Nearly-tight Bounds for Deep Kernel Learning

Yi-Fan Zhang [1 2] **Min-Ling Zhang** [2 3]

Abstract

The generalization analysis of deep kernel learning (DKL) is a crucial and open problem of kernel methods for deep learning. The implicit nonlinear mapping in DKL makes existing methods of capacity-based generalization analysis for deep learning invalid. In an attempt to overcome this challenge and make up for the gap in the generalization theory of DKL, we develop an analysis method based on the composite relationship of function classes and derive capacity-based bounds with mild dependence on the depth, which generalizes learning theory bounds to deep kernels and serves as theoretical guarantees for the generalization of DKL. In this paper, we prove novel and nearly-tight generalization bounds based on the uniform covering number and the Rademacher chaos complexity for deep (multiple) kernel machines. In addition, for some common classes, we estimate their uniform covering numbers and Rademacher chaos complexities by bounding their pseudo-dimensions and kernel pseudo-dimensions, respectively. The mild bounds without strong assumptions partially explain the good generalization ability of deep learning combined with kernel methods.

1. Introduction

Recent work in machine learning has given a revival of attention to deep learning due to its impressive empirical advances across a wide range of tasks. Deep models are typically heavily over-parametrized, while they still achieve good generalization performance, Zhang et al. (2017) showed that deep neural networks can almost perfectly fit the training data even with random labels, this sparked a rush to explain this phenomenon through generalization analysis based on complexity measures. The problem of understanding deep learning theoretically remains relatively under-explored. Meanwhile, kernel machines have a perfect fit of training data while guaranteeing that they generalize well (Bartlett & Mendelson, 2002; Belkin et al., 2018). Therefore, the progress on understanding deep learning is also greatly affected by the new theoretical research of kernel methods.

In recent years, efforts to explain why deep models generalize well have become a hot and important open problem in learning theory. Uniform convergence is a powerful tool in learning theory for understanding the generalization ability of learners, and it is also widely used in the generalization analysis of deep learning. Although ongoing endeavors have developed a considerable amount of non-vacuous generalization error bounds that reflect weak dependence or even independence on the network width and depth, these theoretical results are elaborate and algorithm-based (i.e. theoretical analysis incorporates the implicit regularization of stochastic gradient descent (SGD)) (Neyshabur et al., 2017; Soudry et al., 2018; He et al., 2019; Foret et al., 2021; Lei & Ying, 2021), which makes such theoretical results not applicable to all deep models. The capacity/complexity-based generalization analysis can provide a general theoretical guarantee for the surprising generalization performance of deep learning. However, capacity-based generalization bounds tend to have a strong dependence on the network depth, which makes theoretical results often of limited significance.

Deep kernel learning (DKL), which combines the representation power and structural prior knowledge of deep learning with the non-parametric flexibility of kernel methods, is an important problem of kernel methods for deep learning. A satisfactory and complete study of deep kernel learning should cover three aspects: 1) the design of deep kernels, 2) the efficiency of training algorithms, and 3) the analysis of generalization property. The design of deep kernels has been developed to a certain extent (Cho & Saul, 2009; Wilson et al., 2016b;a; Al-Shedivat et al., 2017; Lee et al., 2018), which mainly benefits from Gaussian processes. In terms of training algorithms, the studies are often related to specific deep kernel machines (Cho & Saul, 2009; Mairal et al., 2014; Al-Shedivat et al., 2017) or involve many heuris-

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China [3]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. Correspondence to: Min-Ling Zhang <zhangml@seu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

tics (Zhuang et al., 2011). As for the theoretical analysis of generalization, the studies are still in the early stage. Almost all existing works are for shallow kernel learning problems (Bartlett & Mendelson, 2002; Ying & Campbell, 2010; Lei & Ding, 2014), and the kernel machines involved do not exceed two layers (Zhuang et al., 2011). Therefore, it is necessary to perform generalization analysis for DKL. As a matter of fact, theoretical research on deep kernel learning can also promote the understanding of deep learning (Jacot et al., 2018).

This paper analyzes the generalization of deep kernel learning from the perspective of capacity, and aims to provide a general theoretical guarantee for the generalization ability of deep kernel machines (DKMs). The DKM is a special deep model based on kernel methods. The most significant difference between the DKM and the deep neural network (DNN) is that the nonlinear mapping (i.e., the kernel mapping) is implicitly induced by the kernel function, i.e., the nonlinear mapping is implicit. Golowich et al. (2018) revealed that existing tight generalization bounds on capacity-based generalization analysis of deep neural networks are accompanied by strong assumptions about the properties of activation functions and the norm constraints on the parameter matrix of each layer. One cannot make assumptions on the implicit nonlinear mapping in DKMs and the Reproducing Kernel Hilbert Space (RKHS) norm, which makes existing methods of generalization analysis for DNNs inapplicable to DKMs. In view of the challenges brought by the implicit nonlinear mapping in deep kernel learning, we need to develop a specific generalization analysis method for deep kernel learning and further derive meaningful generalization error bounds.

In this paper, we derive novel and nearly-tight capacity-based generalization bounds based on the uniform covering number and the Rademacher chaos complexity for DKMs and deep multiple kernel machines (DMKMs), respectively. Specifically, we first obtain the composite relationship of uniform covering numbers, which is used to deal with the difficulties in theoretical analysis caused by the implicit nonlinear mapping of DKL, then derive bounds for DKMs based on the uniform covering number which can be bounded using pseudo-dimensions, and derive bounds for DMKMs based on the Rademacher chaos complexity which can be bounded using kernel pseudo-dimensions. Furthermore, we bound pseudo-dimensions and kernel pseudo-dimensions for some specific classes, and further estimate their uniform covering numbers and Rademacher chaos complexities, respectively. Finally, we provide a lower bound for DKMs based on the Rademacher complexity.

To our knowledge, this is the first formal attempt to extend generalization analysis to the case of deep kernels and derive capacity-based generalization bounds for DKL with mild dependence on the depth. Major contributions of the paper include:

- We prove novel and nearly-tight capacity-based generalization bounds based on the complexity of the whole hypothesis space of deep kernel models, which provides general theoretical guarantees for DKL.
- We introduce and formally describe the composite relationship between layers of DKMs/DMKMs for the generalization analysis of DKL, which overcomes the difficulties brought by the implicit nonlinear mapping in DKL for theoretical analysis and leads to bounds with square-root dependence on the depth (outside of log terms).
- We further show how to estimate the uniform covering number and the Rademacher chaos complexity for the function class of DKMs/DMKMs by bounding the pseudo-dimension and the kernel pseudo-dimension.

We structure our work as follows. We first introduce the related work in Section 2, followed by an overview of the definitions of related complexities, the problem setting on DKL and the notation in Section 3. We then present our main theoretical results in Sections 4 and 5, which are the generalization bound based on the uniform covering number for DKL and the generalization bound based on the Rademacher chaos complexity for deep multiple kernel learning (DMKL), respectively. In Section 6, we show how to estimate the relative complexities for DKMs and DMKMs. In Section 7, we provide a lower bound based on the Rademacher complexity for DKL. In Section 8, we provide a discussion of the implications and inspirations of our theoretical results. Finally, we give a conclusion of our work in Section 9.

2. Related Work

In this section, we introduce the related work about kernel methods for deep learning and capacity-based generalization bounds for deep learning.

2.1. Kernel Methods for Deep Learning

A considerable amount of deep kernels and DKMs have been proposed to link kernel methods with deep learning. Various deep kernels were designed to simulate the computation of deep neural networks based on Gaussian processes (Cho & Saul, 2009; Hazan & Jaakkola, 2015; Wilson et al., 2016a;b; Al-Shedivat et al., 2017; Lee et al., 2018). Convolution kernels and convolution kernel networks were used to encode the invariance of image representations Mairal et al. (2014); Mairal (2016). Furthermore, SVM-based deep stacking networks (Wang et al., 2019), deep spectral kernel

networks (Xue et al., 2019; Li et al., 2020; 2022) and deep models based on multiple kernel fusion (Song et al., 2017) were all designed to introduce kernels into deep learning. In terms of theoretical analysis, The relationship between the representer theorem and DKMs was established (Bohn et al., 2019). Regularizing deep neural networks with Reproducing Kernel Hilbert Space norm was proposed (Bietti et al., 2019) and generalization error bounds of specific deep networks were derived (Suzuki, 2018) for the generalization theory. Similarity indexes and kernel PCA were used to measure the relationship between representations in deep networks (Kornblith et al., 2019; Montavon et al., 2011). The neural tangent kernel obtained at initialization was proven to control the learning dynamics of gradient descent in the over-parameterized regime (Jacot et al., 2018), and its inductive bias was also studied (Bietti & Mairal, 2019). Nevertheless, there is still a lack of new theories for kernel methods to analyze deep learning (Belkin et al., 2018). To make up for the gap in the generalization theory, in this paper, we derive capacity-based generalization bounds for DKMs and DMKMs, respectively.

