Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks

Tengyuan Liang & Hai Tran-Bach

To cite this article: Tengyuan Liang & Hai Tran-Bach (2021): Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1853547

To link to this article: https://doi.org/10.1080/01621459.2020.1853547

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2020.1853547

Mehler's Formula, Branching Process, and Compositional Kernels of Deep Neural Networks

Tengyuan Liang[a] and Hai Tran-Bach[b]

[a]Booth School of Business, University of Chicago, Chicago, IL; [b]Department of Statistics, University of Chicago, Chicago, IL

ABSTRACT

We use a connection between compositional kernels and branching processes via Mehler's formula to study deep neural networks. This new probabilistic insight provides us a novel perspective on the mathematical role of activation functions in compositional neural networks. We study the unscaled and rescaled limits of the compositional kernels and explore the different phases of the limiting behavior, as the compositional depth increases. We investigate the memorization capacity of the compositional kernels and neural networks by characterizing the interplay among compositional depth, sample size, dimensionality, and nonlinearity of the activation. Explicit formulas on the eigenvalues of the compositional kernel are provided, which quantify the complexity of the corresponding reproducing kernel Hilbert space. On the methodological front, we propose a new random features algorithm, which compresses the compositional layers by devising a new activation function. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2020
Accepted November 2020

KEYWORDS
Compositional kernels; Deep neural networks; Galton–Watson process; Random features regression; Random weights initialization

1. Introduction

Kernel methods and deep neural networks (DNNs) are arguably two representative methods that achieved the state-of-the-art results in regression and classification tasks (Shankar et al. 2020). However, unlike the kernel methods where both the statistical and computational aspects of learning have been understood reasonably well, there are still many theoretical puzzles around the generalization, computation, and representation aspects of DNNs (Zhang et al. 2017). One hopeful direction to resolve some of the puzzles in neural networks is through the lens of kernels (Rahimi and Recht 2008, 2009; Cho and Saul 2009; Belkin, Ma, and Mandal 2018). Such a connection can be readily observed in a two-layer infinite-width network with random weights, see the pioneering work by Neal (1996a) and Rahimi and Recht (2008, 2009). For deep networks with hierarchical structures and randomly initialized weights, compositional kernels (Poole et al. 2016; Daniely, Frostig, and Singer 2017) are proposed to rigorously characterize such a connection, with promising empirical performances (Cho and Saul 2009). A list of simple algebraic operations on kernels (Stitson et al. 1999; Shankar et al. 2020) are introduced to incorporate specific data structures that contain bag-of-features, such as images and time series.

In this article, we continue to study DNNs and their dual compositional kernels, furthering the aforementioned mathematical connection, based on the foundational work of Rahimi and Recht (2008, 2009), Daniely, Frostig, and Singer (2017), and Daniely et al. (2017). We focus on a standard multilayer perceptron (MLP) architecture with Gaussian weights and study the role of the activation function and its effect on composition, data memorization, spectral properties, algorithms, among others. Our main results are based on a simple yet elegant connection between *compositional kernels* and *branching processes* via Mehler's formula (Lemma 3.1). This new connection, in turn, opens up the possibility of studying the mathematical role of activation functions in compositional DNNs, using the probabilistic tools in branching processes (Theorem 3.1). Specifically, the new probabilistic insight allows us to answer the following questions:

Limits and phase transitions. Given an activation function, one can define the corresponding compositional kernel (Daniely, Frostig, and Singer 2017; Daniely et al. 2017). How to classify the activation functions according to the limits of their dual compositional kernels, as the compositional depth increases? What properties of the activation functions govern the different phases of such limits? How do we properly rescale the compositional kernel such that there is a limit unique to the activation function? The above questions will be explored in Section 3.

Memorization capacity of compositions: tradeoffs. DNNs and kernel machines can have a good out-of-sample performance even in the interpolation regime (Zhang et al. 2017; Belkin, Ma, and Mandal 2018), with perfect memorization of the training dataset. What is the memorization capacity of the compositional kernels? What are the tradeoffs among compositional depth, number of samples in the dataset, input dimensionality, and properties of the nonlinear activation functions? Section 4 studies such interplay explicitly.

CONTACT Tengyuan Liang *tengyuan.liang@chicagobooth.edu* Booth School of Business, University of Chicago, Chicago, IL 60637.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Spectral properties of compositional kernels. Spectral properties of the kernel (and the corresponding integral operator) affect the statistical rate of convergence, for kernel regressions (Caponnetto and Vito 2007). What is the spectral decomposition of the compositional kernels? How do the eigenvalues of the compositional kernel depend on the activation function? Section 5 is devoted to answering the above questions.

New randomized algorithms. Given a compositional kernel with a finite depth associate with an activation, can we devise a new "compressed" activation and new randomized algorithms, such that the DNN (with random weights) with the original activation is equivalent to a shallow neural network with the "compressed" activation? Such algorithmic questions are closely related to the seminal random Fourier features (RFF) algorithm in Rahimi and Recht (2008, 2009), yet different. Section 6 investigates such algorithmic questions by considering compositional kernels and, more broadly, the inner-product kernels. Differences to the RFF are also discussed in detail therein.

Borrowing the insight from branching process, we start with studying the role of activation function in the compositional kernel, memorization capacity, and spectral properties, and conclude with the converse question of designing new activations and random nonlinear features algorithm based on kernels, thus contributing to a strengthened mathematical understanding of activation functions, compositional kernel classes, and DNNs.

1.1. Related Work

The connections between neural networks (with random weights) and kernel methods have been formalized by researchers using different mathematical languages. Instead of aiming to provide a complete list, here we only highlight a few that directly motivate our work. Neal (1996a, 1996b) advocated using Gaussian processes to characterize the neural networks with random weights from a Bayesian viewpoint. For two-layer neural networks, such correspondence has been strengthened mathematically by the work of Rahimi and Recht (2008, 2009). By Bochner's theorem, Rahimi and Recht (2008) showed that any positive definite (PD) translation-invariant kernel could be realized by a two-layer neural network with a specific distribution on the weights, via trigonometric activations. Such insights also motivated the well-known random features algorithm, random kitchen sinks (Rahimi and Recht 2009). One highlight of such an algorithm is that in the first layer of weights, sampling is employed to replace the optimization. Later, several works extended along the line, see, for instance, Kar and Karnick (2012) on the rotation-invariant kernels, Pennington, Yu, and Kumar (2015) on the polynomial kernels, and Bach (2016) on kernels associated to ReLU-like activations (using spherical harmonics). Recently, Mei and Montanari (2019) investigated the precise asymptotics of the random features model using random matrix theory. For DNNs, compositional kernels are proposed to carry such connections further. Cho and Saul (2009) introduced the compositional kernel as the inner-product of compositional features. Daniely, Frostig, and Singer (2017) and Daniely et al. (2017) described the compositional kernel through the language of the computational skeleton, and introduced the duality between the activation function and compositional kernel. We refer the readers to Poole et al. (2016), Yang (2019), and Shankar et al. (2020) for more information on the connection between kernels and neural networks.

One might argue that neural networks with static random weights may not fully explain the success of neural networks, noticing that the evolution of the weights during training is yet another critical component. On this front, Chizat and Bach (2018b), Mei, Montanari, and Nguyen (2018), Sirignano and Spiliopoulos (2018), and Rotskoff and Vanden-Eijnden (2018) employed the mean-field characterization to describe the distribution dynamics of the weights, for two-layer networks. Rotskoff and Vanden-Eijnden (2018) and Dou and Liang (2020) studied the favorable properties of the dynamic kernel due to the evolution of the weight distribution. Nguyen and Pham (2020) carried the mean-field analysis to multilayer networks rigorously. On a different tread (Chizat and Bach 2018a; Du et al. 2018; Jacot, Gabriel, and Hongler 2019; Woodworth et al. 2019), researchers showed that under specific scaling, training over-parameterized networks could be viewed as a kernel regression with perfect memorization of the training data, using a tangent kernel (Jacot, Gabriel, and Hongler 2019) built from a linearization around its initialization. For a more recent resemblance between the kernel learning and the deep learning on the empirical side, we refer the readers to Belkin, Ma, and Mandal (2018).

2. Preliminary

2.1. Mehler's Formula

We will start with reviewing some essential background on the Hermite polynomials that is of direct relevance to our article.

Definition 2.1 (Hermite polynomials). The probabilists' Hermite polynomials $He_k(x)$ for nonnegative integers $k \in \mathbb{Z}^{\geq 0}$ follows the recursive definition with $He_0(x) = 1$ and

$$He_{k+1}(x) = xHe_k(x) - He_k'(x). \tag{2.1}$$

We define the normalized Hermite polynomials as

$$h_k(x) := \frac{1}{\sqrt{k!}} He_k(x), \quad \text{with } \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,1)}[h_k^2(\mathbf{g})] = 1. \tag{2.2}$$

The set $\{h_k : k \in \mathbb{Z}^{\geq 0}\}$ forms an orthogonal basis of L^2_ϕ under the Gaussian measure $\phi \sim \mathcal{N}(0,1)$ as

$$\mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,1)}[h_k(\mathbf{g})h_{k'}(\mathbf{g})] = \mathbb{1}_{k=k'}. \tag{2.3}$$

Proposition 2.1 (Mehler's formula). Mehler's formula establishes the following equality on Hermite polynomials: for any $\rho \in (-1,1)$ and $x, y \in \mathbb{R}$

$$\frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho^2(x^2+y^2) - 2\rho xy}{2(1-\rho^2)}\right) = \sum_{k=0}^{\infty} \rho^k h_k(x) h_k(y). \tag{2.4}$$

2.2. Branching Process

Now, we will describe the branching process and the compositions of probability generating functions (PGF).

Definition 2.2 (Probability generating function). Given a random variable Y on nonnegative integers with the following probability distribution

$$\mathbb{P}(Y = k) = p_k, \quad \forall k \in \mathbb{Z}^{\geq 0}, \tag{2.5}$$

define the associated generating function as

$$G_Y(s) := \mathbb{E}[s^Y] = \sum_{k\geq 0} p_k s^k. \tag{2.6}$$

It is clear that $G_Y(0) = 0$, $G_Y(1) = 1$ and $G_Y(s)$ is nondecreasing and convex on $s \in [0, 1]$.

Definition 2.3 (Galton–Watson branching process). The Galton–Watson (GW) branching process is defined as a Markov chain $\{Z_L : L \in \mathbb{Z}^{\geq 0}\}$, where Z_L denotes the size of the Lth generation of the initial family. Let Y be a random variable on nonnegative integers describing the number of direct children, that is, it has k children with probability p_k with $\sum_{k\geq 0} p_k = 1$. Begin with one individual $Z_0 \equiv 1$, and let it reproduce according to the distribution of Y, and then each of these children then reproduce independently with the same distribution as Y. The generation sizes $\{Z_L : L \in \mathbb{Z}^{\geq 0}\}$ are then defined by

$$Z_{L+1} = \sum_{i=1}^{Z_L} Y_i^{(L)}, \tag{2.7}$$

where $Y_i^{(L)}$ denotes the number of children for the ith individual in generation L.

Proposition 2.2 (Extinction criterion, Lyons and Peres 2016, Proposition 5.4). Given $p_1 \neq 1$ (in Definition 2.2), the extinction probability $\xi := \lim_{L\to\infty} \mathbf{P}(Z_L = 0)$ satisfies

(i) $\xi = 1$ if and only if $\mu := G_Y'(1) \leq 1$.
(ii) ξ is the unique fixed point of $G_Y(s) = s$ in $[0, 1)$.

