Journal of Machine Learning Research 24 (2023) 1-30 Submitted 4/23; Revised 6/23; Published 12/23

Hierarchical Kernels in Deep Kernel Learning

Wentao Huang HUANGWT55@MAIL2.SYSU.EDU.CN

Houbao Lu LUHB6@MAIL2.SYSU.EDU.CN

Haizhang Zhang[*] ZHHAIZH2@MAIL.SYSU.EDU.CN

School of Mathematics (Zhuhai)
Sun Yat-sen University
Zhuhai 519082, P. R. China

Editor: Lorenzo Rosasco

Abstract

Kernel methods are built upon the mathematical theory of reproducing kernels and reproducing kernel Hilbert spaces. They enjoy good interpretability thanks to the solid mathematical foundation. Recently, motivated by deep neural networks in deep learning, which construct learning functions by successive compositions of activation functions and linear functions, a class of methods termed as deep kernel learning has appeared in the literature. The core of deep kernel learning is hierarchical kernels that are constructed from a base reproducing kernel by successive compositions. In this paper, we characterize the corresponding reproducing kernel Hilbert spaces of hierarchical kernels, and study conditions ensuring that the reproducing kernel Hilbert space will be expanding as the layer of hierarchical kernels increases. The results will answer whether the expressive power of hierarchical kernels will be improving as the layer increases, and give guidance to the construction of hierarchical kernels for deep kernel learning.

Keywords: hierarchical kernels, reproducing kernels, deep learning, compositional kernels, reproducing kernel Hilbert spaces

1. Introduction

Machine learning has played an important role in recent advances of artificial intelligence. There are two major categories of learning methods: the classical kernel methods (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998) and deep learning methods (Goodfellow et al., 2016; LeCun et al., 2015). In many applications, the target of these two kinds of learning methods is the same, which is to learn a prediction function from given training data. The target can be approached by minimizing a functional of the form:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^{n} L(f(x_j), y_j), \tag{1}$$

subject to a constraint on the complexity of model f, where $(x_j, y_j), 1 \le j \le n$ are prescribed training data from $X \times Y$ with X being the input space and Y being the output space, L is a chosen loss function, and $\mathcal{F}$ is a set of candidate prediction functions. An essential

[*] Corresponding author

difference between kernel learning and deep learning lies in the choice of $\mathcal{F}$. In the classical kernel methods, $\mathcal{F}$ is generated via a reproducing kernel K (also called a Mercer kernel (Mercer, 1909)) on X by

$$\mathcal{F} = \left\{ \sum_{j=1}^{n} c_j K(x_j, \cdot) : c_j \in \mathbb{R}, 1 \leq j \leq n \right\}.$$

In other words, the candidate functions are linear combinations of the reproducing kernel at the sampling points x_j, $1 \leq j \leq n$. In deep learning methods, $\mathcal{F}$ is generated from a deep neural network by consecutive compositions of linear functions and an activation function. Also, some other techniques such as pooling and batch normalization are adopted in deep neural networks.

A great advantage of kernel learning is that it is built on the solid mathematical foundation of reproducing kernels and reproducing kernel spaces. In fact, long before the emergence of machine learning, renowned mathematicians including Aronszajn (Aronszajn, 1950), Bochner (Bochner, 1959), and Schoenberg (Schoenberg, 1938, 1942) had been studying positive-definite functions. These functions were latter found to be identical to reproducing kernels. Reproducing kernels and reproducing kernel spaces have been extensively studied since then (see, for example, Berlinet and Thomas-Agnan (2004); Cucker and Smale (2002); Cucker and Zhou (2007); Evgeniou et al. (2000); Fukumizu et al. (2004); FitzGerald et al. (1995); Wendland (2005); Wu (1995); Zhang et al. (2009) and references therein). Such a solid mathematical foundation endows good interpretability of kernel methods. For instance, we are able to analyze the generalization ability of many kernel methods by estimating the learning rates from an approximation theory viewpoint (Cucker and Smale, 2002; Cucker and Zhou, 2007).

A disadvantage of kernel methods is that they cannot well handle many challenging learning tasks. Recently, motivated by the success of deep learning, a class of learning methods termed as deep kernel learning has appeared in the literature (Anselmi et al., 2015; Bohn et al., 2019; Chen et al., 2017; Cho and Saul, 2009; Wilson et al., 2016). The essence of deep kernel learning is the usage of hierarchical kernels (also known as compositional kernels) that are constructed from existing kernels by successive function compositions. Such constructions are mainly motivated by composition structure of deep neural networks. Function compositions are able to generate high-dimensional complicated functions from relatively low-dimensional simple functions. This can be explained by the well-known Kolmogorov-Arnold representation theorem in mathematics (Morris, 2021). The applications and implications of the Kolmogorov-Arnold representation theorem to neural networks are discussed in Schmidt-Hieber (2021); Yarotsky (2017).

Deep kernel learning based on hierarchical kernels have shown comparable performances in a few challenging learning problems (Chen et al., 2017; Cho and Saul, 2009; Wilson et al., 2016) . However, two theoretical questions remain unanswered for these hierarchical kernels. The first question concerns the understanding of the reproducing kernel Hilbert spaces of hierarchical kernels. The second question is about what conditions ensure that the expressive power of the hierarchical kernels increases as the number of layers of hierarchical kernels increases. In this paper, we aim to answer these two questions for a type of hierarchical kernels generated by

$$K_0 = K, \quad K_n = g(K_{n-1}),\ n \in \mathbb{N},$$

where the initial kernel K is a commonly-used kernel in machine learning (such as the Gaussian kernel, the exponential kernels, or a polynomial kernel), and g is a chosen univariate function such as the exponential function or a fixed polynomial. Our objective is to characterize the reproducing kernel Hilbert space $\mathcal{H}_{K_n}$ of each K_n, and to investigate conditions ensuring that $\mathcal{H}_{K_{n+1}}$ is strictly larger than $\mathcal{H}_{K_n}$. These results will contribute to the mathematical foundation of deep kernel learning.

The rest of the paper is organized as follows. In Section 2, we introduce some basic facts about reproducing kernels and reproducing kernel Hilbert spaces. Note that our study will build upon the existing results on inclusion relation of the reproducing kernel Hilbert spaces. In Section 3, we present some general results on compositional kernels. Sections 4-6 are devoted to hierarchical kernels generated from the compositions of the Gaussian kernel, the exponential kernel, the polynomial kernel with the exponential function or a fixed polynomial, respectively. Initial experiments on hierarchical kernels are conducted in Section 7 and the paper is concluded in Section 8. We shall see that the main analysis of hierarchical kernels is closely related to the high order Bell numbers. Moreover, the answer to the second question is not always affirmative, as we shall see in the case of hierarchical exponential kernels that the related reproducing kernel Hilbert spaces remain unchanged as the number of layers increases, indicating that one should not use the exponential kernels in deep kernel learning.

Before we enter the formal investigation, we would like to discuss the differences between our mathematical study on hierarchical kernels and those studies aiming to give explanation to deep neural networks by kernels. Such studies include understanding the training process and mechanism of generalization of deep neural networks by neural tangent kernels (see, for example, Cho and Saul (2009); Daniely et al. (2016); Huang and Yau (2020); Huang et al. (2021); Lee et al. (2018); Jacot et al. (2018); Neal (1996)). Compositional kernels related to neural networks and neural tangent kernels are investigated in Bietti and Bach (2021); Chen and Xu (2021); Geifman et al. (2020). These kernels are dot-product kernels on the sphere and are different from the hierarchical kernels on the Euclidean spaces in this paper. There is another line of of recent work aiming to understand shallow neural networks by reproducing kernel Banach spaces (Bach, 2017; Bartolucci et al., 2021; Ongie et al., 2019; Parhi and Nowak, 2021; Spek et al., 2022). In particular, a function composition structure in Banach spaces was proposed in Parhi and Nowak (2022) to understand the functions learned by deep neural networks.

The main purpose of the above researches is to give explanation to neural networks by neural tangent kernels or by reproducing kernel Banach spaces. Hierarchical kernels in the current work are constructed from existing kernels by successive function compositions. Such constructions are stimulated by the composition structure of deep neural networks. Hierarchical kernels can then be used for feature extraction or used in classical kernel methods. Hierarchical kernels are not to compete with deep learning. Therefore, the theme of the paper is different from those in the former paragraph. In the future, we plan to investigate compositional kernels on the sphere and hierarchical kernels involving compositions with multiple kernels.

2. Preliminaries

In this paper, we use $\mathbb{R}, \mathbb{C}, \mathbb{N}, \mathbb{Z}, \mathbb{Z}_+$ to denote the set of real numbers, the set of complex numbers, the set of positive integers, the set of integers, and the set of nonnegative integers, respectively.

Let X be a prescribed input space. A *reproducing kernel* (or kernel for short) K on X is a function from $X \times X$ to $\mathbb{C}$ such that for all finite pairwise distinct inputs $x_j \in X$, $1 \leq j \leq n$, the kernel matrix

$$[K(x_j, x_k) : 1 \leq j, k \leq n],$$

is hermitian and positive semi-definite. A reproducing kernel K on X corresponds to a unique reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}_K$, such that $K(x, \cdot) \in \mathcal{H}_K$ for all $x \in X$ and

$$f(x) = \langle f, K(x, \cdot)\rangle_{\mathcal{H}_K} \quad \text{for all } f \in \mathcal{H}_K,\ x \in X, \tag{2}$$

where $\langle \cdot, \cdot\rangle_{\mathcal{H}_K}$ denotes the inner product on $\mathcal{H}_K$.

Let g be a chosen univariate function. We shall concentrate on characterizing $\mathcal{H}_{g(K)}$ and examining the inclusion relationship between $\mathcal{H}_K$ and $\mathcal{H}_{g(K)}$ in the paper. To this end, we will recall some useful notations and related results in this section.

Given two kernels K, G on X, the inclusion relation $\mathcal{H}_K \subseteq \mathcal{H}_G$ was first investigated by Aronszajn in Aronszajn (1950) and then extensively studied in Xu and Zhang (2007, 2009); Zhang and Zhao (2013). The following result is well-known. **Denote $K \ll G$ if $G - K$ remains a kernel on X.**

Lemma 1 *(Aronszajn, 1950) Let K, G be two kernels on X. Then $\mathcal{H}_K \subseteq \mathcal{H}_G$ if and only if there exists a non-negative constant λ such that $K \ll \lambda G$.*

When $\mathcal{H}_K \subseteq \mathcal{H}_G$, it was observed in Aronszajn (1950) by the closed graph theorem that the identity operator from $\mathcal{H}_K$ into $\mathcal{H}_G$ is bounded. The operator norm of this embedding is denoted by $\beta(K, G)$, (Zhang and Zhao, 2013). We also denote

$$\lambda(K, G) = \inf\{\lambda \geq 0 | K \ll \lambda G\}.$$

The relation between these two important constants was discovered in Zhang and Zhao (2013).

Lemma 2 *(Zhang and Zhao, 2013) When $\mathcal{H}_K \subseteq \mathcal{H}_G$, it holds $\beta(K, G) = \sqrt{\lambda(K, G)}$ and $K \ll \lambda(K, G)G$.*

The following result from Aronszajn (1950) will also be needed. Denote by $\|\cdot\|_{\mathcal{B}}$ the norm on a Banach space $\mathcal{B}$.

