arxiv:2405.00642

Bridging Empirics and Theory: Unveiling Asymptotic Universality Across Gaussian and Gaussian Mixture Inputs in Deep Learning

Jaeyong Bae, Hawoong Jeong† *Department of Physics, KAIST*

qowodyd0116@kaist.ac.kr

I. Introduction

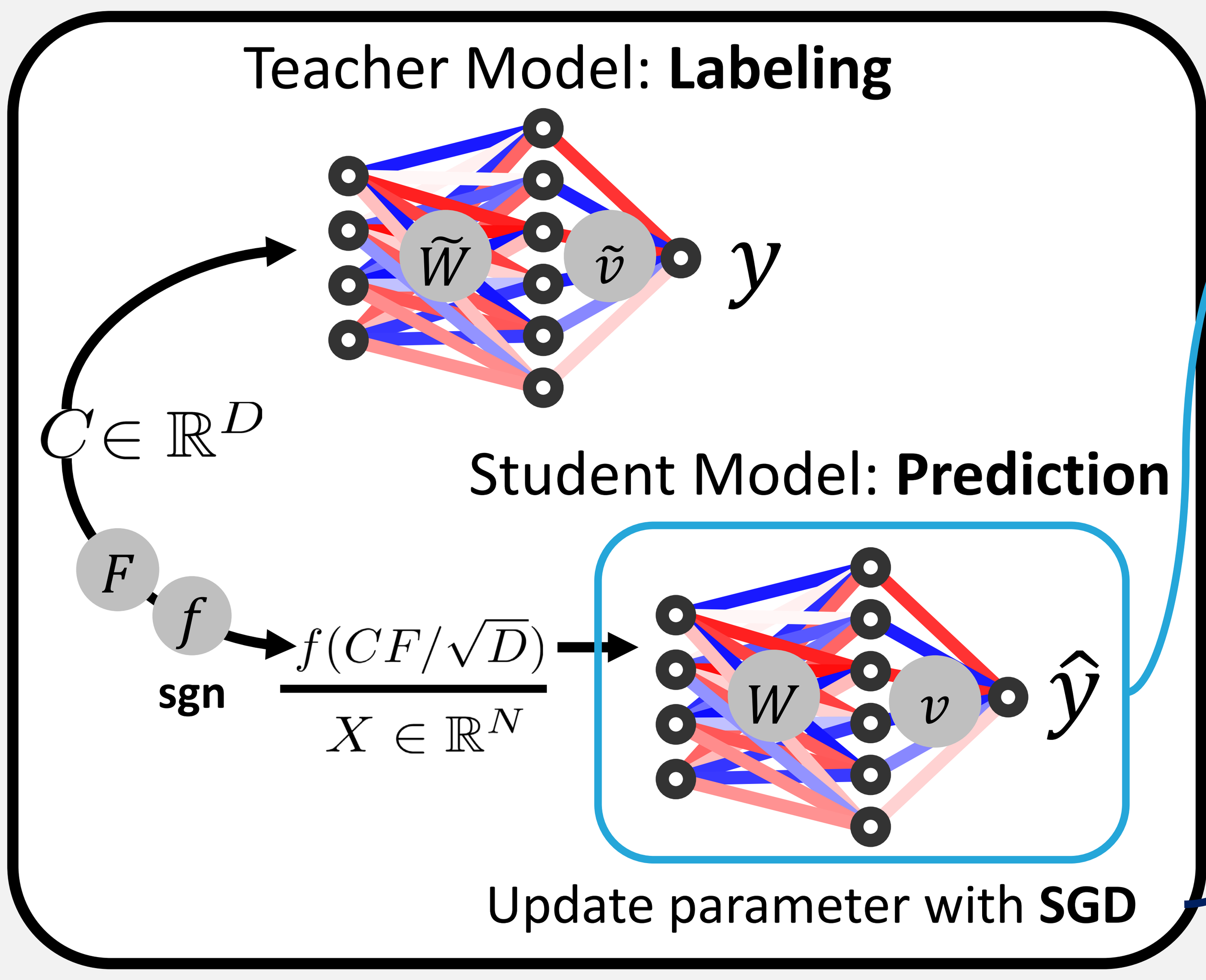
PhysRevX.10.041044

Previous Findings: Dynamics of Hidden Manifold Teacher-Student Model

The **core property C** is based on the **latent space(*manifold*)** dimension D.

In real world, this core property becomes **wrapped, resulting to X**, on the **real space** dimension N.

$C \sim \mathcal{N}(0, I)$



Gaussian Equivalence Property

In the Limit of $N \to \infty, D \to \infty$, the preactivations of the teacher and student model $\lambda = XW^\top/\sqrt{N}$, $\nu = C\widetilde{W}^\top/\sqrt{D}$ conform jointly Gaussian variables.