2.3. Multilayer Perceptrons

We now define the fully connected MLPs, which is among the standard architectures in DNNs.

Definition 2.4 (Activation function). Throughout the article, we will only consider the activation functions $\sigma(\cdot) \in L^2_\phi : \mathbb{R} \to \mathbb{R}$ that are L^2-integrable under the Gaussian measure ϕ. The Hermite expansion of $\sigma(\cdot)$ is denoted as

$$\sigma(x) := \sum_{k\geq 0} a_k h_k(x). \tag{2.8}$$

We will explicitly mention the following two assumptions when they are assumed. Otherwise, we will work with the activation $\sigma(\cdot)$ in Definition 2.4.

Assumption 1 (Normalized activation function). Assume that the activation function $\sigma(\cdot) \in L^2_\phi(1)$ is normalized under the Gaussian measure ϕ, in the following sense

$$\mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,1)}[\sigma^2(\mathbf{g})] = 1, \tag{2.9}$$

with the Hermite coefficients satisfying

$$\sum_{k\geq 0} a_k^2 = 1. \tag{2.10}$$

Assumption 2 (Centered activation function). Assume that the activation function $\sigma(\cdot) \in L^2_\phi$ is centered under the Gaussian measure ϕ, in the following sense

$$\mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,1)}[\sigma(\mathbf{g})] = 0, \tag{2.11}$$

or equivalently the Hermite coefficient $a_0 = 0$.

Remark 2.1. Any activation $\sigma(\cdot) \in L^2_\phi$ that are L^2 integrable under the Gaussian measure ϕ can be recentered and rescaled as $\tilde{\sigma}(\cdot)$ to satisfy Assumption 2.4 and Assumption 2 w.l.o.g.,

$$\tilde{\sigma}(\cdot) = \frac{\sigma(\cdot) - a_0}{(\sum_{k\geq 1} a_k^2)^{\frac{1}{2}}}.$$

Common activation functions such as ReLU, GELU, Sigmoid, and Swish all live in L^2_ϕ.

Definition 2.5 (Fully connected MLPs with random weights). Given an activation function $\sigma(\cdot)$, the number of layers L, and the input vector $x^{(0)} := x \in \mathbb{R}^{d_0}$, we define a multilayer feedforward neural network which inductively computes the output for each intermediate layer

$$x^{(\ell+1)} = \sigma\left(W^{(\ell)} x^{(\ell)} / \|x^{(\ell)}\|\right) \in \mathbb{R}^{d_{l+1}}, \text{ for } 0 \leq \ell < L, \text{ with} \tag{2.12}$$

$$W^{(\ell)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell}, \quad W^{(\ell)} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_{d_{\ell+1}} \otimes \mathbf{I}_{d_\ell}). \tag{2.13}$$

Here $\mathbf{I}_{d_\ell}$ denotes the identity matrix of size d_ℓ, and $\otimes$ denotes the Kronecker product between two matrices. The activation $\sigma(\cdot)$ is applied to each component of the vector input, and the weight matrix $W^{(\ell)}$ in the ℓth layer is sampled from a multivariate Gaussian distribution $\mathcal{MN}(\cdot, \cdot)$. For a vector v and a scalar s, the notation v/s denotes the component-wise division of v by scalar s.

Remark 2.2. We remark that the scaling in (2.12) matches the standard weight initialization scheme in practice, since in the current setting $\|x^{(\ell)}\| \asymp \sqrt{d_\ell}$ and (2.13) is effectively saying that each row $W_{i\cdot}^{(\ell)}/\sqrt{d_\ell} \sim \mathcal{N}(0, 1/d_\ell \cdot \mathbf{I}_{d_\ell})$.

3. Compositional Kernel and Branching Process

3.1. Warm Up: Duality

We start by describing a simple duality between the activation function in MLPs (that satisfies Assumption 1) and the PGF in branching processes. This simple yet essential duality allows us to study DNNs, and compare different activation functions borrowing tools from branching processes. This duality in Lemma 3.1 can be readily established via the Mehler's formula (Proposition 2.1). To the best of our knowledge, this probabilistic interpretation (Lemma 3.2) is new to the literature.

Lemma 3.1 (Duality: activation and generating functions). Let $\sigma(\cdot) \in L^2_\phi(1)$ be an activation function under Assumption 1, with the corresponding Hermite coefficients $\{a_k : k \in \mathbb{Z}^{\geq 0}\}$ satisfying $\sum_{k\geq 0} a_k^2 = 1$. Define the corresponding random variable Y_σ with

$$\mathbf{P}(Y_\sigma = k) = a_k^2, \quad \forall k \in \mathbb{Z}^{\geq 0}. \tag{3.1}$$

We denote the PGF of Y_σ (from Definition 2.2) as $G_\sigma(\cdot) : [-1,1] \to [-1,1]$ to be the dual generating function of the activation $\sigma(\cdot)$. Then, for any $x, z \in \mathbb{R}^d$, with $\rho := \langle x/\|x\|, z/\|z\| \rangle \in [-1,1]$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(0,\mathbf{I}_d)} \left[\sigma\left(\boldsymbol{\theta}^\top x/\|x\|\right) \sigma\left(\boldsymbol{\theta}^\top z/\|z\|\right) \right] = G_\sigma(\rho). \tag{3.2}$$

Sketch of the proof. We can rewrite Equation (3.2) in terms of the bivariate normal distribution with correlation ρ as $\mathbb{E}_{(\tilde{x},\tilde{z})\sim\mathcal{N}_\rho}\left[\sigma(\tilde{x})\sigma(\tilde{z})\right]$. Then, by expanding the density of $\mathcal{N}_\rho$ using Mehler's formula we would retrieve the Hermite coefficients $a_k = \mathbb{E}_{\tilde{x}\sim N(0,1)}[\sigma(\tilde{x})h_k(\tilde{x})]$. □

Based on the above Lemma 3.1, it is easy to define the compositional kernel associated with a fully connected MLP with activation $\sigma(\cdot)$. The compositional kernel approach of studying DNNs has been proposed in Daniely, Frostig, and Singer (2017); Daniely et al. (2017). To see why, let us recall the MLP with random weights defined in Definition 2.5. Then, for any fixed data input $x, z \in \mathbb{R}^{d_0}$, the following holds almost surely for random weights $W^{(\ell)}$

$$\begin{aligned} &\lim_{d_{\ell+1}\to\infty} \left\langle x^{(\ell+1)}/\|x^{(\ell+1)}\|, z^{(\ell+1)}/\|z^{(\ell+1)}\| \right\rangle \\ &\quad = G_\sigma\left(\left\langle x^{(\ell)}/\|x^{(\ell)}\|, z^{(\ell)}/\|z^{(\ell)}\| \right\rangle\right). \end{aligned} \tag{3.3}$$

Motivated by the above equation, one can introduce the asymptotic *compositional kernel* defined by a DNN with activation $\sigma(\cdot)$, in the following way.

Definition 3.1 (Compositional kernel). Let $\sigma(\cdot) \in L^2_\phi(1)$ be an activation function that satisfies Assumption 1. Define the L-layer compositional kernel $K_\sigma^{(L)}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be the (infinite-width) compositional kernel associated with the fully connected MLPs (from Definition 2.5), such that for any $x, z \in \mathcal{X}$, we have

$$K_\sigma^{(L)}(x,z) := \underbrace{G_\sigma \circ \cdots \circ G_\sigma}_{\text{composite } L \text{ times}} \left(\langle x/\|x\|, z/\|z\| \rangle\right). \tag{3.4}$$

Since the kernel only depends on the inner-product $\rho = \langle x/\|x\|, z/\|z\| \rangle$, when there is no confusion, we denote for any $\rho \in [-1,1]$

$$K_\sigma^{(L)}(\rho) = \underbrace{G_\sigma \circ \cdots \circ G_\sigma}_{\text{composite } L \text{ times}}(\rho). \tag{3.5}$$

We will now point out the following connection between the compositional kernel for DNNs and the GW branching process. Later, we will study the (rescaled) limits, phase transitions, memorization capacity, and spectral decomposition of such compositional kernels.

Lemma 3.2 (Duality: MLP and branching process). Let $\sigma(\cdot) \in L^2_\phi(1)$ be an activation function that satisfies Assumption 1, and $G_\sigma(\cdot)$ be the dual generating function as in Lemma 3.1. Let $\{Z_{\sigma,L} : L \in \mathbb{Z}^{\geq 0}\}$ be the GW branching process with offspring distribution Y_σ. Then for any $L \in \mathbb{Z}^{\geq 0}$, the compositional kernel has the following interpretation using the GW branching process

$$K_\sigma^{(L)}(\rho) = \mathbb{E}[\rho^{Z_{\sigma,L}}]. \tag{3.6}$$

Sketch of the proof. We prove by induction on L using $Z_{\sigma,L} \overset{d}{=} \sum_{i=1}^{Z_{\sigma,L-1}} Y_{i,L}$, where $Y_{L,i} \overset{\text{iid}}{\sim} Y_\sigma^{(L-1)}$. □

The above duality can be extended to study other network architectures. For instance, in the residual network, the duality can be defined as follows: for $x, z \in \mathbb{S}^{d-1}$, $r \in [0,1]$, and a centered activation function $\sigma(\cdot)$ (Assumption 2), define the dual residual network PGF G_σ^{res} as

$$\begin{aligned} &G_\sigma^{\text{res}}(\langle x,z\rangle) \\ &:= \mathbb{E}_{\{\boldsymbol{\theta}_j\sim\mathcal{N}(0,\mathbf{I}_d), j\in[d]\}}\Big[\sum_{j=1}^{d}\left(\sqrt{1-r}\sigma(\boldsymbol{\theta}_j^\top x)/\sqrt{d} + \sqrt{r}x_j\right) \\ &\quad \times \left(\sqrt{1-r}\sigma(\boldsymbol{\theta}_j^\top z)/\sqrt{d} + \sqrt{r}z_j\right)\Big] \\ &= (1-r)\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\sigma(\boldsymbol{\theta}^\top x)\sigma(\boldsymbol{\theta}^\top z)\right] + r\langle x,z\rangle \\ &= (1-r)G_\sigma^{\text{mlp}}(\langle x,z\rangle) + r\langle x,z\rangle. \end{aligned} \tag{3.7}$$

In Sections 3.2 and 4 and later in experiments, we will elaborate on the costs and benefits of adding a linear component to the PGF in the corresponding compositional behavior, both in theory and numerics. The above simple calculation sheds light on why in practice, residual network can tolerate a larger compositional depth.

3.2. Limits and Phase Transitions

In this section, we will study the properties of the compositional kernel, in the lens of branching process, utilizing the duality established in the previous section. One important result in branching process is the Kesten–Stigum theorem (Kesten and Stigum 1966), which can be employed to assert the rescaled limit and phase transition of the compositional kernel in Theorem 3.1.