Lemma 3 *(Aronszajn, 1950) Let K_1, K_2 be two kernels on X and $K = K_1 + K_2$. Then*

$$\mathcal{H}_K = \{f_1 + f_2 :\ f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K_2}\},$$

and

$$\|f\|^2_{\mathcal{H}_K} = \inf\{\|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}} : f = f_1 + f_2,\ f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K_2}\}.$$

In particular, if additionally, $\mathcal{H}_{K_1} \cap \mathcal{H}_{K_2} = \{0\}$, then

$$\|f_1 + f_2\|^2_{\mathcal{H}_K} = \|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}},\ f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K_2}.$$

One of the most commonly-used classes of reproducing kernels on the Euclidean spaces is the class of translation-invariant kernels. A kernel K on $\mathbb{R}^d$ is said to be *translation-invariant* if

$$K(x-a, y-a) = K(x,y) \text{ for all } x, y, a \in \mathbb{R}^d.$$

It is clear that K on $\mathbb{R}^d \times \mathbb{R}^d$ is translation-invariant if and only if there exists a function k on $\mathbb{R}^d$ such that

$$K(x,y) = k(x-y), \quad x, y \in \mathbb{R}^d.$$

A celebrated result due to Bochner states that continuous translation-invariant kernels on $\mathbb{R}^d$ are exactly the Fourier transform of finite positive Borel measures on $\mathbb{R}^d$. The Fourier transform and its inverse are defined by

$$\hat{f}(\xi) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\xi} dx, \quad \xi \in \mathbb{R}^d,$$

and

$$\check{f}(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} f(\xi) e^{ix\cdot\xi} d\xi, \quad x \in \mathbb{R}^d,$$

where $x \cdot \xi$ is the standard inner product of x and ξ on $\mathbb{R}^d$. We shall later use $\|x\| = \sqrt{x \cdot x}$ to denote the Euclidean norm on $\mathbb{R}^d$.

Denote by $\mathcal{B}(\mathbb{R}^d)$ the set of finite positive Borel measures on $\mathbb{R}^d$. By the Bochner theorem (Bochner, 1959), a continuous function K on $\mathbb{R}^d \times \mathbb{R}^d$ is a translation-invariant kernel on $\mathbb{R}^d$ if and only if there exists a $\mu \in \mathcal{B}(\mathbb{R}^d)$ such that

$$K(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi}\, d\mu(\xi), \quad x, y \in \mathbb{R}^d. \tag{3}$$

For two translation-invariant kernels K, G on $\mathbb{R}^d$, the inclusion relation $\mathcal{H}_K \subseteq \mathcal{H}_G$ was extensively studied in Zhang and Zhao (2013). We shall need two of the results. By the Jordan decomposition of measures, every measure $\mu \in \mathcal{B}(\mathbb{R}^d)$ can be factored into the sum of two positive measures μ_c and μ_s, which are absolutely continuous and singular with respect to the Lebesgue measure, respectively. For the absolutely continuous measure μ_c, the Radon-Nikodym theorem ensures the existence of a nonnegative function $u \in L^1(\mathbb{R}^d)$ such that

$$\mu_c(A) = \int_A u(\xi) d\xi \ \text{ for every Lebegue measurable } A \subseteq \mathbb{R}^d.$$

Consequently, K defined by (3) can be factored as

$$K = K_c + K_s,$$

where

$$K_c(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi}\, d\mu_c(\xi), \quad K_s(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} d\mu_s(\xi). \tag{4}$$

The two required results are as follows.

Lemma 4 *(Zhang and Zhao, 2013) Let u_c, u_s be two nonnegative Borel measures on $\mathbb{R}^d$ that are absolutely continuous and singular with respect to the Lebesgue measure, respectively. And let K_c, K_s be the associated translation-invariant kernels given by (4). Then $\mathcal{H}_{K_c} \cap \mathcal{H}_{K_s} = \{0\}$.*

Lemma 5 *(Zhang and Zhao, 2013) Let u, v be nonnegative functions in $L^1(\mathbb{R}^d)$ and let K, G be defined by*

$$K(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} u(\xi) d\xi, \quad G(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} v(\xi) d\xi, \quad x, y \in \mathbb{R}^d. \tag{5}$$

Then $\mathcal{H}_K \subseteq \mathcal{H}_G$ if and only if there exists a nonnegative constant C such that $u(\xi) \le Cv(\xi)$ almost everywhere on $\mathbb{R}^d$, in which case

$$\lambda(K,G) = \|u/v\|_{L^\infty(\mathbb{R}^d)}.$$

We next introduce a characterization of the RKHS of a translation-invariant kernel. For a nonnegative function $u \in L^1(\mathbb{R}^d)$, denote by $L^2_u(\mathbb{R}^d)$ the Hilbert space of Borel measurable functions f on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} |f(t)|^2 u(t) dt < +\infty.$$

The inner product and norm on $L^2_u(\mathbb{R}^d)$ are given by

$$\langle f, g\rangle_{L^2_u(\mathbb{R}^d)} = \int_{\mathbb{R}^d} f(t)\overline{g(t)} u(t) dt, \quad \|f\|_{L^2_u(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} |f(t)|^2 u(t) dt\right)^{1/2}.$$

Lemma 6 *(Wendland, 2005) Let u be a nonnegative functions in $L^1(\mathbb{R}^d)$ and let K be defined by*

$$K(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} u(\xi) d\xi, \quad x, y \in \mathbb{R}^d. \tag{6}$$

Then

$$\begin{aligned}\mathcal{H}_K &= \left\{ f(x) = \int_{\mathbb{R}^d} f_u(t) e^{ix\cdot\xi} u(\xi) d\xi : \ f_u \in L^2_u(\mathbb{R}^d)\right\} \\ &= \left\{ f \in C(\mathbb{R}^d) : \ \int_{\mathbb{R}^d} \frac{|\hat{f}(\xi)|^2}{u(\xi)} d\xi < +\infty \right\},\end{aligned}$$

with inner product

$$\langle f, g\rangle_{\mathcal{H}_K} = \langle f_u, g_u\rangle_{L^2_u(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \frac{\hat{f}(\xi)\overline{\hat{g}(\xi)}}{u(\xi)} d\xi.$$

Another important class of reproducing kernels in machine learning is the polynomial kernel, which is a special form of the Hilbert-Schmidt kernel. By Mercer's theorem (Mercer, 1909), any continuous kernel on a compact metric space is a Hilbert-Schmidt kernel. For this sake, we shall first present the general form of Hilbert-Schmidt kernels.

Let a be a nonnegative function on $\mathbb{N}$ and set $a_n := a(n)$, $n \in \mathbb{N}$. Denote by $\ell^2_a(\mathbb{N})$ the Hilbert space of functions c on $\mathbb{N}$ such that

$$\|c\|_{\ell^2_a(\mathbb{N})} := \left(\sum_{n=1}^{\infty} a_n |c_n|^2\right)^{1/2} < +\infty.$$

The inner product on $\ell^2_a(\mathbb{N})$ is

$$\langle c, d\rangle_{\ell^2_a(\mathbb{N})} := \sum_{n=1}^{\infty} a_n c_n \overline{d_n}, \quad c, d \in \ell^2_a(\mathbb{N}).$$

Suppose that ϕ_n, $n \in \mathbb{N}$ is a sequence of functions on the input space X, such that for each $x \in X$ the function $\Phi(x)$ defined on $\mathbb{N}$ as

$$\Phi(x)(n) := \phi_n(x), \quad n \in \mathbb{N}, \tag{7}$$

belongs to $\ell^2_a(\mathbb{N})$. The Hilbert-Schmidt kernel K_a associated with a is given as

$$K_a(x, y) := (\Phi(x), \Phi(y))_{\ell^2_a(\mathbb{N})} = \sum_{n=1}^{\infty} a_n \phi_n(x)\overline{\phi_n(y)}, \quad x, y \in X. \tag{8}$$

Now suppose that there exists another nonnegative function b on $\mathbb{N}$ such that $\Phi(x) \in \ell^2_b(\mathbb{N})$ for all $x \in X$. Set

$$K_b(x, y) := (\Phi(x), \Phi(y))_{\ell^2_b(\mathbb{N})} = \sum_{n=1}^{\infty} b_n \phi_n(x)\overline{\phi_n(y)}, \quad x, y \in X. \tag{9}$$

The inclusion relation $\mathcal{H}_{K_a} \subseteq \mathcal{H}_{K_b}$ was characterized in Zhang and Zhao (2013).

Lemma 7 *Suppose that b is nontrivial, and $\operatorname{span}\{\Phi(x) : x \in X\}$ is dense in both $\ell^2_a(\mathbb{N})$ and $\ell^2_b(\mathbb{N})$. Then $\mathcal{H}_{K_a} \subseteq \mathcal{H}_{K_b}$ if and only if there is a constant $\lambda > 0$ such that $a_n \leq \lambda b_n$ for all $n \in \mathbb{N}$. In this case,*

$$\lambda(K_a, K_b) = \sup\left\{\frac{a_n}{b_n} : n \in \mathbb{N},\ b_n > 0\right\}. \tag{10}$$

We shall also need a characterization of the RKHS of a Hilbert-Schmidt kernel, which is well-known (Cucker and Smale, 2002). Note that when $\operatorname{span}\{\Phi(x) : x \in X\}$ is dense in $\ell^2_a(\mathbb{N})$, if $c \in \ell^2_a(\mathbb{N})$ satisfies

$$\sum_{n=1}^{\infty} c_n a_n \overline{\phi_n(x)} = 0 \text{ for all } x \in X,$$

then $(c, \Phi(x))_{\ell^2_a(\mathbb{N})} = 0$ for all $x \in X$, which implies by the denseness condition that $c = 0$ in $\ell^2_a(\mathbb{N})$.

Lemma 8 *Let K_a be the kernel defined by (8) and suppose that $\operatorname{span}\{\Phi(x) : x \in X\}$ is dense in $\ell^2_a(\mathbb{N})$. Then*

$$\mathcal{H}_{K_a} = \left\{ f_c(x) := (c, \Phi(x))_{\ell^2_a(\mathbb{N})} = \sum_{n=1}^{\infty} c_n a_n \overline{\phi_n(x)}, \quad x \in X,\ c \in \ell^2_a(\mathbb{N}) \right\},$$

with the inner product

$$\langle f_c, f_d\rangle_{\mathcal{H}_{K_a}} = \langle c, d\rangle_{\ell^2_a(\mathbb{N})}, \quad c, d \in \ell^2_a(\mathbb{N}).$$

3. General Characterizations

Let K be a kernel on an input space X. The hierarchical kernels constructed from K are defined recursively via compositing with a function g on $\mathbb{R}$ by

$$K_n(x,y) = g(K_{n-1}(x,y)), \quad x,y \in X,\ n \geq 1,$$

where $K_0 := K$. The first question is for what g, K_n would always remain a reproducing kernel. This can be answered by a classical result on positive-definite functions (FitzGerald et al., 1995).

Lemma 9 *Let g be a function on $\mathbb{C}$. Then for any reproducing kernel K, $g(K)$ remains a reproducing kernel if and only if g is holomorphic on $\mathbb{C}$ and all the coefficients in its Maclaurin series are nonnegative.*

By the above lemma, typical choices of g in deep kernel methods for machine learning including $g(x) = e^x$ and $g(x) = P(x)$, where P is a polynomial with nonnegative coefficients. We first investigate such hierarchical kernels. A couple of simple observations are in order. We shall need the well-known fact that the product of two reproducing kernels remains a reproducing kernel. This follows directly from the Schur product theorem (Horn and Johnson, 1991) that the component-wise product (Hadamard product) of two positive semi-definite matrices is still a positive semi-definite matrix.