2.2. Capacity-based Generalization Bounds for Deep Learning

The capacity-based generalization bounds established by traditional statistical learning theory aim to provide general theoretical guarantees for deep learning. Goldberg & Jerrum (1995); Bartlett & Williamson (1996); Bartlett et al. (1998) proposed upper bounds based on the VC dimension for DNNs. Unfortunately, these theoretical results heavily rely on both the depth and the width of the network, which makes these bounds less attractive once the model size is extremely large, even for the tightest upper bounds based on the VC dimension by Bartlett et al. (2019). There are many bounds that can alleviate the dependence on the width, but often still have a strong dependence on the depth. Neyshabur et al. (2015) used Rademacher complexity to prove the bound with exponential dependence on the depth. Neyshabur et al. (2018) and Bartlett et al. (2017) used the PAC-Bayesian analysis and the covering number to obtain bounds with polynomial dependence on the depth, respectively. Golowich et al. (2018) provided bounds with (sublinear) square-root dependence on the depth and further improved the bounds to be depth-independent. However, these results are all accompanied by strong assumptions about the properties of activation functions and the norm constraints on the parameter matrix of each layer, making these methods of generalization analysis invalid for DKL.

3. Preliminaries

In the context of classification, given a dataset $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with n samples which are identically and independently distributed (i.i.d.) from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. Let $[n] := \{1, \ldots, n\}$ for any natural number n. For any function $f : \mathcal{X} \to \mathcal{Y}$, we denote the expected risk $\mathbb{E}(f)$ as $\mathbb{E}_{(\boldsymbol{x},y)\sim P}[\ell(f(\boldsymbol{x}), y)]$ and denote the empirical risk $\widehat{E}_D(f)$ with respect to the training dataset D as $\frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a kernel function and denote the RKHS by $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$. In this paper, we assume that all kernels fulfill $r := \sup \sqrt{K(\boldsymbol{x}, \boldsymbol{x})} < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$.

3.1. Deep Kernel Learning

Although the nonlinear mapping induced by a kernel is implicit, we can avoid the limitation and directly construct DKMs by combining nonlinear mappings in a composite way:

$$\begin{aligned} &K^l(\boldsymbol{x}, \boldsymbol{x}') = \left\langle \Phi^l(\boldsymbol{x}), \Phi^l(\boldsymbol{x}') \right\rangle \\ =&\left\langle \phi^l\left(\phi^{l-1}(\cdots \phi^1(\boldsymbol{x}))\right), \phi^l\left(\phi^{l-1}(\cdots \phi^1(\boldsymbol{x}'))\right) \right\rangle, \end{aligned}$$

where K^l denotes the l-layer deep kernel, ϕ^l denotes the l-th layer kernel mapping. With the method of constructing deep kernels, we can denote the l-layer DKM as

$$f(\boldsymbol{x}) = {\boldsymbol{W}^l}^\top \Phi^l(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K^l(\boldsymbol{x}, \boldsymbol{x}_i).$$

Let $\mathcal{H}^l$ represent the RKHS induced by the deep kernel K^l and $\|\cdot\|_{\mathcal{H}^l}$ denote the norm in $\mathcal{H}^l$. We define a class of i-th layer deep kernels as follows:

$$\mathcal{K}^i_{sin} = \left\{ K^i(\cdot,\cdot) = g^i\left(\left[K^{i-1}(\cdot,\cdot)\right]\right) \right\},$$

where $K^{i-1}(\cdot,\cdot) \in \mathcal{K}^{i-1}_{sin}$, $i \in \{2, \cdots, l\}$, g^i is a function produced by $K^i(\cdot,\cdot)$ which composites $K^{i-1}(\cdot,\cdot)$, while ensuring that the composite result is still a valid kernel.

For ease of understanding, we consider an example of a two-layer Gaussian kernel. A Gaussian kernel is typically defined as $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right)$, where $\gamma > 0$ is the kernel parameter. We can construct a two-layer Gaussian kernel by compositing the nonlinear mapping corresponding to the Gaussian kernel:

$$\begin{aligned} &K^2(\boldsymbol{x}, \boldsymbol{x}') \\ =&\left\langle \Phi^2(\boldsymbol{x}), \Phi^2(\boldsymbol{x}') \right\rangle = \left\langle \phi\left(\phi(\boldsymbol{x})\right), \phi\left(\phi(\boldsymbol{x}')\right) \right\rangle \\ =&e^{-\gamma\left\|\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}')\right\|^2} = e^{-2\gamma\left(1 - K(\boldsymbol{x}, \boldsymbol{x}')\right)} = \kappa e^{2\gamma K(\boldsymbol{x}, \boldsymbol{x}')}, \end{aligned}$$

where κ is a constant that can be omitted. Hence, the corresponding g function (operation) induced by the composition of the Gaussian kernel to other kernels is $g(\cdot) = \kappa e^{2\gamma(\cdot)}$.

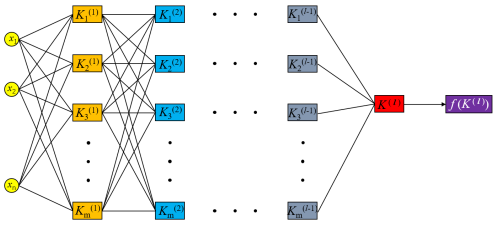

Figure 1. The architecture of DMKL framework. We show the deep multiple kernel machine with l layers, where kernels in i-th layer are from the kernel domain of i-th layer.

The function class of DKMs can be denoted as

$$\mathcal{F} = \{\boldsymbol{x} \mapsto \sum_{i=1}^{n} \alpha_i K^l(\boldsymbol{x}, \boldsymbol{x}_i) : \sum_{i,j} \alpha_i \alpha_j K^p(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 1, \\ K^p(\boldsymbol{x}_i, \boldsymbol{x}_i) \leq A^2, K^p(\cdot,\cdot) \in \mathcal{K}^p_{sin}, \\ \boldsymbol{x}_i \in \mathcal{X}, \forall p \in [l]\}.$$

Then the DKL framework can be cast as the following minimization problem:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i f\left(\boldsymbol{x}_i\right)\right) + \lambda \|f\|_{\mathcal{H}^l}, \tag{1}$$

where $\ell(\cdot)$ is the loss function and λ is a positive regularization parameter. When the loss function is hinge loss, then (1) is reduced to deep support vector machines (SVMs).

However, there are some limitations in the DKMs mentioned above, they are often specific, non-automated and inflexible since the type of kernels involved is single. Therefore, we have to further consider allowing a combination of a variety of different kernels when designing deep kernels, which we refer to as DMKL.

To this end, we first consider the initial set of basis kernel functions as $\mathcal{K}^1_{mul} = \{K_1, \ldots, K_m\}$, then we define a domain (or class) of i-th layer deep multiple kernels as follows:

$$\mathcal{K}^i_{mul} \\ = \left\{K^i(\cdot,\cdot) = g^i \circ h^{i-1}\left(\left[K^{i-1}_1(\cdot,\cdot), \ldots, K^{i-1}_m(\cdot,\cdot)\right]\right)\right\}, \tag{2}$$

where $K^{i-1}_t(\cdot,\cdot) \in \mathcal{K}^{i-1}_{mul}(\forall t \in [m])$, $i \in \{2, \cdots, l\}$, h^{i-1} is a function that combines multiple $(i-1)$-layer deep kernels (such as the linear or convex combination of several base kernels in multiple kernel learning), g^i is a function which composites the results of h^{i-1}, while ensuring that the composite result is still a valid kernel. The function class of DMKMs is similar to that of DKMs, the main difference is that $K^p(\cdot,\cdot) \in \mathcal{K}^p_{mul}$ instead of $\mathcal{K}^p_{sin}$. Hence, we can formulate the DMKL framework as the following two-layer minimization problem:

$$\min_{K^l \in \mathcal{K}^l_{mul}} \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i f\left(\boldsymbol{x}_i\right)\right) + \lambda \|f\|_{\mathcal{H}^l}, \tag{3}$$

where $\ell(\cdot)$ is the loss function, λ is a positive regularization parameter. As a matter of fact, if the h^{i-1} operation is the identity transformation and the initial kernel domain contains only a single type of kernels, it degenerates into DKL. The architecture of DMKL is shown in Figure 1.

3.2. Related Complexity Measures

In this paper, we use the uniform covering number to bound the sample complexity for DKL. The uniform covering number can be bounded by the pseudo-dimension:

Definition 3.1 (uniform covering number)**.** A subset $\mathcal{C} \subseteq \mathcal{G}$ is an ε-cover of a function class $\mathcal{G}$ under the metric d if for any $g \in \mathcal{G}$ and $\varepsilon > 0$ there exists $c \in \mathcal{C}$ with $d(g,c) \leq \varepsilon$. The covering number $\mathcal{N}(\varepsilon, \mathcal{G}, d)$ is the size of the smallest ε-cover of $\mathcal{G}$. Given a dataset D of size n, we define the **uniform covering number** corresponding to the metric d_p for a function class $\mathcal{G}$:

$$\mathcal{N}_p(\varepsilon, \mathcal{G}, n) = \max_{|D|=n} \left\{\mathcal{N}\left(\varepsilon, \mathcal{G}_{|D}, d_p\right)\right\},$$

where $\mathcal{G}_{|D}$ denotes the restriction of the function class $\mathcal{G}$ to the dataset D (that is, the set of restrictions to D of all functions in $\mathcal{G}$).