Theorem 3.1 (Rescaled nontrivial limits and phase transitions: compositional kernels). Let $\sigma(\cdot) \in L^2_\phi(1)$ be an activation function that satisfies Assumption 1, with $\{a_k : k \in \mathbb{Z}^{\geq 0}\}$ be the corresponding Hermite coefficients that satisfy $\sum_{k\geq 0} a_k^2 = 1$. Define two quantities that depend on $\sigma(\cdot)$,

$$\mu := \sum_{k\geq 0} a_k^2 k, \quad \mu_\star := \sum_{k>2} a_k^2 k \log k. \tag{3.8}$$

Recall the MLP compositional kernel $K_\sigma^{(L)}(\cdot) : \mathbb{R} \to \mathbb{R}$ with activation $\sigma(\cdot)$ in Definition 3.1, and the dual PGF $G_\sigma(\cdot)$ in Lemma 3.1. For any $t \geq 0$, the following results hold, depending on the value of μ and $\mu_\star$:

(i) $\mu \le 1$. Then, if $a_1^2 \neq 1$, we have

$$\lim_{L\to\infty} K_\sigma^{(L)}(e^{-t}) = 1 \quad \text{for } t \ge 0; \tag{3.9}$$

and, if $a_1^2 = 1$, we have $K_\sigma^{(L)}(e^{-t}) = e^{-t}$ for all $L \in \mathbb{Z}^{\ge 0}$;

(ii) $\mu > 1$ and $\mu_\star < \infty$. Then, there exists $0 \le \xi < 1$ with $G_\sigma(\xi) = \xi$ and a unique positive random variable W_σ (that depends on σ) with a continuous density on $\mathbb{R}^+$. And, the nontrivial rescaled limit is

$$\lim_{L\to\infty} K_\sigma^{(L)}(e^{-t/\mu^L}) = \xi + (1-\xi)\cdot \mathbb{E}\left[e^{-tW_\sigma}\right]; \tag{3.10}$$

(iii) $\mu > 1$ and $\mu_\star = \infty$. Then, for any positive number $m > 0$, we have

$$\lim_{L\to\infty} K_\sigma^{(L)}(e^{-t/m^L}) = \begin{cases} 1, & \text{if } t = 0, \\ 0, & \text{if } t > 0. \end{cases} \tag{3.11}$$

The above shows that when looking at the compositional kernel at the rescaled location e^{-t/μ^L} for a fixed $t > 0$, the limit can be characterized by the moment generating function associated with a negative random variable $M_{-W_\sigma}(t) = \mathbb{E}[e^{-tW_\sigma}]$ individual to the activation function σ. The intuition behind such a rescaled location is that the limiting kernels witness an abrupt change of value at $K_\sigma^{(L)}(\rho)$ near $\rho = 1$ for large L (see Corollary 3.1). In the case $\mu > 1$, the proper rescaling in Theorem 3.1 stretches out the curve and zooms in the narrow window of width $O(1/\mu^L)$ local to $\rho = 1$ to inspect the detailed behavior of the compositional kernel $K_\sigma^{(L)}(\rho)$. Conceptually, the above theorem classifies the rescaled behavior of the compositional kernel into three phases according to μ and $\mu_\star$, functionals of the activation $\sigma(\cdot)$. One can also see that the unscaled limit for the compositional kernel has the following simple behavior.

Corollary 3.1 (Unscaled limits and phase transitions). Under the same setting as in Theorem 3.1, the following results hold:

(i) $\mu \le 1$. Then, for all $\rho \in [0,1]$, if $a_1^2 \neq 1$, we have

$$\lim_{L\to\infty} K_\sigma^{(L)}(\rho) = 1; \tag{3.12}$$

and $K_\sigma^{(L)}(\rho) = \rho$ for all $L \in \mathbb{Z}^{\ge 0}$ if $a_1^2 = 1$;

(ii) $\mu > 1$. Then, there exists a unique $0 \le \xi < 1$ with $G_\sigma(\xi) = \xi$

$$\lim_{L\to\infty} K_\sigma^{(L)}(\rho) = \begin{cases} 1, & \text{if } \rho = 1, \\ \xi, & \text{if } \rho \in [0,1). \end{cases} \tag{3.13}$$

Under additional assumptions of $G_\sigma(\cdot)$ on $(-1,0)$ such as no fixed points or nonnegativity, we can extend the above results to $[-1,1]$

$$\lim_{L\to\infty} K_\sigma^{(L)}(\rho) = \begin{cases} 1, & \text{if } \rho = 1, \\ \xi, & \text{if } \rho \in (-1,1). \end{cases} \tag{3.14}$$

Under the additional Assumption 2 on $\sigma(\cdot)$, for nonlinear activation $\sigma(\cdot)$, we have $\mu > 1$ and $\xi = 0$. Therefore, the unscaled limit for nonlinear compositional kernel is

$$\lim_{L\to\infty} K_\sigma^{(L)}(\rho) = \begin{cases} 1, & \text{if } \rho = 1, \\ 0, & \text{if } \rho \in (-1,1). \end{cases} \tag{3.15}$$

We remarks that the fact (ii) in the above corollary is not new and has been observed by Daniely, Frostig, and Singer (2017). On the one hand, they use the fact (ii) to shed light on why more than five consecutive fully connected layers are rare in practical architectures. On the other hand, the phase transition at $\mu = 1$ corresponds to the edge-of-chaos and exponential expressiveness of DNNs studied in Poole et al. (2016), using physics language.

4. Memorization Capacity of Compositions: Tradeoffs

One advantage of DNNs is their exceptional data memorization capacity. Empirically, researchers observed that DNNs with large depth and width could memorize large datasets (Zhang et al. 2017), while maintaining good generalization properties. Pioneered by Belkin, a list of recent work contributes to a better understanding of the interpolation regime (Belkin et al. 2018; Belkin, Rakhlin, and Tsybakov 2018; Bartlett et al. 2019; Belkin, Hsu, and Xu 2019; Feldman 2019; Hastie et al. 2019; Liang and Rakhlin 2020; Liang, Rakhlin, and Zhai 2020; Liang and Sur 2020; Montanari et al. 2020; Nakkiran et al. 2020). With the insights gained via branching process, we will investigate the memorization capacity of the compositional kernels corresponding to MLPs, and study the interplay among the *sample size*, *dimensionality*, *properties of the activation*, and the *compositional depth* in a nonasymptotic way.

In this section, we denote $\mathcal{X} = \{x_i \in \mathbb{S}^{d-1} : i \in [n]\}$ as the dataset with each data point lying on the unit sphere. We denote by $\rho := \max_{i\neq j} |\rho_{ij}|$ with $\rho_{ij} := \langle x_i, x_j\rangle$ as the maximum absolute value of the pairwise correlations. Specifically, we consider the following scaling regimes on the sample size n relative to the dimensionality d:

1. Small correlation: We consider the scaling regime $\frac{\log n}{d} < c$ with some small constant $c < 1$, where the dataset $\mathcal{X}$ is generated from a probabilistic model with uniform distribution on the sphere.
2. Large correlation: We consider the scaling regime $\frac{\log n}{d} > C$ with some large constant $C > 1$, where the dataset $\mathcal{X}$ forms a certain packing set of the sphere. The results also extend to the case of iid samples with uniform distribution on the sphere.

We name it "small correlation" since $\sup_{i\neq j} |\rho_{ij}|$ can be vanishingly small, in the special case $\frac{\log n}{d} \to 0$. Similarly, we call it "large correlation" as $\sup_{i\neq j} |\rho_{ij}|$ can be arbitrarily close to 1, in the special case $\frac{\log n}{d} \to \infty$.

For the results in this section, we make the Assumptions 1 and 2 on the activation function $\sigma(\cdot)$, which are guaranteed by a simple rescaling and centering of any L^2 activation function. Let $\mathbf{K}^{(L)} \in \mathbb{R}^{n\times n}$ be the empirical kernel matrix for the compositional kernel at depth L, with

$$\mathbf{K}^{(L)}[i,j] = K_\sigma^{(L)}\left(\langle x_i, x_j\rangle\right). \tag{4.1}$$

For kernel ridge regression, the spectral properties of the empirical kernel matrix affect the memorization capacity: when $\mathbf{K}^{(L)}$ has full rank, the regression function without explicit regularization can interpolate the training dataset. Specifically, the following spectral characterization on the empirical kernel matrix

determines the rate of convergence in terms of optimization to the min-norm interpolated solution, thus further determines memorization. The κ in following definition of κ-memorization can be viewed as a surrogate to the condition number of the empirical kernel matrix, as the condition number is bounded by $\frac{1+\kappa}{1-\kappa}$.

Definition 4.1 (κ-memorization). We call that a symmetric kernel matrix $\mathbf{K} = [K(x_i, x_j)]_{1\le i,j\le n}$ associated with the dataset $\mathcal{X} = \{x_i\}_{i\in[n]}$ has a κ-memorization property if the eigenvalues $\lambda_i(K), i \in [n]$ of $\mathbf{K}$ are well behaved in the following sense

$$1 - \kappa \le \lambda_i(\mathbf{K}) \le 1 + \kappa, \quad i \in [n]. \tag{4.2}$$

We denote by L_κ the minimum compositional depth such that the empirical kernel matrix $\mathbf{K}$ has the κ-memorization property.

Definition 4.2. (ϵ-closeness) We say that a kernel matrix $\mathbf{K} = [K(x_i, x_j)]_{1\le i,j\le n}$ associated with the dataset $\mathcal{X} = \{x_i\}_{i\in[n]}$ satisfies the ϵ closeness property if

$$\max_{i\neq j} |K(x_i, x_j)| \le \epsilon. \tag{4.3}$$

We denote by $\tilde{L}_\epsilon$ the minimum compositional depth such that the empirical kernel matrix $\mathbf{K}$ satisfies the ϵ-closeness property.

We will assume throughout the rest of this section that

Assumption 3. (Symmetry of PGF) $|G_\sigma(s)| = G_\sigma(|s|)$ for all $s \in (-1, 1)$.

4.1. Small Correlation Regime

To study the small correlation regime, we consider a typical instance of the dataset that are generated iid from a uniform distribution on the sphere.

Theorem 4.1 (Memorization capacity: small correlation regime). Let $\mathcal{X} = \{x_i \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^{d-1}) : i \in [n]\}$ be a dataset with random instances. Consider the regime $\frac{\log n}{d(n)} < c$ with some absolute constant $c < 1$ small enough that only depends on the activation $\sigma(\cdot)$. For any $\kappa \in (0, \rho)$, with probability at least $1 - 4n^{-1/2}$, the minimum compositional depth L_κ to obtain κ-memorization satisfies

$$0.5 \cdot \frac{\log \frac{\log n}{d}}{\log a_1^{-2}} + \frac{\log(0.5\kappa^{-1})}{\log a_1^{-2}} \le L_\kappa \le \frac{\log \frac{\log n}{d}}{\log a_1^{-2}} + \frac{2\log(3n\kappa^{-1})}{\log a_1^{-2}} + 1. \tag{4.4}$$

The proof is due to the sharp upper and lower estimates obtained in the following lemma.