Proposition 10 *Let g be a holomorphic function on $\mathbb{C}$ of the form*

$$g(z) = \sum_{n=0}^{\infty} a_n z^n, \quad z \in \mathbb{C}, \quad a_n \geq 0,\ n \in \mathbb{Z}, \tag{11}$$

where $a_1 > 0$. Then $\mathcal{H}_K \subseteq \mathcal{H}_{g(K)}$. In particular, $\mathcal{H}_K \subseteq \mathcal{H}_{e^K}$.

Proof By Lemma 9, $g(K) - a_1 K$ is a kernel on X. Then by Lemma 1,

$$\mathcal{H}_K = \mathcal{H}_{a_1 K} \subseteq \mathcal{H}_{g(K)}.$$

Clearly, the exponential function satisfies the requirements on g. ■

Theorem 11 *Let K, G be two kernels on X such that $\mathcal{H}_K \subseteq \mathcal{H}_G$, and let g be given by (11). Then $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(\lambda G)}$, where $\lambda = \lambda(K,G)$. In particular, if $\lambda(K,G) \leq 1$ then $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$.*

Proof As $\mathcal{H}_K \subseteq \mathcal{H}_G$, by Lemma 2, $K \ll \lambda G$ where $\lambda = \lambda(K,G)$. Then there exists a kernel L such that $\lambda G = K + L$. By (11),

$$\begin{aligned} g(\lambda G) &= \sum_{n=0}^{\infty} a_n (K+L)^n \\ &= \sum_{n=0}^{\infty} a_n K^n + \sum_{n=1}^{\infty} a_n \sum_{j=0}^{n-1} \binom{n}{j} K^j L^{n-j} \\ &= g(K) + \sum_{n=1}^{\infty} a_n \sum_{j=0}^{n-1} \binom{n}{j} K^j L^{n-j}. \end{aligned}$$

Since the product of two kernels remains a kernel, each of the K^jL^{n-j} is a kernel. Therefore, $g(K) \ll g(\lambda G)$. By Lemma 1, $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(\lambda G)}$. When $\lambda(K,G) \leq 1$, one can choose $\lambda = 1$ and apply the above arguments to show that $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$. ■

Corollary 12 *If g given by (11) is a polynomial and $\mathcal{H}_K \subseteq \mathcal{H}_G$ then $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$.*

Proof If $\lambda(K,G) \leq 1$ then the result is true by Theorem 11. Now suppose

$$g(z) = \sum_{n=0}^{m} a_n z^n, \quad a_n \geq 0,\ 0 \leq n \leq m,$$

and $\lambda = \lambda(K,G) \geq 1$. Then we have

$$g(K) = \sum_{n=0}^{m} a_n K^n \ll \sum_{n=0}^{m} a_n (\lambda G)^n \ll \sum_{n=0}^{m} a_n \lambda^m G^n = \lambda^m g(G),$$

which implies that $g(K) \ll \lambda^m g(G)$. By Lemma 1, $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$. ■

We make two remarks about the above results. The first one is that when g is not a polynomial then $\mathcal{H}_K \subseteq \mathcal{H}_G$ may not imply $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$. To see a counterexample, we shall use Lemma 7 on the inclusion relation of RKHSs of Hilbert-Schmidt kernels. Let $K(x,y) = 2xy, G(x,y) = xy$ be two polynomial kernels on $\mathbb{R}$. Then $\mathcal{H}_K = \mathcal{H}_G$. But

$$e^K = e^{2xy} = \sum_{k=0}^{\infty} \frac{2^k (xy)^k}{k!}, e^G = e^{xy} = \sum_{k=0}^{\infty} \frac{(xy)^k}{k!}.$$

Clearly,

$$\sup\left\{\frac{\frac{2^k}{k!}}{\frac{1}{k!}} = 2^k : k \in \mathbb{N}\right\} = +\infty,$$

which implies by Lemma 7 that $\mathcal{H}_{e^K} \nsubseteq \mathcal{H}_{e^G}$.

The second remark is that $\lambda(K,G) \leq 1$ is not necessary to imply $\mathcal{H}_{g(K)} \subseteq \mathcal{H}_{g(G)}$ from $\mathcal{H}_K \subseteq \mathcal{H}_G$ even when g is not a polynomial. An example is given by the following two kernels

$$K(x,y) = e^{-(x-y)^2} = \int_{\mathbb{R}} u(\xi) e^{-i(x-y)\xi} d\xi,\ G(x,y) = e^{-|x-y|} = \int_{\mathbb{R}} v(\xi) e^{-i(x-y)\xi} d\xi,\ x,y \in \mathbb{R},$$

where

$$u(\xi) = \frac{1}{2\sqrt{\pi}} e^{-\frac{\xi^2}{4}}, v(\xi) = \frac{1}{\pi(1+\xi^2)}.$$

We compute that

$$\lambda(K,G) = \sup_{\xi \in \mathbb{R}} \frac{u(\xi)}{v(\xi)} = \frac{\sqrt{\pi}}{2} \sup_{\xi \in \mathbb{R}} \frac{\xi^2+1}{e^{\frac{\xi^2}{4}}} = 2\sqrt{\pi} e^{-\frac{3}{4}} > 1.$$

By Lemma 5, $\mathcal{H}_K \subseteq \mathcal{H}_G$. Now consider

$$e^{K(x,y)} = 1 + \sum_{k=1}^{\infty} \frac{e^{-k(x-y)^2}}{k!} := 1 + K_1(x,y) := 1 + \int_{\mathbb{R}} e^{-i(x-y)\xi} u_1(\xi) d\xi,$$

and

$$e^{G(x,y)} = 1 + \sum_{k=1}^{\infty} \frac{e^{-k|x-y|}}{k!} := 1 + G_1(x,y) := 1 + \int_{\mathbb{R}} e^{-i(x-y)\xi} v_1(\xi) d\xi,$$

where

$$u_1(\xi) = \frac{1}{2\sqrt{\pi}} \sum_{k=1}^{\infty} \frac{1}{k!\sqrt{k}} e^{-\frac{\xi^2}{4k}}, v_1(\xi) = \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{1}{k!} \frac{k}{\xi^2 + k^2}.$$

It holds

$$\begin{aligned} \lambda(K_1, F_1) = \sup_{\xi \in \mathbb{R}} \frac{u_1(\xi)}{v_1(\xi)} &= \frac{\sqrt{\pi}}{2} \sup_{\xi \in \mathbb{R}} \frac{\sum_{k=1}^{\infty} \frac{1}{k!\sqrt{k}} e^{-\frac{\xi^2}{4k}}}{\sum_{k=1}^{\infty} \frac{1}{k!} \frac{k}{\xi^2+k^2}} \\ &\leq \frac{\sqrt{\pi}}{2} \sup_{\xi \in \mathbb{R}} \frac{\sum_{k=1}^{\infty} \frac{1}{k!\sqrt{k}} \frac{1+\xi^2}{e^{\frac{\xi^2}{4k}}}}{\sum_{k=1}^{\infty} \frac{1}{k!k}} \leq \frac{4\sqrt{\pi}}{e} \frac{\sum_{k=1}^{\infty} e^{\frac{1}{4k}} \frac{1}{(k-1)!\sqrt{k}}}{\sum_{k=1}^{\infty} \frac{1}{k!k}} < +\infty. \end{aligned}$$

By Lemma 5, $\mathcal{H}_{e^K} \subseteq \mathcal{H}_{e^G}$ despite that $\lambda(K, G) > 1$.

4. Hierarchical Gaussian Kernels

The purpose of this section is to study the characterization and inclusion relation of RKHSs corresponding to hierarchical kernels generated from composition of the Gaussian kernel and a fixed function g. Popular choices of g including the exponential function and a polynomial. We start with the exponential function.

4.1 Composition with Exponential Function

The Gaussian kernel is given by

$$G_0(x,y) = \exp(-\lambda \|x-y\|^2) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} g_0(\xi) d\xi, \quad x, y \in \mathbb{R}^d, \quad \lambda > 0, \tag{12}$$

where

$$g_0(\xi) := \frac{1}{(2\sqrt{\lambda\pi})^d} \exp\left(- \frac{\|\xi\|^2}{4\lambda} \right), \quad \xi \in \mathbb{R}^d. \tag{13}$$

The hierarchical Gaussian kernels from composition with the exponential function are given by

$$G_n(x,y) = \exp\left(G_{n-1}(x,y)\right), \quad n \in \mathbb{N}. \tag{14}$$

In particular,

$$G_1(x,y) = \exp(G_0(x,y)) = \sum_{k=0}^{\infty} \frac{G_0(x,y)^k}{k!} := 1 + K_1(x,y), \quad x, y \in \mathbb{R}^d, \tag{15}$$

where

$$K_1(x,y) := \sum_{k=1}^{\infty} \frac{\exp(-\lambda k\|x-y\|^2)}{k!}. \tag{16}$$

One computes that

$$K_1(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} g_1(\xi) d\xi, \quad x, y \in \mathbb{R}^d,$$

where

$$g_1(\xi) := \frac{1}{(2\sqrt{\lambda\pi})^d} \sum_{k=1}^{\infty} \frac{1}{k!k^{d/2}} \exp\Big(-\frac{\|\xi\|^2}{4\lambda k}\Big), \quad \xi \in \mathbb{R}^d. \tag{17}$$

Theorem 13 *It holds* $\mathcal{H}_{G_0} \subseteq \mathcal{H}_{G_1}$*, but* $\mathcal{H}_{G_1} \nsubseteq \mathcal{H}_{G_0}$.

Proof Firstly, by Lemmas 3 and 4

$$\mathcal{H}_{G_1} = \{c + f : \ c \in \mathbb{C}, f \in \mathcal{H}_{K_1}\},$$

and

$$||c+f||^2_{\mathcal{H}_{G_1}} = |c|^2 + ||f||^2_{\mathcal{H}_{K_1}}.$$

Apparently, $g_0(\xi) \leqslant g_1(\xi)$. It follows by Lemma 4 that $\mathcal{H}_{G_0} \subseteq \mathcal{H}_{K_1} \subseteq \mathcal{H}_{G_1}$.

On the other hand, one sees

$$\begin{aligned}\frac{g_1(\xi)}{g_0(\xi)} &= \frac{\sum_{k=1}^{\infty} \frac{1}{k!k^{d/2}} \exp\big(-\frac{\|\xi\|^2}{4\lambda k}\big)}{\exp\big(-\frac{\|\xi\|^2}{4\lambda}\big)} \\ &= 1 + \sum_{k=2}^{\infty} \frac{1}{k!k^{d/2}} \exp\Big[\frac{\|\xi\|^2}{4\lambda}\Big(1-\frac{1}{k}\Big)\Big] \\ &\geqslant 1 + \exp\Big(\frac{\|\xi\|^2}{8\lambda}\Big) \sum_{k=2}^{\infty} \frac{1}{k!k^{d/2}}.\end{aligned}$$

Therefore, $\frac{g_1(\xi)}{g_0(\xi)}$ is unbounded on $\mathbb{R}^d$. By Lemma 5, $\mathcal{H}_{K_1} \nsubseteq \mathcal{H}_{G_0}$. Consequently, $\mathcal{H}_{G_1} \nsubseteq \mathcal{H}_{G_0}$. ■

Next we want to study the inclusion relation between $\mathcal{H}_{G_n}$ and $\mathcal{H}_{G_{n+1}}$ for general n. By Lemma 5, the key problem is to estimate a dominating relation between the Fourier transforms of G_n and G_{n+1}. A sequence of coefficients from the exponential generating functions

$$e_0(x) = x, \quad e_n(x) = \exp\left(e_{n-1}(x)\right), \ n \in \mathbb{N}, \tag{18}$$

will play an important role. For this reason, we first define and study them (Asai et al., 2001; Bell, 1938).

