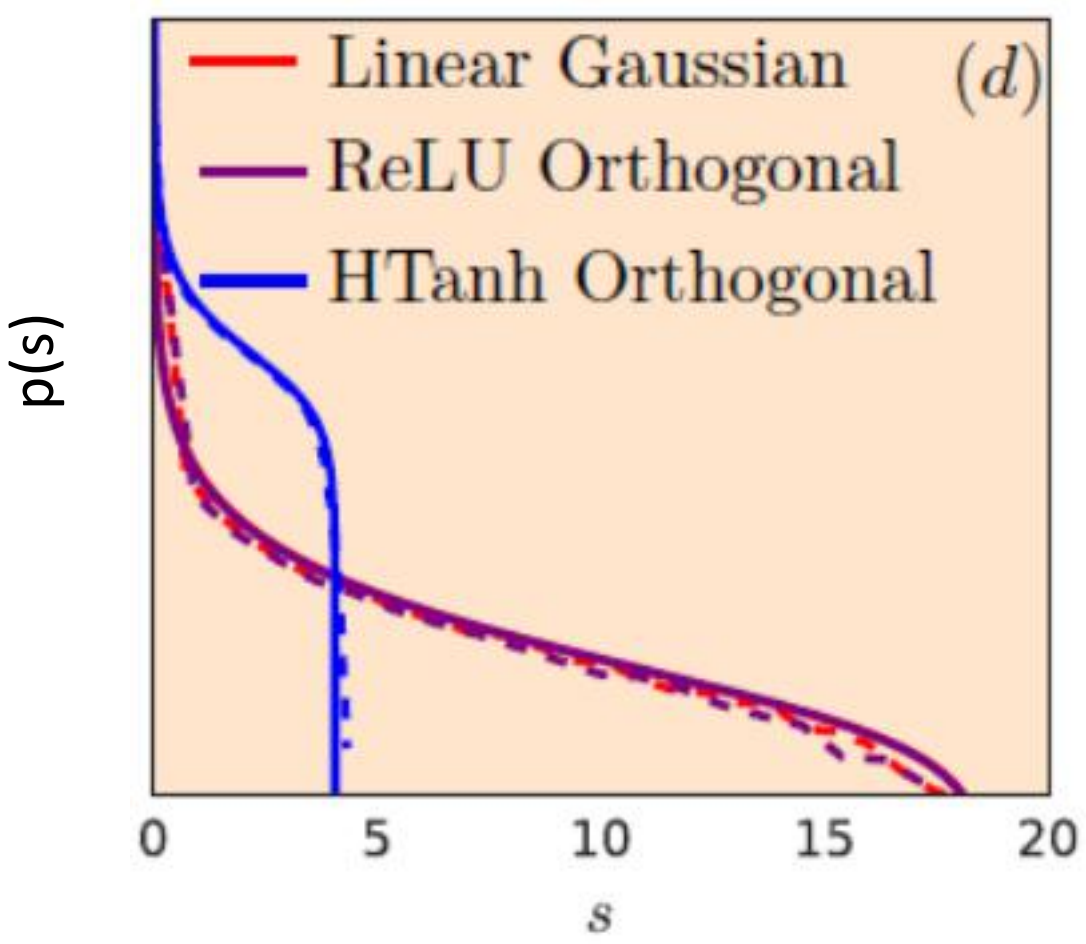


dashed line: Gaussian
solid line: orthogonal

tanh with $\sigma_w^2 = 1.05$ (close to EOC)
tanh with $\sigma_w^2 = 2$ (far away from EOC)
ReLU with $\sigma_w^2 = 2$ (at critical point)

