2026/05/02 AISTATS 2026 Spotlight Talk

Precise Dynamics of Diagonal Linear Networks

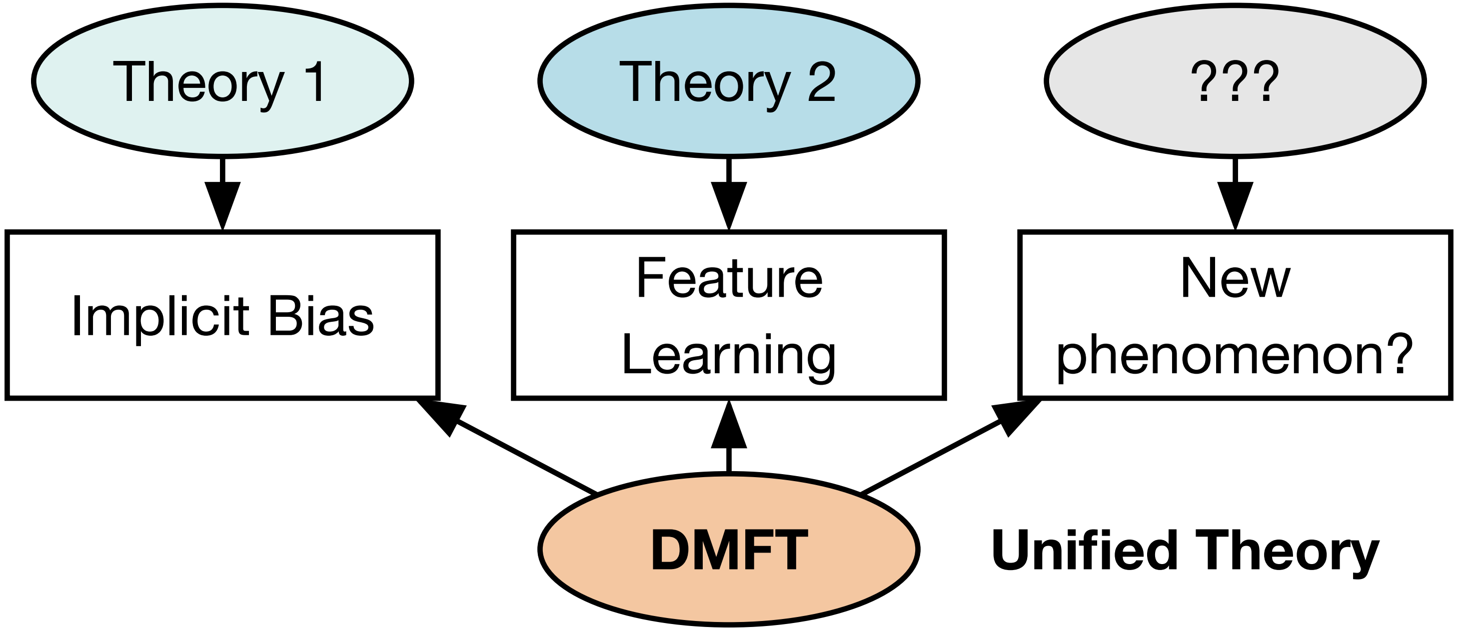
A Unifying Analysis by Dynamical Mean-Field Theory

Sota Nishiyama[1,2] **Masaaki Imaizumi**[1,2,3]

The University of Tokyo[1] RIKEN AIP[2] Kyoto University[3]

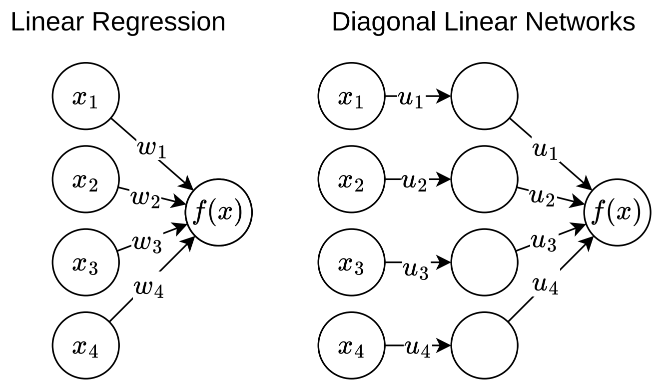
Quick Overview | Unifying Analysis of NN Dynamics