Definition 14 *The* **high order Bell numbers** $\beta_{n,k}, n, k \in \mathbb{Z}_+$, *are the coefficients satisfying*

$$e_n(x) = e_n(0) \sum_{k=0}^{\infty} \frac{\beta_{n,k}}{k!} x^k, \ x \in \mathbb{R}. \tag{19}$$

Note that $\beta_{2,k}, \ k \geq 0$ *are refered as the Bell numbers.*

We next look at the properties of the high order Bell numbers. For $T \in \mathbb{Z}_+^d$, we shall denote $|T| = T_1 + T_2 + \cdots + T_d$. Also, for latter use, we set for $\beta_{n,k}$, $x \in \mathbb{R}^d$ and $T \in \mathbb{Z}_+^d$,

$$\binom{k}{T} = \frac{k!}{\prod_{i=1}^d T_i}, \ \beta_{n,T} = \prod_{i=1}^d \beta_{n,T_i}, \ x^T = x_1^{T_1} \cdots x_d^{T_d}. \tag{20}$$

Proposition 15 *The high order Bell numbers satisfy*

$$\beta_{1,k} = 1, \ k \geq 0, \quad \beta_{n,0} = 1, \ n \in \mathbb{N}, \tag{21}$$

and

$$\beta_{n+1,k} = \sum_{j=1}^{k} \frac{e_n^j(0)}{j!} \sum_{\substack{T \in \mathbb{N}^j \\ |T|=k}} \binom{k}{T} \beta_{n,T}, \ n, k \geq 1. \tag{22}$$

Proof Equation (21) is obvious. Since $e_{n+1}(x) = e_{n+1}(0) \exp(e_n(x) - e_n(0))$, we could expand $e_{n+1}(x)$ as follows

$$\begin{aligned} e_{n+1}(x) &= e_{n+1}(0) \exp\Big(e_n(x) - e_n(0) \Big) \\ &= e_{n+1}(0) \exp\Big(e_n(0) \sum_{l=1}^{\infty} \frac{\beta_{n,l}}{l!} x^l \Big) \\ &= e_{n+1}(0) \Big(1 + \sum_{j=1}^{\infty} \frac{e_n^j(0)}{j!} \Big(\sum_{l=1}^{\infty} \frac{\beta_{n,l}}{l!} x^l \Big)^j \Big). \end{aligned} \tag{23}$$

The term $(\sum_{l=1}^{\infty} \frac{\beta_{n,l}}{l!} x^l)^j$ in the last equation above can be expanded into an infinite polynomial by

$$\Big(\sum_{l=1}^{\infty} \frac{\beta_{n,l}}{l!} x^l \Big)^j = \sum_{k=j}^{\infty} a_k x^k.$$

Using the notations (20), one observes that

$$a_k = \sum_{\substack{T \in \mathbb{N}^j \\ |T|=k}} \prod_{l=1}^{j} \frac{\beta_{n,T_l}}{T_l!} = \sum_{\substack{T \in \mathbb{N}^j \\ |T|=k}} \frac{1}{k!} \frac{k!}{\prod_{l=1}^j T_l!} \beta_{n,T} = \frac{1}{k!} \sum_{\substack{T \in \mathbb{N}^j \\ |T|=k}} \binom{k}{T} \beta_{n,T}.$$

Combining the above equation with (23), we have

$$\begin{aligned} e_{n+1}(x) &= e_{n+1}(0)\Bigg(1+\sum_{j=1}^{\infty}\frac{e_n^j(0)}{j!}\Bigg(\sum_{k\geq j}^{\infty}\frac{1}{k!}\sum_{\substack{T\in\mathbb{N}^j\\|T|=k}}\binom{k}{T}\beta_{n,T}x^k\Bigg)\Bigg)\\ &= e_{n+1}(0)\Bigg(1+\sum_{k=1}^{\infty}\frac{x^k}{k!}\Bigg(\sum_{j=1}^{k}\frac{e_n^j(0)}{j!}\sum_{\substack{T\in\mathbb{N}^j\\|T|=k}}\binom{k}{T}\beta_{n,T}\Bigg)\Bigg). \end{aligned} \tag{24}$$

By comparing the coefficients in (19) and (24), we could see that (22) is true. ■

We next estimate the ratio $\beta_{n+1,k}/\beta_{n,k}$ with the results in Asai et al. (2001).

Lemma 16 *(Asai et al., 2001) The high order Bell numbers satisfy the inequality*

$$2^{-k_1-k_2}\beta_{n,k_1+k_2}\leq\beta_{n,k_1}\beta_{n,k_2},\ \ k_1,k_2\geq 0. \tag{25}$$

Lemma 17 *It holds*

$$\beta_{n+1,k}\geq\frac{\beta_{2,k}}{2^k}\beta_{n,k},\ \ n\geq 1,\ \ k\geq 1, \tag{26}$$

and for some positive constant C that

$$\beta_{n+1,k}\geq C\left(2k^{\frac{3}{2}}+\log(2k)\right)\beta_{n,k},\ \ n\geq 1,\ \ k\geq 1. \tag{27}$$

Proof By Proposition 15 and Lemma 16, we have

$$\beta_{n+1,k}=\sum_{j=1}^{k}\frac{e_n^j(0)}{j!}\sum_{\substack{T\in\mathbb{N}^j\\|T|=k}}\binom{k}{T}\beta_{n,T}\geq\Bigg(\sum_{j=1}^{k}\frac{1}{j!}\sum_{\substack{T\in\mathbb{N}^j\\|T|=k}}\binom{k}{T}\Bigg)2^{-k}\beta_{n,k}=\frac{\beta_{2,k}}{2^k}\beta_{n,k}, \tag{28}$$

where we have used the identity $\beta_{1,k}=1$ in equation (21). We then recall an asymptotic formula for the Bell numbers given in Bruijin (1981)

$$\frac{\log\beta_{2,k}}{k}=\log k-\log\log k-1+\frac{\log\log k}{\log k}+\frac{1}{\log k}+\frac{1}{2}\left(\frac{\log\log k}{\log k}\right)^2+O\left[\frac{\log\log k}{(\log k)^2}\right].$$

Thus

$$\frac{\log\beta_{2,k}}{k}=\log k+O(\log\log k),$$

which implies that for k large enough,

$$\frac{\log\beta_{2,k}}{k}\geq\frac{1}{2}\log k.$$

Consequently, for k large enough,

$$\beta_{2,k}\geq k^{k/2}=k^{\frac{k}{2}-2}k^2\geq k^{\frac{k}{2}-2}(2k^{\frac{3}{2}}+\log(2k)). \tag{29}$$

Notice that

$$\lim_{k\to\infty}\frac{2^k}{k^{\frac{k}{2}-2}}=\lim_{k\to\infty}\frac{2^kk^2}{k^{\frac{k}{2}}}=\lim_{k\to\infty}\frac{2^kk^2}{\sqrt{k}^k}\le\lim_{k\to\infty}\frac{2^kk^2}{3^k}=\lim_{k\to\infty}\frac{k^2}{(\frac{3}{2})^k}=0.$$

Thus, $k^{\frac{k}{2}-2}\ge 2^k$ for k large enough. This inequality together with (29) implies that for k large enough,

$$\beta_{2,k}\ge 2^k(2k^{\frac{3}{2}}+\log(2k)).$$

Inequality (27) now follows from the above equation and equation (28). ■

With the preparations above, we are ready to characterize the reproducing kernel Hilbert space $\mathcal{H}_{G_n}$ of the hierarchical Gaussian kernel G_n, and to show that $\mathcal{H}_{G_n}$ are strictly expanding as n increases. Below we shall use $A\subsetneqq B$ to denote that A is a proper subset of B. We shall also denote by $\lceil x\rceil$ and $\lfloor x\rfloor$ the least integer greater than or equal to x and the largest integer small than or equal to x, respectively.

Theorem 18 *Given the hierarchical Gaussian kernels defined by (12) and (14), it holds*

$$G_n(x,y)=e_n(0)\sum_{k=0}^{\infty}\frac{\beta_{n,k}}{k!}\exp(-k\lambda\|x-y\|^2),\ n\in\mathbb{N},\ x,y\in\mathbb{R}^d, \tag{30}$$

and

$$G_n(x,y)=e_n(0)+K_n(x,y)\ \textit{with}\ K_n(x,y)=\int_{\mathbb{R}^d}e^{-i(x-y)\cdot\xi}g_n(\xi)d\xi,$$

where

$$g_n(\xi)=\frac{e_n(0)}{(2\sqrt{\lambda\pi})^d}\sum_{k=1}^{\infty}\frac{\beta_{n,k}}{k!k^{d/2}}\exp\left(-\frac{\|\xi\|^2}{4\lambda k}\right),\ \xi\in\mathbb{R}^d. \tag{31}$$

Consequently,

$$\mathcal{H}_{G_n}=\left\{c+f:c\in\mathbb{R},\ f\in C(\mathbb{R}^d)\ \textit{satisfying}\ \int_{\mathbb{R}^d}\frac{|\hat{f}(\xi)|^2}{g_n(\xi)}d\xi<+\infty\right\}, \tag{32}$$

and

$$\mathcal{H}_{G_n}\subsetneqq\mathcal{H}_{G_{n+1}},\ n\in\mathbb{N}. \tag{33}$$

Proof Identity (31) is obtained by applying (19) and the Fourier transform (13) of Gaussian kernels. Thus, equation (32) is a direct consequence of Lemmas 3 and 6. Clearly,

$$g_n(\xi)=\frac{e_n(0)}{(2\sqrt{\lambda\pi})^d}\sum_{k=1}^{\infty}\frac{\beta_{n,k}}{k!k^{d/2}}\exp\left(-\frac{\|\xi\|^2}{4\lambda k}\right)\le\frac{e_{n+1}(0)}{(2\sqrt{\lambda\pi})^d}\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^{d/2}}\exp\left(-\frac{\|\xi\|^2}{4\lambda k}\right)=g_{n+1}(\xi).$$

Thus, $\mathcal{H}_{G_n}\subseteq\mathcal{H}_{G_{n+1}}$. On the other hand, it holds

$$\frac{g_{n+1}(\xi)}{g_n(\xi)}=\frac{e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^{d/2}}\exp\left(-\frac{\|\xi\|^2}{4\lambda k}\right)}{e_n(0)\sum_{k=1}^{\infty}\frac{\beta_{n,k}}{k!k^{d/2}}\exp\left(-\frac{\|\xi\|^2}{4\lambda k}\right)}.$$

We shall show that $g_{n+1}(\xi)/g_n(\xi)$ is unbounded by showing that there exists a positive constant C such that for $\|\xi\| = 8m\sqrt{\lambda}$,

$$\frac{g_{n+1}(\xi)}{g_n(\xi)} \geq C\log(2m) \text{ for sufficiently large } m. \tag{34}$$

The above equation can be rewritten as

$$\sum_{k=1}^{\infty} \frac{1}{k!k^{d/2}} \left(e_{n+1}(0)\beta_{n+1,k} - Ce_n(0)\beta_{n,k}\log(2m)\right)\exp\left(-\frac{16m^2}{k}\right) \geq 0. \tag{35}$$