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice | arXiv:1711.04735

Activation functions

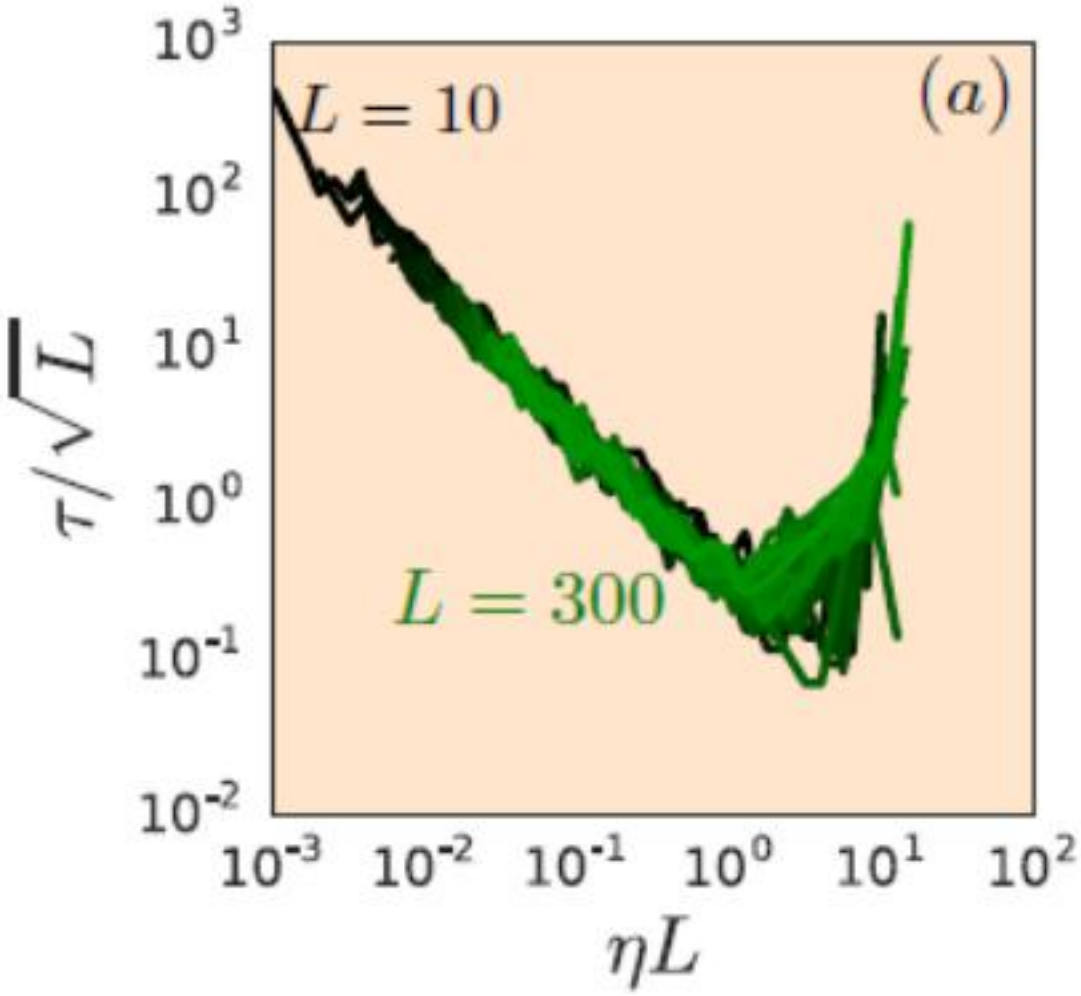


singular values of *J*

- ReLU has no dynamical isometry
- tanh is better to be initialized by Orthogonal scheme with a slightly small σ_b^2

