From Empirical Observations to Universality: Dynamics of Deep Learning with Inputs Built on Gaussian mixture

Anonymous Authors[1]

Abstract

This study broadens the scope of theoretical frameworks in deep learning by delving into the dynamics of neural networks with inputs that demonstrate the structural characteristics to Gaussian Mixture (GM). We analyzed how the dynamics of neural networks under GM-structured inputs diverge from the predictions of conventional theories based on simple Gaussian structures. A revelation of our work is the observed convergence of neural network dynamics towards conventional theory even with standardized GM inputs, highlighting an unexpected universality. We found that standardization, especially in conjunction with certain nonlinear functions, plays a critical role in this phenomena. Consequently, despite the complex and varied nature of GM distributions, we demonstrate that neural networks exhibit asymptotic behaviors in line with predictions under simple Gaussian frameworks.

1. Introduction

The endeavor to connect practical deep learning applications with theoretical understanding is a growing field of research (Seung et al., 1992; Engel, 2001; Zdeborová & Krzakala, 2016; Bahri et al., 2020; Zdeborová, 2020). Recent studies in this area have begun to study on neural network characteristics under structured nature of input data (Bartlett et al., 2020; Korada & Montanari, 2011; Candès & Sur, 2020; Goldt et al., 2019; 2020; 2022; Dandi et al., 2023).

The notion of structured data posits that despite the high-dimensional nature of typical datasets (such as MNIST in LeCun (1998) has 28x28, CIFAR in Saad & Solla (1995) has 3x32x32), these can often be distilled into lower-dimensional representations. This concept is exemplified in the MNIST dataset, where digits, rather than being random pixel assemblies, exhibit structured patterns like lines and circles. This phenomenon of low-dimensional structural features in data has been corroborated by numerous studies (Pope et al., 2021; Fefferman et al., 2016; Hinton & Salakhutdinov, 2006; Peyré, 2009; Creswell et al., 2018; Goodfellow et al., 2020).

From discussions on the presence of these low-dimensional structural features, considerable research has delved into the intriguing aspects manifested in deep learning when inputs with such characteristics are presented. Particularly, recent studies have found that when inputs inherently exhibit single Gaussian distribution traits, these characteristics are preserved even after passing through the first layer, facilitating a theoretical understanding of deep learning dynamics, such as weight and loss evolution, in simple two-layer models (Goldt et al., 2019; 2020; 2022).

However, these significant theoretical discoveries fall short of fully mirroring the complexities and heterogeneities of real-world data. Typically, Gaussian mixtures are preferred over single Gaussian distributions for modeling general data characteristics (Carreira-Perpinan, 2000; Scott, 2015; Goodfellow et al., 2016), with recent findings suggesting that real-world data often follows Gaussian mixture-like distributions (Seddik et al., 2020; Dandi et al., 2023). This gap underscores the necessity for our investigation into Gaussian mixture-based models to capture the complexity of real-world data more accurately

Given this context, our research extends the current theoretical discourse by examining the neural network dynamics when inputs are characterized not by simple Gaussian but by Gaussian mixtures. More specifically, we analyze how neural network dynamics changes as the inherent distribution deviate from simple Gaussian to Gaussian mixture.

Our main findings from investigation into neural network dynamics reflecting Gaussian mixture structural properties are follow:

- Applying "standardization" to input datasets with inherent Gaussian mixture properties reveals convergence to the predicted dynamics outcomes of existing theories

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Goldt et al., 2020).

- Gaussian mixtures, even without standardization and significantly diverging from simple Gaussian, do not show an infinite divergence in their dynamics from those predicted in simple Gaussian settings.
- This observed non-divergence is attributed to the distinctive characteristics of nonlinear functions utilized in deep learning network and dataset modeling process, makes deep learning dynamics are predominantly influenced by the distribution's lower-order cumulants.

The subsequent section, Background, will present a concise overview of the relevant research. The Methods section will introduce the Gaussian mixture settings employed in our research and explain the methodology developed to analyze the shift in dynamics as the distribution changes from a simple Gaussian to a Gaussian mixture. In the Results section, we delineate a sequence of experiments showcasing intriguing patterns of convergence, even when the distribution markedly deviates from the typical simple Gaussian distribution. In the Discussion section, we present a mathematical proof that clarifies the observed phenomena.

2. Background

This section offers a overview of the teacher-student model framework and introduces its evolved variant, the Hidden Manifold Teacher-Student model, as proposed by Goldt et al. (2020). We explain how preceding research efforts have approached the task of understanding theoretical dynamics within the manifold model context.

2.1. Hidden Manifold Teacher-Student Model

The teacher-student model is well-regarded method in the study of high-dimensional problems (Gabrié, 2020; Baity-Jesi et al., 2018; Bahri et al., 2020; Zdeborová, 2020). This model framework consists of a teacher model, which generates dataset labels, and a student model that learns these labels. Our study specifically zeroes in on the dynamics of a fully-connected two-layer neural network.

The weights of the first and second layers of the teacher model are represented by matrices $\tilde{W} \in \mathbb{R}^{M\times D}$ and $\tilde{v} \in \mathbb{R}^{1\times M}$, respectively. We define the activation function of the teacher model as $\tilde{g}$. Similarly, the weights of the first and second layers of the student model are denoted by $W \in \mathbb{R}^{K\times N}$ and $v \in \mathbb{R}^{1\times K}$, with the activation function represented as g.

In the canonical teacher-student model, the input $X \in \mathbb{R}^N$ is typically an element-wise i.i.d. from a Gaussian distribution. However, our desired input characteristics are not inherently Gaussian but should reflect intrinsic properties as seen in datasets like the Swiss-roll, which exists on a specific manifold. To embed these intrinsic properties into the input X, they utilize a D-dimensional vector $C \in \mathbb{R}^D$ that follows an element-wise i.i.d. Gaussian distribution. This is achieved through a feature matrix $F \in \mathbb{R}^{D\times N}$ and a nonlinear function f as follows:

$$X = f(CF/\sqrt{D}) = f(U) \in \mathbb{R}^N \tag{1}$$

By modeling the dataset in this manner, the input X intrinsically follows the characteristics of C which is distributed as a Gaussian. Furthermore, the labels generated by the teacher model are derived not directly from X, but from C, which reflects intrinsic characteristics. In essence, the teacher model generates labels, and the student model learns these labels as follows:

$$y = \tilde{g}(C\tilde{W}^\top/\sqrt{D})\tilde{v}^\top, \hat{y} = g(XW^\top/\sqrt{N})v^\top \tag{2}$$

This model, recognizing a hidden structure in lower dimensions, is termed the Hidden Manifold Teacher-Student model (Goldt et al., 2020).

Additionally, for the convenience of subsequent discussions, let's denote the preactivations of the teacher and student models as ν and λ, respectively:

$$\begin{aligned}\nu &= C\tilde{W}^\top/\sqrt{D} \in \mathbb{R}^M \\ \lambda &= XW^\top/\sqrt{N} = f(U)W^\top/\sqrt{N} \in \mathbb{R}^K\end{aligned} \tag{3}$$

When input C spans beyond a singular data point to represent a dataset of size P, it assumes a matrix form, denoted as $C \in \mathbb{R}^{P\times D}$. Consequently, each preactivation is expressed through matrices $\nu \in \mathbb{R}^{P\times M}$ and $\lambda \in \mathbb{R}^{P\times K}$.
The notation $\mathsf{M}_{i,j}$ will represent the element located at the i-th row and j-th column of any given matrix M, and v_i will denote the i-th component of any vector v.
Furthermore, in the theoretical analysis to follow, we frequently consider the limit where the dataset dimensions N and intrinsic dimension D become infinitely large ($N \to \infty, D \to \infty$), a scenario often referred to as the thermodynamic limit from the perspective of statistical physics. To maintain consistency with prior theoretical studies, this research also adopts the term thermodynamic limit to describe the $N \to \infty, D \to \infty$ scenario.

2.2. Dynamics of Neural Network are dominant by correlation of function

In this study, we update weights through a simple scaled stochastic gradient descent (SGD) with a batch size of one:

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

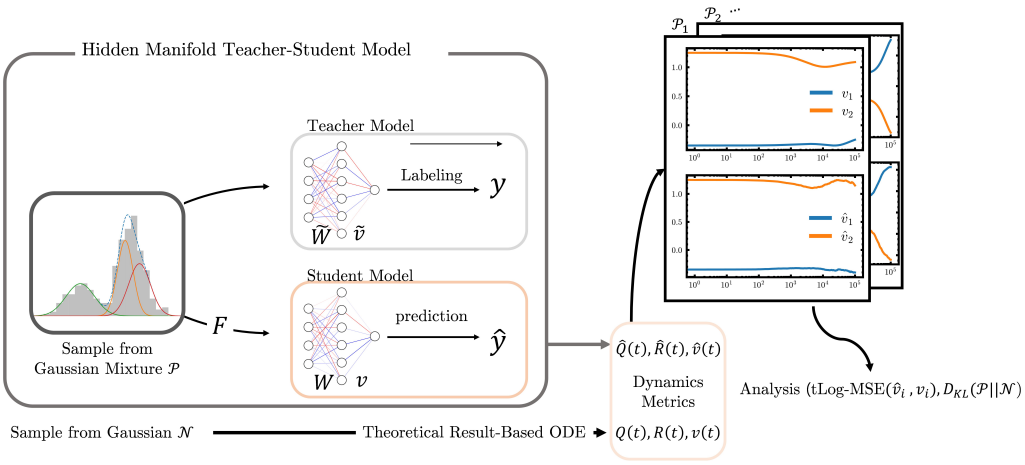

Figure 1. Schematic Representation of the Hidden Manifold Teacher-Student Model and Discrepancies in Dynamics, Gaussian to Gaussian Mixture Inputs. The right graphs showcases an example of the evolution of dynamic metrics $v(t)$ and $\hat{v}(t)$. We investigate the behavior by calculating the discrepancy measure tLog-MSE (18, 19), and examining its response to changes in $D_{KL}(P\|\mathcal{N})$.