NN training dynamics exhibits various (nonlinear) phenomena



Contribution: We provide a unifying analysis of various dynamical phenomena in a simple NN model using a tool from physics

Setup | Diagonal Linear Networks



- Linear function with nonlinear parameterization

$$f(\boldsymbol{x}; \boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{w}^\top \boldsymbol{x}, \quad \boldsymbol{w} = \frac{1}{2}(\boldsymbol{u}^{\odot 2} - \boldsymbol{v}^{\odot 2})$$

(trainable parameter $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, input data $\boldsymbol{x} \in \mathbb{R}^d$)

- Although linear in input, the gradient descent dynamics is nonlinear

DLN is a simple theoretical model of **nonlinear learning dynamics** in NNs

- It captures nontrivial nonlinear learning phenomena

e.g. implicit bias (Woodworth et al., 2020) and incremental learning (Berthier, 2023; Pesme & Flammarion, 2023)

Method | Dynamical Mean-Field Theory (DMFT)

DMFT allows us to analyze the **typical macroscopic behavior** of high-dimensional systems in a tractable way

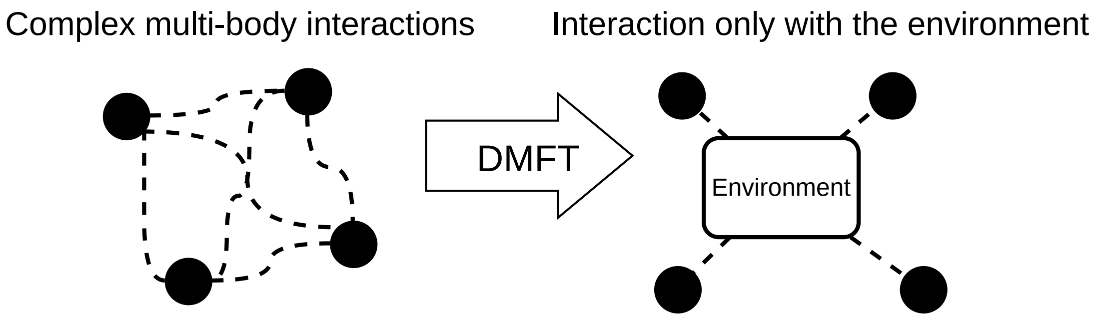
High-dimensional random dynamics ❌:

many interacting components, intractable to analyze directly

↓ **Dynamical Mean-Field Theory (DMFT)**

Low-dimensional effective dynamics ✓:

interaction between a single effective component and a mean-field

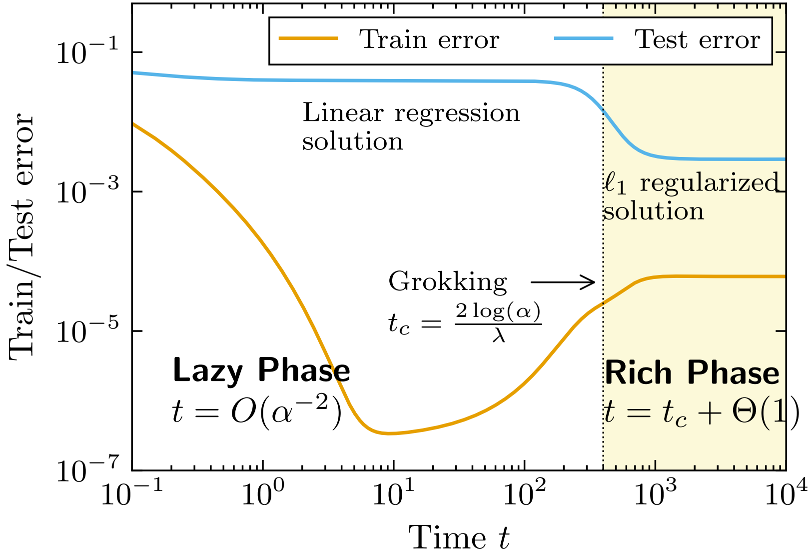


Result 1 | Timescale Structure of Training Dynamics

Qualitatively different behaviors
for large vs small initialization scale α

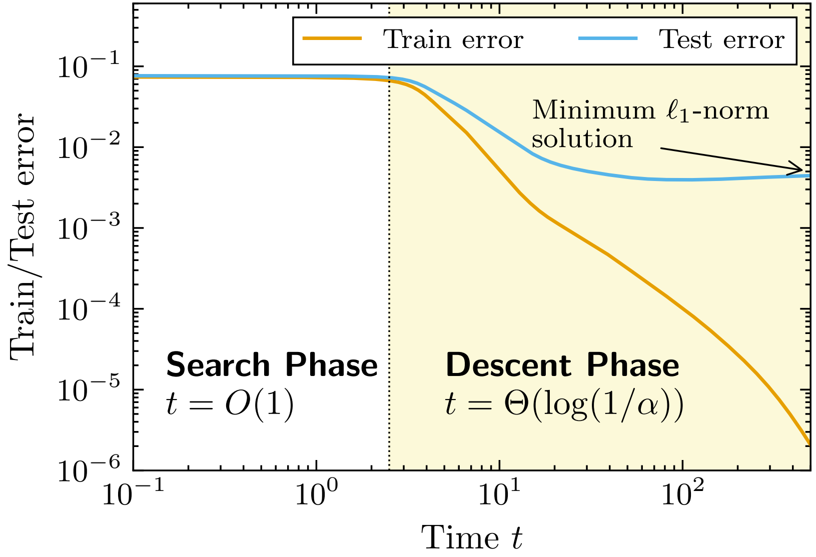
Large initialization $\alpha \gg 1$

Fast fitting, late generalization



Small initialization $\alpha \ll 1$

Slow incremental learning



Result 2 | Long-Time Behavior

Fixed point

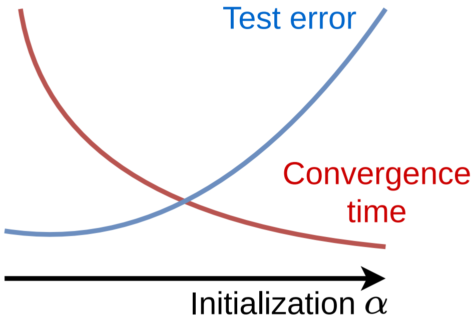
Smaller initialization α

$\rightarrow$ sparser implicit bias, **better generalization** 👍

Convergence rate

Smaller initialization α

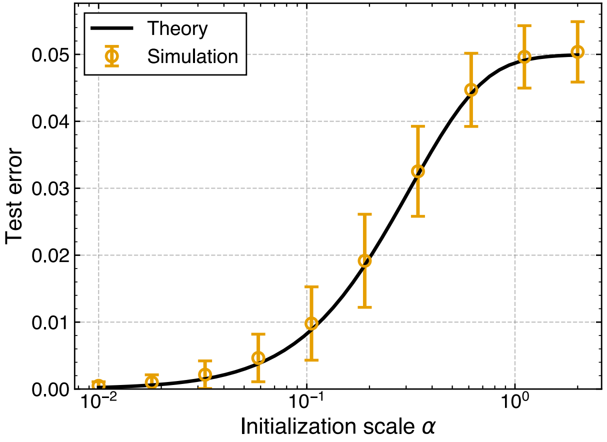
$\rightarrow$ **slower convergence** 👎

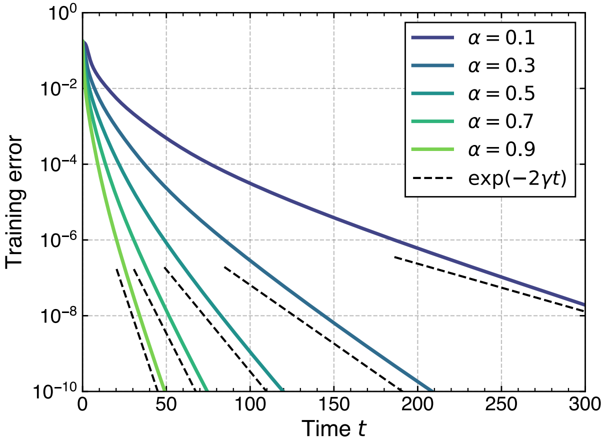


Trade-off between

generalization vs optimization

Better solutions are harder to find





Summary

What we did

- Derived a DMFT equation for DLNs to capture its training dynamics
- Analyzing the DMFT equation, we revealed…
  - **Initialization scale dependent timescale structure, implicit bias, convergence speed**
  - **Trade-off between generalization vs optimization**

Take-home message: DMFT is a powerful tool that uncovers a rich structure of nonlinear NN training dynamics