Definition 3.2 (pseudo-dimension)**.** Let $\mathcal{F}$ be a set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. We say that $D_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$ is **pseudo-shattered** by $\mathcal{F}$ if there are real numbers $\{r_i \in \mathbb{R} : i \in [n]\}$ such that for any $b \in \{-1, 1\}^n$ there is a function $f \in \mathcal{F}$ with property $\operatorname{sgn}\left(f(\boldsymbol{x}_i) - r_i\right) = b_i$ for any $i \in [n]$. Then, we define a **pseudo-dimension** $d^{\mathrm{p}}_{\mathcal{F}}$ of $\mathcal{F}$ to be the maximum cardinality of D_n that is pseudo-shattered by $\mathcal{F}$.

In this paper, we also use the Rademacher complexity to provide a lower bound for DKL.

Definition 3.3 (Rademacher complexity)**.** Let $\mathcal{F}$ be a class of real-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. The empirical **Rademacher complexity** over $\mathcal{F}$ is defined by

$$\hat{R}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f\left(\boldsymbol{x}_i\right)\right|\right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables, and we refer to the expectation $R_n(\mathcal{F}) = \mathbb{E}_D[\hat{R}_D(\mathcal{F})]$ as the Rademacher complexity of $\mathcal{F}$.

The Rademacher chaos complexity was introduced into the discussion on learning rates of MKL machines (Ying &

Campbell, 2010). Some comprehensive studies and significant results were provided to justify that the Rademacher chaos complexity is appropriate to treat the learning rates (Ying & Campbell, 2009; 2010; Lei & Ding, 2014), which also demonstrate that the Rademacher chaos complexity has inherited the advantage of the Rademacher complexity. Here we use the Rademacher chaos complexity to perform generalization analysis for DMKL.

Definition 3.4 (Rademacher chaos complexity)**.** Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. The empirical **Rademacher chaos complexity** over $\mathcal{F}$ is defined by

$$\hat{U}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i,j\in[n],i<j}\epsilon_i\epsilon_j f\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\right|\right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables, and we refer to the expectation $U_n(\mathcal{F}) = \mathbb{E}_D[\hat{U}_D(\mathcal{F})]$ as the Rademacher chaos complexity of $\mathcal{F}$.

The Rademacher chaos complexity can be bounded by the kernel pseudo-dimension of the set of candidate kernels. The kernel pseudo-dimension measures the complexity of a kernel domain.

Definition 3.5 (kernel pseudo-dimension)**.** Let $\mathcal{K}$ be a set of reproducing kernel functions mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. We say that $S_n = \{(\boldsymbol{x}_i, \boldsymbol{x}_i') \in \mathcal{X} \times \mathcal{X} : i \in [n]\}$ is **pseudo-shattered** by $\mathcal{K}$ if there are real numbers $\{r_i \in \mathbb{R} : i \in [n]\}$ such that for any $b \in \{-1, 1\}^n$ there is a function $K \in \mathcal{K}$ with property $\operatorname{sgn}\left(K\left(\boldsymbol{x}_i, \boldsymbol{x}_i'\right) - r_i\right) = b_i$ for any $i \in [n]$. Then, we define a **kernel pseudo-dimension** $d_{\mathcal{K}}^{\mathrm{k}}$ of $\mathcal{K}$ to be the maximum cardinality of S_n that is pseudo-shattered by $\mathcal{K}$.

4. Generalization Bounds Based on the Uniform Covering Number

In this section, we are committed to establishing a generalization error bound based on the uniform covering number for DKMs. Since the composite relationship between layers of DKMs, to derive tight generalization error bounds, we first have to reveal the relationship between uniform covering numbers of composite function classes, and then bound these uniform covering numbers with the pseudo-dimension.

Suppose that the input space is $\mathcal{X} = \mathbb{R}^d$, a l-layer DKM can be simplified to the composition of a series of mappings:

$$\begin{aligned} &f = f_l \circ \cdots \circ f_1 \circ f_0(x) \\ &f_0 : \mathbb{R}^d \to \mathcal{H}^1, \\ &f_i : \mathcal{H}^i \to \mathcal{H}^{i+1}, 1 \le i \le l-1, \\ &f_l : \mathcal{H}^l \to \{-1, 1\}. \end{aligned}$$

$\mathcal{H}^i$ represents the i-th layer RKHS, the class of functions on the i-th layer can be denoted as $\mathcal{F}^{(i)}$, then the class of l-layer DKMs can be expressed as

$$\mathcal{F} = \mathcal{F}^{(l)} \circ \cdots \circ \mathcal{F}^{(2)} \circ \mathcal{F}^{(1)},$$

we ignore $\mathcal{F}^{(0)}$ since it is induced by kernels implicitly. Then, we first have the following lemma:

Lemma 4.1. *Let $\mathcal{F}^{(1)} \subset \mathcal{Y}_1^{\mathcal{X}}$ and $\mathcal{F}^{(2)} \subset \mathcal{Y}_2^{\mathcal{Y}_1}$ be function classes of kernels, $\mathcal{F} = \mathcal{F}^{(1)} \circ \mathcal{F}^{(2)}$ be the class of composite functions, we have*

$$\mathcal{N}_p(2\varepsilon, \mathcal{F}, n) \le \mathcal{N}_p(\varepsilon, \mathcal{F}^{(1)}, n) \cdot \mathcal{N}_p(\varepsilon, \mathcal{F}^{(2)}, n).$$

Proof Sketch. We first bound the cover of $\mathcal{F}$ on the dataset D through appropriate scaling, and then complete the proof with the arbitrariness of D. □

Lemma 4.1 reveals the relationship between uniform covering numbers of composite function classes, similar results exist in the deep neural networks literature (Nagarajan & Kolter, 2019; Wei & Ma, 2019; Ledent et al., 2021; Graf et al., 2022).

We denote the pseudo-dimension of the i-th layer function class as $d_i^{\mathrm{p}} \ge 1, i \in [l]$, Let $d_{max}^{\mathrm{p}} = \max_i\{d_i^{\mathrm{p}}\}$, then we have the following lemma:

Lemma 4.2. *Let $\mathcal{F}$ be a class of functions, which represents l-layer DKMs. Let $\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n)$ be the uniform covering number defined on $\mathcal{F}$. Then we have*

$$\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \le \left(\frac{lenA}{\varepsilon}\right)^{ld_{max}^{\mathrm{p}}}.$$

Proof Sketch. First, we have to use d_i^{p} to bound $\mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(i)}, n)$, $i \in [l]$. Then, with Lemma 4.1, $\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n)$ can be bounded using d_{max}^{p}. □

With these conclusions about the uniform covering number, the generalization error bound for DKMs can be derived as follows:

Theorem 4.3. *Let $\mathcal{F}$ be a class of functions, which represents l-layer DKMs, taking values in $\{-1, 1\}$. Let $\gamma > 0, 0 < \delta < 1$ and define the γ-margin cost function by*

$$\psi(x) = \begin{cases} 0, & \gamma \le x \\ 1 - x/\gamma, & 0 \le x \le \gamma \\ 1, & x \le 0 \end{cases}.$$

Given a dataset D of size n. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for any $f \in \mathcal{F}$:

$$\mathbb{E}(f) \le \widehat{E}_D(f) + \sqrt{\frac{8\ln\frac{2}{\delta} + 8ld_{max}^{\mathrm{p}}\ln(\frac{4leAn}{\gamma})}{n}}.$$

Remark 4.4. The generalization error bounds for traditional kernel learning problem do not apply to deep kernel learning since the impact of "deep" is not considered. The explicit nonlinear mapping in deep learning makes one might often be constrained to perform generalization analysis using a peeling argument, but this is not feasible in DKL. We introduce the composite relationship between layers to eliminate the challenges posed by the implicit nonlinear mapping in DKL, which makes our theoretical results non-trivial extensions of kernel learning. The bound in Theorem 4.3 is represented by the pseudo dimension of kernels. It is obvious that the generalization error bounds we derived are only related to the depth of DKMs and the number of samples, and the convergence rate is $\widetilde{O}(\sqrt{l/n})$.

5. Generalization Bounds Based on the Rademacher Chaos Complexity

In this section, we are committed to establishing a tight generalization error bound based on the Rademacher chaos complexity for DMKMs. When considering the case of DMKL, the pseudo-dimension cannot capture the complexity information of multiple deep kernels, so we use the kernel pseudo-dimension to measure the complexity of a kernel domain (or class).