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice | arXiv:1711.04735

Scaling of training time

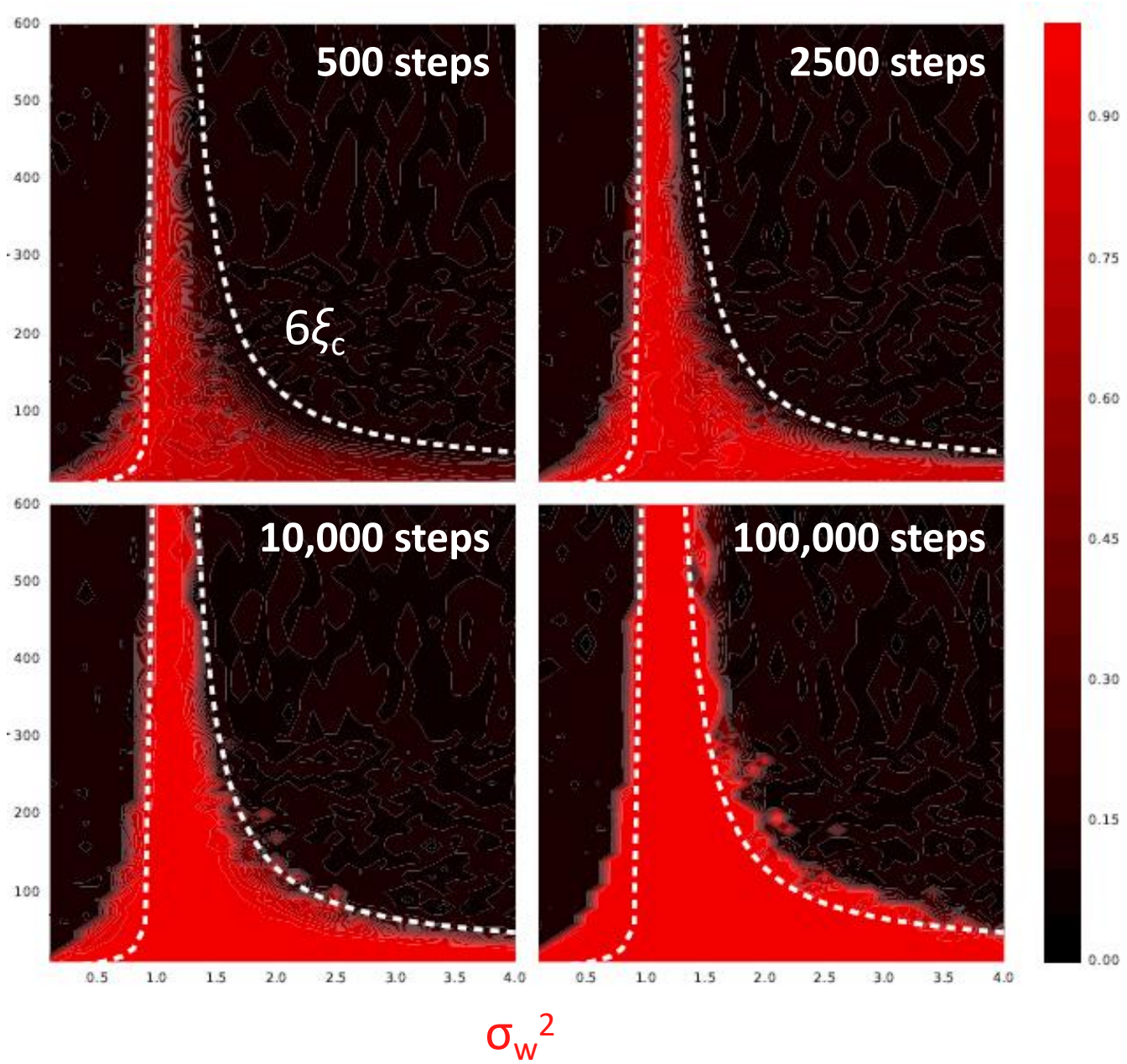


Under dynamical isometry

- training time grows slower than *L*
- optimal training rate η ~ 1/*L*

Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice | arXiv:1711.04735

Deep ConvNet

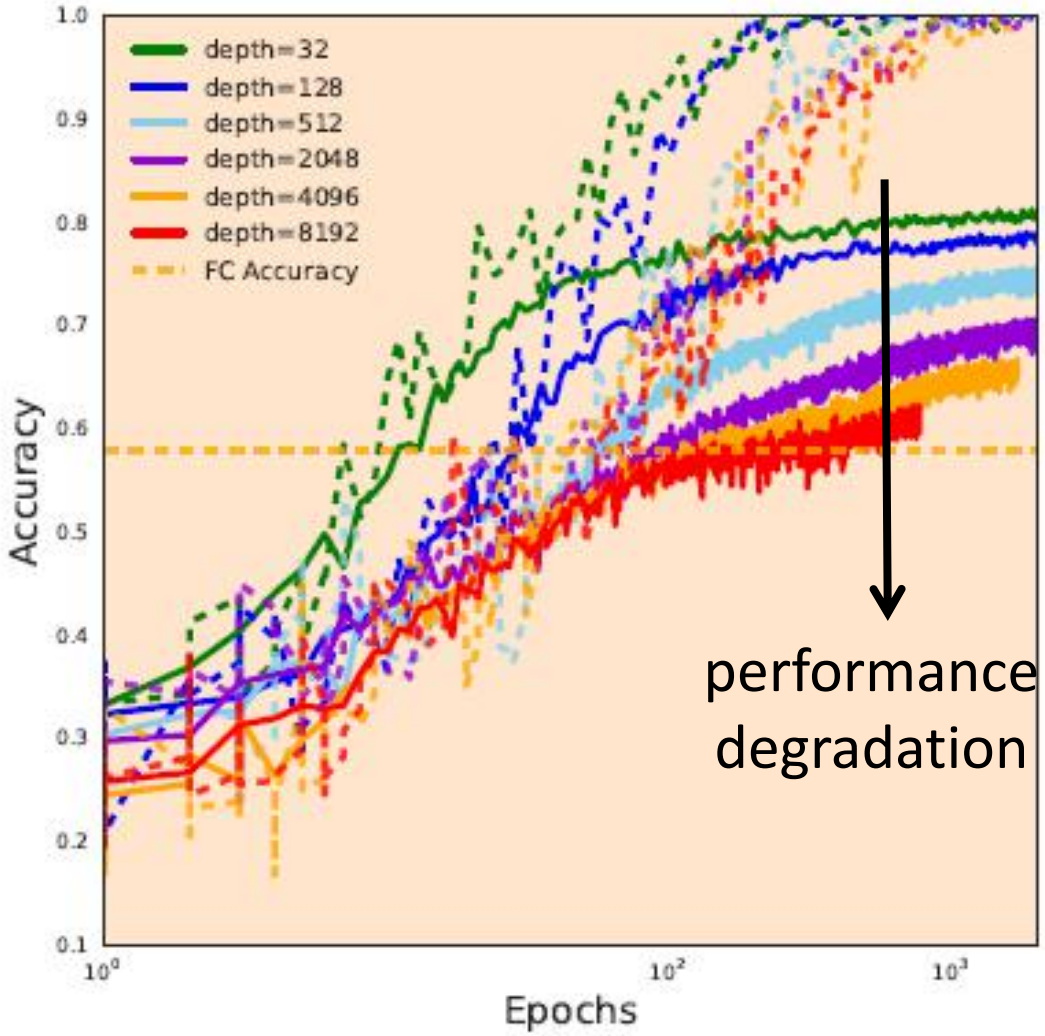


training accuracy

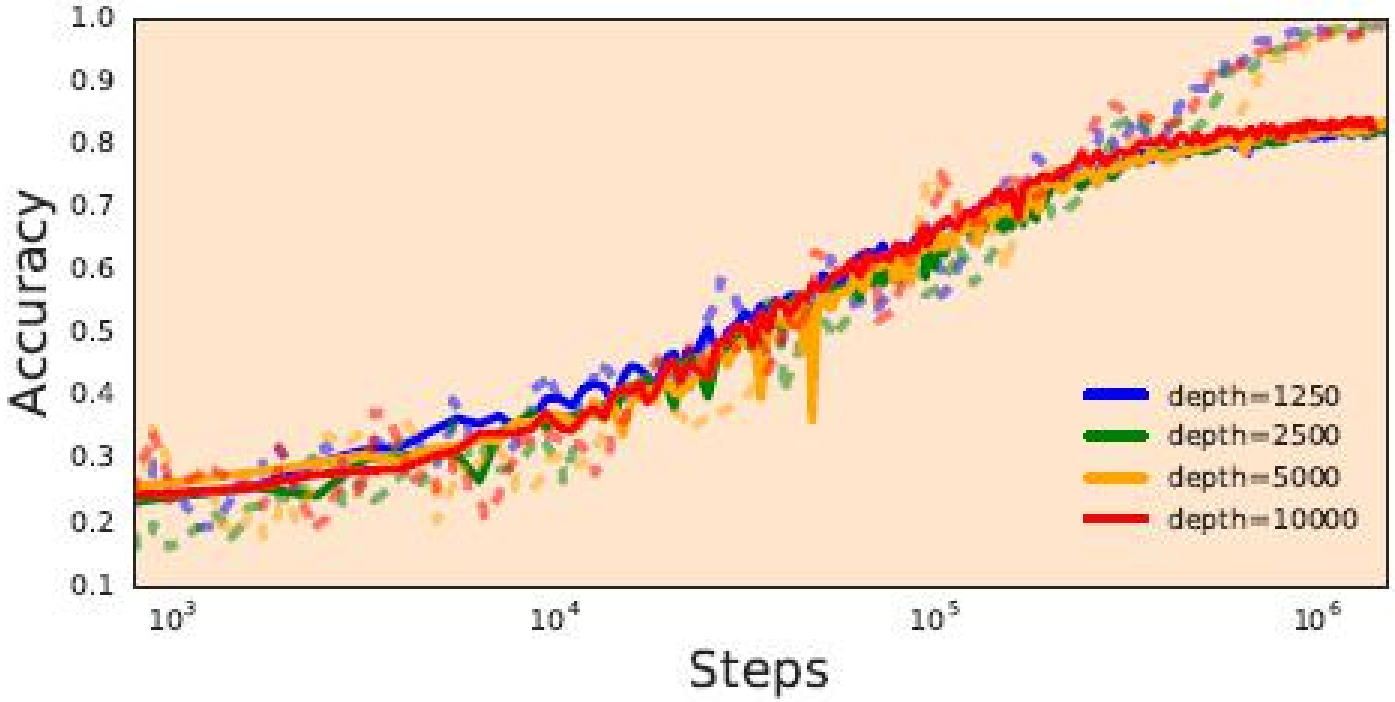
Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks | arXiv:1806.05393