To illustrate a straightforward example, let's examine the dynamics through the explicit ordinary differential equation (ODE) form of the second weight, v. Defining the normalized number of steps as $t = 1/N$ in the thermodynamic limit $N \to \infty$, which can be interpreted as a continuous time-like variable. Consequently, v_k satisfies the following ODE.

$$\frac{dv_k}{dt} = \eta \left[\sum_n^M \tilde{v}_n I_2(k,m) - \sum_j^K v_j I_2(k,j) \right] \tag{4}$$

where $I_2(k,m) = \mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)]$ and $I_2(k,j) = \mathbb{E}[g(\lambda_k)g(\lambda_j)]$ represent the correlations of function. Using a similar approach for v, we can derive the dynamics of our teacher-student model in the form of ODEs, as detailed in the Appendix B. The dynamics are predominantly influenced by the correlations of specific functions, like $\mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)], \mathbb{E}[g(\lambda_k)g(\lambda_j)]$. To calculate these *function correlation* values, the expectation values under the variables λ, ν, requires information on the underlying distribution of $\{\lambda, \nu\}$.

2.3. The Gaussian Equivalence Property

Previous analyses have delved into understanding the distribution of preactivations $\{\lambda, \nu\}$ under certain assumptions. To summarize the findings, $\{\lambda, \nu\}$ adhere to a Gaussian distribution characterized by a specific covariance matrix. To provide a simplified derivation, we first explore how the *function correlation* approximately follows. Suppose random variables from a joint Gaussian distribution $\{x_1, x_2\}$ are weakly correlated ($\mathbb{E}[x_1 x_2] \sim \mathcal{O}(\epsilon)$), and arbitrary functions u, v are regular enough to guarantee the existence of an expectation value, then the following lemma holds:

Lemma 2.1 (Function correlation approximation)**.**

$$\mathbb{E}[u(x_1)v(x_2)] \approx \mathbb{E}[u(x_1)]\mathbb{E}[v(x_2)] + \mathcal{O}(\epsilon) \tag{5}$$

If the feature matrix F and the student weight matrix W are sufficiently bounded, fulfilling the following assumption:

Assumption 2.2 (Bounded assumption)**.** For all $p, q \geq 1$ and any indices $k_1, \cdots, k_p, r_1, \cdots, r_p$:

$$\frac{1}{\sqrt{N}} \sum_i W_{k_1,i} \cdots W_{k_p,i} \times F_{r_1,i} \cdots F_{r_q,i} = \mathcal{O}(1) \tag{6}$$

we can approximately calculate the covariances of $\{\lambda, \nu\}$, i.e., $\mathbb{E}[\lambda\lambda], \mathbb{E}[\nu\lambda], \mathbb{E}[\nu\nu]$. By referring to the definition of $\{\lambda, \nu\}$ and employing the lemma 2.1 to decompose each *function correlation*, and then sorting out the terms that vanish in the thermodynamic limit ($N \to \infty, D \to \infty$) according to the bounded assumption (Appendix A.3). This analysis enables the derivation of the asymptotic form of all covariance matrices, where higher-order correlation vanish in the thermodynamic limit (Goldt et al., 2020). Consequently, the preactivations follow a Gaussian distribution, a result summarized as the Gaussian Equivalence Property (GEP).

Property 2.3 (Gaussian Equivalence Property (GEP))**.** *In the thermodynamic limit ($N \to \infty$, $D \to \infty$), with finite K, M, D/N, and under the assumption 2.2, if the C follows a normal Gaussian distribution $\mathcal{N}(0, I)$, then $\{\lambda, \nu\}$ conform to $K + M$ jointly Gaussian variables. This means that statistics involving $\{\lambda, \nu\}$ are entirely represented by their mean and covariance.*

Just as the dynamics of v significantly depend on the distribution characteristics of $\{\lambda, \nu\}$, the neural network's dynamics can be analyzed through these characteristics (Appendix

B). This property 2.3 allows for an understanding of the student and teacher models' dynamics, via the mean and covariance of the joint Gaussian distribution of preactivations. A comprehensive derivation and discussion are provided in the appendix A.3 and B.

For convenience, let's redefine $\bar{\lambda}_k$ as:

$$\bar{\lambda}_k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_{k,i}(f(U_i) - \mathbb{E}_{u\sim\mathcal{N}(0,1)}[f(u)]) \quad (7)$$

$\bar{\lambda}_k$ also follows a jointly Gaussian distribution, and its expectation value satisfies $\mathbb{E}[\bar{\lambda}_k] = 0$.

Consequently, the new distribution $\{\bar{\lambda}, \nu\}$ follows a more straightforward distribution with the mean

$$\mathbb{E}\left[\bar{\lambda}_k\right] = \mathbb{E}\left[\nu_m\right] = 0 \quad (8)$$

and the covariance

$$Q_{k,\ell} \equiv \mathbb{E}\left[\bar{\lambda}_k \bar{\lambda}_\ell\right] = \left(c - a^2 - b^2\right) \Omega_{k,\ell} + b^2 \Sigma_{k,\ell} \quad (9)$$

$$R_{k,m} \equiv \mathbb{E}\left[\bar{\lambda}_k \nu_m\right] = b \frac{1}{D} \sum_{r=1}^{D} S_{k,r} \tilde{W}_{m,r} \quad (10)$$

$$T_{m,n} \equiv \mathbb{E}\left[\nu_m \nu_n\right] = \frac{1}{D} \sum_{r=1}^{D} \tilde{W}_{m,r} \tilde{W}_{n,r}. \quad (11)$$

Here, a, b, and c represent the statistical properties of the nonlinear function f, used in the transformation of student model inputs X, $X = f(CF/\sqrt{D})$,

$$a = \mathbb{E}[f(u)], \quad b = \mathbb{E}[uf(u)], \quad c = \mathbb{E}[f(u)^2] \quad (12)$$

under $u \sim \mathcal{N}(0, 1)$. The newly defined matrices satisfy the following relations:

$$S_{k,r} \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_{k,i} F_{r,i} \quad (13)$$

$$\Omega_{k,\ell} \equiv \frac{1}{N} \sum_{i=1}^{N} W_{k,i} W_{\ell,i} \quad (14)$$

$$\Sigma_{k,\ell} \equiv \frac{1}{D} \sum_{r=1}^{D} S_{k,r} S_{\ell,r} \quad (15)$$

These defined covariances capture essential characteristics inherent to the teacher-student model dynamics. Given that the weights of the student model, v and W, evolve under the stochastic gradient descent (SGD) dynamics, these values, and consequently the covariances Q, R, T, inherently depend on the progression of training steps (time).

The student model learns by attempting to emulate the teacher model's outputs. Each covariance matrix holds a distinct significance in relation to the dynamics of the model. Specifically, R signifies the correlation between the preactivations of the student and teacher models, reflecting the student model's accuracy in mirroring the teacher. Conversely, Q relates to the correlation among the student model's own preactivations, shedding light on the dynamics of the student model's first layer. While T remains constant throughout the learning process and thus stands apart from Q and R, it serves as a mirror to the teacher model's inherent characteristics.

To summarize, within the context of a hidden manifold teacher-student model characterized by simple Gaussian distribution properties, the learning dynamics of the student model are primarily influenced by terms like $\mathbb{E}[g(\lambda)\tilde{g}(\nu)]$. Such terms, which depend on the distributional properties of $\{\lambda, \nu\}$, can be determined once these properties are understood. Under certain assumptions, the GEP elucidates that the preactivations follow a Gaussian distribution. Consequently, this allows us to analytically dissect the dynamics of the student model.

3. Method

In this section, we detail our approach to configuring Gaussian mixture inputs and explain the methodology developed to analyze the shift in dynamics as the distribution changes from a simple Gaussian to a Gaussian mixture. Our study's dynamic metrics of interest are Q, R, and v.

3.1. Gaussian mixture setting

In this study, instead of the simple Gaussian input ($N(0, I)$) used in previous research, we employ a more generalized Gaussian mixture distribution as the input of teacher model C. Here, we describe the Gaussian mixture setting utilized in our analysis. Consider a random variable r emerging from a D-dimensional Gaussian mixture with m mixture components. r can be represented as:

$$r = r_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \in \mathbb{R}^D \text{ with probability } p_i$$

where the sum of the probabilities $\sum^m p_i = 1$.

To gauge the divergence of our Gaussian mixture from a standard Gaussian distribution, we utilize the Kullback–Leibler (KL) divergence as the metric for distribution distance. The KL divergence quantifies the statistical discrepancy between two distributions, defined for the divergence of P from Q as:

$$D_{\mathrm{KL}}(P\|Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx \quad (16)$$

For our Gaussian mixture's components, we fix the deviation $\sigma_i = 1$ for all and assign the means μ_i by uniformly distributing them within the interval $[-\alpha, \alpha)$. By defining the Gaussian mixture model in this manner, we ascertain that increasing α leads to a monotonically increasing KL

divergence from a single Gaussian (Appendix D).
Hence, a random variable r from Gaussian mixture distribution comprising m Gaussian components is formalized as follows:

$$r = r_i \sim \mathcal{N}(\mu_i, I), \quad \mu_i \sim [-\alpha, \alpha), \text{ with } p_i \tag{17}$$

For convenience, we denote this specific Gaussian mixture distribution as $\mathcal{P}$.

This methodology allows us to undertake empirical investigations across a spectrum of Gaussian mixtures by adjusting the parameters α and m.