Let $\mathcal{K}_{mul}$ be a domain of (multiple) kernel functions on $\mathcal{X} \times \mathcal{X}$ and $D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, we define pseudo-metrics on $\mathcal{K}_{mul}$ as follows:

$$d_\infty(f,g) := \max_{1\le i,j\le n} |f(\boldsymbol{x}_i,\boldsymbol{x}_j) - g(\boldsymbol{x}_i,\boldsymbol{x}_j)|,$$
$$d_2(f,g) := \sqrt{\frac{1}{n^2}\sum_{1\le i<j\le n} |f(\boldsymbol{x}_i,\boldsymbol{x}_j) - g(\boldsymbol{x}_i,\boldsymbol{x}_j)|^2}.$$

The uniform covering number $\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}, n)$ and $\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n)$ correspond to the pseudo-metric $d_\infty(f,g)$ and $d_2(f,g)$. Let $\mathcal{K}^i_{mul}$ be the domain of i-th layer deep multiple kernels as defined in (2), that is, the kernels in $\mathcal{K}^i_{mul}$ are transformed by multiple $(i-1)$-th layer deep kernels in $\mathcal{K}^{i-1}_{mul}$, then the domain of l-layer deep multiple kernels can be expressed as

$$\mathcal{K}_{mul} = \mathcal{K}^{(l)}_{mul} \circ \cdots \circ \mathcal{K}^{(2)}_{mul} \circ \mathcal{K}^{(1)}_{mul}.$$

We denote the kernel pseudo-dimension of the i-th layer kernel domain $\mathcal{K}^i_{mul}$ as $d^{\mathrm{k}}_i \ge 1, i \in [l]$. Let $d^{\mathrm{k}}_{max} = \max_i\{d^{\mathrm{k}}_i\}$, then we have:

Lemma 5.1. *Let $\mathcal{K}_{mul}$ be a domain of l-layer deep multiple kernels defined by (2). Let $\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}, n)$ and $\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n)$ be uniform covering numbers defined on $\mathcal{K}_{mul}$. Then we have*

$$\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n) \le \left(\frac{8l^2eA^4}{\varepsilon^2}\right)^{ld^{\mathrm{k}}_{max}},$$
$$\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}, n) \le \left(\frac{len^2A^2}{\varepsilon}\right)^{ld^{\mathrm{k}}_{max}}.$$

Proof Sketch. First, we use d^{k}_i to bound $\mathcal{N}_2(\varepsilon, \mathcal{K}^{(i)}_{mul}, n)$ and $\mathcal{N}_\infty(\varepsilon, \mathcal{K}^{(i)}_{mul}, n)$, $i \in [l]$, respectively. Then, with Lemma 4.1, $\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n)$ and $\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}, n)$ can be bounded using d^{k}_{max}. □

We then establish a generalization error bound based on the Rademacher chaos complexity as follows:

Lemma 5.2. *Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\{-1,1\}$, $\mathcal{K}_{mul}$ be a class of multiple kernel functions defined like (2). Given a dataset D of size n. Let $\gamma > 0, 0 < \delta < 1$ and the loss function be the γ-margin cost function. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for any $f \in \mathcal{F}$:*

$$\mathbb{E}(f) \le \hat{E}_D(f) + \frac{4}{\lambda\gamma}\left(\frac{2U_n(\mathcal{K}_{mul})}{n}\right)^{\frac{1}{2}} + \frac{4A}{\lambda\gamma\sqrt{n}} + \left(\frac{\ln\left(\frac{1}{\delta}\right)}{2n}\right)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.$$

Proof Sketch. Let the union of the unit balls of RKHSs be $\mathcal{B}_{\mathcal{K}_{mul}}$, $\mathcal{B}_\lambda = \frac{1}{\lambda}\mathcal{B}_{\mathcal{K}_{mul}}$, $\Phi(D) = \sup_{f\in\mathcal{B}_\lambda} |\mathbb{E}(f) - \hat{E}_D(f)|$. We first have $\Phi(D) \le \mathbb{E}_D[\Phi(D)] + \sqrt{\ln(1/\delta)/2n}$ by McDiarmid's inequality. Then, with the contraction property of Rademacher complexities and the reproducibility of kernels, we can derive the upper bound on $\mathbb{E}_D[\Phi(D)]$. Finally, substituting the upper bound into the above inequality, the desired bound is immediate. □

With the composite relationship between layers, we can bound the Rademacher chaos complexity with the kernel pseudo-dimension for DMKMs:

Lemma 5.3. *Let $\mathcal{K}_{mul}$ be a domain of l-layer deep multiple kernels. For any $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, there holds*

$$U_n(\mathcal{K}_{mul}) \le 220eA^2ld^{\mathrm{k}}_{max}\ln l.$$

Proof Sketch. Combining with Lemma 5.1 and the relationship between the uniform covering number and the Rademacher chaos complexity, the desired bound can be derived by appropriate scaling. □

With Lemma 5.2 and Lemma 5.3, the generalization error bound based on the Rademacher chaos complexity for DMKMs is given as follows:

Theorem 5.4. *Let $\mathcal{F}$ be a class of functions, which represents l-layer DMKMs. Given a dataset D of size n. Let $\gamma > 0, 0 < \delta < 1$ and the loss function be γ-margin cost function. Then, with probability at least $1-\delta$, there holds for any $f \in \mathcal{F}$:*

$$\mathbb{E}(f) \leq \hat{E}_D(f) + \frac{8}{\lambda\gamma}\left(\frac{110eA^2ld^{\mathrm{k}}_{max}\ln l}{n}\right)^{\frac{1}{2}} + \frac{4A}{\lambda\gamma\sqrt{n}} + \left(\frac{\ln\left(\frac{1}{\delta}\right)}{2n}\right)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.$$

Remark 5.5. The theorems above reveal that the generalization analysis of deep kernel learning is not only to simply bound the (kernel) pseudo-dimension of each layer, but also to clarify the composite relationship of uniform covering numbers, which makes our bounds with square-root dependence (outside of log terms) on the depth non-trivial. It is obvious that the bound for DMKMs is tighter than DKMs with a faster convergence rate $O(\sqrt{l\ln l/n})$, since the kernel pseudo-dimension is used to bound the Rademacher chaos complexity. Next we will further show that the kernel pseudo-dimension of DMKMs is related to the number of basis kernel functions. Compared with the results of (Ying & Campbell, 2009), we extend the theory bounds of multiple kernel learning to the deep case (i.e., deep multiple kernel learning), and give the generalization bound corresponding to the deep multiple kernel function class produced by multiple combinations and compositions of multiple kernels, and show how to estimate the complexity of the deep multiple kernel function class. Although the bound has some dependence on the parameter λ, our focus here is to show that at least some compromise assumption leads to bounds with square-root dependence (outside of log terms) on the depth, and hope this dependence can be improved in further work.

6. Estimating the Uniform Covering Number and the Rademacher Chaos Complexity

In this section, for both DKMs and DMKMs, we further estimate the uniform covering number which can be bounded by the pseudo-dimension and the Rademacher chaos complexity which can be bounded by the kernel pseudo-dimension for some common classes, respectively.

6.1. Estimating the Uniform Covering Number

For l-layer DKMs, we will bound the pseudo-dimension for each layer, then further bound the uniform covering number. The pseudo-dimension of kernels is as follows:

Theorem 6.1. *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function and let $\Phi : \mathcal{X} \to \mathcal{H}$ be a feature mapping associated to K. Let $D \subseteq \left\{\boldsymbol{x} : K(\boldsymbol{x},\boldsymbol{x}) \leq r^2\right\}$ be a dataset of size n, and let*

$$\mathcal{F} = \{\boldsymbol{x} \mapsto \langle \mathbf{w}, \Phi(\boldsymbol{x})\rangle : \min_{\boldsymbol{x}} |\mathbf{w}^\top \Phi(\boldsymbol{x})| = 1 \bigwedge \|\mathbf{w}\|_{\mathcal{H}} \leq \Lambda\}$$

for some $\Lambda \geq 0$. Then the pseudo-dimension d^{p} of the function class $\mathcal{F}$ satisfies

$$d^{\mathrm{p}} \leq r^2\Lambda^2.$$

Proof Sketch. We first obtain the relationship between d and Λ from the definition of the pseudo-dimension and introduce the expectation of $\boldsymbol{y} = (y_1, \ldots, y_d) \in \{-1, 1\}^d$. Then, using Jensen's inequality and the property of convex functions, the desired bound is derived. □

6.2. Estimating the Rademacher Chaos Complexity

The above theoretical results reveal that the Rademacher chaos complexity of DMKMs is mainly affected by the depth and the kernel pseudo-dimension which measures the complexity of a kernel domain $\mathcal{K}_{mul}$. Moreover, different kernel domains correspond to different kernel pseudo-dimensions. Here, for some common kernel domains, we will bound the kernel pseudo-dimension and further estimate the Rademacher chaos complexity corresponding to these specific classes of DMKMs.