Dynamical isometry for deep ConvNet

Orthogonal initialization



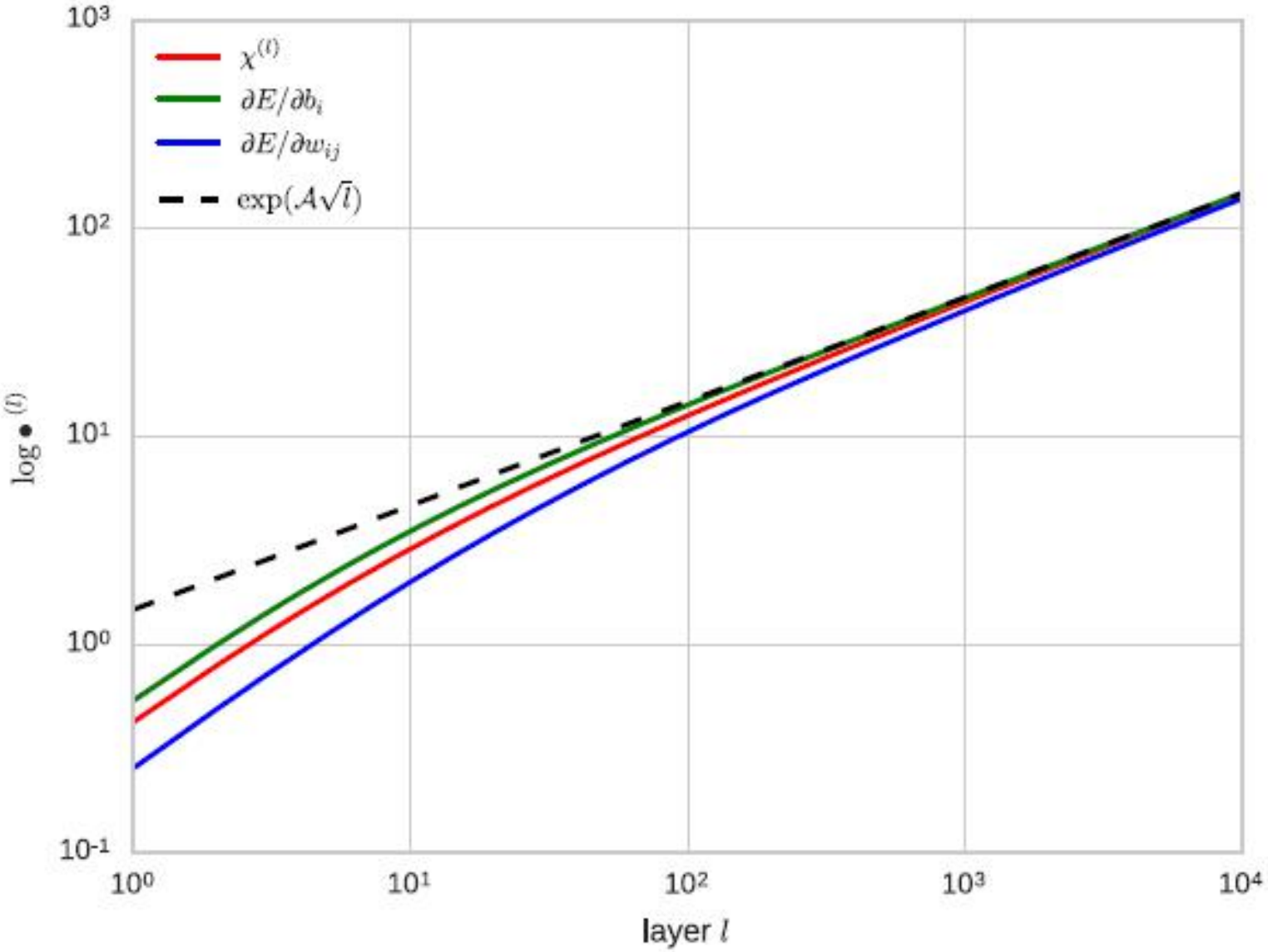
Delta-Orthogonal initialization



Dynamical isometry prevents performance degradation caused by increasing depth.

Reminder: ResNet is an alternative

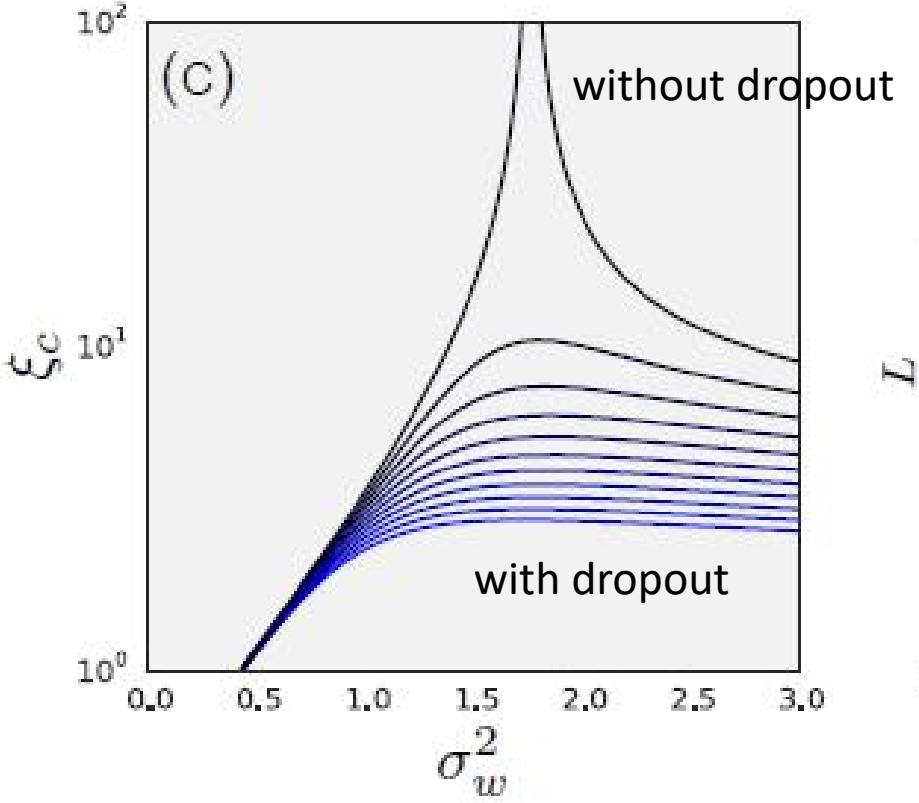
Why ResNet?