Let C be the constant in Lemma 17, by $e_{n+1}(0) \geq e_n(0)$, Lemma 17 and (35), it suffices to show that for large enough m

$$\sum_{k=1}^{\infty} \frac{\beta_{n,k}}{k!k^{d/2}} \left(2k^{\frac{3}{2}} + \log(2k) - \log(2m)\right)\exp\left(-\frac{16m^2}{k}\right) \geq 0.$$

First note that for all $k \geq \lfloor \log(2m) \rfloor$,

$$2k^{\frac{3}{2}} + \log(2k) - \log(2m) \geq 0.$$

Therefore, for sufficiently large m, using $2m^{\frac{3}{2}} \geq e$ and $m^{d/2} \leq m!$, we get

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{\beta_{n,k}}{k!k^{d/2}} \left(2k^{\frac{3}{2}} + \log(2k) - \log(2m)\right)\exp\left(-\frac{16m^2}{k}\right) \\
&\geq \frac{\beta_{n,m}}{m!m^{d/2}} 2m^{\frac{3}{2}}\exp(-16m) + \sum_{k=1}^{\lfloor \log(2m) \rfloor} \frac{\beta_{n,k}}{k!k^{d/2}} \left(2k^{\frac{3}{2}} + \log(2k) - \log(2m)\right)\exp\left(-\frac{16m^2}{k}\right) \\
&\geq \frac{2\beta_{n,m}m}{(m!)^2} e^{-16m+1} - \sum_{k=1}^{\lfloor \log(2m) \rfloor} \frac{\beta_{n,k}}{k!k^{d/2}} \log(2m) e^{-\frac{16m^2}{k}} \\
&\geq \frac{2\beta_{n,m}m}{(m!)^2} e^{-16m+1} - \sum_{k=1}^{\lfloor \log(2m) \rfloor} \frac{\beta_{n,m}}{k!} \log(2m) e^{-\frac{8m^2}{\log(2m)}} \\
&\geq \frac{\beta_{n,m}\log(2m)}{(m!)^2} e^{-16m+1} - \beta_{n,m}\log(2m) e^{-\frac{8m^2}{\log(2m)}} \sum_{k=0}^{\infty} \frac{1}{k!} \\
&= e\beta_{n,m}\log(2m)\left(\frac{e^{-16m}}{(m!)^2} - e^{-\frac{8m^2}{\log(2m)}}\right) \\
&= e\beta_{n,m}\log(2m)e^{-\frac{8m^2}{\log(2m)}}\left(\frac{\exp\left(-16m + \frac{8m^2}{\log(2m)}\right)}{(m!)^2} - 1\right) \\
&= e\beta_{n,m}\log(2m)e^{-\frac{8m^2}{\log(2m)}}\left(\left(\frac{\exp\left(\frac{4m^2}{\log(2m)} - 8m\right)}{m!}\right)^2 - 1\right).
\end{aligned}$$

Therefore, it suffices to show that for m large enough,

$$\exp\left(\frac{4m^2}{\log(2m)} - 8m\right) \geq m!.$$

By $m! \leq m^m$, the above equation is true if

$$\frac{4m^2}{\log(2m)} - 8m \geq m \log m,$$

which is clearly true for sufficiently large m. We conclude that inequality (34) holds, which implies that $g_{n+1}(\xi)/g_n(\xi)$ is unbounded. By Lemma 5, $\mathcal{H}_{G_{n+1}} \nsubseteq \mathcal{H}_{G_n}$. ■

4.2 Composition with a Polynomial

We consider hierarchical kernels generated from the composition of the Gaussian kernel and a fixed polynomial in this subsection. Let P be a given polynomial

$$P(x) = \sum_{k=1}^{N} a_k x^k, \tag{36}$$

where $a_k \geq 0, a_N > 0, N \geq 2$. The hierarchical kernels under investigation are defined recursively by

$$\mathcal{G}_n(x,y) = P(\mathcal{G}_{n-1}(x,y)), \quad x, y \in \mathbb{R}^d,\ n \in \mathbb{Z}_+, \tag{37}$$

with $\mathcal{G}_0 = \exp(-\lambda\|x-y\|^2)$. To characterize $\mathcal{H}_{\mathcal{G}_n}$ and their inclusion relations, we shall need a well-known result on the inclusion relation between RKHSs of Gaussian kernels (Steinwart et al., 2006). It can also be viewed as a direct consequence of Lemma 5 and the Fourier transform (13) of the Gaussian kernel.

Lemma 19 *(Steinwart et al., 2006) Given Gaussian kernels*

$$\mathbb{G}_\gamma(x,y) = \exp(-\gamma\|x-y\|^2), \quad x, y \in \mathbb{R}^d,$$

it holds $\mathcal{H}_{\mathbb{G}_{\gamma_1}} \subsetneqq \mathcal{H}_{\mathbb{G}_{\gamma_2}}$ whenever $\gamma_1 < \gamma_2$.

Another result in need follows directly from Lemma 3.

Lemma 20 *Given two kernels K and G, if $\mathcal{H}_K \subseteq \mathcal{H}_G$ then $\mathcal{H}_{K+G} = \mathcal{H}_G$.*

We are ready to present the main result on hierarchical Gaussian kernels from composition with a fixed polynomial.

Theorem 21 *Let P be a fixed polynomial given by (36) and define the hierarchical Gaussian kernels recursively by (37). Then, for every $n \geq 0$,*

$$\mathcal{H}_{\mathcal{G}_n} = \mathcal{H}_{\mathbb{G}_{N^n\lambda}}, \tag{38}$$

$$\mathcal{H}_{\mathcal{G}_n} \subsetneqq \mathcal{H}_{\mathcal{G}_{n+1}}, \tag{39}$$

and

$$\mathcal{H}_{\mathcal{G}_n} = \left\{ f \in C(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\xi)|^2 \exp\left(\frac{\|\xi\|^2}{4N^n\lambda}\right) d\xi < +\infty \right\}. \tag{40}$$

Proof Notice that each $\mathcal{G}_n$ is a linear combination of finitely many Gaussian kernels with positive coefficients. Assume that

$$\mathcal{G}_n = \sum_{k=1}^{m} c_k \mathbb{G}_{\gamma_k},$$

where $c_k > 0$ and $\gamma_1 < \gamma_2 < \cdots < \gamma_m$. One observes from the definition of $\mathcal{G}_n$ that $\gamma_m = N^n\lambda$. By Lemmas 19 and 20, for $0 < \gamma < \gamma'$ and $a, b > 0$,

$$\mathcal{H}_{a\mathbb{G}_\gamma + b\mathbb{G}_{\gamma'}} = \mathcal{H}_{b\mathbb{G}_{\gamma'}} = \mathcal{H}_{\mathbb{G}_{\gamma'}}.$$

Therefore,

$$\mathcal{H}_{\mathcal{G}_n} = \mathcal{H}_{\mathbb{G}_{\gamma_m}} = \mathcal{H}_{\mathbb{G}_{N^n\lambda}}.$$

Equation (39) follows from the above equation and Lemma 19. And equation (40) follows from (38) and Lemma 6. ■

5. Hierarchical Exponential Kernels

We investigate hierarchical kernels generated from the composition of the exponential kernel with the exponential function or a polynomial in this section.

For $x = (x_1, x_2, \cdots, x_d) \in \mathbb{R}^d$, denote $\|x\|_1 = \sum_{i=1}^d |x_i|$ and $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$. Also let Γ denote the Gamma function

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt, \quad s > 0.$$

The exponential kernel is given by

$$E_{p,0}(x,y) = \exp(-\lambda\|x-y\|_p) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} \phi_{p,0}(\xi) d\xi, \quad x, y \in \mathbb{R}^d, \ \lambda > 0, \ p = 1, 2, \quad (41)$$

where

$$\phi_{1,0}(\xi) = \frac{1}{\pi^d} \prod_{j=1}^{d} \frac{\lambda}{\xi_j^2 + \lambda^2}, \quad \xi \in \mathbb{R}^d, \quad (42)$$

and

$$\phi_{2,0}(\xi) = \frac{\Gamma(\frac{d+1}{2})\lambda}{\pi^{\frac{d+1}{2}} \left(\lambda^2 + \|\xi\|_2^2\right)^{\frac{d+1}{2}}}, \quad \xi \in \mathbb{R}^d. \quad (43)$$

5.1 Composition with the Exponential Function

Recall the exponential generating functions e_n defined in (18). The hierarchical exponential kernels via consecutively compositing with the exponential function are defined by

$$E_{p,n}(x,y) = e_n(E_{p,0}(x,y)), \quad n \in \mathbb{N}, \ p = 1, 2. \quad (44)$$

We present the RKHS of the hierarchical exponential kernels and a surprise result on their inclusion relations.

Theorem 22 *Given the hierarchical exponential kernels defined by (41) and (44), it holds*

$$\mathcal{H}_{E_{p,n}} = \left\{c+f: \ c \in \mathbb{C}, f \in C(\mathbb{R}) \ \textit{satisfying} \ \int_{\mathbb{R}^d} \frac{|\hat{f}(\xi)|^2}{\phi_{p,n}(\xi)} d\xi < +\infty \right\}, \ n \in \mathbb{N}, \ p = 1, 2, \tag{45}$$

where

$$\phi_{1,n}(\xi) = \frac{e_n(0)}{\pi^d} \sum_{k=1}^{\infty} \frac{\beta_{n,k}}{k!} \prod_{j=1}^{d} \frac{k\lambda}{\xi_j^2 + k^2\lambda^2}, \ \xi \in \mathbb{R}^d, \tag{46}$$

and

$$\phi_{2,n}(\xi) = \frac{e_n(0)\Gamma(\frac{d+1}{2})}{\pi^{\frac{d+1}{2}}} \sum_{k=1}^{\infty} \frac{\beta_{n,k} k\lambda}{k! \left(k^2\lambda^2 + \|\xi\|_2^2\right)^{\frac{d+1}{2}}}, \ \xi \in \mathbb{R}^d. \tag{47}$$

Moreover,

$$\mathcal{H}_{E_{p,n}} = \mathcal{H}_{E_{p,n+1}}, \quad n \in \mathbb{N}, \ p = 1, 2. \tag{48}$$

Proof We first write

$$E_{p,n}(x,y) = e_{p,n}(0) + F_{p,n}(x,y), \quad n \in \mathbb{N}, \ n \geq 1, \ p = 1, 2,$$

where

$$F_{p,n}(x,y) = \int_{\mathbb{R}^d} e^{-i(x-y)\cdot\xi} \phi_{p,n}(\xi) d\xi, \quad x, y \in \mathbb{R}^d, \ p = 1, 2,$$

and by the expansion (19),

$$\phi_{p,n}(\xi) = e_n(0) \sum_{k=1}^{\infty} \frac{\beta_{n,k}}{k! k^d} \phi_{p,0}(\frac{\xi}{k}), \ \xi \in \mathbb{R}^d, \ p = 1, 2. \tag{49}$$

Combing the above equation and equations (42), (43), we obtain (46) and (47). The first result (45) thus follows directly from Lemmas 3 and 6.