According to the definition of kernel domains, it is obvious that the architecture of DMKMs is controlled by combinations of multiple kernels in the $(i-1)$-layer and composite methods of the i-layer to the combined results of $(i-1)$-layer, i.e., h^{i-1} and g^i. For a kernel set S with k basic kernels, the common types of combination, i.e., common-used h^{l-1} operations, are mainly span, linear, and convex:

$$\mathcal{K}_{\text{span}}(S) \overset{\text{def}}{=} \left\{K_{\boldsymbol{\lambda}} = \sum_{i=1}^{k}\lambda_i K_i \mid \lambda_i \in \mathbb{R}\right\},$$

$$\mathcal{K}_{\text{line}}(S) \overset{\text{def}}{=} \left\{K_{\boldsymbol{\lambda}} = \sum_{i=1}^{k}\lambda_i K_i \mid K_{\boldsymbol{\lambda}} \succcurlyeq 0,\ \sum_{i=1}^{k}\lambda_i = 1\right\},$$

$$\mathcal{K}_{\text{conv}}(S) \overset{\text{def}}{=} \left\{K_{\boldsymbol{\lambda}} = \sum_{i=1}^{k}\lambda_i K_i \mid \lambda_i \geq 0,\ \sum_{i=1}^{k}\lambda_i = 1\right\}.$$

It can be known that $\mathcal{K}_{\text{conv}}(S) \subseteq \mathcal{K}_{\text{line}}(S) \subseteq \mathcal{K}_{\text{span}}(S)$.

Then we consider the composite method produced by the common-used kernel functions, such as Polynomial kernel $K(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}^\top\boldsymbol{y})^d$, Gaussian kernel $K(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2\right)$, and Laplacian kernel $K(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\mu\|\boldsymbol{x}-\boldsymbol{y}\|\right)$. The compositions of these kernels on other kernels yield corresponding g^i operations in the defi-

nition of kernel domains as follows:

$$\begin{array}{ll} g^i_{\text{Pol}}(K) = (K)^d, & d > 1 \\ g^i_{\text{Gau}}(K) = \kappa e^{2\gamma K}, & \kappa, \gamma > 0 \\ g^i_{\text{Lap}}(K) = e^{\sqrt{2\mu^2(K-1)}}, & \mu > 0 \end{array} . \quad (4)$$

Obviously, these g^i operations are monotonous (non-decreasing) with respect to their input kernels.

With the above discussion on the methods of combination and composition, we can bound the Rademacher chaos complexity for DMKMs as follows:

Theorem 6.2. *Let $\mathcal{K}_{mul}$ be a domain of l-layer deep kernels. Given the initial set of basis kernels $\mathcal{K}^1 = \{K_1, \ldots, K_m\}$. For any $i \in \{2, \ldots, l\}$, $\mathcal{K}^i_{mul}$ is the kernel domain of i-th layer deep kernels. If the h^{i-1} operation is the type of span combination and the g^i operation is non-decreasing with respect to its input kernels, the kernel pseudo-dimension of the domain $\mathcal{K}^i_{mul}$ satisfies*

$$d^{\text{k}}(\mathcal{K}^i_{mul}) \leq m$$

and the Rademacher chaos complexity of $\mathcal{K}_{mul}$ is bounded by

$$U_n(\mathcal{K}_{mul}) \leq 220eA^2 l(\ln l)m.$$

Proof Sketch. According to the definition of kernel domains, combined with the monotonicity of the kernel pseudo-dimension, the desired bounds can be derived. □

Remark 6.3. If the given initial basic kernel set contains only a single type of kernel, i.e., $m = 1$, DMKMs degenerate into DKMs. Then we can immediately bound the Rademacher chaos complexity for DKMs:

$$U_n(\mathcal{K}_{sin}) \leq 220eA^2 l \ln l.$$

We can find that the Rademacher chaos complexity for DKMs here will lead to the generalization bound with a convergence rate $O(\sqrt{l \ln l / n})$, which is tighter than Theorem 4.3, The main reason is that they use different complexity measures, the Rademacher chaos complexity takes into account the data distribution to some extent.

7. A Lower Bound Based on the Rademacher Complexity

In this section, we provide a lower bound based on the Rademacher complexity for DKMs studied here. The formal result is the following:

Theorem 7.1. *Let $\mathcal{F}$ be a class of functions, which represents l-layer DKMs, taking values in $\{-1, 1\}$. Given a dataset D of size n. Then, there exists a $c > 0$ such that for any $f \in \mathcal{F}$:*

$$\hat{R}_D(\mathcal{F}) \geq \frac{cA}{\sqrt{n}}.$$

Remark 7.2. We can see that the lower bound of $\hat{R}_D(\mathcal{F})$ is $\Omega(\frac{1}{\sqrt{n}})$, which will result in a lower generalization bound of order $\Omega(\frac{1}{\sqrt{n}})$. Our derived upper bounds avoid exponential dependence on the depth, although not depth-independent but are square-root dependent on the depth. The depth-independent lower bound in Theorem 7.1 demonstrates that the capacity-based upper bounds on the uniform covering number are nearly-tight.

8. Discussion

Our capacity-based generalization bounds provide general worst-case theoretical guarantees for the generalization of DKL. The reason our generalization bounds do not appear to be ideally non-vacuous (i.e. less than 1) is that our bounds characterize the complexity of the whole hypothesis space rather than the effective hypothesis space learned by the algorithm. The effective hypothesis space is significantly smaller than the whole hypothesis space, so it is not surprising to expect much tighter generalization bounds. Moreover, such tight generalization bounds are algorithm-based, which reflect the implicit regularization of the optimization algorithm, and these generalization bounds cannot provide general theoretical guarantees for deep learning.

A major challenge in obtaining non-trivial capacity-based theoretical results is that when the deep model goes beyond 2 layers, the generalization bounds, which originally get smaller with increasing width, become vacuous since the introduction of the depth, so the heavy dependence on the depth should be at least reduced to be weaker than linear dependence. The common capacity-based generalization analysis method is to use a "peeling" argument, i.e., the complexity bound for l-layer networks is reduced to a complexity bound for $(l-1)$-layer networks. In this method, a product factor of the constant related to the Lipschitz property of the activation function and the upper bound of the norm of the weight metrix will be introduced for each reduction due to scaling. After applying the reduction l times, the multiplication of product factors with exponential dependence on the depth makes the bound vacuous. Only strong mathematical skills and assumptions can obtain tight bounds with weak dependence on the depth. The success of the peeling argument is premised on having 1) the explicit nonlinear mapping, and in addition, 2) some norm-based assumptions, and 3) some assumptions on the properties of activation functions are sufficient conditions for obtaining non-vacuous generalization bounds. Obviously, the implicit mapping of DKL directly isolates the peeling argument. However, our generalization analysis based on the composite relationship of hypothesis spaces provides a good demonstration for overcoming this challenge.

Since existing training algorithms for DKMs are difficult to generalize, how to train DKMs/DMKMs automatically and

efficiently is still an urgent issue to be studied. However, there is a very coincidental example of DMKMs that can be used to illustrate the difficulty of generalization analysis. Xue et al. (2019) and Li et al. (2020) proposed the (deep) spectral kernel network (DSKN) which use the inverse Fourier transform to make the implicit mapping explicit, the proposed method makes it possible to use the gradient descent algorithm to train DKMs easily. Li et al. (2020) gave the bound of spectral kernel models based on the Rademacher complexity. Li et al. (2022) extended the bound to deep (convolutional) spectral kernel models, which is depth-independent, with certain assumptions on the kernels. When these assumptions are not satisfied, or even if we can use some methods to make kernel mappings explicit, but the kernels involved in DKMs are not spectral kernels, the theoretical results in (Li et al., 2022) will no longer apply. Although DSKN satisfies the premise of the peeling argument, this explicit mapping is relatively fixed, and it is still difficult to constrain 2) and 3). At this time, in order to upper bound the Rademacher complexity of the deep kernel classes, we need to use an analysis method similar to that in deep learning, i.e., the "peeling" argument, which will still result in the bound that is exponentially dependent on the depth. Fortunately, our generalization analysis based on the composite relationship of hypothesis spaces provides a good demonstration for overcoming this challenge, which reduces such exponential depth dependency to the square-root one. We provide general and efficient theoretical guarantees with square-root dependence on the depth for DKL/DMKL.

9. Conclusion

In this paper, we propose novel and nearly-tight capacity-based bounds on the uniform covering number and the Rademacher chaos complexity for DKL. We then bound pseudo-dimensions and kernel pseudo-dimensions for some common classes and further estimate their uniform covering numbers and Rademacher chaos complexities, respectively.

In future work, we will extend our bounds to more general settings (especially for more types of combinations and composite methods of multiple kernels), and derive tighter capacity-based generalization bounds for DKL with weaker dependence on the depth (i.e., log-dependent or even depth-independent), and further design general and efficient deep kernel training algorithms to construct end-to-end DKMs/DMKMs with good generalization performance.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62225602).

References

Al-Shedivat, M., Wilson, A. G., Saatchi, Y., Hu, Z., and Xing, E. P. Learning scalable deep kernels with recurrent structure. *Journal of Machine Learning Research*, 18(82): 1–37, 2017.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 1999.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L. and Williamson, R. C. The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. *Neural Computation*, 8(3):625–628, 1996.

Bartlett, P. L., Maiorov, V., and Meir, R. Almost linear VC dimension bounds for piecewise polynomial networks. *Advances in Neural Information Processing Systems*, 11 (NIPS 1998):190–196, 1998.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30 (NIPS 2017):6240–6249, 2017.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. *Proceedings of Machine Learning Research*, 80(ICML 2018): 540–548, 2018.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32(NeurIPS 2019):12873–12884, 2019.