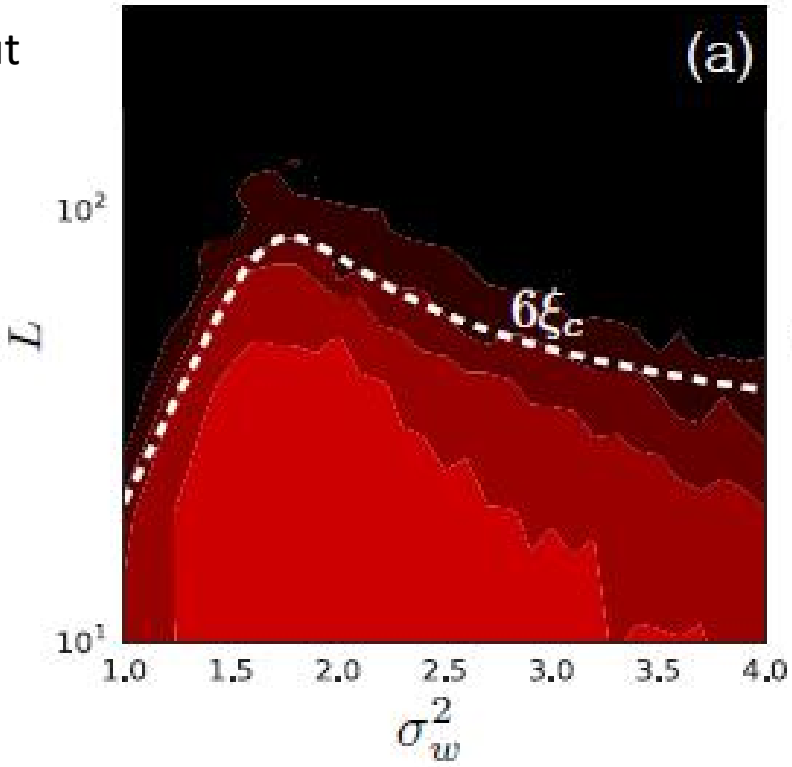


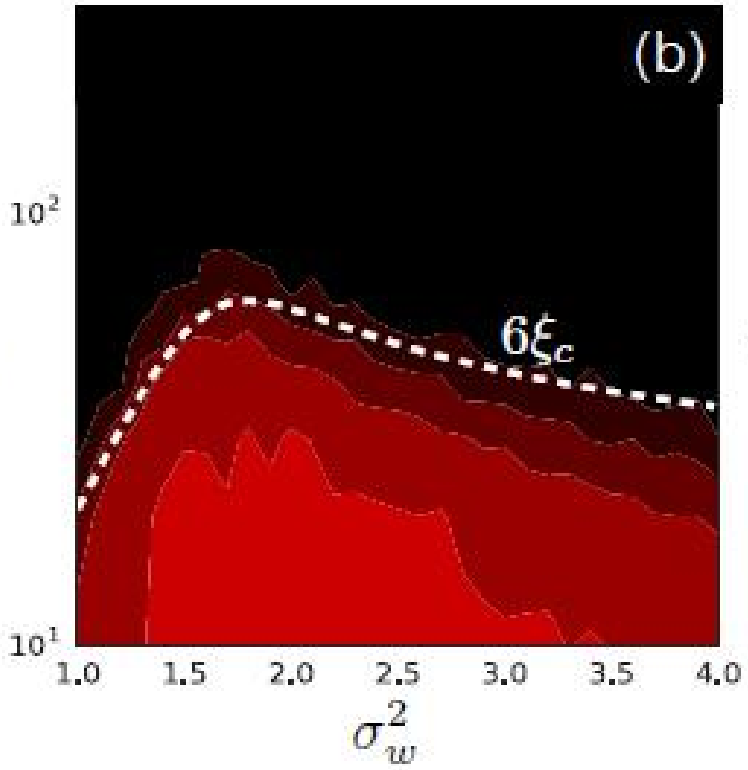
- **Xavier or the He initializations are not optimal for residual networks**
- **the optimal initialization variances depend on the depth**