Lemma 4.1 (ϵ-closeness: small correlation regime). Consider the same setting as in Theorem 4.1. For any $\epsilon \in (0, \rho)$, with probability at least $1 - 4n^{-1/2}$, the minimum depth $\tilde{L}_\epsilon$ to obtain ϵ-closeness satisfies

$$\frac{0.5\log \frac{\log n}{d} + \log(0.5\epsilon^{-1})}{\log a_1^{-2}} \le \tilde{L}_\epsilon \le \frac{\log \frac{\log n}{d} + 2\log(3\epsilon^{-1})}{\log a_1^{-2}} + 1. \tag{4.5}$$

In this small correlation regime, Theorem 4.1 states that in order for us to memorize a size n-dataset $\mathcal{X}$, the depth L for the compositional kernel $K_\sigma^{(L)}$ scales with three quantities: the linear component in the activation function a_1^{-2}, a factor between $\log(\kappa^{-1})$ and $\log(n\kappa^{-1})$, and the logarithm of the regime scaling $\log(\frac{\log n}{d})$. Two remarks are in order. First, as the quantity $\frac{\log n}{d}$ becomes larger, we need a larger depth for the compositional kernel to achieve the same memorization. However, such an effect is mild since the regime scaling enters *precisely* logarithmically in the form of $\log(\frac{\log n}{d})$. In other words, for iid data on the unit sphere with $\frac{\log n}{d} \to 0$, it is indeed easy for a shallow compositional kernel (with depth at most $\log \frac{n\log n}{d}$) to memorize the data. In fact, consider the proportional high-dimensional regime $d(n) \asymp n$, then a very shallow network with $1 \precsim L_\kappa^{\text{easy}} \precsim \log\log n$ is sufficient and necessary to memorize. Second, $a_1^{-2} = 1 + (\sum_{k\ge 2} a_k^2)/a_1^2$ can be interpreted as the amount of nonlinearity in the activation function. Therefore, when the nonlinear component is larger, we will need fewer compositions for memorization. This explains the necessary large depth of an architecture such as ResNet (Equation (3.7)), where a larger linear component is added in each layer to the corresponding kernel. A simple contrast should be mentioned for comparison: memorization is only possible for linear models when $d \ge n$, whereas with composition and nonlinearity, $d \gg \log n$ suffices for good memorization.

4.2. Large Correlation Regime

To study the large correlation regime, we consider a natural instance of the dataset that falls under such a setting. The construction is based on the sphere packing.

Definition 4.3 (r-polarized packing). For a compact subset $V \subset \mathbb{R}^d$, we say $\mathcal{X} = \{x_i\}_{i\in[n]} \subset V$ is a r-polarized packing of V if for all $x_i \neq x_j \in \mathcal{X}$, we have $\|x_i - x_j\| > r$ and $\|x_i + x_j\| > r$. We define the polarized packing number of V as $\mathcal{P}_r(V)$, that is

$$\mathcal{P}_r(V) = \max\{n\colon \text{ exists } r\text{-polarized packing of } V \text{ of size } n\}. \tag{4.6}$$

Theorem 4.2 (Memorization capacity: large correlation regime). Let a size-n dataset $\mathcal{X} = \{x_i \in \mathbb{S}^{d-1} : i \in [n]\}$ be a maximal polarized packing set of the sphere $\mathbb{S}^{d-1}$. Consider the regime $\frac{\log n}{d(n)} > C$ with some absolute constant $C > 1$ that only depends on the activation $\sigma(\cdot)$. For any $\kappa \in (0, \min\{\rho, ns_\star\})$, the minimum depth $\tilde{L}_\epsilon$ to obtain ϵ-closeness satisfies

$$1.5 \cdot \frac{\frac{\log n}{d}}{\log \mu} + \frac{\log(0.5\kappa^{-1})}{\log a_1^{-2}} - 1 \le L_\kappa \le 3 \cdot \frac{\frac{\log n}{d-1}}{\log \frac{\mu+1}{2}} + \frac{\log(s_\star n\kappa^{-1})}{\log \frac{s_\star}{1-(1-s_\star)\frac{1+\mu}{2}}} + 2. \tag{4.7}$$

Here $s_\star := \inf\left\{s \in (0,1) : \frac{1-G_\sigma(s)}{1-s} \ge \frac{1+\mu}{2}\right\}$ is a constant that only depends on $\sigma(\cdot)$.

Remark 4.1. One can carry out an identical analysis in the large correlation regime for the iid random samples case $x_i \sim$

$\mathrm{Unif}(\mathbb{S}^{d-1}), i \in [n]$ with $\frac{\log n}{d} > C$, as in the sphere packing case. Here the constant C only depends on $\sigma(\cdot)$. Exactly the same bounds on L_κ hold with high probability.

The proof is due to the sharp upper and lower estimates in the following lemma.

Lemma 4.2 (ϵ-closeness: large correlation regime). Consider the same setting as in Theorem 4.2. For any $\epsilon \in (0, \rho)$, the minimum depth $\tilde{L}_\epsilon$ to obtain ϵ-closeness satisfies

$$\tilde{L}_\epsilon \geq \max_{s\in(\epsilon,\rho)} \left\{ \frac{\log(\epsilon^{-1}) + \log(s)}{\log a_1^{-2}} + \frac{2\frac{\log n}{d} + \log\left(\frac{1-s}{18.2}\right)}{\log \mu} \right\} - 1, \tag{4.8}$$

$$\tilde{L}_\epsilon \leq \min_{s\in(\epsilon,\rho)} \left\{ \frac{\log(\epsilon^{-1}) + \log(s)}{\log \frac{s}{G_\sigma(s)}} + \frac{2\frac{\log n}{d-1} + \log\left(\frac{1-s}{0.06}\right)}{\log \frac{1-G_\sigma(s)}{1-s}} \right\} + 2. \tag{4.9}$$

In this large correlation regime, to memorize a dataset, the behavior of the compositional depth is rather different from the small correlation regime. By Theorem 4.2, we have that the depth L scales with following quantities: a factor between $\log(\kappa^{-1})$ and $\log(n\kappa^{-1})$ same as before, the regime scaling $\frac{\log n}{d}$, and functionals of the activation μ, a_1^{-2}, and $s_\star$. Few remarks are in order. First, in this large correlation regime $n \gg \exp(d)$, memorization is indeed possible. However, the compositional depth needed increases *precisely* linearly as a function of the regime scaling $\frac{\log n}{d}$. The above is in contrast to the small correlation regime, where the dependence on the regime scaling is logarithmic as $\log(\frac{\log n}{d})$. For hard dataset instances on the sphere with $\frac{\log n}{d} \to \infty$, one needs at least $\frac{\log n}{d}$ depth for the compositional kernel to achieve memorization. In fact, consider the fixed dimensional regime with $d(n) \asymp 1$, then a deeper network with depth $L_\kappa^{\mathrm{hard}} \asymp \log n$ is sufficient and necessary to memorize, which is much larger than the depth needed in the proportional high-dimensional regime with $L_\kappa^{\mathrm{easy}} \precsim \log\log n$. Second, for larger values of a_1^{-2} and μ we will need less compositional depths, as the amount of nonlinearity is larger. To sum up, with nonlinearity and composition, even in the $d \ll \log n$ regime with a hard data instance, memorization is possible but with a DNN.

5. Spectral Decomposition of Compositional Kernels

In this section, we investigate the spectral decomposition of the compositional kernel function. We study the case where the base measure is a uniform distribution on the unit sphere, denoted by τ_{d-1}. Let S_{d-1} be the surface area of $\mathbb{S}^{d-1}$. To state the results, we will need some background on the spherical harmonics. We consider the dimension $d \geq 2$, and will use $k \in \mathbb{Z}^{\geq 0}$ to denote an integer.

Definition 5.1 (Spherical harmonics, Atkinson and Han 2012, chap. 2.8.4). Let $\mathbb{Y}_k^d$ be the space of kth degree spherical harmonics in dimension d, and let $\{Y_{k,j}(x) : j \in [N_{k,d}]\}$ be an orthonormal basis for $\mathbb{Y}_k^d$, with

$$\int Y_{k,i}(x) Y_{k,j}(x) d\tau_{d-1}(x) = \mathbb{1}_{i=j}. \tag{5.1}$$

Then, the sets form an orthogonal basis for the space $L^2_{\tau_{d-1}}$ of L^2-integrable functions on $\mathbb{S}^{d-1}$ with the base measure τ_{d-1}, noted below

$$L^2_{\tau_{d-1}} = \bigoplus_{k=0}^{\infty} \mathbb{Y}_k^d. \tag{5.2}$$

Moreover, the dimensionality $\dim \mathbb{Y}_k^d = N_{k,d}$ are the coefficients of the generating function

$$\sum_{k=0}^{\infty} N_{k,d} s^k = \frac{1+s}{(1-s)^{d-1}}, \ |s| < 1. \tag{5.3}$$

Definition 5.2 (Legendre polynomial, Atkinson and Han 2012, chap. 2.7). Define the Legendre polynomial of degree k with dimension d to be

$$P_{k,d}(t) = (-1)^k \frac{\Gamma(\frac{d-1}{2})}{2^k \Gamma(k + \frac{d-1}{2})} (1-t^2)^{-\frac{d-3}{2}} \left(\frac{d}{dt}\right)^k (1-t^2)^{k+\frac{d-3}{2}}. \tag{5.4}$$

The following orthogonality holds

$$\int_{-1}^{1} P_{k,d}(t) P_{\ell,d}(t) (1-t^2)^{\frac{d-3}{2}} dt = \begin{cases} 0, & \text{if } k \neq \ell, \\ \frac{S_{d-1}}{N_{k,d} S_{d-2}}, & \text{if } k = \ell. \end{cases} \tag{5.5}$$

Recall that the compositional kernel $K_\sigma^{(L)}(\cdot)$ and the random variable $Z_{\sigma,L}$, which denotes the size of the Lth generation, as in Lemma 3.2. Then, we have the following theorem describing the spectral decomposition of the compositional kernel function and the associated integral operator.

Theorem 5.1 (Spectral decomposition of compositional kernel). Consider any $x, z \in \mathbb{S}^{d-1}$. Then, the following spectral decomposition holds for the compositional kernel $K_\sigma^{(L)}$ with any fixed depth $L \in \mathbb{Z}^{\geq 0}$:

$$K_\sigma^{(L)}(\langle x, z\rangle) = \sum_{k=0}^{\infty} \lambda_k \sum_{j=1}^{N_{k,d}} Y_{k,j}(x) Y_{k,j}(z), \tag{5.6}$$

where the eigenfunctions $Y_{k,j}(\cdot)$ form an orthogonal basis of $L^2_{\tau_{d-1}}$, and the eigenvalues λ_k satisfy the following formula

$$\lambda_k := \frac{S_{d-2}}{S_{d-1}} \Gamma\left(\frac{d-1}{2}\right) \sum_{\ell \geq 0} \mathbf{P}(Z_{\sigma,L} = k + \ell) \times \frac{(k+\ell)!}{\ell!} \frac{1 + (-1)^\ell}{2^{k+1}} \frac{\Gamma(\frac{\ell+1}{2})}{\Gamma(k + \frac{\ell+d}{2})}. \tag{5.7}$$

The associated integral operator $\mathcal{T}_\sigma^{(L)} : L^2_{\tau_{d-1}} \to L^2_{\tau_{d-1}}$ with respect to the kernel $K_\sigma^{(L)}$ is defined as

$$\left(\mathcal{T}_\sigma^{(L)} f\right)(x) := \int K_\sigma^{(L)}(\langle x, z\rangle) f(z) d\tau_{d-1}(z) \quad \text{for any } f \in L^2_{\tau_{d-1}}. \tag{5.8}$$

From Theorem 5.1, we know that the eigenfunctions of the operator $\mathcal{T}_\sigma^{(L)}$ are the spherical harmonic basis $\{Y_{k,j} : k \in \mathbb{Z}^{\geq 0}, j \in [N_{k,d}]\}$, with $N_{k,d}$ identical eigenvalues λ_k such that

$$\mathcal{T}_\sigma^{(L)} Y_{k,j} = \lambda_k Y_{k,j}. \tag{5.9}$$

The above spectral decompositions are important because it helps us study the generalization error (in the fixed dimensional setting) of regression methods with the compositional kernel $K_\sigma^{(L)}$. More specifically, understanding the eigenvalues of the compositional kernels means that we can employ the classical theory on reproducing kernel Hilbert spaces regression (Caponnetto and Vito 2007) to quantify generalization error, when the dimension is fixed. In the case when dimensionality grows with the sample size, several attempts have been made to understand the generalization properties of the inner-product kernels (Liang and Rakhlin 2020; Liang, Rakhlin, and Zhai 2020) in the interpolation regime, which includes these compositional kernels as special cases.