Bietti, A., Mialon, G., Chen, D., and Mairal, J. A kernel perspective for regularizing deep neural networks. *Proceedings of Machine Learning Research*, 97(ICML 2019): 664–674, 2019.

Bohn, B., Rieger, C., and Griebel, M. A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, 20(64):1–32, 2019.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. *Advances in Neural Information Processing Systems*, 22 (NIPS 2009):342–350, 2009.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving

generalization. In *Proceedings of 9th International Conference on Learning Representations*, number ICLR 2021, 2021.

Goldberg, P. W. and Jerrum, M. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148, 1995.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. *International Conference on Computational Learning Theory*, 75(COLT 2018):297–299, 2018.

Graf, F., Zeng, S., Niethammer, M., and Kwitt, R. On measuring excess capacity in neural networks. *arXiv:2202.08070v2*, 2022.

Hazan, T. and Jaakkola, T. S. Steps toward deep kernel methods from infinite neural networks. *arXiv:1508.05133v2*, 2015.

He, F., Liu, T., and Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32(NeurIPS 2019):1141–1150, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31 (NeurIPS 2018):8571–8580, 2018.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. E. Similarity of neural network representations revisited. *Proceedings of Machine Learning Research*, 97(ICML 2019): 3519–3529, 2019.

Ledent, A., Mustafa, W., Lei, Y., and Kloft, M. Norm-based generalisation bounds for multi-class convolutional neural networks. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 35(AAAI 2021):8279–8287, 2021.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. In *Proceedings of the 6th International Conference on Learning Representations*, number ICLR 2018, 2018.

Lei, Y. and Ding, L. Refined rademacher chaos complexity bounds with applications to the multikernel learning problem. *Neural Computation*, 26(4):739–760, 2014.

Lei, Y. and Ying, Y. Sharper generalization bounds for learning with gradient-dominated objective functions. In *Proceedings of 9th International Conference on Learning Representations*, number ICLR 2021, 2021.

Li, J., Liu, Y., and Wang, W. Automated spectral kernel learning. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 34(AAAI 2020):4618–4625, 2020.

Li, J., Liu, Y., and Wang, W. Convolutional spectral kernel learning with generalization guarantees. *Artificial Intelligence*, 313:103803, 2022.

Mairal, J. End-to-end kernel learning with supervised convolutional kernel networks. *Advances in Neural Information Processing Systems*, 29(NIPS 2016):1399–1407, 2016.

Mairal, J., Koniusz, P., Harchaoui, Z., and Schmid, C. Convolutional kernel networks. *Advances in Neural Information Processing Systems*, 27(NIPS 2014):2627–2635, 2014.

Montavon, G., Braun, M. L., and Müller, K. Kernel analysis of deep networks. *Journal of Machine Learning Research*, 12:2563–2581, 2011.

Nagarajan, V. and Kolter, J. Z. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *Proceedings of 6th International Conference on Learning Representations*, number ICLR 2019, 2019.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. *International Conference on Computational Learning Theory*, 40(COLT 2015):1376–1401, 2015.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30(NIPS 2017): 5947–5956, 2017.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of 6th International Conference on Learning Representations*, number ICLR 2018, 2018.

Song, H., Thiagarajan, J. J., Sattigeri, P., Ramamurthy, K. N., and Spanias, A. A deep learning approach to multiple kernel fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, number ICASSP 2017, pp. 2292–2296, 2017.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. In *Proceedings of 6th International Conference on Learning Representations*, number ICLR 2018, 2018.

Srebro, N. and Ben-David, S. Learning bounds for support vector machines with learned kernels. *International Conference on Computational Learning Theory*, (COLT 2006):169–183, 2006.

Suzuki, T. Fast generalization error bound of deep learning from a kernel perspective. *Proceedings of Machine Learning Research*, 84(AISTATS 2018):1397–1406, 2018.

Wang, J., Feng, K., and Wu, J. Svm-based deep stacking networks. *The Thirty-Third AAAI Conference on Artificial Intelligence*, (AAAI 2019):5273–5280, 2019.

Wei, C. and Ma, T. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *Advances in Neural Information Processing Systems*, 32 (NeurIPS 2019):9722–9733, 2019.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Stochastic variational deep kernel learning. *Advances in Neural Information Processing Systems*, 29(NIPS 2016): 2586–2594, 2016a.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. *Proceedings of Machine Learning Research*, 51(AISTATS 2016):370–378, 2016b.

Xue, H., Wu, Z., and Sun, W. Deep spectral kernel learning. *International Joint Conference on Artificial Intelligence*, (IJCAI 2019):4019–4025, 2019.

Ying, Y. and Campbell, C. Generalization bounds for learning the kernel problem. In *International Conference on Computational Learning Theory*, number COLT 2009, 2009.

Ying, Y. and Campbell, C. Rademacher chaos complexities for learning the kernel problem. *Neural Computation*, 22 (11):2858–2886, 2010.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th International Conference on Learning Representations*, number ICLR 2017, 2017.

Zhuang, J., Tsang, I. W., and Hoi, S. C. Two-layer multiple kernel learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, number AISTATS 2011, pp. 909–917, 2011.

A. Appendix Outline

In the appendix, we give the detailed proofs of those theoretical results which we only provide proof sketches in the main paper. Our main proofs include:

- The relationship between uniform covering numbers of composite function classes (Lemma 4.1).
- The bound of the uniform covering number for DKMs (Lemma 4.2).
- The generalization bound based on the uniform covering number for DKMs (Theorem 4.3).
- Bounds of uniform covering numbers for DMKMs (Lemma 5.1).
- The generalization bound based on the Rademacher chaos complexity for kernels (Lemma 5.2).
- The bound of the Rademacher chaos complexity for DMKMs (Lemma 5.3).
- The pseudo-dimension of kernels (Theorem 6.1).
- The kernel pseudo-dimension of multiple deep kernels (Theorem 6.2).
- The lower bound of the Rademacher complexity for DKMs (Theorem 7.1).

B. Generalization Bounds Based on the Uniform Covering Number

B.1. Proof of Lemma 4.1

Proof. Given a dataset D with n training samples drawn independently from a probability distribution P on $\mathcal{X} \times \mathbb{R}$. According to the definition of $\mathcal{F}$, we have

$$\begin{aligned}
&\mathcal{F}_{|D} \\
=&\left\{\left(f_2\left(f_1\left(\boldsymbol{x}_1\right)\right), \ldots, f_2\left(f_1\left(\boldsymbol{x}_n\right)\right)\right) \mid f_1 \in \mathcal{F}^{(1)}, f_2 \in \mathcal{F}^{(2)}\right\} \\
=&\bigcup_{\boldsymbol{z}_i \in \mathcal{F}_{|D}^{(1)}}\left\{\left(f_2\left(\boldsymbol{z}_1\right), \ldots, f_2\left(\boldsymbol{z}_n\right)\right) \mid f_2 \in \mathcal{F}^{(2)}\right\}.
\end{aligned}$$

Hence, we have $\mathcal{F}_{|D} \subseteq \mathcal{F}_{|D}^{(1)} \circ \mathcal{F}_{|D}^{(2)}$. Let $\mathcal{C}_{\mathcal{F}_{|D}^{(1)}} \subset \mathcal{F}_{|D}^{(1)}$ be a smallest ε-cover of $\mathcal{F}_{|D}^{(1)}$, i.e. $\operatorname{card}(\mathcal{C}_{\mathcal{F}_{|D}^{(1)}}) = \mathcal{N}\left(\varepsilon, \mathcal{F}_{|D}^{(1)}, d_{\mathcal{H}^1}\right)$. For any element $c_1 \in \mathcal{C}_{\mathcal{F}_{|D}^{(1)}}$, let $\mathcal{C}_{\mathcal{F}_{|D}^{(2)}}(c_1) \subset \mathcal{F}_{|D}^{(2)}$ be a smallest ε-cover of $\mathcal{F}_{|D}^{(2)}$, i.e., $\operatorname{card}(\mathcal{C}_{\mathcal{F}_{|D}^{(2)}}(c_1)) = \mathcal{N}\left(\varepsilon, \mathcal{F}_{|D}^{(2)}, d_{\mathcal{H}^2}\right)$.

According to the reproducing property of RKHS, combined with the Cauchy-Schwartz inequality, we have $|f(\boldsymbol{x}) - f\left(\boldsymbol{x}'\right)| = |\langle f, \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} d_{\mathcal{H}}\left(\boldsymbol{x}, \boldsymbol{x}'\right)$. This means that functions in the RKHS fulfill a Lipschitz-like condition, with Lipschitz constant given by the norm $\|f\|_{\mathcal{H}}$. Since $\|f\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j K\left(\cdot, \cdot\right) \leq 1$, without loss of generality, we can obtain that the function $f_2 \in \mathcal{F}_{|D}^{(2)}$ is 1-Lipschitz.