3.2. Analyzing Discrepancies in Dynamics: Gaussian to Gaussian Mixture Inputs

This section explains the methods to explore how neural network dynamics deviates as input (C) originally assumed Gaussian, transitions towards a Gaussian mixture. Our examination centers around metrics such as Q, R, and the second-layer weight v of the student model. While the generalization error is undoubtedly a critical feature, the behavior of error decreasing is a common characteristic across all trainable datasets and network sets, distinguishing it slightly from Q, R, and v (4, 9, 10)

The methodology initiates with the assignment of initial values to the weights $(\tilde{W}, \tilde{v}, W, v)$ and the feature matrix F. Firstly, with C drawn from a Gaussian mixture distribution $\mathcal{P}$, we obtain $\hat{Q}_{k,\ell}(t)$, $\hat{R}_{k,m}(t)$ and $\hat{v}_k(t)$ at each step of the training process via SGD.
Subsequently, using identical initial values $(\tilde{W}, \tilde{v}, W, v)$ and the feature matrix F, we obtain $Q_{k,\ell}(t)$, $R_{k,m}(t)$, and $v_k(t)$ through ODEs based on theoretical analysis (Appendix B). This is theoretically equivalent to the dynamics observed with SGD when C originates from a simple Gaussian distribution $\mathcal{N}(0, I)$.
Theoretical ODE results ($Q_{k,\ell}(t)$, $R_{k,m}(t)$, $v_k(t)$) and empirical SGD results ($\hat{Q}_{k,\ell}(t)$, $\hat{R}_{k,m}(t)$, $\hat{v}_k(t)$) are collected at each time step. Given that both sets of dynamics commence from identical starting conditions, significant discrepancies are not initially observed. To evaluate the extent of divergence in network dynamics, we employ the temporal Log-sampled Mean Squared Error (tLog-MSE) as our metric. We temporally log-sampled both ODE results and SGD results then computed the Mean Squared Error (MSE) for these sampled data points. For instance, the tLog-MSE for $v_k(t)$ and $\hat{v}_k(t)$ can be formulated as follows:

$$\text{tLog-MSE}(v_k(t), \hat{v}_k(t)) = \sqrt{\sum_{\mathbf{t}} (v_k(\mathbf{t}) - \hat{v}_k(\mathbf{t}))} \tag{18}$$

where $\mathbf{t}$ represents the log-sampled time points derived from the original t time steps. Furthermore, we examine two distinct scenarios concerning the teacher model's input C. In the unstandardized scenario, C is directly utilized from the distribution $C \sim \mathcal{P}$. Conversely, the standardized scenario implements a simple normalization method, where the random variable C is rescaled as $C := (C - \mathbb{E}[C])/\sqrt{\mathbb{E}[(C - \mathbb{E}[C])^2]}$. The expectation value is computed by empirical samples $\{C\}$. For the sake of clarity and convenience in notation, we refer to the standardized setting as $C \sim \bar{\mathcal{P}}$.

3.3. Additional Detailed Experimental Conditions

In this study, the C dimension was set to $D = 500$ and the X dimension for student input to $N = 1000$. The dimensions of the hidden layers for both teacher and student models were uniformly set to $K = M = 2$. For the activation functions, both the teacher and student models utilized the same function, $g(x) = \tilde{g}(x) = \text{erf}(x/\sqrt{2})$. The nonlinear function $f(x) = \text{sgn}(x)$ was employed to generate the student input. The learning rate was set at $\gamma = 0.2$, and training was conducted using the Mean Squared Error (MSE) loss. The SGD update used was a single batch SGD with layerwise learning rate scaling, as mentioned earlier 2.2. For training, the neural network was updated for a total of 100×1000 steps.

4. Results

In this section, we delve into the empirical exploration of how the dynamics shift as the input distribution C transitions from a simple Gaussian to a Gaussian mixture. Through our investigation, we evaluated 20 distinct Gaussian mixtures, each undergoing 100×1000 training steps. The results provided the average tLog-MSE value across these varied mixtures. The dimensions of matrices and vectors are $Q, R \in \mathbb{R}^{2\times 2}$, and $v \in \mathbb{R}^2$. For instance, since the dynamic metric Q consists of a total of four values ($Q_{k,\ell}$), the average Log-sampled Mean Squared Error (tLog-MSE) for Q is calculated as follows:

$$\text{tLog-MSE}(Q) = \mathbb{E}_{\{k,\ell\}}[\text{tLog-MSE}(Q_{k,\ell}(t), \hat{Q}_{k,\ell}(t))] \tag{19}$$

We calculate the KL divergence $D_{\text{KL}}(\mathcal{P}\|\mathcal{N})$, across different $\mathcal{P}$ that parametrized by α as mixture setting outlined earlier (17). It then assessed the influence of these $D_{\text{KL}}(\mathcal{P}|\mathcal{N})$ values on the discrepancies, tlog-MSE. Note the spectrum of KL divergence varies across different components, depending on α. In our Gaussian mixture setting, since the components have mean values constrained within $[-\alpha, \alpha)$, an increase in the number of components can lead to overlapping phenomena, thereby reducing the KL divergence. This effect can be observed in subsequent results, where different lengths along the KL divergence axis are shown.

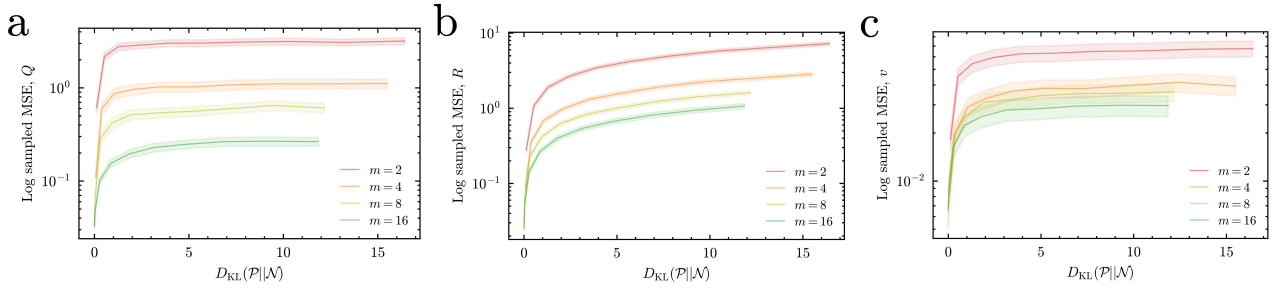

Figure 2. Log Sampled MSE under KL divergence in un-standardized Gaussian mixture setting. a, b, and c display the results corresponding to metrics Q, R, and v, respectively. The shaded areas around each curve indicate the standard error, across varying numbers of Gaussian mixture components m.

4.1. Results from un-standardized Gaussian mixture

Figure [2] intriguingly demonstrates that despite significant deviation from $\mathcal{N}(0, I)$ (increasing KL divergence), certain dynamic metrics— Q and v—appear to converge towards an asymptotic limit. Here, Q pertains to the correlation of the student's first-layer weights, and v denotes the weight of the second layer.

This phenomenon suggests that even as the intrinsic structure of the hidden manifold (C) markedly diverges from a simple Gaussian, the student model's dynamics might still adhere to bounded limits, proposing the possibility of incorporating very general Gaussian mixtures within the same theoretical framework.

4.2. Results from standardized Gaussian mixture

Subsequently, we explored the dynamics under teacher model inputs $C \sim \bar{\mathcal{P}}$, derived from the standardization of Gaussian mixture samples. The analysis of dynamics with standardized Gaussian mixtures yielded notably insightful outcomes. As illustrated in Figure [3], the tLog-MSE for various mixtures appears minimal relative to the metrics' scale, indicated by a dotted line. These mixtures do not introduce any significant discrepancies between theoretical predictions and empirical observations. To further elucidate the minimal error magnitude, an inset in the figure provides a example of the dynamic metrics' evolution, showing the almost negligible error.

The subsequent Discussion section will provide explanation of these phenomena, why mixture based experiment result conceptually converge with conventional theory.

5. Discussion

In this section, we delve into the mathematical analysis of the convergence properties of the standardized Gaussian mixture. As discussed in the Background section, the dynamics of our teacher-student model setting are influenced by the distribution followed by $\{\lambda, \nu\}$. Therefore, to analyze how network dynamics change when C is a Gaussian mixture, it is crucial to examine the distribution of $\{\lambda, \nu\}$ under such a mixture setting.

Following the proof sequence for the Gaussian Equivalence Property 2.3 where λ, ν adhere to a Gaussian distribution, we first investigate how the expectation of function correlations converges in the context of a Gaussian mixture, then derive a modified version of GEP applicable to this scenario.

5.1. Convergence in Standardized Gaussian Mixtures

5.1.1. Correlations between Functions with Gaussian mixture

Consider $I + J$ random variables represented as $x = (x_1, x_2, \cdots, x_I)^\top$ and $y = (y_1, y_2, \cdots, y_J)^\top$. Each variable x_i and y_j originates from a Gaussian mixture: $x_i = X_k$ with probability p_k where $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, and similarly, $y_j = Y_l$ with probability p_l where $Y_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$.