We now turn to the proof of identity (48). On one hand, since the high order Bell numbers satisfy

$$\beta_{n,k} \leq \beta_{n+1,k}, \quad k \in \mathbb{N},$$

it implies $\mathcal{H}_{E_{p,n}} \subseteq \mathcal{H}_{E_{p,n+1}}$ for $p = 1, 2$. On the other hand, since

$$\frac{\phi_{1,0}(\frac{\xi}{k})}{\phi_{1,0}(\xi)} = \prod_{j=1}^{d} \frac{\lambda^2 + \xi_j^2}{\lambda^2 + \xi_j^2/k^2} \leq \prod_{j=1}^{d} k^2 = k^{2d}, \ \xi \in \mathbb{R}^d, \ k \in \mathbb{N}_+,$$

and

$$\frac{\phi_{2,0}(\frac{\xi}{k})}{\phi_{2,0}(\xi)} = \left(\frac{\lambda^2 + \|\xi\|_2^2}{\lambda^2 + \|\xi\|_2^2/k^2}\right)^{\frac{d+1}{2}} \leq (k^2)^{\frac{d+1}{2}} \leq k^{2d}, \ \xi \in \mathbb{R}^d, \ k \in \mathbb{N}_+,$$

with (49) we get

$$\begin{aligned}\frac{\phi_{p,n+1}(\xi)}{\phi_{p,n}(\xi)} &= \frac{e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^d}\phi_{p,0}(\frac{\xi}{k})}{e_n(0)\sum_{k=1}^{\infty}\frac{\beta_{n,k}}{k!k^d}\phi_{p,0}(\frac{\xi}{k})}\\ &\le \frac{e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^d}\phi_{p,0}(\frac{\xi}{k})}{e_n(0)\beta_{n,1}\phi_{p,0}(\xi)}\\ &\le e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^d}\frac{\phi_{p,0}(\frac{\xi}{k})}{\phi_{p,0}(\xi)}\\ &\le e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!k^d}k^{2d}\\ &= e_{n+1}(0)\sum_{k=1}^{\infty}\frac{\beta_{n+1,k}}{k!}k^{d}\\ &\le e_{n+1}(0)\sum_{k=0}^{\infty}\frac{\beta_{n+1,k}}{k!}e^{kd}\\ &= e_{n+1}(e^d) < +\infty,\ p=1,2,\end{aligned}$$

which implies by Lemma 5 that $\mathcal{H}_{E_{p,n+1}} \subseteq \mathcal{H}_{E_{p,n}}$ for $p = 1, 2$. We conclude that (48) holds. ■

The above theorem reveals that the hierarchical structure of reproducing kernels does not necessarily yield RKHSs with increasing expressive power.

5.2 Composition with a Polynomial

We consider hierarchical exponential kernels generated from compositions with a fixed polynomial in this subsection. Let P be a polynomial as described in (36). We generate the hierarchical exponential kernels by

$$\mathcal{E}_{p,n}(x,y) = P(\mathcal{E}_{p,n-1}(x,y)),\ n\in\mathbb{N},\ x,y\in\mathbb{R}^d,\ p=1,2. \tag{50}$$

where $\mathcal{E}_{p,0}$ is the exponential kernel $E_{p,0}$ given by (41).

The following result was proved in Zhang and Zhao (2013).

Lemma 23 *Given exponential kernels*

$$\mathbb{E}_{p,\lambda}(x,y) = \exp(-\lambda\|x-y\|_p),\quad x,y\in\mathbb{R}^d,\ p=1,2,$$

it holds $\mathcal{H}_{\mathbb{E}_{p,\lambda_1}} = \mathcal{H}_{\mathbb{E}_{p,\lambda_2}}$ for all $\lambda_1,\lambda_2>0$ and $p=1,2$.

We show that composition of the exponential kernel with a polynomial will not enlarge the corresponding RKHS either.

Theorem 24 *Let P be a fixed polynomial given by (36) and define the hierarchical exponential kernels recursively by (50). Then for every $n\ge 0$ and $p=1,2$, $\mathcal{H}_{\mathcal{E}_{p,n}} = \mathcal{H}_{\mathcal{E}_{p,0}}$.*

Proof Notice that each $\mathcal{E}_{p,n}$ is a linear combination of finitely many exponential kernels with positive coefficients. Assume that

$$\mathcal{E}_{p,n} = \sum_{k=1}^{m} c_k \mathbb{E}_{p,\lambda_k}.$$

By Lemma 23

$$\mathcal{H}_{\mathbb{E}_{p,\lambda_1}} = \mathcal{H}_{\mathbb{E}_{p,\lambda_2}} = \cdots = \mathcal{H}_{\mathbb{E}_{p,\lambda_k}}.$$

Then Lemma 20 implies that

$$\mathcal{H}_{\mathcal{E}_{p,n}} = \mathcal{H}_{\mathbb{E}_{p,\lambda_1}} = \mathcal{H}_{\mathbb{E}_{p,\lambda}} = \mathcal{H}_{E_{p,0}} = \mathcal{H}_{\mathcal{E}_{p,0}},$$

which proves the result. ■

6. Hierarchical Polynomial Kernels

We study hierarchical kernels generated from the compositions of a given polynomial kernel and the exponential function in this section. A polynomial kernel is a reproducing kernel of the form

$$K(x,y) = \sum_{k=0}^{\infty} a_k (x \cdot y)^k, \quad x, y \in \mathbb{R}^d, \tag{51}$$

where $a_k \geq 0$ for every $k \in \mathbb{Z}_+$. Let r be the radius of convergence of the associated polynomial

$$P(z) = \sum_{k=0}^{\infty} a_k z^k. \tag{52}$$

Then the kernel K in (51) is well-defined on $\{x \in \mathbb{R}^d : \|x\| < \sqrt{r}\}$. The hierarchical kernels via compositions of K and the exponential function are generated by

$$K_n = e_n(K), \quad n \in \mathbb{N}, \tag{53}$$

where e_n are the exponential generating functions given in (18). We first show that under a mild condition, the RKHS $\mathcal{H}_{K_n}$ is indeed expanding as n increases.

Theorem 25 *Suppose the radius of convergence of the polynomial P given in (52) is infinity and that P is not a constant. Then for each $n \in \mathbb{N}$, $\mathcal{H}_{K_{n-1}} \subsetneqq \mathcal{H}_{K_n}$.*

Proof By Proposition 10, $\mathcal{H}_{K_{n-1}} \subseteq \mathcal{H}_{K_n}$ for each $n \in \mathbb{N}$. Assume that $\mathcal{H}_{K_1} \subseteq \mathcal{H}_{K_0}$. By Lemma 1, there exists a constant $\lambda > 0$ such that $\lambda K_0 - K_1$ is a kernel. Consequently,

$$\lambda K_0(x,x) - \exp(K_0(x,x)) \geq 0 \text{ for all } x \in \mathbb{R}^d, \tag{54}$$

which implies that

$$K_0(x,x) = \sum_{k=0}^{\infty} a_k (x \cdot x)^k,$$

is bounded on $\mathbb{R}^d$. As a result, $P(z)$ is bounded on $\mathbb{C}$. By Liouville's theorem, P must be a constant, a contradiction. Similarly, if for some $n \in \mathbb{N}$, $\mathcal{H}_{K_n} \subseteq \mathcal{H}_{K_{n-1}}$ then P_{n-1} must be a constant. Here,

$$P_0 = P, \;\; P_n = \exp(P_{n-1}), \;\; n \geq 1.$$

But P_{n-1} is constant if and only if P is constant. Therefore, we conclude that $\mathcal{H}_{K_{n-1}} \subsetneqq \mathcal{H}_{K_n}$ for every $n \in \mathbb{N}$. ■

In the rest of this section, we focus on the most popular polynomial kernel in machine learning, which is of the form

$$K(x,y) = (x \cdot y)^q, \quad x, y \in \mathbb{R}^d, \tag{55}$$

where $q \in \mathbb{N}$ is fixed. Thus K corresponds to the polynomial

$$P(z) = z^q.$$

Let K_n be the hierarchical kernels (53). Theorem 25 applies to this special case. Thus, $\mathcal{H}_{K_n}$ are expanding as n increases. As the final theoretical task of this paper, we desire to characterize $\mathcal{H}_{K_n}$.

Theorem 26 *Let $n \in \mathbb{N}$ and K_n be the hierarchical polynomial kernels defined by (53) with K be given in (55). Then*

$$K_n(x,y) = e_n(0) \sum_{k=0}^{\infty} \frac{\beta_{n,k}}{k!} (x \cdot y)^{qk}, \quad x, y \in \mathbb{R}^d, \tag{56}$$

and

$$\mathcal{H}_{K_n} = \left\{ f_a(x) = \sum_{k=0}^{\infty} \frac{e_n(0)\beta_{n,k}}{k!} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha| = qk}} \binom{qk}{\alpha} a_{k,\alpha} x^\alpha : \; \sum_{k=0}^{\infty} \frac{e_n(0)\beta_{n,k}}{k!} \sum_{\substack{\alpha \in \mathbb{Z}_+^d \\ |\alpha| = qk}} \binom{qk}{\alpha} |a_{k,\alpha}|^2 < +\infty \right\}, \tag{57}$$

with inner product

$$\langle f_a, f_b \rangle_{\mathcal{H}_{K_n}} = e_n(0) \sum_{k=0}^{\infty} \frac{\beta_{n,k}}{k!} \sum_{\alpha \in \mathbb{Z}_+^d, |\alpha| = qk} \binom{qk}{\alpha} a_{k,\alpha} b_{k,\alpha}. \tag{58}$$

Proof Recall the exponential generating functions e_n given in (18). Apparently,

$$K_n(x,y) = e_n(K(x,y)) = e_n((x \cdot y)^q), \quad x, y \in \mathbb{R}^d.$$

Therefore, by (19),

$$\begin{aligned} K_n(x,y) &= e_n(0)\sum_{k=0}^{\infty}\frac{\beta_{n,k}}{k!}(x\cdot y)^{qk} \\ &= e_n(0)\sum_{k=0}^{\infty}\frac{\beta_{n,k}}{k!}(x_1y_1+x_2y_2+\cdots+x_dy_d)^{qk} \\ &= e_n(0)\sum_{k=0}^{\infty}\frac{\beta_{n,k}}{k!}\sum_{\substack{\alpha\in\mathbb{Z}_+^d \\ |\alpha|=qk}}\binom{qk}{\alpha}x^{\alpha}y^{\alpha}. \end{aligned}$$

Equations (57), (58) now follow directly from the above equation and Lemma 8. ■

7. Experiments

To verify the theoretical results in the paper, we shall conduct initial experiments with hierarchical Gaussian kernels and hierarchical exponential kernels in this section. Specifically, we shall evaluate the hierarchical Gaussian kernels from consecutive composition with the exponential function on various tasks, including classification on the Scikit-learn (Pedregosa et al., 2011) moon scattering dataset, classification on CIFAR-10, and regression on UCI datasets. Considering the numerical stability, we shall only evaluate hierarchical kernels with at most three layers, and shall slightly modify the hierarchical Gaussian kernels as

$$G_{k+1}(x,y)=\exp\left(e_k(1)\left(G_k(x,y)-1\right)\right),\quad G_0(x,y)=\exp(-\lambda\|x-y\|^2),\quad x,y\in\mathbb{R}^d. \tag{59}$$

To further confirm the theoretical results of the paper, we shall also evaluate the ℓ_1 norm hierarchical exponential kernels on three regression tasks from LIBSVM (Chang and Lin, 2011). These ℓ_1 norm hierarchical exponential kernels are also slightly modified as

$$H_{k+1}(x,y)=\exp\left(e_k(1)\left(H_k(x,y)-1\right)\right),\quad H_0(x,y)=\exp(-\lambda\|x-y\|_1),\quad x,y\in\mathbb{R}^d. \tag{60}$$

For classification tasks, we shall use the C-support vector classification (Boser et al., 1992; Cortes and Vapnik, 1995) method

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}}\quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i=1}^{n}\xi_i \\ \text{subject to}\quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b)\geq 1-\xi_i \\ & \xi_i\geq 0, i=1,\cdots,n, \end{aligned}$$

where $\mathbf{x_i}, 1\leq i\leq n$ are the given training data in two classes, $y_i\in\{-1,1\}, 1\leq i\leq n$ are the corresponding labels, $\phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ into a feature space, and $C>0$ is the regularization parameter. For multi-class classification, we adapt the “one-against-one” strategy that trains $\frac{N(N-1)}{2}$ classifiers, where N is the number of classes, and predict the label of a new input through majority voting.