Let $c_2 \in \mathcal{C}_{\mathcal{F}_{|D}^{(2)}}(c_1)$, according to the definition, for any $f_1 \in \mathcal{F}_{|D}^{(1)}$ and any $f_2 \in \mathcal{F}_{|D}^{(2)}$, we have $\|f_1 - c_1\|_{\mathcal{H}^1} \leq \varepsilon$ and $\|f_2 - c_2\|_{\mathcal{H}^2} \leq \varepsilon$. Since

$$\begin{aligned}
&\|f_2 \circ f_1 - c_2 \circ c_1\|_{\mathcal{H}^1} \\
=&\|(f_2 \circ f_1 - f_2 \circ c_1) + (f_2 \circ c_1 - c_2 \circ c_1)\|_{\mathcal{H}^1} \\
\leq&\|f_2 \circ f_1 - f_2 \circ c_1\|_{\mathcal{H}^1} + \|f_2 \circ c_1 - c_2 \circ c_1\|_{\mathcal{H}^1} \\
\leq&\|f_1 - c_1\|_{\mathcal{H}^1} + \|f_2 - c_2\|_{\mathcal{H}^2} \\
\leq&2\varepsilon,
\end{aligned}$$

than we have that $\mathcal{C}_{\mathcal{F}_{|D}} = \{c_2 \circ c_1 : c_1 \in \mathcal{C}_{\mathcal{F}_{|D}^{(1)}}, c_2 \in \mathcal{C}_{\mathcal{F}_{|D}^{(2)}}(c_1)\} \subset \mathcal{F}_{|D}$ is a 2ε-cover of $\mathcal{F}_{|D}$. Hence,

$$
\begin{aligned}
&\mathcal{N}\left(2\varepsilon, \mathcal{F}_{|D}, d_{\mathcal{H}^1}\right) \\
\leq &\operatorname{card}\left(\mathcal{C}_{\mathcal{F}_{|D}}\right) \\
= &\sum_{c_1 \in \mathcal{C}_{\mathcal{F}^{(1)}_{|D}}} \operatorname{card}\left(\mathcal{C}_{\mathcal{F}^{(2)}_{|D}}(c_1)\right) \\
= &\operatorname{card}\left(\mathcal{C}_{\mathcal{F}^{(2)}_{|D}}(c_1)\right) \operatorname{card}\left(\mathcal{C}_{\mathcal{F}^{(1)}_{|D}}\right) \\
\leq &\operatorname{card}\left(\mathcal{C}_{\mathcal{F}^{(2)}_{|D}}(f_1)\right) \operatorname{card}\left(\mathcal{C}_{\mathcal{F}^{(1)}_{|D}}\right) \\
\leq &\mathcal{N}\left(\varepsilon, \mathcal{F}^{(2)}_{|D}, d_{\mathcal{H}^2}\right) \mathcal{N}\left(\varepsilon, \mathcal{F}^{(1)}_{|D}, d_{\mathcal{H}^1}\right) \\
\leq &\max_{|D|=n}\left\{\mathcal{N}(\varepsilon, \mathcal{F}^{(1)}_{|D}, d_{\mathcal{H}^1})\right\} \cdot \max_{|D|=n}\left\{\mathcal{N}(\varepsilon, \mathcal{F}^{(2)}_{|D}, d_{\mathcal{H}^2})\right\} \\
= &\mathcal{N}_p(\varepsilon, \mathcal{F}^{(1)}, n) \cdot \mathcal{N}_p(\varepsilon, \mathcal{F}^{(2)}, n).
\end{aligned}
$$

Since D is arbitrary, which completes the proof. □

B.2. Proof of Lemma 4.2

Proof. We start with the following lemma that bounds the d_∞-covering numbers by a quantity involving the pseudo-dimension.

Lemma B.1 (Theorem 12.2 of (Anthony & Bartlett, 1999))**.** *Let $\mathcal{F}$ be a set of real functions from a domain $\mathcal{X}$ to the bounded interval $[0, r]$. Let $\varepsilon > 0$ and suppose that the pseudo-dimension of $\mathcal{F}$ is d^{p}. Then the following holds for $n \geq d^{\mathrm{p}}$*

$$
\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \left(\frac{enr}{\varepsilon d^{\mathrm{p}}}\right)^{d^{\mathrm{p}}}.
$$

Since $\mathcal{N}_p(2\varepsilon, \mathcal{F}, n) \leq \mathcal{N}_p(\varepsilon, \mathcal{F}^{(1)}, n) \cdot \mathcal{N}_p(\varepsilon, \mathcal{F}^{(2)}, n)$ holds for $\mathcal{F} = \mathcal{F}^{(2)} \circ \mathcal{F}^{(1)}$, let $\mathcal{F}^{(i)}_{\downarrow} = \mathcal{F}^{(i-1)} \circ \cdots \circ \mathcal{F}^{(2)} \circ \mathcal{F}^{(1)}$, we denote the pseudo-dimension of the i-th layer function class $\mathcal{F}^{(i)}$ as $d_i^{\mathrm{p}} \geq 1, i \in [l]$, let $d^{\mathrm{p}}_{max} = \max_i\{d_i^{\mathrm{p}}\}$, then for the function class of l-layer DKMs, we have

$$
\begin{aligned}
\mathcal{N}_\infty(l\varepsilon, \mathcal{F}, n) = &\mathcal{N}_\infty(l\varepsilon, \mathcal{F}^{(l)} \circ \mathcal{F}^{(l)}_{\downarrow}, n) \\
\leq &\mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(l)}, n) \cdot \mathcal{N}_\infty((l-1)\varepsilon, \mathcal{F}^{(l)}_{\downarrow}, n) \\
\leq &\mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(l)}, n) \cdot \mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(l-1)}, n) \cdot \mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(l-1)}_{\downarrow}, n) \\
\leq &\cdots \\
\leq &\prod_{i=1}^{l} \mathcal{N}_\infty(\varepsilon, \mathcal{F}^{(i)}, n) \leq \prod_{i=1}^{l} \left(\frac{enr_i}{\varepsilon d_i^{\mathrm{p}}}\right)^{d_i^{\mathrm{p}}} \leq \prod_{i=1}^{l} \left(\frac{enr_i}{\varepsilon}\right)^{d_i^{\mathrm{p}}}.
\end{aligned}
$$

Therefore, $\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \prod_{i=1}^{l} \left(\frac{lenr_i}{\varepsilon}\right)^{d_i^{\mathrm{p}}}$. Taking the logarithm of the above inequality on both sides,

$$
\ln \mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq l \ln \left(\frac{lenA}{\varepsilon}\right)^{d^{\mathrm{p}}_{max}} = \ln \left(\frac{lenA}{\varepsilon}\right)^{ld^{\mathrm{p}}_{max}},
$$

so

$$
\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \left(\frac{lenA}{\varepsilon}\right)^{ld^{\mathrm{p}}_{max}}.
$$

□

B.3. Proof of Theorem 4.3

Proof. We start with the following lemma that proves a uniform convergence result for a class of real-valued functions.

Lemma B.2 (Theorem 10.1 of (Anthony & Bartlett, 1999)). *Suppose that $\mathcal{F}$ is a set of real-valued functions defined on the domain $\mathcal{X}$. Let P be any probability distribution on $\mathcal{X} \times \{-1, 1\}$. Given a dataset D of size n. Let ε be any real number between 0 and 1, the loss function be the γ-margin cost function. Then for $f \in \mathcal{F}$*

$$P\left\{\mathbb{E}(f) \geq \hat{E}_D(f) + \epsilon\right\} \leq 2\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2n) \exp\left(-\frac{\epsilon^2 n}{8}\right).$$

For the function class of l-layer DKMs, since $\mathcal{N}_\infty(\varepsilon, \mathcal{F}, n) \leq \left(\frac{lenA}{\varepsilon}\right)^{ld^{\mathrm{p}}_{max}}$, we have

$$2\mathcal{N}_\infty(\gamma/2, \mathcal{F}, 2n) \exp\left(-\frac{\epsilon^2 n}{8}\right) \leq 2\left(\frac{4lenA}{\gamma}\right)^{ld^{\mathrm{p}}_{max}} \exp\left(-\frac{\epsilon^2 n}{8}\right),$$

Let $2\left(\frac{4lenA}{\gamma}\right)^{ld^{\mathrm{p}}_{max}} \exp\left(-\frac{\epsilon^2 n}{8}\right) = \delta$, then

$$\epsilon = \sqrt{\frac{8\ln\frac{2}{\delta} + 8ld^{\mathrm{p}}_{max}\ln(\frac{4leAn}{\gamma})}{n}}. \tag{5}$$

Substituting equation (5) into the above lemma, we complete the proof. □

C. Generalization Bounds Based on the Rademacher Chaos Complexity

C.1. Proof of Lemma 5.1

Proof. For any $0 < \varepsilon \leq r_i^2$, we have $d_i^{\mathrm{k}} \geq 1$, $\frac{r_i^4}{\varepsilon^2} \geq 1$ and $\frac{en^2 r_i^2}{\varepsilon} \geq 1$, where d_i^{k} represents the kernel pseudo-dimension of the domain (class) $\mathcal{K}_{mul}^{(i)}$ of the i-th layer kernels and $i \in [l]$. Let $d^{\mathrm{k}}_{max} = \max_i\{d_i^{\mathrm{k}}\}$. Then, with the following lemmas,