Despite utilizing Gaussian mixtures, each random variable can essentially be considered as following a single Gaussian distribution with a certain probability, allowing us to straightforwardly implement the existing function correlation approximation 2.1. In the Gaussian mixture setting, we derive the following lemma for function correlation approximation under mixture distribution:

Lemma 5.1 (Function correlation approximation with mixture distribution)**.** *For the $I = J = 1$ case with two Gaussian mixture variables u_1 and u_2 standardized to mean zero and variance one, and assuming weakly correlated covariance ($\mathbb{E}[u_1^2] = 1$, $\mathbb{E}[u_2^2] = 1$, $\mathbb{E}[u_1 u_2] = \epsilon m_{12}$), the approximation of function correlation for the Gaussian mixture in the limit as $\epsilon \to 0$ is given by:*

$$\mathbb{E}[f(u_1)g(u_2)] = \sum_{\{p_k, p_l\}} p_k p_l [\langle f(u_1)\rangle_k \langle g(u_2)\rangle_l] + \mathcal{O}(\epsilon) \tag{20}$$

with

$$\langle f(u_1)\rangle_k \approx \mathbb{E}_{u_1 \sim \mathcal{N}_k}[f(u_1)], \quad \langle g(u_2)\rangle_l = \mathbb{E}_{u_2 \sim \mathcal{N}_l}[g(u_2)]$$

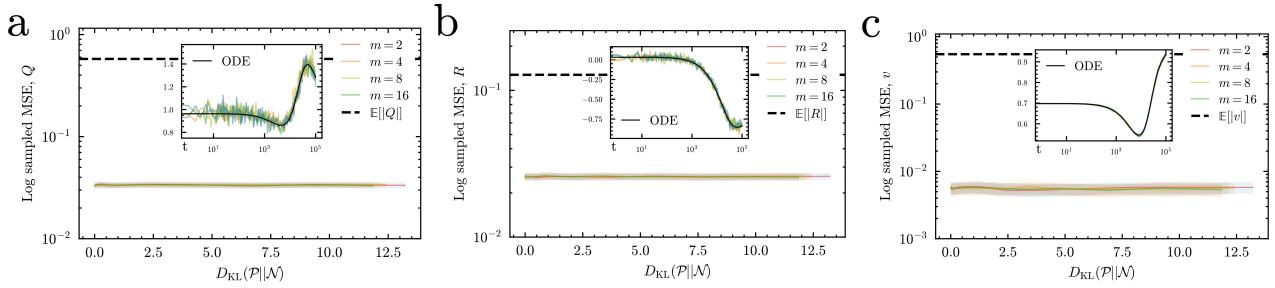

Figure 3. log sampled MSE(tLog-MSE) under KL divergence in standardized Gaussian mixture setting. a, b, and c display the results corresponding to metrics Q, R, and v, respectively, with the standard error represented by the shaded regions around each curve, across different numbers of Gaussian mixture components m. To illustrate the minimal scale of the error, a dotted line representing the mean of the each dynamic metric values $\mathbb{E}[Q]$, $\mathbb{E}[R]$, $\mathbb{E}[v]$ is added, highlighting the small error magnitude in comparison to the overall scale of the values. Additionally, a randomly sampled example of evolution of dynamic metrics, demonstrating almost negligible error, is included as an inset.

Therefore, even with Gaussian mixtures, an approximation of function correlation can achieve a similar form.

5.1.2. Dominance of Moments in the Expectation Value of Functions

In deriving the original Gaussian Equivalence Property 2.3 we utilized the approximation form of function correlation 2.1 to determine the covariance matrix of $\{\lambda, \nu\}$. Since the function correlation approximation under Gaussian mixture 5.1 has equivalent form with lemma 2.1, remaining question pertains to how the expectation values of functions under random variables $\{u_1, u_2\}$ ($\mathbb{E}[f(u_1)]$ and $\mathbb{E}[g(u_2)]$), following a Gaussian mixture distribution, differ from those assuming random variables $\{u_1, u_2\}$ following simple Gaussian distribution.

In our study, we particularly employed the sgn function as our function f within the $\lambda = f(U)W^\top/\sqrt{N}$. Dissecting the sgn function into differentiable regions reveals that higher-order derivative terms become negligible. This insight, coupled with Taylor expansion, allows us to probe the characteristics of the expectation value $\mathbb{E}[f(x)]$ for a random variable x following an arbitrary distribution.

For simplicity, we introduce the following notation:

Definition 5.2. Given an arbitrary distribution denoted by $\mathcal{P}$ and another distribution denoted by $\mathcal{D}$, if $\mathcal{P}$ shares identical cumulants with $\mathcal{D}$ up to order 2, we represent $\mathcal{P}$ as $\mathcal{P}_{\mathcal{D}_2}$, $\mathcal{P} \equiv \mathcal{P}_{\mathcal{D}_2}$.

Under this setting, for specific functions f where higher-order differential terms are insignificant, the following lemma is derived:

Lemma 5.3 (Function Expectation Approximation)**.** *Let x be a random variable with mean μ and variance σ^2 under distribution $\mathcal{P}_{\mathcal{D}2}$. Suppose f is a C^∞ function almost everywhere ($\mathbb{R}\backslash\mu$), with the condition that for $x \sim \mathcal{P}_{\mathcal{D}2}$ or $x \sim \mathcal{D}$, $\mathbb{E}[f(x)]_{x\in(\mu-\epsilon,\mu+\epsilon)}$ approaches $f(\mu)$, and $(x-\mu)^n f^{(n)}(\mu)$ for $n > 2$ is negligible in the certain limit of our interest. Then, function expectation possesses the following approximate property:*

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

The derivation hinges on the function f's property, which erase the effect of high-order related terms, leading to convergence to the same expectation value. Detailed derivation can be found in the appendix C.

This suggests that for any random variable x standardized to mean zero and variance one, the expected value of its function, $\mathbb{E}_{x\sim\mathcal{P}}[f(x)]$, can be closely approximated by $\mathbb{E}_{x\sim\mathcal{N}}[f(x)]$.

Thus, for a standardized Gaussian mixture distribution $\mathcal{P}$, since the nonlinear function $f \equiv \mathrm{sgn}$ satisfies the condition, the expectations of covariance components in the Gaussian mixture align closely with their Gaussian counterparts:

$$Q_{k,\ell} \equiv \mathbb{E}_{C\sim\mathcal{N}}\left[\bar{\lambda}_k\bar{\lambda}_\ell\right] \approx \mathbb{E}_{C\sim\mathcal{P}}\left[\bar{\lambda}_k\bar{\lambda}_\ell\right] \tag{21}$$

$$R_{k,m} \equiv \mathbb{E}_{C\sim\mathcal{N}}\left[\bar{\lambda}_k\nu_m\right] \approx \mathbb{E}_{C\sim\mathcal{P}}\left[\bar{\lambda}_k\nu_m\right] \tag{22}$$

$$T_{m,n} \equiv \mathbb{E}_{C\sim\mathcal{N}}\left[\nu_m\nu_n\right] \approx \mathbb{E}_{C\sim\mathcal{P}}\left[\nu_m\nu_n\right] \tag{23}$$

The summary of above convergence can summarized under Modified Equivalence Property:

Property 5.4 (Modified Equivalence Property)**.** *With the same condition of 2.3 except and the distribution of C is standardized Gaussian mixture $\bar{\mathcal{P}}$, take the Gaussian Equivalence Property's results $\{\lambda, \nu\}$ distribution as $\mathcal{G}$. Then the under standardized Gaussian mixture random variable C, current preactivation distribution $\{\lambda, \nu\}$ follows $\tilde{\mathcal{G}}$ such that*

$$\tilde{\mathcal{G}} \equiv \tilde{\mathcal{G}}_{\mathcal{G}_2} \tag{24}$$

It is important to note that $\tilde{\mathcal{G}}$ is not follow Gaussian distribution but only shares identical cumulants with $\mathcal{G}$ up to order 2.

Surprisingly, if the neural network's activation function g, $\tilde{g}$ also satisfy the condition for function expectation approximation 5.3, results to their correlation with the random variables λ, ν, such as $\mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)]$, used in the ODE formulation of v, also exhibits an approximated equivalence.

Fortunately, since our activation function erf is bounded $\in (-1, 1)$[1], the higher order term in Taylor expansion is less dominant then the lower order term. Thus, a loose approximation is viable:

$$\mathbb{E}_{\{\lambda,\nu\}\sim\mathcal{G}}[g(_k)\tilde{g}(\nu_n)] \approx \mathbb{E}_{\{\lambda,\nu\}\sim\tilde{\mathcal{G}}_{\mathcal{G}_2}}[g(_k)\tilde{g}(\nu_n)] \quad (25)$$

Thus, even without higher order equivalence between $\mathcal{G}$ and $\tilde{\mathcal{G}} = \tilde{\mathcal{G}}_{\mathcal{G}_2}$, the core dynamics governing the neural network exhibit approximate equivalence.
In summary, our findings articulate the following points:

1. Under a Gaussian mixture model, Lemma 2.1 transitions smoothly to Lemma 5.1, adapting to the mixture context.
2. Specific functions f that meet the criteria outlined, predominantly influence the function expectation by the first and second cumulants, 5.3.
3. This adaptation and the conditions specified lead to an equivalence in the covariance of preactivations $\{\lambda, \nu\}$, achieve equivalence up to second cumulants from a distribution perspective, 5.4.
4. The dynamics of the neural network, as dictated by function correlations involving the activation function and the random variables $\{\lambda, \nu\}$. 5.4 and 5.3 with activation function, yielding even dynamic equivalence under standardized Gaussian mixture.

Ultimately, despite the input distributions deviating from a Gaussian form, within the thermodynamic limit ($N \to \infty$, $D \to \infty$) and for specific functions that adhere to the conditions of lemma 5.3—where the expectation value is dominated by the first and second moments—the dynamics of neural networks asymptotically converge to those anticipated under simple Gaussian inputs.This convergence facilitates the incorporation of dynamics observed under Gaussian mixtures into conventional theoretical frameworks originally devised for single Gaussian inputs.

This broadens the applicability of these theories, extending their relevance to encompass scenarios not previously considered in analyses confined to simple Gaussian inputs, thereby enhancing our theoretical understanding of neural networks across a more diverse array of input distributions.

[1]If ReLU were used as the activation function, the conditions for lemma 5.3 would be more readily satisfied, and a similar discussion could be applicable.

6. Conclusion

This study advances our understanding of deep learning dynamics, by evolving and occasionally deviating from previous theoretical frameworks.