Statistics involving $\{\lambda, \nu\}$ are entirely represented by their mean and convariances

$Q_{k,\ell} \equiv \mathbb{E}[\lambda_k \lambda_\ell]$, $R_{k,m} \equiv \mathbb{E}[\lambda_k \nu_m]$, $\cdots$

SGD to ODE + Change of Basis

Dynamics are defined by statisitics of preactivation distribution

" The Dynamics of Student Model are fully theoretically tractable"

Ex) v_k dynamics

$$W_{k,i} := W_{k,i} - \frac{\eta}{\sqrt{N}} v_k (\hat{y} - y) g'(\lambda_k) f(U_i)$$

$$v_k := v_k - \frac{\eta}{N} g(\lambda_k)(\hat{y} - y)$$

Scaled SGD

$\Delta t \to dt$

$$\frac{dv_k}{dt} = \eta \left[\sum_n^M \tilde{v}_n I_2(k,m) - \sum_j^K v_j I_2(k,j) \right]$$

Expectation of functions

$$I_2(k,m) = \mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)]$$

$$I_2(k,j) = \mathbb{E}[g(\lambda_k)g(\lambda_j)]$$

Tractable under GEP;

$\because$ ***$\{\lambda, \nu\}$ Statistics are tractable***

Q : The Simple Gaussian $C \sim \mathcal{N}(0, I)$ setting is enough?

How dynamics changes as the inherent distribution deviate from simple Gaussian to Gaussian mixture?

II. Setting

Gaussian Mixture Setting

- m component Gaussian mixture with moderating parameter α

$$r = r_i \sim \mathcal{N}(\mu_i, I), \quad \mu_i \sim \mathcal{U}[-\alpha, \alpha), \text{ with } p_i, \sum^m p_i = 1$$

Empirical investigations across a spectrum (m, α) of Gaussian mixture

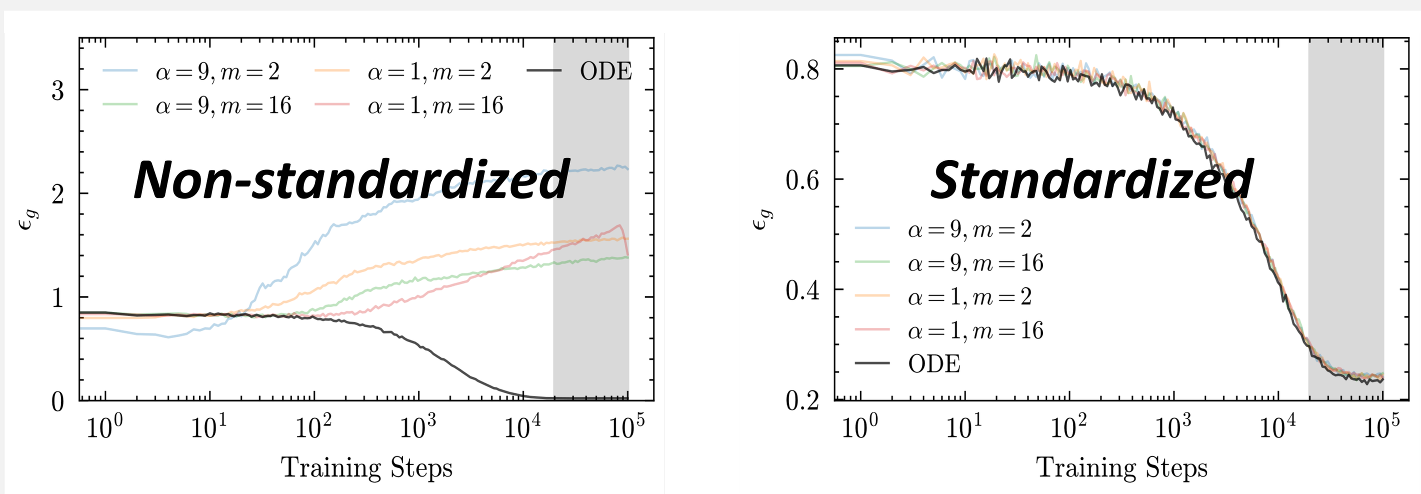
Dynamic Metrics

- Covariance of preactivations $Q_{k,\ell}, R_{k,m}$,
- generalization error ϵ_g and 2nd layer weight v_k