6. Kernels to Activations: New Random Features Algorithms

Given any L_ϕ^2 activation function $\sigma(\cdot)$ (as in Definition 2.4), we can define a sequence of PD compositional kernels $K_\sigma^{(L)}$, with $L \geq 0$, whose spectral properties have been studied in the previous section, utilizing the duality established in Section 3.1. Such compositional kernels are nonlinear functions on the inner-product $\langle x, z\rangle$ (rotation-invariant), and we will call them the *inner-product kernels* (Kar and Karnick 2012). In this section, we will investigate the *converse question*: given an arbitrary PD inner-product kernel, can we identify an activation function associated with it? We will provide a positive answer in this section. Direct algorithmic implications are new random features algorithms that are distinct from the well-known RFF and random kitchen sinks algorithms studied in Rahimi and Recht (2008, 2009).

Define an inner-product kernel $K(\cdot,\cdot) : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, with $d \geq 2$,

$$K(x, z) := f(\langle x, z\rangle), \tag{6.1}$$

where $f\colon [-1, 1] \to \mathbb{R}$ is a continuous function. Denote the expansion of f under the Legendre polynomials $P_{k,d}$ (see Definition 5.2) as

$$f(t) = \sum_{k=0}^{\infty} \alpha_k P_{k,d}(t). \tag{6.2}$$

To define the activations corresponding to an arbitrary PD inner-product kernel, we require the following theorem due to Schoenberg (1942).

Proposition 6.1 (Schoenberg 1942, Theorem 1). For a fixed $d \geq 2$, the inner-product kernel $K(\cdot,\cdot)$ in (6.1) is PD on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ if and only if $\alpha_k \geq 0$ for all $k \in \mathbb{Z}^{\geq 0}$ in Equation (6.2).

Now, we are ready to state the activation function $\sigma_f(\cdot)$ defined based on the inner-product kernel function $f(\cdot)$.

Algorithm 1: New random features algorithm for PD inner-product kernels $K(x, z) = f(\langle x, z\rangle)$ based on Theorem 6.1.

Result: Given a normalized dataset $\mathcal{X} = \{x_i \in \mathbb{S}^{d-1}, i \in [n]\}$, an integer m, and a PD inner-product kernel $K(x, z) = f(\langle x, z\rangle)$, return a randomized nonlinear feature matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times m}$ for kernel ridge regression.

Step 1: Calculate the Legendre coefficients of $f(t) = \sum_{k=0}^{\infty} \alpha_k P_{k,d}(t)$;

Step 2: Obtain the dual activation function $\sigma_f(t) = \sum_{k=0}^{\infty} \sqrt{\alpha_k N_{k,d}} P_{k,d}(t)$, $t \in [-1, 1]$;

Step 3: Sample m-iid isotropic Gaussian vectors $t_j \sim \mathcal{N}(0, \mathbf{I}_d)$, and then define the nonlinear feature matrix

$$\mathbf{\Phi}[i, j] = \sigma_f\left(\langle x_i, \boldsymbol{\theta}_j/\|\boldsymbol{\theta}_j\|\rangle\right), \ i \in [n], j \in [m]. \tag{6.5}$$

The feature matrix satisfies $\lim_{m\to\infty}\left(1/m \cdot \mathbf{\Phi}\mathbf{\Phi}^\top\right)[i, \ell] = f(\langle x_i, x_\ell\rangle)$ for all $i, \ell \in [n]$ a.s.

Theorem 6.1 (Kernels to activations). Consider any PD inner product kernel $K(x, z) := f(\langle x, z\rangle)$ on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ associated with the continuous function $f : [-1, 1] \to \mathbb{R}$. Assume without loss of generality that $f(1) = 1$, and recall the definition of $N_{k,d}$ in (5.3). Due to Proposition 6.1, the Legendre coefficients $\{\alpha_k\colon k \in \mathbb{Z}^{\geq 0}\}$, defined in Equation (6.2), of $f(\cdot)$ are nonnegative.

One can define the following dual activation function $\sigma_f : [-1, 1] \to \mathbb{R}$

$$\sigma_f(t) = \sum_{k=0}^{\infty} \sqrt{\alpha_k N_{k,d}} P_{k,d}(t). \tag{6.3}$$

Then, the following statements hold:

(i) $\sigma_f(\cdot)$ is L^2 in the following sense $\int_{-1}^{1} \left(\sigma_f(t)\right)^2 \frac{S_{d-2}}{S_{d-1}}(1 - t^2)^{\frac{d-3}{2}} dt = 1$.

(ii) For any $x, z \in \mathbb{S}^{d-1}$

$$\mathbb{E}_{\boldsymbol{\xi}\sim\tau_{d-1}}\left[\sigma_f(\boldsymbol{\xi}^\top x)\sigma_f(\boldsymbol{\xi}^\top z)\right] = f(\langle x, z\rangle) = K(x, z), \tag{6.4}$$

where ξ is sampled from a uniform distribution on the sphere $\mathbb{S}^{d-1}$.

The above theorem naturally induces a new random features algorithm for kernel ridge regression, described below. Note that the kernel K can be any compositional kernel, which is PD and of an inner-product form.

It is clear from Theorem 5.1 that all compositional kernels $K_\sigma^{(L)}$ are PD, though the converse statement is not true. A notable example is the kernel $P_{k,d}(\langle x, z\rangle) : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, which is PD kernel but with negative Taylor coefficients, thus cannot be a compositional kernel. For the special case of compositional kernels with depth L and activation $\sigma(\cdot)$, it turns out one can define a new "compressed" activation $\sigma^{(L)}(\cdot) \in L_\phi^2$ to represent the depth-L compositional kernel. We propose the following algorithm:

Algorithm 2: New random features algorithm for compositional kernels $K_\sigma^{(L)}$.

Result: Given a normalized dataset $\mathcal{X} = \{x_i \in \mathbb{S}^{d-1}, i \in [n]\}$, an integer m, and a compositional kernel $K_\sigma^{(L)}(x,z) = K_\sigma^{(L)}(\langle x,z\rangle)$, return a randomized nonlinear feature matrix $\mathbf{\Psi} \in \mathbb{R}^{n\times m}$ for kernel ridge regression.

Step 1: Calculate the Taylor coefficients of $K_\sigma^{(L)}(t) = \sum_{k=0}^\infty \alpha_k t^k$ with $\alpha_k = \mathbf{P}(Z_{\sigma,L} = k)$;

Step 2: Obtain the "compressed" activation function

$$\sigma^{(L)}(t) = \sum_{k=0}^{\infty} \sqrt{\alpha_k} h_k(t),\ t \in \mathbb{R}; \tag{6.6}$$

Step 3: Sample m-iid isotropic Gaussian vectors $\boldsymbol{\theta}_j \sim \mathcal{N}(0, \mathbf{I}_d)$, and then define the nonlinear feature matrix

$$\mathbf{\Psi}[i,j] = \sigma^{(L)}\left(\langle x_i, \boldsymbol{\theta}_j\rangle\right),\ i \in [n], j \in [m]. \tag{6.7}$$

The feature matrix satisfies $\lim_{m\to\infty}\left(1/m \cdot \mathbf{\Psi}\mathbf{\Psi}^\top\right)[i,\ell] = K_\sigma^{(L)}(\langle x_i, x_\ell\rangle)$ for all $i, \ell \in [n]$ a.s.

First, let us discuss the relationship between our random features algorithms above and that in Rahimi and Recht (2008, 2009). Rahimi and Recht (2008) employed the duality between the PD shift-invariant kernel $k(x-z)$ and a positive measure that corresponds to the inverse Fourier transform of k: the random features are constructed based on the specific positive measure where sampling could be a nontrivial task. In contrast, we utilize the duality between the PD inner-product kernel and an activation function, where the random features are always generated based on the uniform distribution on the sphere (or the isotropic Gaussian), but with different activations $\sigma(\cdot)$. It is clear that sampling uniformly from the sphere can be easily done by sampling $\boldsymbol{\theta} \sim \mathcal{N}(0, I_d)$, and returning $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$.

We now conclude this section by stating another property of Algorithm 2, implied by the Mehler's formula. It turns out that under a certain high-dimensional scaling regime, say $d(n) \asymp n^{\frac{2}{\iota+1}}$ for some fixed integer $\iota \in \mathbb{Z}^{>0}$, the compressed activation $\sigma^{(L)}$ in (6.6) can be truncated without loss using only the low degree components $k \le \iota$. For notation simplicity, in the statement below, we drop the superscript (L) in the kernel and the compressed activation.

Theorem 6.2 (Empirical kernel matrix: truncation and implicit regularization). Consider the compositional kernel function $K_\sigma(t) = \sum_{k=0}^\infty \alpha_k t^k$ as in Algorithm 2, and the dataset $\mathbf{X} = [x_1^\top, \ldots, x_n^\top]^\top \in \mathbb{R}^{n\times d}$ with column $x_i \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^{d-1})$. Consider the high-dimensional regime where the dimensionality $d(n)$ scales with n, satisfying

$$d(n) \gtrsim n^{\frac{2}{\iota+1}+\delta},$$

for some fixed integer $\iota \in \mathbb{Z}^{>0}$ and any fixed small $\delta > 0$. Define a truncated activation based on (6.6), at degree level ι

$$\sigma_{\le\iota}(t) := \sum_{k=0}^{\iota} \sqrt{\alpha_k} h_k(t). \tag{6.8}$$

Then the empirical kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ satisfies the following decomposition

$$\mathbf{K} = \underbrace{\left(\sum_{k>\iota} \alpha_k\right)\cdot \mathbf{I}_n}_{\text{implicit regularization}} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\sigma_{\le\iota}(\mathbf{X}\boldsymbol{\theta})\sigma_{\le\iota}(\mathbf{X}\boldsymbol{\theta})^\top\right]}_{\text{degree truncated random features}} + \underbrace{\mathbf{R}}_{\text{remainder}} \tag{6.9}$$

with the remainder matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ satisfies

$$\|\mathbf{R}\|_{\text{op}} \to 0, \text{ as } n, d(n) \to \infty. \tag{6.10}$$

A direct consequence of the above theorem is that, the empirical kernel matrix $\mathbf{K}$ shares the same eigenvalues and empirical spectral density as the matrix $(\sum_{k>\iota}\alpha_k)\cdot\mathbf{I}_n + \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{N}(0,\mathbf{I}_d)}[\sigma_{\le\iota}(\mathbf{X}\boldsymbol{\theta})\sigma_{\le\iota}(\mathbf{X}\boldsymbol{\theta})^\top]$, asymptotically. The implicit regularization matrix is contributed by the high degree components of the activation function collectively. Therefore, in Algorithm 2 with truncation level ι the random features in Equation (6.7) are generated according to

$$\mathbf{\Psi}[i,j] = \sigma_{\le\iota}^{(L)}\left(\langle x_i, \boldsymbol{\theta}_j\rangle\right) + \left(1 - \sum_{k=0}^{\iota}\alpha_k\right)^{1/2}\cdot z_{ij},\ i \in [n], j \in [m], \tag{6.11}$$

where $z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ are Gaussian noise.