Our research focused on broadening the scope of conventional theoretical models, which predominantly assumed a simple Gaussian distribution in the teacher model input C that depict intrinsic structure of input dataset. We delved into investigating dynamics neural network when exposed to Gaussian mixtures—a more comprehensive representation of distributions. We scrutinized the discrepancies between dynamics under simple Gaussian based input and dynamics under Gaussian mixture based input with SGD update rule, yielding insightful findings.

The key takeaway from our investigation is the pivotal role of standardization. Applying standardization to Gaussian mixtures facilitates an alignment with the predictions posited by pre-existing conventional theories (Goldt et al., 2020). Furthermore, our rigorous mathematical substantiation enabled the extension of these conventional theories to accommodate a broader spectrum of distributions. In instances involving non-standardized Gaussian mixtures, we observed a remarkable asymptotic limiting behavior. This observation underscores that, notwithstanding variations in the input structure from a Gaussian baseline, the dynamics of the student model essentially adhere to conventional theoretical models.

In essence, our study unveils a newfound universality in which the inferences drawn from function correlation and the Gaussian Equivalence Property retain their validity within the domain of Gaussian mixtures. Given the generalized Gaussian mixture model assumption and its substantial expressive capacity, it stands to reason that a similar universality might be observed across other distributions. Moreover, our mathematical justification, predicated on the significance of the first and second moments (5.3, 5.4), paves the way for extending this framework to encompass a more diverse array of distributions, both traditional and beyond.

The inherent structural properties of data offer an intriguing and insightful foundation for exploration. By incorporating these characteristics into the analysis of deep learning dynamics, we aim to bridge the gap between the impressive practical successes of applied deep learning and the still-emerging theoretical research. Ultimately, we hope our research seeks to deepen the understanding of deep learning, contributing to further breakthroughs and insights in the field.

7. Reproducibility

For a comprehensive understanding of our numerical SGD implementation and ODE update mechanisms, along with ensuring reproducibility, please consult our code repository at the following link: https://anonymous.4open.science/r/_GaussianMixture-44EC/.

8. Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11:501–528, 2020.

Baity-Jesi, M., Sagun, L., Geiger, M., Spigler, S., Arous, G. B., Cammarota, C., LeCun, Y., Wyart, M., and Biroli, G. Comparing dynamics: Deep neural networks versus glassy systems. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2018.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Candès, E. J. and Sur, P. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27 – 42, 2020. doi: 10.1214/18-AOS1789. URL https://doi.org/10.1214/18-AOS1789.

Carreira-Perpinan, M. A. Mode-finding for mixtures of gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Dandi, Y., Stephan, L., Krzakala, F., Loureiro, B., and Zdeborová, L. Universality laws for gaussian mixtures in generalized linear models. *arXiv preprint arXiv:2302.08933*, 2023.

Engel, A. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Gabrié, M. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Korada, S. B. and Montanari, A. Applications of the lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Marchenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Peyré, G. Manifold models for signals and images. *Computer vision and image understanding*, 113(2):249–260, 2009.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Saad, D. and Solla, S. A. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Seddik, M. E. A., Louart, C., Tamaazousti, M., and Couillet, R. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning*, pp. 8573–8582. PMLR, 2020.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Zdeborová, L. Understanding deep learning is also a job for physicists. *Nature Physics*, 16(6):602–604, 2020.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

A. Derivation of Gaussian Equivalent Property

A.0.1. Correlation of Two Functions

It is important to consider how to express the correlation of functions, such as $\mathbb{E}[f(x)g(y)]$, for the analysis of neural network dynamics. Let's consider random variables following a $\mathcal{N}(0,1)$ distribution and examine the correlation of functions taking these random variables as inputs.

Represent two random variables, adhering to a Joint Gaussian Distribution, as vectors,

$$x = (x_1, \cdots, x_I)^\top, \quad y = (y_1, \cdots, y_J)^\top. \tag{26}$$

The assumption of joint Gaussian distribution for these random variables implies that the vectors have the following mean and covariance.

$$\mathbb{E}[x_i] = \mathbb{E}[y_j] = 0, \quad \mathbb{E}[x_i x_j] = Q_{ij}, \mathbb{E}[y_i y_j] = R_{ij}, \mathbb{E}[x_i y_j] = \epsilon S_{ij} \tag{27}$$

The joint distribution of x and y can be represented as:

$$P(x,y) = \frac{1}{Z} \exp\left[-\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} Q & \epsilon S \\ \epsilon S^\top & R \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right] \tag{28}$$

Considering a first-order approximation in ϵ, the inverse matrix part becomes,

$$\begin{pmatrix} Q & \epsilon S \\ \epsilon S^\top & R \end{pmatrix}^{-1} \approx \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}^{-1} - \epsilon \begin{pmatrix} 0 & Q^{-1}SR^{-1} \\ \left[Q^{-1}SR^{-1}\right]^\top & 0 \end{pmatrix}. \tag{29}$$

Inserting this back into the joint distribution and approximating again with respect to ϵ, we obtain following results.

$$\begin{aligned} P(x,y) = \frac{1}{Z} \exp\left[-\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} Q^{-1} & 0 \\ 0 & R^{-1} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right] \\ \times \left[1 + \varepsilon \sum_{i=1}^{I} \sum_{j=1}^{J} x_i \left(Q^{-1}SR^{-1}\right)_{ij} y_j + \mathcal{O}\left(\varepsilon^2\right) \right] \end{aligned} \tag{30}$$

To directly apply the aforementioned equation to the correlation of two functions, consider $f(x)$ and $g(y)$ as functions of x and y, respectively. Provided these functions are sufficiently regular to possess expectations $\mathbb{E}_x[x_i f(x)]$, $\mathbb{E}y[y_j g(y)]$, $\mathbb{E}x[x_i x_j f(x)]$, and $\mathbb{E}y[y_i y_j g(y)]$, the correlation between the two functions $\mathbb{E}[f(x)g(y)]$ can be expressed as:

$$\mathbb{E}[f(x)g(y)] = \mathbb{E}[f(x)]\mathbb{E}[g(y)] + \epsilon \sum_{i=1}^{I} \sum_{j=1}^{J} \mathbb{E}[x_i f(x)](Q^{-1}SR^{-1})_{ij} \mathbb{E}[y_j g(y)] + \mathcal{O}(\epsilon^2) \tag{31}$$

A.0.2. Gaussian Equivalence Property

From the function correlation approximations, it becomes clear that for functions of sufficient regularity, their correlations are primarily dictated by the function's mean, distribution characteristics such as $\mathbb{E}[uf(u)]$, and the covariance of the original random variables. This underscores the pivotal role of function correlation in dissecting the dynamics within neural networks.

In our investigation, the weight update mechanism is facilitated by employing a straightforward stochastic gradient descent (SGD) strategy, with the batch size set to one.

$$W_{k,i} := W_{k,i} - \frac{\eta}{\sqrt{N}} v_k (\hat{y} - y) g'(\lambda_k) f(U_i) \tag{32}$$

$$v_k := v_k - \frac{\eta}{N} g(\lambda_k) (\hat{y} - y) \tag{33}$$

By defining the normalized number of steps as $t = 1/N$ within the thermodynamic limit as $N \to \infty$, which analogously functions as a continuous time-like variable, we are equipped to elucidate the dynamics of the second layer weight in the

student model by examining the function correlations of the preactivations from an averaged standpoint. Consequently, the dynamics of v_k adhere to the following ODE formulation.

$$\frac{dv_k}{dt} = \eta \left[\sum_n^M \tilde{v}_n \mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)] - \sum_j^K v_j \mathbb{E}[g(\lambda_k)g(\lambda_j)] \right] \tag{34}$$

Given the crucial role of function correlation in unpacking the dynamics prompted by weight updates, it is imperative to understand the distribution characterizing λ, ν to compute expectation values such as $\mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)]$. This analytical approach enables a deeper understanding of the underlying mechanics governing the behavior of neural networks, particularly in how weight adjustments influence overall learning and adaptation processes.

Unlike the earlier discussion on simple function correlation, where the variable x of the function was assumed to be a simple Gaussian, in the context of deep learning SGD updates, the random variable entering the function is not just an assumable random variable but the preactivations.

Therefore, it's essential to ascertain the distribution of these preactivations. Let's make the following assumptions:

Assumption A.1. In the thermodynamic limit $N \to \infty$, $D \to \infty$, matrices W, $\tilde{W}$, and F possess explicit bounds:

$$\frac{1}{\sqrt{D}} \sum_{r=1}^{D} F_{r,i} F_{r,j} = \mathcal{O}(1), \quad \sum_{r=1}^{D} (F_{r,i})^2 = D \tag{35}$$

Assumption A.2. Even when considering matrices F and W together, they maintain explicit bounds. For all $p, q \geq 1$ and any indices $k_1, \cdots, k_p, r_1, \cdots, r_p$:

$$\frac{1}{\sqrt{N}} \sum_i W_{k_1,i} \times \cdots \times W_{k_p,i} \times F_{r_1,i} \times \cdots \times F_{r_q,i} = \mathcal{O}(1) \tag{36}$$

In typical deep learning scenarios, activations that address gradient vanishing or explosiveness involve gradients directly influencing weight updates in a non-vanishing limit. Thus, considering bounds for student weights during initialization is sufficient. Since the remaining teacher weights and feature matrix F are constant, ensuring proper bounds for teacher and student weights during initialization, and setting the feature matrix F to be sufficiently bounded, these assumptions can be adequately met.

With these assumptions and the result of function correlation, the Gaussian Equivalence Property holds as follows:

Property A.3 (Gaussian Equivalence Property (GEP))**.** *In the thermodynamic limit ($N \to \infty$, $D \to \infty$), with finite K, M, D/N, and under the assumption A.1 and A.2, if the C follows a normal Gaussian distribution $\mathcal{N}(0, I)$, then $\{\lambda, \nu\}$ conform to $K + M$ jointly Gaussian variables. This means that statistics involving $\{\lambda, \nu\}$ are entirely represented by their mean and covariance.*

This property allows us to representing characteristics of the student and teacher models, generalization error and dynamics of the student model's second layer weights, through the mean and covariance of the joint Gaussian distribution of preactivations.