For regression tasks, we shall use the ϵ-support vector regression method

$$\begin{array}{ll}\min\limits_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} & \frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum\limits_{i=1}^{n}(\xi_i+\xi_i^*) \\ \text{subject to} & \mathbf{w}^T\phi(\mathbf{x}_i)+b-y_i\leq\epsilon+\xi_i \\ & y_i-\mathbf{w}^T\phi(\mathbf{x}_i)-b\leq\epsilon+\xi_i^* \\ & \xi_i,\xi_i^*\geq 0, i=1,\cdots,n,\end{array}$$

where $\mathbf{x}_i$ and y_i, $1\leq i\leq n$ are the input and output data, respectively, and $C>0,\epsilon>0$ are given regularization parameters. Note that we shall use the Root Mean Square Error (RMSE) to measure the performance of regression tasks.

Both the classification and regression tasks will be solved with the Sequential Minimal Optimization (SMO) solver. The computational cost of this method is analyzed in (Platt, 1998). According to (Platt, 1998), this algorithm does not involve matrix computations with the kernel matrices. This together with existence of regularization in C-support vector classification and ϵ-support vector regression ensure the stability of the numerical experiments in this section. As our focus is on the hierarchical kernels, we shall fix $C=1$ and $\varepsilon=10^{-3}$ in all the tasks for fair comparison. For each task and each hierarchical kernel, the hyperparameter λ in the hierarchical kernels (59) and (60) will be optimally chosen.

For evaluations with hierarchical Gaussian kernels, we implement with the thundersvm (Wen et al., 2018) to accelerate the training process. And for evaluations with hierarchical exponential kernels, we directly implement the SVM module in Scikit-learn since the datasets are relatively small. Notice that thundersvm only supports precomputed mode for custom kernels, which is inefficient and memory consuming for large scale data. Thus, we directly modify their source codes for hierarchical kernels. More details about the implementation of thundersvm could be found in (Wen et al., 2017), and our codes could be accessed via the github repository `https://github.com/SaebaHuang/Hierarchical-Kernel-in-Deep-Kernel-Learning`.

For hierarchical Gaussian kernels, we will see that in all tasks, the best results are obtained with G_3 as the layer increases from 0 to 3. While for hierarchical exponential kernels, we will see that the result is not improving as the number of layers increases. These confirm our results in previous sections that as the number of layers increases, the RKHS of hierarchical Gaussian kernels is expanding while the RKHS of hierarchical exponential kernels remains the same. We present detailed results of the experiments as follows.

7.1 Scikit-learn moon scattering dataset with hierarchical Gaussian kernels

We generate a set of points representing moon scattering using Scikit-learn, and perform evaluations with different hierarchical Gaussian kernels. We randomly choose 500 points as the training set which is equally divided into two classes. For each hierarchal Guassian kernel G_k, $0\leq k\leq 3$, the hyperparameter λ will be optimally chosen from $[2^{-5},2^{-4},\cdots,2^9,2^{10}]$. The results are tabulated in Table 1. We also plot the decision boundaries of corresponding to the best λ in Figure 1. One could see that the decision boundary becomes tighter as the number of layers increases.

Kernel \ $\log_2(\lambda)$	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
G_0	86.3%	87.2%	89.8%	98.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.9%	99.8%	99.0%
G_1	86.4%	88.6%	95.7%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.9%	99.7%	99.3%
G_2	90.4%	98.6%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.8%	99.5%	98.8%
G_3	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.8%	99.3%	99.4%	96.1%	90.1%	85.4%	83.2%

Table 1: Accuracy on randomly generated set of moon scattering points with different λ.

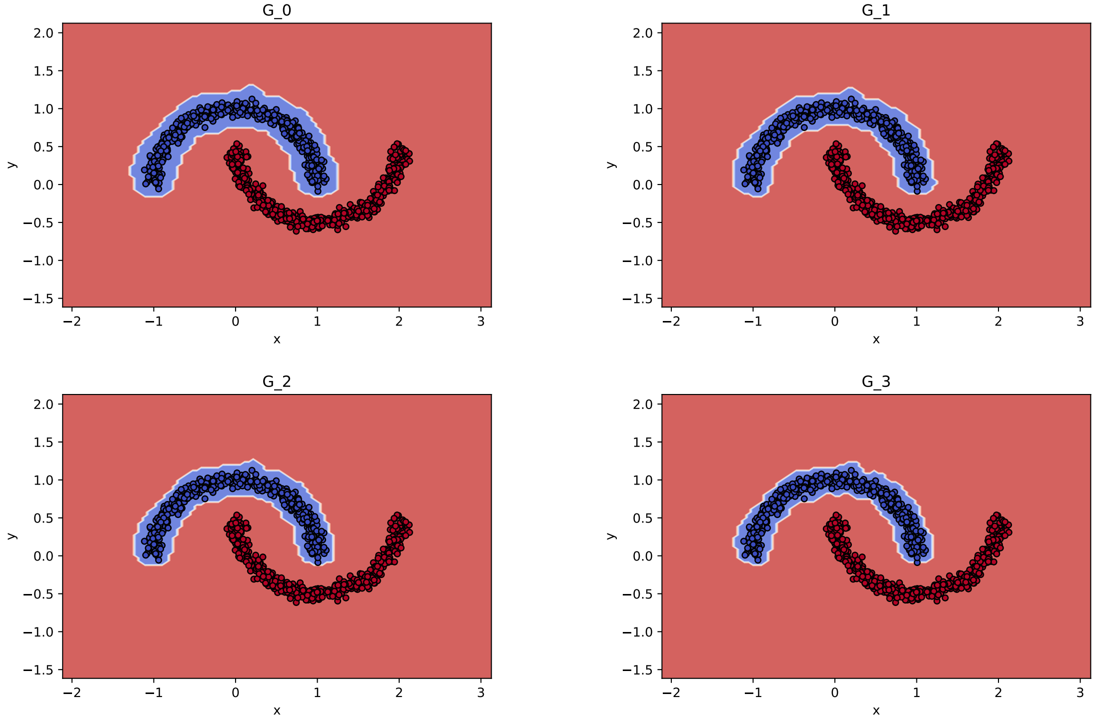

Figure 1: Decision boundaries of different hierarchical Gaussian kernels with their own optimal λ.

7.2 CIFAR-10 dataset with hierarchical Gaussian kernels

For evaluations on CIFAR-10 dataset with hierarchical Gaussian kernels, the λ of each layer is optimally chosen from $[2^{-19}, , 2^{-18}, \cdots, 2^{-8}]$. As shown in Table 2, the best result is obtained with G_3.

$\log_2(\lambda)$ Kernel	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8
G_0	36.48%	38.03%	40.38%	43.01%	45.28%	48.13%	50.55%	53.73%	**55.16%**	54.30%	45.43%	29.84%
G_1	36.48%	38.19%	40.88%	43.59%	46.00%	48.96%	51.16%	52.97%	**53.29%**	50.32%	35.68%	16.45%
G_2	39.17%	42.04%	44.65%	46.91%	49.56%	52.15%	53.66%	**54.13%**	52.85%	48.45%	34.83%	16.81%
G_3	49.09%	51.77%	54.41%	**55.55%**	53.97%	47.63%	38.05%	32.45%	31.93%	31.96%	20.76%	11.19%

Table 2: Accuracy of hierarchical Gaussian kernels on CIFAR-10 with different λ.

7.3 UCI Regression tasks with hierarchical Gaussian kernels

We also evaluate the hierarchical Gaussian kernels on the UCI regression datasets elevators, kin40k, and servo. To read these datasets, we utilize code from the repository `https://github.com/treforevans/uci_datasets`. For each evaluation, we normalize the data to be between -1 and 1, and perform 5-fold cross-validation. The results are presented in Tables 3-5. One sees that the best test RMSE are obtained with G_3 for each dataset.

$\log_2(\lambda)$ Kernel	-8	-7	-6	-5	-4	-3	-2	-1	0	1
G_0	0.1320±0.0030	0.1227±0.0028	0.1165±0.0029	0.1119±0.0028	0.1077±0.0025	0.1051±0.0024	0.1032±0.0023	**0.1021±0.0022**	0.1021±0.0024	0.1068±0.0025
G_1	0.1296±0.0030	0.1205±0.0029	0.1148±0.0029	0.1098±0.0027	0.1062±0.0025	0.1039±0.0024	0.1024±0.0023	**0.1019±0.0024**	0.1038±0.0025	0.1092±0.0026
G_2	0.1181±0.0029	0.1130±0.0029	0.1081±0.0026	0.1052±0.0025	0.1032±0.0023	0.1020±0.0023	**0.1017±0.0025**	0.1037±0.0025	0.1077±0.0024	0.1163±0.0019
G_3	0.1041±0.0023	0.1025±0.0023	**0.1017±0.0023**	0.1030±0.0024	0.1084±0.0026	0.1204±0.0021	0.1441±0.0023	0.1835±0.0025	0.2273±0.0028	0.2548±0.0035

Table 3: RMSE of hierarchical Gaussian kernels on elevators dataset with different λ.

$\log_2(\lambda)$ Kernel	-8	-7	-6	-5	-4	-3	-2	-1	0	1
G_0	0.2540±0.0024	0.2409±0.0030	0.2397±0.0036	0.2148±0.0036	0.1738±0.0062	0.1102±0.0031	0.0715±0.0013	0.0460±0.0006	**0.0310±0.0005**	0.0340±0.0007
G_1	0.2450±0.0026	0.2378±0.0033	0.2208±0.0033	0.1739±0.0024	0.1096±0.0019	0.0672±0.0012	0.0430±0.0007	0.0300±0.0007	**0.0285±0.0006**	0.0436±0.0009
G_2	0.2357±0.0034	0.2074±0.0028	0.1576±0.0025	0.0922±0.0014	0.0578±0.0014	0.0384±0.0007	**0.0282±0.0007**	0.0292±0.0006	0.0400±0.0008	0.0668±0.0014
G_3	0.0829±0.0011	0.0515±0.0012	0.0361±0.0007	**0.0279±0.0005**	0.0386±0.0008	0.0755±0.0016	0.1580±0.0023	0.2472±0.0022	0.2734±0.0022	0.2770±0.0022

Table 4: RMSE of hierarchical Gaussian kernels on kin40k dataset with different λ.