7. Numerical Investigation

In this section, we will study numerically the theory established in the previous sections. We will experiment with four common activations used in practice as a proof of concept, namely ReLU, GeLU, Swish, and Sigmoid, and four PGFs associated with nonnegative discrete probability distributions including Poisson(λ), Binomial(n,p), Geometric(p), and Uniform(0, n). To execute the theory numerically, we introduce a simple and general Algorithm 3 for estimating the Hermite coefficients of *any* activation function $\sigma(\cdot)$, with provable guarantees. The numerical stability of this algorithm to estimate the Hermite coefficients can be seen in Figure 1. The truncation level considered is $\iota = 20$. We will use this level throughout the rest of the experiments in this section.

7.1. Duality: Activations and PGFs

According to Lemma 3.1, there is a duality between activation functions $\sigma(\cdot)$ (normalized as in 1) and PGFs $G_\sigma(\cdot)$ (of a discrete nonnegative probability distribution). In the first two plots in Figure 2, we start from a probability distribution, and construct its corresponding activation function. Reversely, in the last two plots in Figure 2, we start from an activation function, and approximate (using Algorithm 3) its corresponding PGF.

7.2. Kernel Limits

In this section, we will illustrate the compositional behavior of the kernels, both in the unscaled (Corollary 3.1) and rescaled

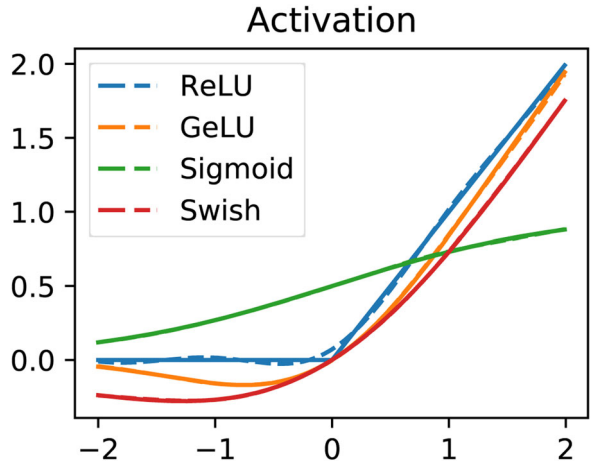

Figure 1. Plot of the numerical stability of Algorithm 3 in approximating activation functions. Solid lines represent the activations, and the dashed lines represent the approximations of the activation functions.

Algorithm 3: Estimating the Hermite coefficients of an activation $\sigma(\cdot)$.

Result: Given an activation function $\sigma(t) = \sum_{k=0}^{\infty} a_k h_k(t) \in L^2_\phi$ and a truncation level ι, return an estimate of the Hermite coefficients $\{a_k\}_{k=0}^{\iota}$.

Step 1: Generate M inputs $\{x_i\}_{i=1}^{M}$ from a standard normal $\mathcal{N}(0,1)$;

Step 2: For each $k = 0, 1, 2, \ldots, \iota$, calculate

$$\hat{a}_k = \frac{1}{M} \sum_{i=1}^{M} \sigma(x_i) h_k(x_i).$$

The coefficients satisfy $\lim_{M\to\infty} \hat{a}_k = a_k$ a.s. for all k.

(Theorem 3.1) cases. We start with Figure 3 on the compositional behavior of unscaled kernels $K_\sigma^{(L)}(\rho)$ as a function of $\rho \in [-1, 1]$, without centering the activation functions. According to Corollary 3.1, we know that the unscaled limits of compositional kernels are determined strictly by a phase transition of μ at 1. On the interval $(-1, 1)$, ReLU and Sigmoid's composite kernels converge to 1, while GeLU and Swish's converge to their corresponding extinction probability. In Figure 4, we plot the compositional behavior of the $K_\sigma^{(L)}(\rho)$ for centered activations σ: For centered ReLU, GeLU, Swish, the limit approaches 0 for $\rho \in [-1, 1)$, and approaches 1 at $\rho = 1$. For centered Sigmoid, $a_1 \approx 1$, which explains the resemblance to the linear kernel.

On the other hand, for rescaled kernels $K_\sigma^{(L)}(e^{-t/\mu^L})$ as a function of $t \in [0, \infty)$, nontrivial limits that depend on $\sigma(\cdot)$ exist, when $\mu_\star < \infty$. In the un-centered and rescaled case (Theorem 3.1), we have that ReLU and Sigmoid's compositional kernels converge to 1, while GeLU and Swish's ones approach a nontrivial limit, shown in Figure 5. If we center the activations, all rescaled kernels will approach nontrivial limits as seen in Figure 6.

In terms of the convergence speed of kernels to their unscaled limit, Theorem 4.1 explains how fast the curve flattens around 0, and Theorem 4.2 on the rate it flattens around 1. The convergence speed, in fact, determines the memorization capacity of composition kernels. To flatten around 0, we need the number of compositions to scale with $\frac{1}{\log a_1^{-2}}$, while to flatten around 1, we need the compositional depth to scale with $\frac{1}{\mu-1}$. For example, Sigmoid has $\mu \approx 1.03$ and $a_1^2 \approx 0.99$, thus explaining slow convergence compared to the other activations. Table 2 summarizes the crucial quantities that determine the compositional behavior for each activation.

7.3. Applications to Datasets: New Random Features Algorithm

In this section, we will investigate the "compressed" activations obtained from compositional kernels to generate random features, as in Algorithm 2. In Figure 7, we plot the shape of the compressed activations, where the initial activations were recentered and rescaled (Assumptions 1 and 2). We will test the validity of the new random features Algorithm 2 on

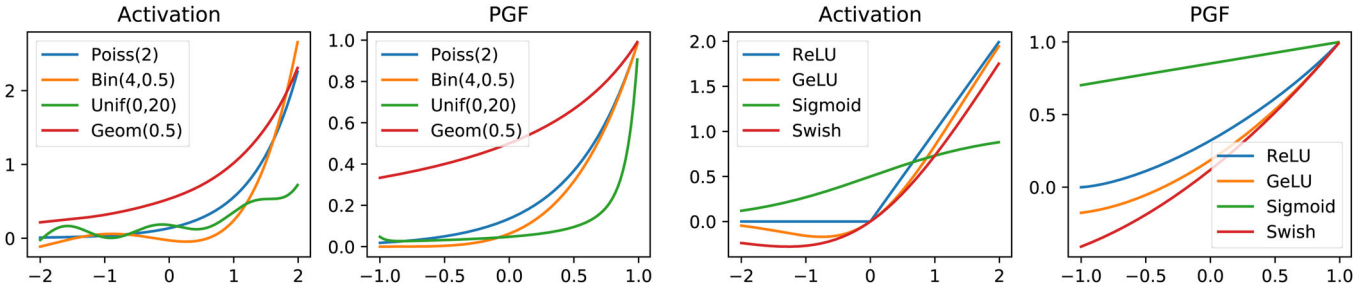

Figure 2. Duality between activations and PGFs.

Table 1. Examples of common activation functions and PGFs.

Activations	ReLU	GeLU	Sigmoid	Swish
$\sigma(t)$	$\max(0, t)$	$t \cdot \Phi(t)$	$1/(1+e^{-t})$	$t/(1+e^{-t})$
PGFs	Poisson(λ)	Binomial(n,p)	Geometric(p)	Uniform(0,n)
$G(s)$	$\exp\{\lambda(s-1)\}$	$(1-p+ps)^n$	$(1-p)/(1-ps)$	$(1-s^{n+1})/(n(1-s))$

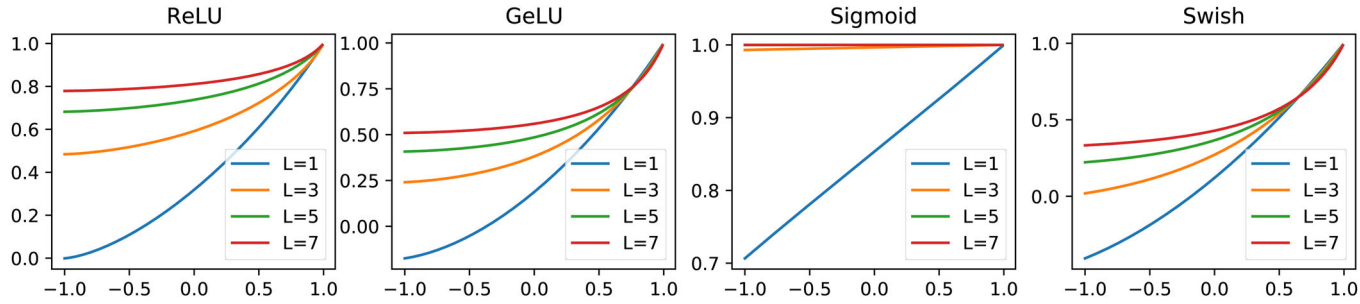

Figure 3. Compositional behavior of unscaled kernels.

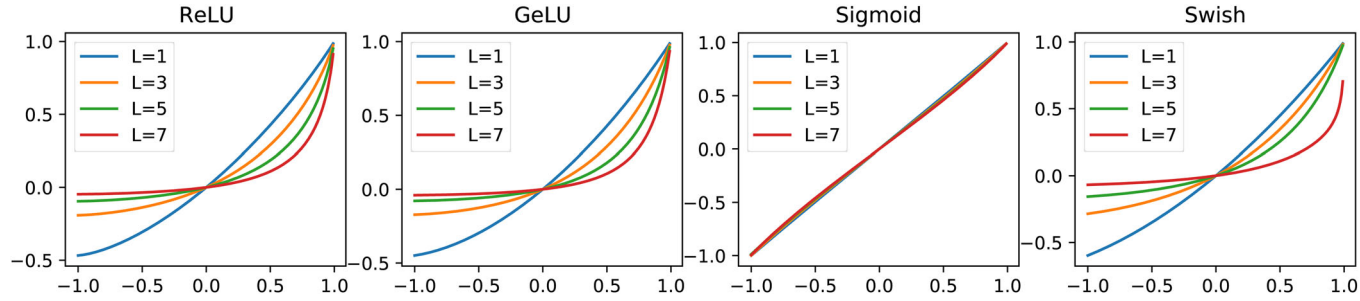

Figure 4. Compositional behavior of unscaled kernels with centered activations.