For convenience, let's redefine $\bar{\lambda}_k$ as:

$$\bar{\lambda}_k = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_{k,i} (f(U_i) - \mathbb{E}_{u \sim \mathcal{N}(0,1)}[f(u)]) \tag{37}$$

$\bar{\lambda}_k$ also follows a jointly Gaussian distribution, and its expectation value satisfies $\mathbb{E}[\bar{\lambda}_k] = 0$ as per function correlation.

In this appendix, we present a concise derivation of $Q_{k,\ell}$. For a additional derivation, we refer the reader to prior research (Goldt et al., 2019). To facilitate the explanation, we first define a, b, and c as statistical properties of the nonlinear function f, which is utilized in transforming the student model inputs X, where $X = f(CF/\sqrt{D})$:

$$a = \mathbb{E}_{u \sim \mathcal{N}(0,1)}[f(u)], \quad b = \mathbb{E}_{u \sim \mathcal{N}(0,1)}[uf(u)], \quad c = \mathbb{E}_{u \sim \mathcal{N}(0,1)}[f(u)^2] \tag{38}$$

With these definitions in place, $Q_{k,\ell}$ can be expressed as follows:

$$Q_{k,\ell} \equiv \mathbb{E}\left[\bar{\lambda}_k \bar{\lambda}_\ell\right] \tag{39}$$

$$= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{(k,i)} W_{\ell,j}(f(U_i - a)(f(U_j) - a))] \tag{40}$$

Considering the case where $i \neq j$, and applying the expectation, we implement the function correlation approximation 31 to derive:

$$\mathbb{E}[f(U_i)f(U_j)] \approx \mathbb{E}[f(U_i)f(U_j)] + \mathbb{E}[U_i U_j]\mathbb{E}[uf(U_i)]\mathbb{E}[uf(U_j)] \tag{41}$$

$$\approx a^2 + \frac{1}{D}\sum_{r=1}^{D} F_{r,i}F_{r,j}b^2 \tag{42}$$

$$\therefore \mathbb{E}[(f(U_i) - a)(f(U_j) - a)] \approx \frac{1}{D}\sum_{r=1}^{D} F_{r,i}F_{r,j}b^2 \tag{43}$$

Hence, $Q_{k,\ell}$ can be succinctly rearranged for both $i \neq j$ and $i = j$ cases as:

$$Q_{k,\ell} = (c - a^2)\frac{1}{N}\sum_{i=j=1}^{N} W_{(k,i)}W_{\ell,j} + \frac{1}{N}\sum_{i\neq j}^{N} W_{(k,i)}W_{\ell,j}[b^2\frac{1}{D}\sum_{r=1}^{D} F_{r,i}F_{r,j}] \tag{44}$$

$$= (c - a^2 - b^2)\frac{1}{N}\sum_{i=j=1}^{N} W_{(k,i)}W_{\ell,j} + \frac{1}{N}\sum_{i,j}^{N} W_{(k,i)}W_{\ell,j}[b^2\frac{1}{D}\sum_{r=1}^{D} F_{r,i}F_{r,j}] \tag{45}$$

A similar approach can be applied to derive the remaining covariance components. Regarding high-order moments, an analogous method is employed by extending the function correlation approximation 31 to more general cases, thereby demonstrating that such preactivations follow a Gaussian distribution in the thermodynamic limit. For a comprehensive explanation of this process, the reader is encouraged to consult the referenced research (Goldt et al., 2019).

Consequently, the new distribution $\{\bar{\lambda}, \nu\}$ follows a more straightforward distribution with the mean

$$\mathbb{E}\left[\bar{\lambda}_k\right] = \mathbb{E}\left[\nu_m\right] = 0 \tag{46}$$

and the covariance

$$Q_{k,\ell} \equiv \mathbb{E}\left[\bar{\lambda}_k \bar{\lambda}_\ell\right] = \left(c - a^2 - b^2\right)\Omega_{k,\ell} + b^2\Sigma_{k,\ell} \tag{47}$$

$$R_{k,m} \equiv \mathbb{E}\left[\bar{\lambda}_k \nu_m\right] = b\frac{1}{D}\sum_{r=1}^{D} S_{k,r}\tilde{W}_{m,r} \tag{48}$$

$$T_{m,n} \equiv \mathbb{E}\left[\nu_m \nu_n\right] = \frac{1}{D}\sum_{r=1}^{D} \tilde{W}_{m,r}\tilde{W}_{n,r}. \tag{49}$$

The newly defined matrices satisfy the following relations:

$$S_{k,r} \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_{k,i}F_{r,i} \tag{50}$$

$$\Omega_{k,\ell} \equiv \frac{1}{N}\sum_{i=1}^{N} W_{k,i}W_{\ell,i} \tag{51}$$

$$\Sigma_{k,\ell} \equiv \frac{1}{D}\sum_{r=1}^{D} S_{k,r}S_{\ell,r} \tag{52}$$

B. Derivation of the ODE for Covariance and Weights

To derive the ODE for our main metrics of interest - the covariances Q, R, and the 2nd layer weight v - we begin with our single batch gradient update.

$$W_{k,i} := W_{k,i} - \frac{\eta}{\sqrt{N}} v_k (\hat{y} - y) g'(\lambda_k) f(U_i), \tag{53}$$

$$v_k := v_k - \frac{\eta}{N} g(\lambda_k)(\hat{y} - y) \tag{54}$$

The preactivations are related to the first layer weights, and thus we consider quantities such as $S_{k,r}$ and $\Sigma_{k,\ell}$ that are proportional to the first layer weights W. The dynamics of the first layer weights are determined by a term involving $(\hat{y} - y)g'(\lambda_k)f(U_i)$, assuming the second layer is constant. The average update of these quantities can be obtained from the following equation:

$$\left[\sum_{j=1}^{K} v_j g(\lambda_j) - \sum_{m=1}^{M} \tilde{v}_m \tilde{g}(v_m) \right] g'(\lambda_k) f(U_i) \tag{55}$$

Starting with $S_{k,r} \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_{k,i} F_{r,i}$, we obtain:

$$\begin{aligned} S_{k,r} := S_{k,r} - \frac{\eta}{\sqrt{N}} v^k \Bigg[& \sum_{j \neq k}^{K} v_j \mathbb{E}[g(\lambda_j) g'(\lambda_k) \beta_r] + v_k \mathbb{E}[g(\lambda_k) g'(\lambda_k) \beta_r] \\ & - \sum_{n}^{M} \tilde{v}_n \mathbb{E}[\tilde{g}(\nu_n) g'(\lambda_k) \beta_r] \Bigg] \end{aligned} \tag{56}$$

with $\beta_r = \frac{1}{\sqrt{N}} \sum_i F_{r,i} f(U_i)$.

Function correlations are employed to express these updates in terms of statistical quantities of the distributions λ, ν. However, the equations for covariances remain coupled. To uncouple them, we need to consider the eigenvectors and eigenvalues, ψ_τ and ρ_τ, of the $D \times D$ matrix $\mathcal{F}$ formed by $\mathcal{F}_{r,s} = 1/N \sum_i F_{r,i} F_{s,i}$. The eigenvectors and eigenvalues are obtained under the following normalization condition:

$$\sum_s \mathcal{F}_{r,s} (\psi_\tau)_s = \rho_\tau (\psi_\tau)_r, \quad \sum_s (\psi_\tau)_s (\psi_{\tau'})_s = D\delta(\tau, \tau'), \quad \sum_\tau (\psi_\tau)_r (\psi_\tau)_s = D\delta(r, s) \tag{57}$$

Using these, we can express the teacher-student overlap covariance $R_{k,m}$ through two projected matrices:

$$\mathcal{S}_{k,r} = \frac{1}{\sqrt{D}} \sum_r S_{k,r} (\psi_\tau)_r, \quad \mathcal{W}_{m,\tau} = \frac{1}{\sqrt{D}} \sum_r \tilde{W}_{m,r} (\psi_\tau)_r \tag{58}$$

and thus:

$$R_{k,m} = \frac{b}{D} \sum_\tau \mathcal{S}_{k,r} \mathcal{W}_{m,\tau} \tag{59}$$

Since the teacher model's matrix is static, its projection matrix $\mathcal{S}$ is given by:

$$\begin{aligned} \mathcal{S}_{k,\tau} := \mathcal{S}_{k,\tau} - \frac{\eta}{\sqrt{DN}} v^k \sum_r (\psi_\tau)_r \Bigg[& \sum_{j \neq k}^{K} v_j \mathbb{E}[g(\lambda_j) g'(\lambda_k) \beta_r] + v_k \mathbb{E}[g(\lambda_k) g'(\lambda_k) \beta_r] \\ & - \sum_{n}^{M} \tilde{v}_n \mathbb{E}[\tilde{g}(\nu_n) g'(\lambda_k) \beta_r] \Bigg] \end{aligned} \tag{60}$$

The update rule for R is then derived using these projections. Explicitly at timestep t, it can be expressed as:

$$(R_{k,m})_{t+1} - (R_{k,m})_t = \frac{b}{D} \sum_\tau \left[(\mathcal{S}_{k,\tau})_{t+1} - (\mathcal{S}_{k,\tau})_t \right] \tilde{W}_{m,r} \tag{61}$$

During the summation over τ, two types of terms emerge:

$$\mathcal{T}_{m,n} \equiv \frac{1}{D}\sum_{\tau} \rho_\tau \tilde{W}_{m,r}\tilde{W}_{n,r}, \quad \frac{1}{D}\sum_{\tau} \rho_\tau \mathcal{S}_{\ell,\tau}\tilde{W}_{n,\tau} \tag{62}$$