$\log_2(\lambda)$ Kernel	-7	-6	-5	-4	-3	-2	-1	0	1	2
G_0	0.3481±0.0347	0.3258±0.0299	0.2813±0.0283	0.2364±0.0269	0.1851±0.0340	0.1632±0.0422	0.1561±0.0461	**0.1440±0.0456**	0.1568±0.0436	0.2007±0.0462
G_1	0.3479±0.0332	0.3187±0.0304	0.2728±0.0284	0.2276±0.0305	0.1795±0.0372	0.1630±0.0454	0.1500±0.0465	**0.1475±0.0489**	0.1726±0.0477	0.2419±0.0465
G_2	0.2991±0.0297	0.2571±0.0293	0.2027±0.0311	0.1732±0.0395	0.1581±0.0455	**0.1430±0.0454**	0.1457±0.0464	0.1717±0.0476	0.2255±0.0483	0.3212±0.0404
G_3	0.1588±0.0437	0.1470±0.0434	**0.1424±0.0461**	0.1670±0.0439	0.2198±0.0483	0.3378±0.0396	0.4288±0.0337	0.4512±0.0329	0.4534±0.0329	0.4536±0.0329

Table 5: RMSE of hierarchical Gaussian kernels on servo dataset with different λ.

7.4 LIBSVM regression datasets with hierarchical exponential kernels

Finally, we shall evaluate the hierarchical exponential kernels on the datasets bodyfat, mpg, and triazines from LIBSVM. Similarly, we normalize the data to be between -1 and 1, and perform 5-fold cross-validation for each evaluation. The results are presented in Tables 6-8. One sees that the best test RMSE is not improving as the number of layers increases, which justifies our results about hierarchical exponential kernels in Section 5.

Kernel \ $\log_2(\lambda)$	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3
H_0	0.3199±0.0201	0.3063±0.0200	0.2842±0.0205	0.2503±0.0214	0.2062±0.0244	0.1564±0.0285	0.1082±0.0331	0.0800±0.0354	0.0686±0.0361	**0.0658±0.0360**	0.0708±0.0362
H_1	0.3199±0.0201	0.3063±0.0199	0.2843±0.0205	0.2504±0.0213	0.2064±0.0245	0.1582±0.0283	0.1110±0.0331	0.0816±0.0353	0.0721±0.0367	**0.0718±0.0367**	0.0792±0.0383
H_2	0.2973±0.0202	0.2712±0.0208	0.2333±0.0222	0.1873±0.0261	0.1332±0.0304	0.0955±0.0346	0.0746±0.0356	**0.0690±0.0359**	0.0713±0.0373	0.0809±0.0375	0.0956±0.0397
H_3	0.1360±0.0300	0.0953±0.0347	0.0743±0.0357	**0.0669±0.0361**	0.0676±0.0355	0.0753±0.0374	0.0914±0.0394	0.1184±0.0397	0.1714±0.0345	0.2605±0.0259	0.3228±0.0211

Table 6: RMSE of hierarchical exponential kernels on bodyfat dataset with different λ.

Kernel \ $\log_2(\lambda)$	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3
H_0	0.3115±0.0723	0.3050±0.0749	0.2922±0.0797	0.2755±0.0859	0.2474±0.0977	0.2281±0.1044	0.2053±0.1120	0.1949±0.1174	0.1936±0.1179	**0.1934±0.1173**	0.1988±0.1177
H_1	0.3115±0.0723	0.3051±0.0749	0.2923±0.0796	0.2758±0.0857	0.2483±0.0974	0.2294±0.1037	0.2070±0.1117	0.1962±0.1169	0.1946±0.1182	**0.1941±0.1191**	0.1992±0.1213
H_2	0.3002±0.0766	0.2865±0.0816	0.2649±0.0906	0.2380±0.1014	0.2197±0.1106	0.1989±0.1127	0.1946±0.1180	**0.1932±0.1184**	0.1951±0.1198	0.2006±0.1211	0.2170±0.1187
H_3	0.2205±0.1105	0.1987±0.1127	0.1942±0.1181	**0.1934±0.1176**	0.1951±0.1181	0.2007±0.1185	0.2172±0.1190	0.2454±0.1121	0.2728±0.1017	0.2931±0.0935	0.3036±0.0883

Table 7: RMSE of hierarchical exponential kernels on pyrim dataset with different λ.

Kernel \ $\log_2(\lambda)$	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3
H_0	0.3969±0.0470	0.3924±0.0473	0.3866±0.0482	0.3758±0.0487	0.3668±0.0466	0.3577±0.0482	0.3495±0.0510	0.3429±0.0499	0.3370±0.0524	0.3382±0.0519	**0.3365±0.0521**
H_1	0.3969±0.0470	0.3924±0.0473	0.3867±0.0482	0.3760±0.0488	0.3671±0.0466	0.3582±0.0486	0.3501±0.0513	0.3455±0.0507	0.3377±0.0528	**0.3351±0.0515**	0.3394±0.0542
H_2	0.3903±0.0478	0.3824±0.0485	0.3716±0.0475	0.3628±0.0472	0.3545±0.0505	0.3478±0.0509	0.3398±0.0519	0.3376±0.0534	**0.3339±0.0514**	0.3358±0.0530	0.3401±0.0513
H_3	0.3550±0.0500	0.3472±0.0513	0.3382±0.0505	0.3365±0.0526	0.3387±0.0513	**0.3365±0.0511**	0.3370±0.0495	0.3496±0.0453	0.3671±0.0481	0.3832±0.0489	0.3891±0.0465

Table 8: RMSE of hierarchical exponential kernels on triazines dataset with different λ.

8. Conclusion

Kernel methods constitute an important category of machine learning methodologies. They enjoy solid mathematical foundations and good interpretability consequently. Motivated by deep neural networks, which generate learning functions through successive composition of activation functions and linear functions, a class of hierarchical kernels has appeared in the literature recently. Such kernels are generated by successive composition of a base kernel and a chosen univariate function. An important theoretical question about hierarchical kernels is whether the expressive power of the kernel will be improving as the number of layer increases. We investigate this question by studying the reproducing kernel Hilbert spaces of hierarchical kernels. It is shown in the paper that the RKHS of hierarchical Gaussian kernels and polynomial kernels is indeed expanding as the number of layer increases, while the RKHS of hierarchical exponential kernels always remains the same. The results reveal that we should not use the exponential kernels as bases kernels in deep kernel learning. In contrast, Gaussian kernels and polynomial kernels are good choices.

Numerical experiments on the Scikit-learn demo datasets, the CIFAR-10, and UCI datasets confirm that the learning ability of the hierarchical Gaussian kernel is improving as the number of layer increases. And experiments on datasets from LIBSVM indicate that the learning ability of the hierarchical exponential kernel is not improving as the number of layer increases. These numerical findings justify the theorems in the paper.

Finally, we remark that the hierarchical kernels considered in the paper has a simple structure. For applications to complicated learning problems, hierarchical kernels with more sophisticated structures should be investigated in the future.

Acknowledgments

We would like to express gratitude to the reviewers for careful reading our manuscript, and insightful comments and suggestions that help improve the quality of this manuscript.

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 12371103, 11971490 and 12126610, and by Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2020B1212060032).

References

F. Anselmi, L. Rosasco, C. Tan, and T. Poggio. Deep convolutional networks are hierarchical kernel machines. *CBMM Memos*, 35, 2015.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68: 337–404, 1950.

N. Asai, I. Kubo, and H. Kuo. Bell numbers, log-concavity, and log-convexity. *Acta Applicandae Mathematicae*, 63: 79–87, 2001.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1): 629-681, 2017.

F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna. Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, 62: 194–236, 2021.

E. T. Bell. The iterated exponential integers. *The Annals of Mathematics*, 39: 539–557, 1938.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

A. Bietti and F. Bach. Deep equals shallow for ReLU networks in kernel regimes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

S. Bochner. *Lectures on Fourier Integrals with an Author's Supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis*. Annals of Mathematics Studies 42, Princeton University Press, New Jersey, 1959.

N. G. de Bruijn. *Asymptotic Methods in Analysis*. Dover Publications, New York, pp. 102-109, 1981.

B. Bohn, C. Rieger, and M. Griebel. A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, 20: 1–32, 2019.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *Annual Conference Computational Learning Theory*, 1992.

C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 1–27, 2011.

J. Chen, H. Avron, and V. Sindhwani. Hierarchically compositional kernels for scalable nonparametric learning. *Journal of Machine Learning Research*, 18: 1–42, 2017.

L. Chen and S. Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Y. Cho and L. Saul. Kernel methods for deep learning. *Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS)*, 2009.

C. Cortes, and V. Vapnik. Support-vector network. *Machine Learning*, 20: 273–297, 1995.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint.* Cambridge Monographs on Applied and Computational Mathematics, 24, Cambridge University Press, Cambridge, 2007.

A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in Neural Information Processing Systems (NIPS)*, 2016.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13: 1–50, 2000.

C. H. FitzGerald, C. A. Micchelli, and A. Pinkus. Functions that preserve families of positive semidefinite matrices. *Linear Algebra and its Applications*, 221: 83–102, 1995.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5: 73–99, 2004.

A. Geifman, A. Yadav, Y. Kasten, M. Galun, D. Jacobs, and R. Basri. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, Cambridge, 2016.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991.

J. Huang and H. T. Yau. Dynamics of deep neural networks and neural tangent hierarchy. *Int. Conf. Mach. Learn.*, PMLR, 2020, 4542–4551.

W. Huang, W. Du and R. Y. Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization. IJCAI, 2021.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

J. Mercer. Functions of positive and negative type and their connection with the theorey of integral equations. *Philosophical Transactions of the Royal Society A*, 209: 415–446, 1909.

S. A. Morris. Hilbert 13: Are there any genuine continuous multivariate real-valued functions? *Bulletin of the American Mathematical Society*, 58(1): 107–118, 2021.

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width ReLU nets: the multivariate case. *International Conference on Learning Representations*, 2019.

R. Parhi and R. D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(1): 1960-1999, 2021.

R. Parhi and R. D. Nowak. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science*, 4(2): 464-489, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, G. Louppe, P. Prettenhofer, R. Weiss, R.J. Weiss, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

J. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Microsoft Research Technical Report*, 1998.

J. Schmidt-Hieber. The Kolmogorov-Arnold representation theorem revisited. *Neural Networks*, 137: 119-126, 2021.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics (2)*, 39: 811–841, 1938.

I. J. Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9: 96–108, 1942.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, Massachusetts, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, 2004.

L. Spek, T. J. Heeringa, and C. Brune. Duality for neural networks through reproducing kernel Banach spaces. arXiv:2211.05020, 2022.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52: 4635–4643, 2006.

V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Z. Wen, J. Shi, Q. Li, B. He, and J. Chen. ThunderSVM: a fast SVM library on GPUs and CPUs. *Journal of Machine Learning Research*, 19: 1–5, 2018.

Z. Wen, J. Shi, Q. Li, B. He, and J. Chen. Supplementary material of ThunderSVM: `https://github.com/zeyiwen/thundersvm/blob/master/thundersvm-full.pdf`, 2017.

H. Wendland. *Scattered Data Approximation.* Cambridge Monographs on Applied and Computational Mathematics **17**, Cambridge University Press, Cambridge, 2005.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. *Proceedings of Machine Learning Research*, 2016.

Z. M. Wu. Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4(3): 283–292, 1995.

Y. Xu and H. Zhang. Refinable kernels. *Journal of Machine Learning Research*, 8: 2083–2120, 2007.

Y. Xu and H. Zhang. Refinement of reproducing kernels. *Journal of Machine Learning Research*, 10: 107–140, 2009.

D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10: 2741–2775, 2009.

H. Zhang and L. Zhao. On the inclusion relation of reproducing kernel Hilbert spaces. *Analysis and Applications*, 11: 1350014, 31 pages, 2013.