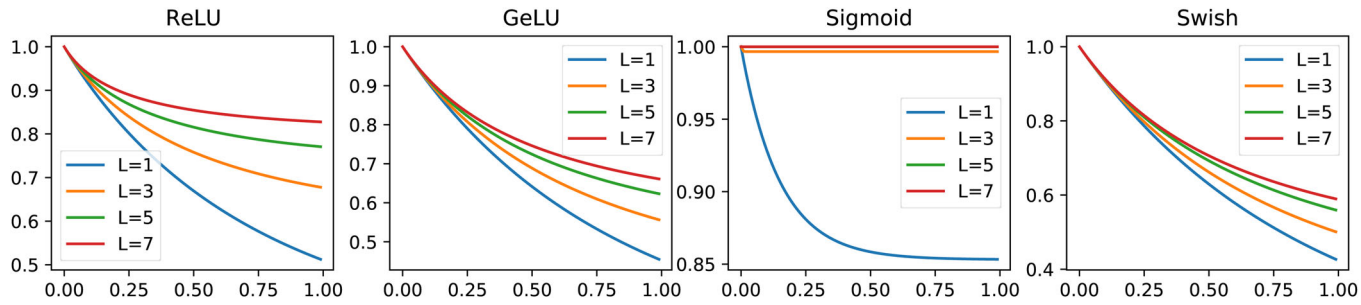

Figure 5. Compositional behavior of rescaled kernels.

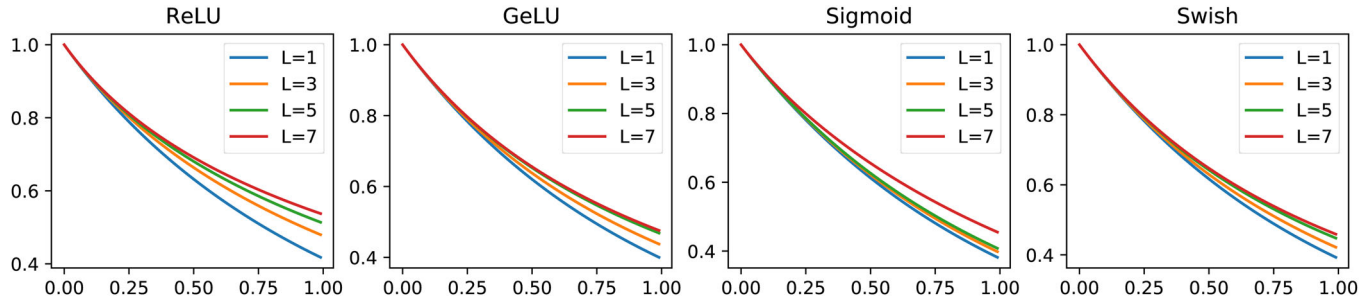

Figure 6. Compositional behavior of rescaled kernels with centered activations.

two available datasets, MNIST and CIFAR10. In addition, we construct a new dataset called VGG11, which takes as input the last convolutional layer of the architecture VGG11 (the 8th layer) trained on CIFAR10, and as outputs the CIFAR10 labels.

We plot in Figure 8 the condition number of the empirical kernel matrix as depth increases. Empirically, we see that depth improves the kernel matrix's condition number, as discussed in Section 4. Note that, unlike the other three activation functions, the Sigmoid activation's kernel has the slowest decay as depth increases. This behavior of the Sigmoid activation results from the fact that the activation has a significant component when projected on the first ι Hermite polynomials (see Figure 9),

Table 2. Approximations of μ, $\mu_\star$ (from Theorem 3.1), and a_1 (from Definition 2.4).

Activation	ReLU		GeLU		Sigmoid		Swish	
μ	0.95	1.39	1.08	1.47	0.15	1.03	1.07	1.22
$\mu_\star$	0.48	0.69	0.39	0.89	0.01	0.05	0.28	0.31
a_1^2	0.50	0.74	0.59	0.71	0.15	0.99	0.80	0.70
ξ	1.00	0.00	0.76	0.00	1.00	0.00	0.66	0.00

NOTE: For each activation, there are two columns: values for un-centered activation (left) and values for centered activation (right).

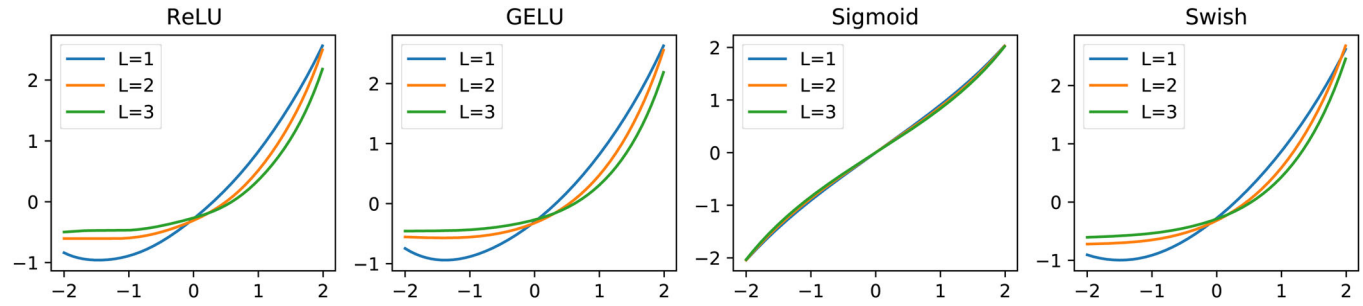

Figure 7. Compressed activation functions truncated at level $\iota = 20$ as in Algorithm 2.

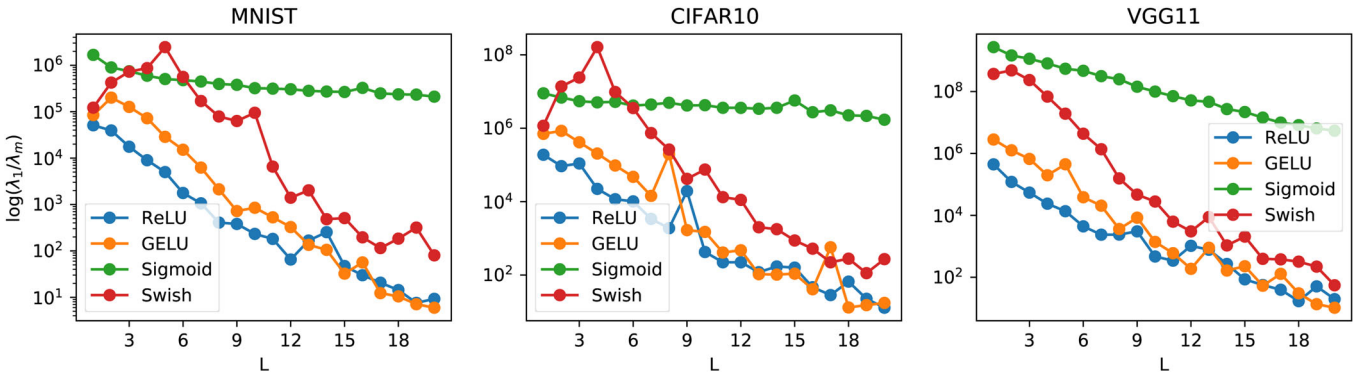

Figure 8. Condition number of the empirical kernel matrix $\lambda_1(\Psi\Psi^T/m)/\lambda_m(\Psi\Psi^T/m)$ as a function of depth, where the random features matrix Ψ is defined in Algorithm 2.

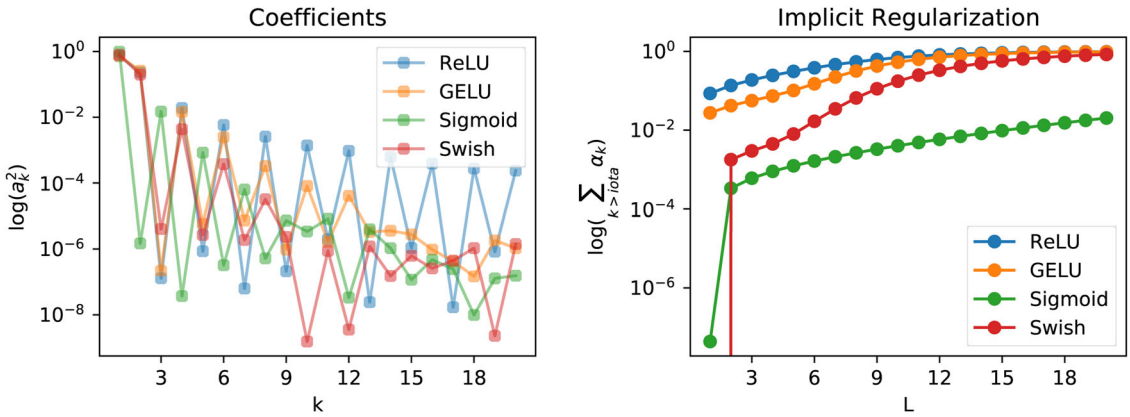

Figure 9. On the left: estimated coefficients of the activation functions as in Algorithm 3. On the right: the amount of implicit regularization $\sum_{k>\iota}\alpha_k$ of the compressed activation function due to the truncation level $\iota = 20$, as defined in Theorem 6.2.

Table 3. Accuracy percentage based on activation and depths.

Activation		ReLU			GeLU			Sigmoid			Swish		
Depth		1	2	3	1	2	3	1	2	3	1	2	3
MNIST	Train	1.00	0.99	0.99	1.00	1.00	0.99	0.97	0.97	0.98	0.99	1.00	0.99
	Test	0.97	0.97	0.97	0.97	0.97	0.96	0.95	0.96	0.96	0.97	0.97	0.97
Depth		1	4	8	1	4	8	1	4	8	1	4	8
CIFAR10	Train	0.55	0.51	0.44	0.55	0.51	0.46	0.50	0.52	0.52	0.55	0.51	0.47
	Test	0.47	0.39	0.30	0.47	0.44	0.32	0.43	0.44	0.44	0.44	0.39	0.33
Depth		1	4	8	1	4	8	1	4	8	1	4	8
VGG11	Train	1.00	0.96	0.34	0.99	0.96	0.66	1.00	1.00	0.99	0.99	0.95	0.83
	Test	0.78	0.74	0.30	0.77	0.75	0.54	0.77	0.77	0.77	0.76	0.74	0.65