The second summation is not readily reducible to a simpler expression. Instead, we introduce the following density function:

$$r_{k,m}(\rho) = \frac{1}{\epsilon_\rho}\frac{1}{D}\sum_{\tau} \tilde{S}_{k,\tau}\tilde{W}_{m,\tau}\mathbf{1}_{\rho_\tau \in [\rho,\rho+\epsilon_\rho]} \tag{63}$$

This density function allows us to express the covariance R in terms of the eigenvalue distribution ρ:

$$R_{k,m} = b\int d\rho p(\rho) r_{k,m}(\rho) \tag{64}$$

Under the assumption that the feature matrix elements are i.i.d. from a normal distribution $\mathcal{N}(0,1)$, this distribution adheres to the Marchenko-Pastur law (Marchenko & Pastur, 1967):

$$p(\rho) = \frac{1}{2\pi D/N}\frac{\sqrt{((1+\sqrt{D/N})^2-\rho)(\rho-(1-\sqrt{D/N})^2)}}{\rho} \tag{65}$$

The update equation for $r_{k,m}(\rho)$ is straightforwardly derived from the update equation and definition of $\mathcal{S}$. Ultimately, in the thermodynamic limit, with $t = 1/N$ transforming into a continuous time-like variable, the equation of motion for $r_{k,m}(\rho,t)$ satisfies the following ODE:

$$\begin{aligned}
\frac{\partial r_{k,m}(\rho,t)}{\partial t} = -\frac{\eta}{D/N} v_k d(\rho)\Bigg(& r_{km}(\rho)\sum_{j\neq k}^{K} v_j \frac{Q_{jj}\mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_j)] - Q_{kj}\mathbb{E}[g'(\lambda_k)\lambda_j g(\lambda_j)]}{Q_{jj}Q_{kk}-(Q_{kj})^2} \\
& + \sum_{j\neq k}^{K} v_j r_{jm}(\rho)\frac{Q_{kk}\mathbb{E}[g'(\lambda_k)\lambda_j g(\lambda_j)] - Q_{kj}\mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_j)]}{Q_{jj}Q_{kk}-(Q_{kj})^2} \\
& + v_k r_{km}(\rho)\frac{1}{Q_{kk}}\mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_k)] - \\
& r_{km}(\rho)\sum_{n}^{M}\tilde{v}_n \frac{T_{nn}\mathbb{E}[g'(\lambda_k)\lambda_k\tilde{g}(\nu_n)] - R_{kn}\mathbb{E}[g'(\lambda_k)\nu_n\tilde{g}(\nu_n)]}{Q_{kk}T_{nn}-(R_{kn})^2} \\
& - \frac{b\rho}{d(\rho)}\sum_{n}^{M}\tilde{v}_n\mathcal{T}_{nm}\frac{Q_{kk}\mathbb{E}[g'(\lambda_k)\nu_n\tilde{g}(\nu_n)] - R_{kn}\mathbb{E}[g'(\lambda_k)\lambda_k\tilde{g}(\nu_n)]}{Q_{kk}T_{nn}-(R_{kn})^2}\Bigg)
\end{aligned} \tag{66}$$

where $d(\rho) = (c-b^2)\dfrac{D}{N} + b^2\rho$. Note that all explicit time dependencies on the right side of the equation are omitted for clarity. In this numerical ODE implementation, the right side corresponds to the immediate preceding time t, and the left side to the updated time $t+1$.

Similarly, the covariance Q associated with the first weight W can be derived in a repetitive manner, starting from:

$$Q_{k,\ell} \equiv \mathbb{E}[\lambda_k\lambda_\ell] = [c-b^2]W_{k,\ell} + b^2\Sigma_{k,\ell} \tag{67}$$

Following a similar process as before, we find that the first term, $W_{k,l}$, adheres to:

$$\begin{aligned}\frac{\mathrm{d}W_{k,\ell}(t)}{\mathrm{d}t} =& -\eta v_k \left(\sum_j^K v_j \mathbb{E}[g'(\lambda_k)\lambda_\ell g(\lambda_j)] - \sum_n \tilde{v}_n \mathbb{E}[g'(\lambda_k)\lambda_\ell \tilde{g}(\nu_n)] \right) \\ & -\eta v_\ell \left(\sum_j^K v_j \mathbb{E}[g'(\lambda_\ell)\lambda_k g(\lambda_j)] - \sum_n \tilde{v}_n \mathbb{E}[g'(\lambda_\ell)\lambda_k \tilde{g}(\nu_n)] \right) \\ & + c\eta^2 v_k v_\ell \Bigg(\sum_{j,\iota}^K v_j v_\iota \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) g(\lambda_j) g(\lambda_\iota)] \\ & - 2\sum_j^K \sum_m^M v_j \tilde{v}_m \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) g(\lambda_j) \tilde{g}(\nu_m)] \\ & + \sum_{n,m}^M \tilde{v}_n \tilde{v}_m \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) \tilde{g}(\nu_n) \tilde{g}(\nu_m)] \Bigg)\end{aligned} \tag{68}$$

The second term, $\Sigma_{k,\ell}$, can be expressed using the rotating basis ψ_τ:

$$\Sigma_{k,\ell} \equiv \frac{1}{D} \sum_r S_{k,r} S_{\ell,r} = \frac{1}{D} \sum_\tau \mathcal{S}_{k,\tau} \mathcal{S}_{\ell,\tau} \tag{69}$$

and thus, integral form for $\Sigma_{k,\ell}(t)$ can be derived:

$$\sigma_{k,\ell}(\rho) = \frac{1}{\epsilon_\rho} \frac{1}{D} \sum_\tau \mathcal{S}_{k,\tau} \mathcal{S}_{\ell,\tau} \mathbf{1}_{\rho_\tau \in [\rho, \rho+\epsilon_\rho]} \tag{70}$$

with

$$\begin{aligned}\frac{\partial \sigma_{k\ell}(\rho,t)}{\partial t} =& -\frac{\eta}{D/N} \Bigg(d(\rho) v_k \sigma_{k\ell}(\rho) \sum_{j\neq k} v_j \frac{Q_{jj}\mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_j)] - Q_{kj}\mathbb{E}[g'(\lambda_k)\lambda_j g(\lambda_j)]}{Q_{jj}Q_{kk} - (Q_{kj})^2} \\ & + v_k \sum_{j\neq k} v_j d(\rho) \sigma_{j\ell}(\rho) \frac{Q_{kk}\mathbb{E}[g'(\lambda_k)\lambda_j g(\lambda_j)] - Q_{kj}\mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_j)]}{Q_{jj}Q_{kk} - (Q_{kj})^2} \\ & + d(\rho) v_k \sigma_{k\ell}(\rho) v_k \frac{1}{Q_{kk}} \mathbb{E}[g'(\lambda_k)\lambda_k g(\lambda_k)] \\ & - d(\rho) v_k \sigma_{k\ell}(\rho) \sum_n \tilde{v}_n \frac{T_{nn}\mathbb{E}[g'(\lambda_k)\lambda_k \tilde{g}(\nu_n)] - R_{kn}\mathbb{E}[g'(\lambda_k)\nu_n \tilde{g}(\nu_n)]}{Q_{kk}T_{nn} - (R_{kn})^2} \\ & - b\rho v_k \sum_n \tilde{v}_n r_{\ell n}(\rho) \frac{Q_{kk}\mathbb{E}[g'(\lambda_k)\nu_n \tilde{g}(\nu_n)] - R_{kn}\mathbb{E}[g'(\lambda_k)\lambda_k \tilde{g}(\nu_n)]}{Q_{kk}T_{nn} - (R_{kn})^2} \\ & + \text{ all of the above with } \ell \to k, k \to \ell). \\ & + \eta^2 v_k v_\ell \left[(c - b^2)\rho + \frac{b^2}{\delta}\rho^2 \right] \left(\sum_{j,\iota}^K v_j v_\iota \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) g(\lambda_j) g(\lambda_\iota)] \right. \\ & \left. -2\sum_j^K \sum_m^M v_j \tilde{v}_m \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) g(\lambda_j) \tilde{g}(\nu_m)] + \sum_{n,m}^M \tilde{v}_n \tilde{v}_m \mathbb{E}[g'(\lambda_k) g'(\lambda_\ell) \tilde{g}(\nu_n) \tilde{g}(\nu_m)] \right)\end{aligned} \tag{71}$$

The weight v and generalization error ϵ_g can be directly obtained from the weight update formula and the definition of generalization error with MSE:

$$\frac{dv_k}{dt} = \eta \left[\sum_n^M \tilde{v}_n \mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)] - \sum_j^K v_j \mathbb{E}[g(\lambda_k) g(\lambda_j)] \right] \tag{72}$$

with

$$\begin{aligned}\epsilon_g(\theta,\widetilde{\theta}) &= \frac{1}{2}\mathbb{E}\left[\left(\sum_k^K v_k g\left(\lambda_k\right) - \sum_m^M \tilde{v}_m \tilde{g}\left(\nu_m\right)\right)^2\right] \\ &= \frac{1}{2}\sum_{k,\ell} v_k v_\ell \mathbb{E}[g(\lambda_k)g(\lambda_\ell)] + \frac{1}{2}\sum_{n,m}\tilde{v}^n\tilde{v}^m\mathbb{E}[\tilde{g}(\nu_n)\tilde{g}(\nu_m)] - \sum_{k,n} v_k\tilde{v}_n\mathbb{E}[g(\lambda_k)\tilde{g}(\nu_n)]\end{aligned} \tag{73}$$

C. Derivation of Function Expectation Approximation Lemma

Lemma's statement is as following:

Lemma C.1. *Expectation Approximation:Let x be a random variable with mean μ and variance σ^2 under distribution $\mathcal{P}_{\mathcal{D}2}$. Suppose f is a C^∞ function almost everywhere ($\mathbb{R}\backslash\mu$), with the condition that for $x \sim \mathcal{P}_{\mathcal{D}2}$ or $x \sim \mathcal{D}$, $\mathbb{E}[f(x)]_{x\in(\mu-\epsilon,\mu+\epsilon)}$ approaches $f(\mu)$, and $(x-\mu)^n f^{(n)}(\mu)$ for $n > 2$ is negligible in the limit of our interest. Then, function expectation possesses the following approximate property:*

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

To derive above lemma, First, to use Taylor expansion, we need to separate the interval. And since the condition of $\mathbb{E}[f(x)]_{x\in(\mu-\epsilon,\mu+\epsilon)}$ approaches $f(\mu)$ yield follwing results.