Mean Field Residual Networks: On the Edge of Chaos | arXiv:1712.08969

Dropout limits the depth

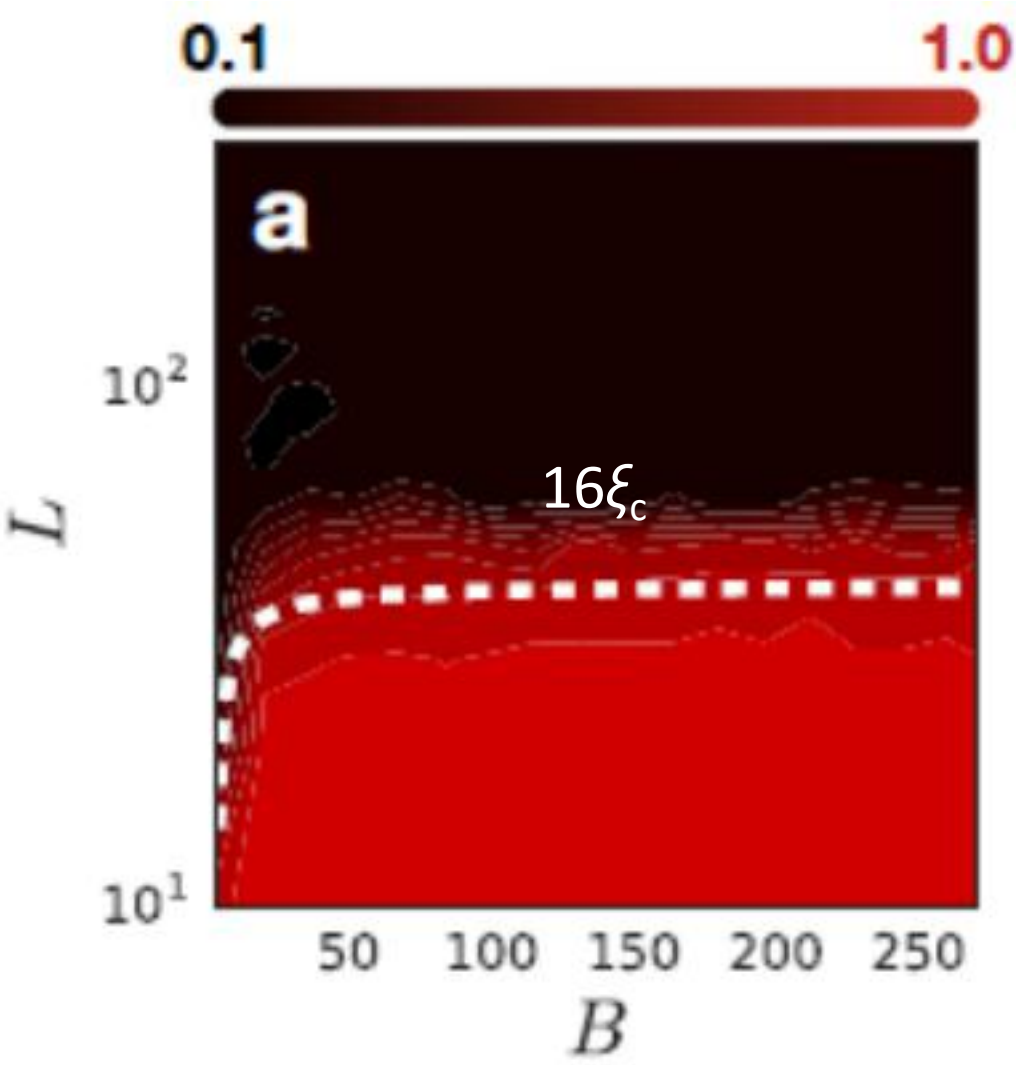






Deep Information Propagation | arXiv:1611.01232

Batch normalization limits the depth



Method to overcome

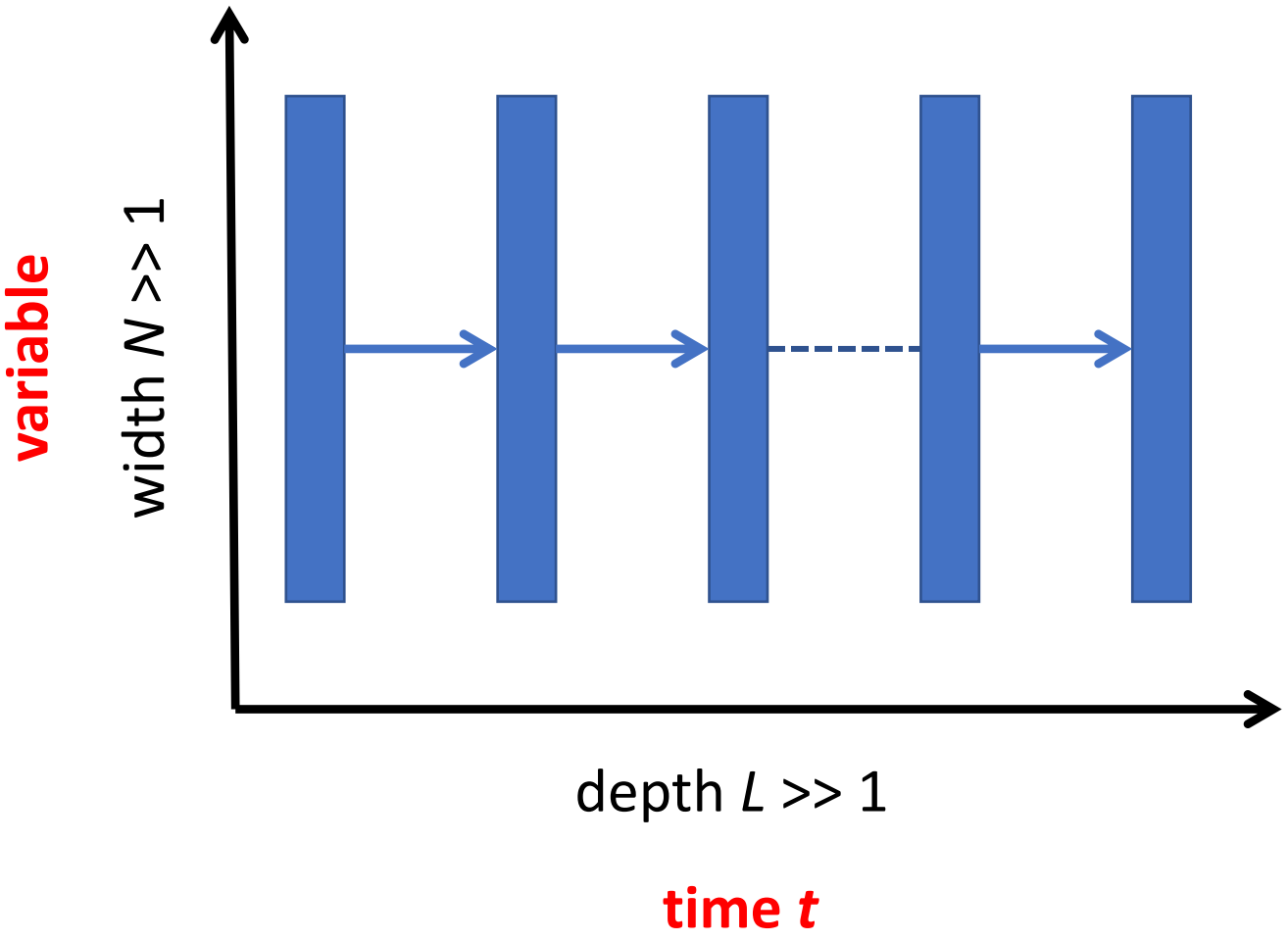
*Gradient explosion can be **reduced** by tuning the network close to the **linear** regime*

A Mean Field Theory of Batch Normalization | [arXiv:1902.08129]

Outlines

- Introduction
- Mean-field theory framework and its predictions
  - Initialization strategies
    - MLP
    - CNN
  - Architectures
    - ResNet
    - Dropout
    - batch normalization
- **Details of the theory**
  - Assumptions
  - Possible pitfalls

Mean-field theory Assumptions



deep uniform neural networks

=

dynamical system

Assumptions

- Uniform layers (constant N)
- independent weights for forward & back propogation
- $N >> 1$
- $L >> 1$
- theory works *only* for untrained model (time t = layer number ≠ epoch)

Mean-field theory limitations/pitfalls

- **Theory works *only* for untrained model**
  (time t = layer number ≠ epoch)

- **Orthogonal initialization ≠ dynamical isometry**
  necessary but not sufficient

- **May not work if width is small**

- **May not work for shallow networks**

- *Not theory on generalization/test performance, training speed*
  only on information propogation

Summary

- Mean-field theory (MFT) describes information propogation.
- MFT determine the maximal training depth.
- Initialization at the edge of chaos unlimits the training depth.
- Dynamical isometry speeds up training, and prevents test accuracy degradation with depth.
- Orthogonal initialization is a powerful scheme to reach dynamical isometry.
- BatchNorm and dropout limit the training depth.
- ReLU has no dynamical isometry.