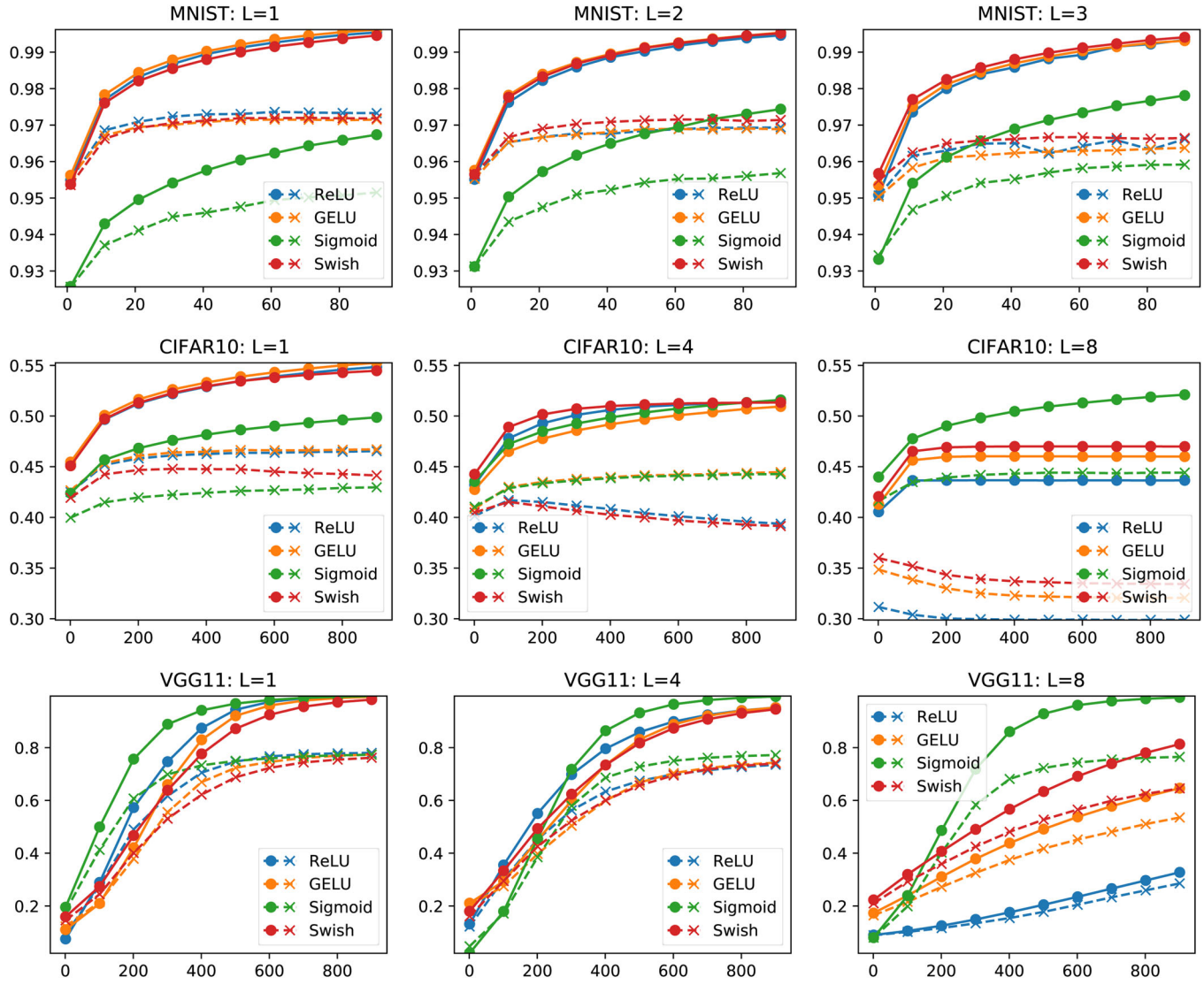

Figure 10. Accuracy for varying activation functions and depths: solid lines represent the training accuracy and dashed lines represent the testing accuracy. We used 100 epochs (lr $= 10^{-2}$), 1000 epochs (lr $= 10^{-3}$), and 1000 epochs (lr $= 10^{-7}$) for MNIST, CIFAR10, and VGG11, respectively. The number of random features in every experiment was set to $m = 2048$ (yielding 10×2048 trainable parameters).

which further leads to a smaller implicit regularization due to truncation at level ι. In contrast, ReLU activation has a much smaller component for the first ι Hermite coefficients, and therefore the condition number decays much faster. With the Sigmoid activation function, the compositional kernel can tolerate much higher depths such that the lower order Hermite coefficients do not vanish.

For each dataset, we run multi-class logistic regression (a simple one-layer neural network) with 10 categories, using the random features generated by Algorithm 2, with truncation level $\iota = 20$. We consider four activations and three compositional depths, in total 12 experiments. We run vanilla stochastic gradient descent as the training algorithm with batches of size 256. The results are plotted in Figure 10 and the numerical results are displayed in Table 3. We remark that only for the Sigmoid activation, the compositional kernels help improve the test accuracy. We postulate that this is because of the slower decay of the lower order Hermite coefficients of the compositional kernel, allowing one to vary the depth L for a favorable trade-off between memorization and generalization.

Supplementary Materials

The proofs of Theorems and Lemmas are designated to the appendix, which is included in the supplementary materials.

Acknowledgments

We thank the anonymous referees and editors for the constructive feedback that significantly improved our article.

Funding

Liang acknowledges the support of George C. Tiao faculty fellowship and thanks Max H. Farrell for early comments on a draft of the article. Tran-Bach would like to thank the Neubauer Family Initiative and Radix Trading LLC for supporting him in his graduate studies.

References

Atkinson, K. E., and Han, W. (2012), *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, Lecture Notes in Mathematics (Vol. 2044), Heidelberg: Springer. [7]

Bach, F. (2016), "Breaking the Curse of Dimensionality With Convex Neural Networks," arXiv no. 1412.8690. [2]

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019), "Benign Overfitting in Linear Regression," arXiv no. 1906.11300. [5]

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2018), "Reconciling Modern Machine Learning and the Bias-Variance Trade-off," arXiv no. 1812.11118. [5]

Belkin, M., Hsu, D., and Xu, J. (2019), "Two Models of Double Descent for Weak Features," arXiv no. 1903.07571. [5]

Belkin, M., Ma, S., and Mandal, S. (2018), "To Understand Deep Learning We Need to Understand Kernel Learning," arXiv no. 1802.01396. [1,2]

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018), "Does Data Interpolation Contradict Statistical Optimality?," arXiv no. 1806.09471. [5]

Caponnetto, A., and De Vito, E. (2007), "Optimal Rates for Regularized Least-Squares Algorithm," *Foundations of Computational Mathematics*, 7, 331–368. [2,8]

Chizat, L., and Bach, F. (2018a), "A Note on Lazy Training in Supervised Differentiable Programming," arXiv no. 1812.07956. [2]

——— (2018b), "On the Global Convergence of Gradient Descent for Over-Parameterized Models Using Optimal Transport," arXiv no. 1805.09545. [2]

Cho, Y., and Saul, L. K. (2009), "Kernel Methods for Deep Learning," in *Advances in Neural Information Processing Systems* (Vol. 22), eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Curran Associates, Inc., pp. 342–350. [1,2]

Daniely, A., Frostig, R., Gupta, V., and Singer, Y. (2017), "Random Features for Compositional Kernels," arXiv no. 1703.07872. [1,2,4]

Daniely, A., Frostig, R., and Singer, Y. (2017), "Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity," arXiv no. 1602.05897. [1,2,4,5]

Dou, X., and Liang, T. (2020), "Training Neural Networks as Learning Data-Adaptive Kernels: Provable Representation and Approximation Benefits," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2020.1745812. [2]

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018), "Gradient Descent Provably Optimizes Over-Parameterized Neural Networks," arXiv no. 1810.02054. [2]

Feldman, V. (2019), "Does Learning Require Memorization? A Short Tale About a Long Tail," arXiv no. 1906.05271. [5]

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019), "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," arXiv no. 1903.08560. [5]

Jacot, A., Gabriel, F., and Hongler, C. (2019), "Freeze and Chaos for DNNs: An NTK View of Batch Normalization, Checkerboard and Boundary Effects," arXiv no. 1907.05715. [2]

Kar, P., and Karnick, H. (2012), "Random Feature Maps for Dot Product Kernels," arXiv no. 1201.6530. [2,8]

Kesten, H., and Stigum, B. P. (1966), "A Limit Theorem for Multidimensional Galton-Watson Processes," *The Annals of Mathematical Statistics*, 37, 1211–1223, DOI: 10.1214/aoms/1177699266. [4]

Liang, T., and Rakhlin, A. (2020), "Just Interpolate: Kernel 'Ridgeless' Regression Can Generalize," *The Annals of Statistics*, 48, 1329–1347, DOI: 10.1214/19-AOS1849. [5,8]

Liang, T., Rakhlin, A., and Zhai, X. (2020), "On the Multiple Descent of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels," in *Proceedings of 33rd Conference on Learning Theory*, Proceedings of Machine Learning Research (Vol. 125), eds. J. Abernethy and S. Agarwal, PMLR, pp. 2683–2711. [5,8]

Liang, T., and Sur, P. (2020), "A Precise High-Dimensional Asymptotic Theory for Boosting and Min-ℓ_1-Norm Interpolated Classifiers," arXiv no. 2002.01586. [5]

Lyons, R., and Peres, Y. (2016), *Probability on Trees and Networks*, Cambridge: Cambridge University Press. [3]

Mei, S., and Montanari, A. (2019), "The Generalization Error of Random Features Regression: Precise Asymptotics and Double Descent Curve," arXiv no. 1908.05355. [2]

Mei, S., Montanari, A., and Nguyen, P.-M. (2018), "A Mean Field View of the Landscape of Two-Layers Neural Networks," arXiv no. 1804.06561. [2]

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2020), "The Generalization Error of Max-Margin Linear Classifiers: High-Dimensional Asymptotics in the Overparametrized Regime," arXiv no. 1911.01544. [5]

Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. (2020), "Optimal Regularization Can Mitigate Double Descent," arXiv no. 2003.01897. [5]

Neal, R. M. (1996a), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics (Vol. 118), New York: Springer . [1,2]

——— (1996b), "Priors for Infinite Networks," in *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, ed. R. M. Neal, New York: Springer, pp. 29–53. [2]

Nguyen, P.-M., and Pham, H. T. (2020), "A Rigorous Framework for the Mean Field Limit of Multilayer Neural Networks," arXiv no. 2001.11443. [2]

Pennington, J., Xinnan, F., Yu, X., and Kumar, S. (2015), "Spherical Random Features for Polynomial Kernels," in *Advances in Neural Information Processing Systems* (Vol. 28), eds. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Curran Associates, Inc., pp. 1846–1854. [2]

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016), "Exponential Expressivity in Deep Neural Networks Through Transient Chaos," in *Advances in Neural Information Processing Systems* (Vol. 29), eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Curran Associates, Inc., pp. 3360–3368. [1,2,5]

Rahimi, A., and Recht, B. (2008), "Random Features for Large-Scale Kernel Machines," in *Advances in Neural Information Processing Systems* (Vol. 20), eds. J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Curran Associates, Inc., pp. 1177–1184. [1,2,8,9]

——— (2009), "Weighted Sums of Random Kitchen Sinks: Replacing Minimization With Randomization in Learning," in *Advances in Neural Information Processing Systems* (Vol. 21), Curran Associates, Inc., pp. 1313–1320. [1,2,8,9]

Rotskoff, G. M., and Vanden-Eijnden, E. (2018), "Trainability and Accuracy of Neural Networks: An Interacting Particle System Approach," arXiv no. 1805.00915. [2]

Schoenberg, I. J. (1942), "Positive Definite Functions on Spheres," *Duke Mathematical Journal*, 9, 96–108, DOI: 10.1215/S0012-7094-42-00908-6. [8]

Shankar, V., Fang, A., Guo, W., Fridovich-Keil, S., Schmidt, L., Ragan-Kelley, J., and Recht, B. (2020), "Neural Kernels Without Tangents," arXiv no. 2003.02237. [1,2]

Sirignano, J., and Spiliopoulos, K. (2018), "Mean Field Analysis of Neural Networks: A Law of Large Numbers," *SIAM Journal on Applied Mathematics*, 80, 725–752. [2]

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. (1999), "Support Vector Regression With ANOVA Decomposition Kernels," in *Advances in Kernel Methods: Support Vector Learning*, pp. 285–291. [1]

Woodworth, B., Gunasekar, S., Savarese, P., Moroshko, E., Golan, I., Lee, J., Soudry, D., and Srebro, N. (2019), "Kernel and Rich Regimes in Overparametrized Models," arXiv no. 1906.05827. [2]

Yang, G. (2019), "Scaling Limits of Wide Neural Networks With Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation," arXiv no. 1902.04760. [2]

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017), "Understanding Deep Learning Requires Rethinking Generalization," arXiv no. 1611.03530. [1,5]