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{P}_{\mathcal{D}2}}[f(x)] &= \int_{-\infty}^{\infty} f(x)P(x)dx \\ &= \int_{-\infty}^{\mu-\epsilon} f(x)P(x)dx + \int_{\mu+\epsilon}^{\infty} f(x)P(x)dx + \int_{\mu-\epsilon}^{\mu+\epsilon} f(x)P(x)dx \\ &= \mathbb{E}[f(x)]_{x\in(-\infty,-\epsilon)} + \mathbb{E}[f(x)]_{x\in(\epsilon,\infty)} + f(\mu)\end{aligned}$$

Let's take $\mathbb{E}[f(x)]_{x\in(-\infty,-\epsilon)}$ part. Since the lemma condition making vanishing of high order derivation, we can directly found expectation of the function approximately converges to the expectation under $\mathcal{D}$ distribution.

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{P}_{\mathcal{D}2}}[f(x)]_{x\in(-\infty,-\epsilon)} &= \mathbb{E}[f(\mu) + \cdots + f^{(n)}(\mu)\frac{(x-\mu)^n}{n!} + \cdots]_{x\in(-\infty,-\epsilon)} \\ &\approx \mathbb{E}[f(\mu) + f''(\mu)\frac{(x-\mu)^2}{2!}] \\ &\approx f(\mu) + f''(\mu)\mathbb{E}[\frac{(x-\mu)^2}{2!}] \\ &\approx \mathbb{E}_{x\sim\mathcal{D}}[f(x)]_{x\in(-\infty,-\epsilon)}\end{aligned}$$

Applying a similar approach to other terms leads to the general result that the expectation value of a function over a random variable x from a distribution $\mathcal{P}_{\mathcal{D}_2}$ can be approximated by the expectation value of the same function over a random variable from a standard Gaussian distribution $\mathcal{D}$:

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

Functions like erf, ReLU, and sgn approximately satisfy the aforementioned condition, allowing for the following equivalency in expectation values. This equivalency can be empirically verified starting from the Central Limit Theorem, which suggests that the sample average of n independently sampled random variables X_i, each with expectation μ and a finite variance σ^2, approaches a normal distribution with mean μ and variance σ^2/n as n increases.

D. KL divergence under Gaussian Mixture setting

Experimentally, by adjusting the distribution bound α of the mixture component means μ_i, we observed a regular correlation in the KL divergence between the mixture distribution and the normal Gaussian distribution [Figure 5].

This approach enabled us to conduct empirical analyses on various Gaussian mixtures by adjusting the values of α and m.

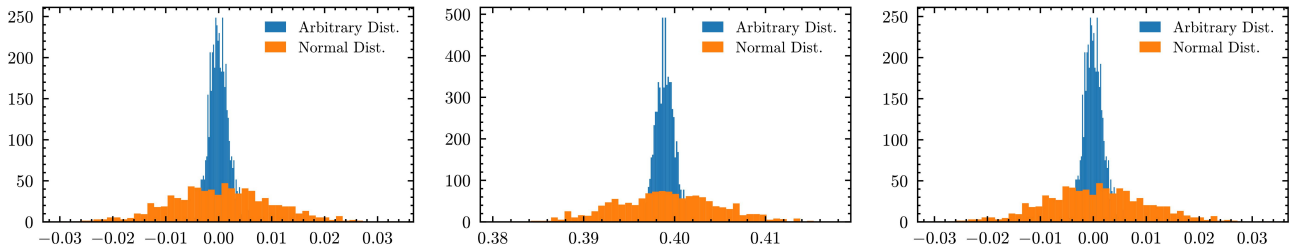

Figure 4. From left to right, the figure compares expectation values under the erf, ReLU, and sgn functions. When a specific random variable X_i follows a distribution with an expectation of μ and a finite variance of σ^2, sampling from this distribution and calculating the sample average $n\hat{X}_n = X_1 + \cdots + X_n$ demonstrates convergence towards a normal distribution with mean μ and variance σ^2/n, as per the Central Limit Theorem. This provides indirect evidence that the function expectation retains the same expectation value across the distributions.

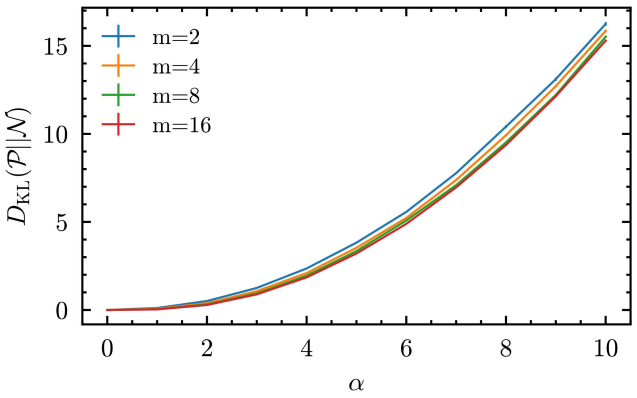

Figure 5. Numerical results depicting the relationship between KL divergence and the parameter α for different numbers of Gaussian mixture components, m.

E. Additional Detailed Experimental Conditions

In this study, the manifold dimension was set to $D = 500$ and the real-world dimension for student input to $N = 1000$. The dimensions of the hidden layers for both teacher and student models were uniformly set to $K = M = 2$. For the activation functions, both the teacher and student models utilized the same function, $g(x) = \tilde{g}(x) = \mathrm{erf}(x/\sqrt{2})$. The nonlinear function $f(x) = \mathrm{sgn}(x)$ was employed to generate the student input.

The learning rate was set at $\gamma = 0.2$, and training was conducted using the Mean Squared Error (MSE) loss. The SGD update used was a single batch SGD with layerwise learning rate scaling, as mentioned earlier. For training, the neural network was updated for a total of 100×1000 steps. For more detailed information on the numerical SGD implementation and ODE update implementation, please refer to the following code repository: `https://anonymous.4open.science/r/_GaussianMixture-44EC/`. Our research was carried out on a computing setup equipped with an AMD Ryzen 7 7700X CPU and an NVIDIA GeForce RTX 3060 12GB GPU.

F. Numerical Asymptotic Behavior in Non-standardized Gaussian mixtures

In scenarios where standardization is not applied, our results show a distinct deviation from the expected asymptotic behavior. However, intriguingly, there seems to exist an upper limit, indicating that distributions considerably divergent from Gaussian can still, to some extent, display Gaussian-like characteristics.

Additionally, we investigated the numerical asymptotic behavior of non-standardized Gaussian mixtures using regression analysis. Without standardization, when the number of components in the Gaussian mixture increases within our mixture framework, the mean of the Gaussian mixture approaches zero. This tendency elucidates the observed reduction in error as

the number of components becomes large.

However, the observed limiting behavior in the context of large deviation from the Gaussian mixture (or high KL divergence) is particularly noteworthy. Employing a rough approximation, we found that the asymptotic error can be described by the following relation:

For KL divergence x, and number of Gaussian mixture components $m \geq 2$, in existence of bounded behavior, Q, v.

$$\text{MSE}(x; m) = \frac{f(m)x}{\sqrt{1+x^2}} + g(m), \quad f(m) = am^{-b}, \quad g(m) = cm^{-d} \tag{74}$$

For KL divergence x, and number of Gaussian mixture components $m \geq 2$, in absence of bounded behavior, R.

$$\text{MSE}(x; m) = \frac{f(m)x}{\sqrt{1+x^{2/m}}} + g(m), \quad f(m) = am^{-b}, \quad g(m) = cm^{-d} \tag{75}$$

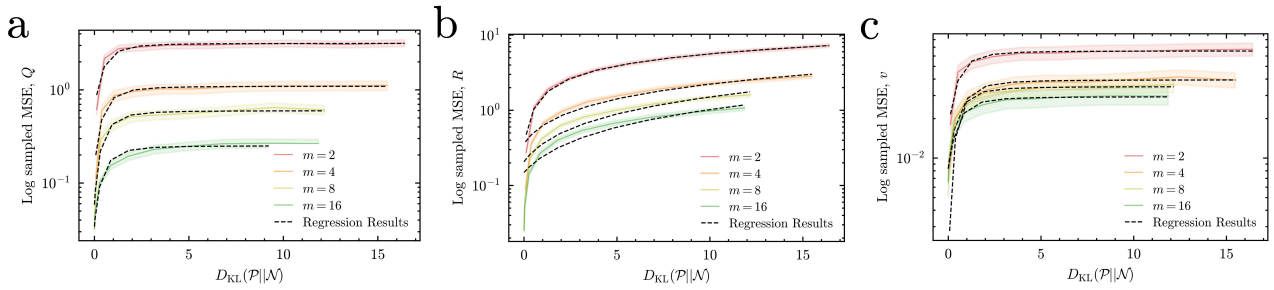

Figure 6. Log Sampled MSE under KL divergence in un-standardized Gaussian mixture setting with numerical regression results. The dashed lines indicate the regression results, offering a predictive view of the MSE's dependence on the KL divergence.

This approximation suggests a functional dependence of the error on both the KL divergence and the number of components in the mixture, providing valuable insights into the dynamics of un-standardized Gaussian mixtures.