III. Results & Discussion

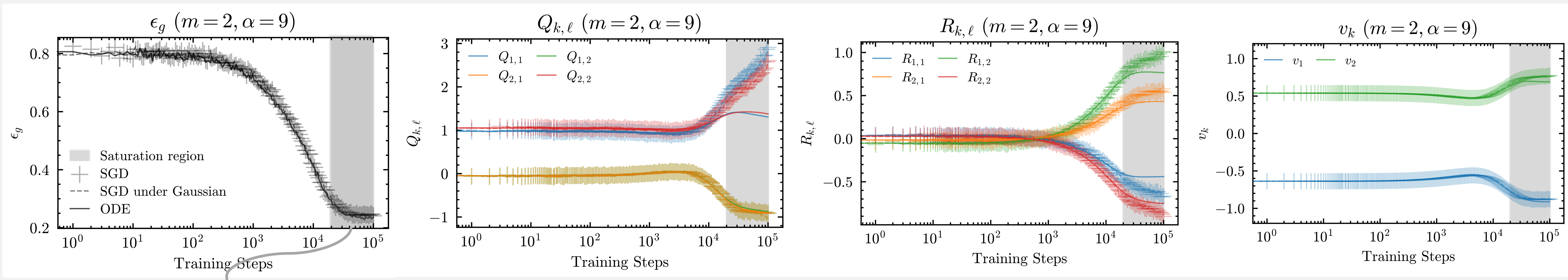
Dynamics across various Gaussian mixture

- With ***non-standardized*** Gaussian mixture, the theoretical prediction completely **collapse** for ***all spectrum of mixture***
- With ***standardized*** Gaussian mixture***, mixture spectrum dose not effect***



Dynamics under standardized Gaussian mixture

Sample-wise $C \sim \mathcal{P}$, $\quad C := (C - \mathbb{E}[C])/\sqrt{\mathbb{E}[(C - \mathbb{E}[C])^2]}$.



Saturation region; Randomness has a significant effect

- Theoretical prediction under the Simple Gaussian **surprisingly aligned** with Mixture dynamics

Convergence in Standardized Gaussian Mixture

Key quantity: "expectation value of functions" in dynamics

Dominance of Moments in the Expectation value of Functions

[Definition] If $\mathcal{P}$ shares identical cumulants with $\mathcal{D}$ up to order 2, represent $\mathcal{P}$ as $\mathcal{P}_{\mathcal{D}_2}$

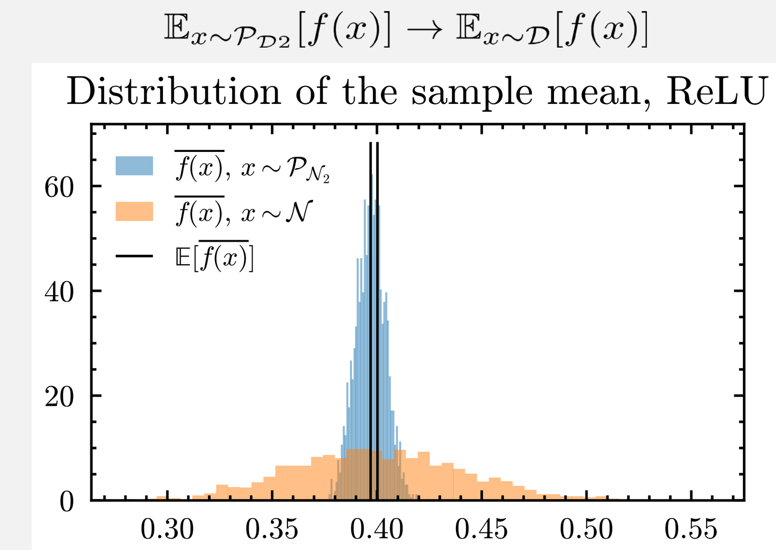
[Lemma] If a function f has the property of **erasing the influence of high-order cumulants**, then the expectation value of the function tends towards the same value.

$$\mathbb{E}_{x \sim \mathcal{P}_{\mathcal{D}2}}[f(x)] \to \mathbb{E}_{x \sim \mathcal{D}}[f(x)]$$

[Proof] Taylor expansion of expectation + Function property vanish high order derivation

$$\mathbb{E}_{x \sim \mathcal{P}_{\mathcal{D}_2}}[f(x)]_{x \in I} = \mathbb{E}[f(\mu) + \cdots + f^{(n)}(\mu)\frac{(x-\mu)^n}{n!} + \cdots]_{x \in I} \approx f(\mu) + f''(\mu)\mathbb{E}[\frac{(x-\mu)^2}{2!}] \to \mathbb{E}_{x \sim \mathcal{D}}[f(x)]_{x \in I}$$

$\mathbb{E}_{x \sim \mathcal{P}_{\mathcal{D}2}}[f(x)] \to \mathbb{E}_{x \sim \mathcal{D}}[f(x)]$



Emprical results, with $f \equiv$ ReLU

[Lemma] ReLU activation function **well erasing** influence of high-order cumulants

[Proof] In piecewise view, ReLU function is zero in 2nd Derivative

IV. Summary

Specific function + standardized ⇒ Make Dynamics dominantly up to 2nd Cumulants.

Even the mixture, the dynamics also tend to approximate same.

Acknowledgement

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF Grant No. 2022R1A2B5B02001752).