Lemma C.1 (Theorem 3 in (Ying & Campbell, 2010)). *If the kernel pseudo-dimension $d_{\mathcal{K}}^{\mathrm{k}}$ of the set of basis kernels is finite, then we have that*

$$\mathcal{N}(\varepsilon, \mathcal{K}, d_2) \leq 2\left(\frac{4er^4}{\varepsilon^2}\right)^{d_{\mathcal{K}}^{\mathrm{k}}}.$$

Lemma C.2 (Lemma 3 in (Srebro & Ben-David, 2006)). *For any kernel family $\mathcal{K}$ with kernel pseudo-dimension $d_{\mathcal{K}}^{\mathrm{k}}$:*

$$\mathcal{N}_\infty(\varepsilon, \mathcal{K}, n) \leq \left(\frac{en^2 r^2}{\varepsilon d_{\mathcal{K}}^{\mathrm{k}}}\right)^{d_{\mathcal{K}}^{\mathrm{k}}}.$$

we have that

$$\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}^{(i)}, n) \leq 2\left(\frac{4er_i^4}{\varepsilon^2}\right)^{d_i^{\mathrm{k}}} \leq \left(\frac{8er_i^4}{\varepsilon^2}\right)^{d_i^{\mathrm{k}}},$$
$$\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}^{(i)}, n) \leq \left(\frac{en^2 r_i^2}{\varepsilon}\right)^{d_i^{\mathrm{k}}}.$$

Since $\mathcal{N}_p(2\varepsilon, \mathcal{F}, n) \leq \mathcal{N}_p(\varepsilon, \mathcal{F}^{(1)}, n) \cdot \mathcal{N}_p(\varepsilon, \mathcal{F}^{(2)}, n)$ holds for $\mathcal{F} = \mathcal{F}^{(1)} \circ \mathcal{F}^{(2)}$, then for a domain $\mathcal{K}_{mul}$ of l-layer deep kernels, we have

$$\mathcal{N}_2(l\varepsilon, \mathcal{K}_{mul}, n) \leq \prod_{i=1}^{l} \mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}^{(i)}, n) \leq \prod_{i=1}^{l}\left(\frac{8er_i^4}{\varepsilon^2}\right)^{d_i^{\mathrm{k}}},$$
$$\mathcal{N}_\infty(l\varepsilon, \mathcal{K}_{mul}, n) \leq \prod_{i=1}^{l} \mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}^{(i)}, n) \leq \prod_{i=1}^{l}\left(\frac{en^2 r_i^2}{\varepsilon}\right)^{d_i^{\mathrm{k}}}.$$

Taking the logarithm of the above inequality on both sides,

$$\ln \mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n) \leq l \ln \left(\frac{8l^2 eA^4}{\varepsilon^2} \right)^{d_{max}^{\mathrm{k}}} = \ln \left(\frac{8l^2 eA^4}{\varepsilon^2} \right)^{l d_{max}^{\mathrm{k}}},$$

so

$$\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n) \leq \left(\frac{8l^2 eA^4}{\varepsilon^2} \right)^{l d_{max}^{\mathrm{k}}}.$$

Similarly, one can bound

$$\mathcal{N}_\infty(\varepsilon, \mathcal{K}_{mul}, n) \leq \left(\frac{len^2 A^2}{\varepsilon} \right)^{l d_{max}^{\mathrm{k}}}.$$

□

C.2. Proof of Lemma 5.2

Proof. We denote the union of the unit balls of RKHSs as

$$\mathcal{B}_{\mathcal{K}_{mul}} := \{ f : f \in \mathcal{H}_K \text{ and } \|f\|_{\mathcal{H}_K} \leq 1, K \in \mathcal{K}_{mul} \}.$$

For some RKHS $\mathcal{H}_K$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \psi \left(y_i f^*(\boldsymbol{x}_i) \right) + \lambda \left\| f^* \right\|_{\mathcal{H}_K} \leq \frac{1}{n} \sum_{i=1}^{n} \psi(0) + \lambda \|0\|_{\mathcal{H}_K} = 1,$$

where

$$f^* = \underset{K \in \mathcal{K}_{mul}, f \in \mathcal{H}_K}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \psi \left(y_i f \left(\boldsymbol{x}_i \right) \right) + \lambda \|f\|_{\mathcal{H}_K}.$$

Hence, $\left\| f^* \right\|_{\mathcal{H}_K} \leq 1/\lambda$. This implies, for any samples in D, that

$$f^* \in \mathcal{B}_\lambda := \frac{1}{\lambda} \mathcal{B}_{\mathcal{K}_{mul}} := \left\{ \frac{f}{\lambda} : f \in \mathcal{B}_{\mathcal{K}_{mul}} \right\}.$$

For any training set $D = \{(\boldsymbol{x}_i, y_i) : i \in [n]\}$, let $D' = \{(\boldsymbol{x}_i, y_i) : i \in [n]\}$ be the training set with only one sample different from D, where the k-th sample is replaced by $(\boldsymbol{x}'_k, y'_k)$. Let $\Phi(D) = \sup_{f \in \mathcal{B}_\lambda} \left| \mathbb{E}(f) - \hat{E}_D(f) \right|$, then

$$\begin{aligned}
& \left| \Phi\left(D'\right) - \Phi(D) \right| \\
= & \left| \sup_{f \in \mathcal{B}_\lambda} \left| \mathbb{E}(f) - \hat{E}_{D'}(f) \right| - \sup_{f \in \mathcal{B}_\lambda} \left| \mathbb{E}(f) - \hat{E}_D(f) \right| \right| \\
\leq & \sup_{f \in \mathcal{B}_\lambda} \left| \hat{E}_D(f) - \hat{E}_{D'}(f) \right| \\
= & \sup_{f \in \mathcal{B}_\lambda} \frac{\left| \psi \left(y_k f \left(\boldsymbol{x}_k \right) \right) - \psi \left(y'_k f \left(\boldsymbol{x}'_k \right) \right) \right|}{n} \\
\leq & \frac{1}{n}.
\end{aligned}$$

According to McDiarmid's inequality, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the training dataset D, the following holds:

$$\Phi(D) \leq \mathbb{E}_D[\Phi(D)] + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{6}$$

Then, we will estimate the upper bound of $\mathbb{E}_D[\Phi(D)]$.

$$\begin{aligned}
&\mathbb{E}_D[\Phi(D)]\\
=&\mathbb{E}_D\left[\sup_{f\in\mathcal{B}_\lambda}\left|\mathbb{E}_{D'}\left[\hat{E}_{D'}(f)-\hat{E}_D(f)\right]\right|\right]\\
\leq&\mathbb{E}_{D,D'}\left[\sup_{f\in\mathcal{B}_\lambda}\left|\hat{E}_{D'}(f)-\hat{E}_D(f)\right|\right]\\
=&\mathbb{E}_{D,D'}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\psi\left(y_i'f\left(\boldsymbol{x}_i'\right)\right)-\psi\left(y_if\left(\boldsymbol{x}_i\right)\right)\right|\right]\\
=&\mathbb{E}_{\boldsymbol{\epsilon},D,D'}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\left(\psi\left(y_i'f\left(\boldsymbol{x}_i'\right)\right)-\psi\left(y_if\left(\boldsymbol{x}_i\right)\right)\right)\right|\right]\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon},D'}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\psi\left(y_i'f\left(\boldsymbol{x}_i'\right)\right)\right|\right]\\
+&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n-\epsilon_i\psi\left(y_if\left(\boldsymbol{x}_i\right)\right)\right|\right]\\
=&2\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\psi\left(y_if\left(\boldsymbol{x}_i\right)\right)\right|\right].
\end{aligned}\tag{7}$$

To this end, we introduce the decomposition of γ-margin cost function that $\bar{\psi}(x)=\psi(x)-\psi(0):\mathbb{R}\to\mathbb{R}$, which has the Lipschitz constant $\frac{1}{\gamma}$ and $\bar{\psi}(0)=0$ ($\psi(0)=1$). Then, applying the contraction property of Rademacher complexities, with probability at least $1-\delta$, the following holds:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\psi\left(y_if\left(\boldsymbol{x}_i\right)\right)\right|\right]\\
\leq&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\bar{\psi}\left(y_if\left(\boldsymbol{x}_i\right)\right)\right|\right]+\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_i\psi(0)\right|\right]\\
\leq&\frac{2}{\gamma}\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_if\left(\boldsymbol{x}_i\right)\right|\right]+\left(\mathbb{E}_{\boldsymbol{\epsilon},D}\frac{1}{n}\left[\sum_{i=1}^n\epsilon_i\epsilon_j\right]\right)^{1/2}\\
\leq&\frac{2}{\gamma}\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_if\left(\boldsymbol{x}_i\right)\right|\right]+\frac{1}{\sqrt{n}}.
\end{aligned}\tag{8}$$

Hence, we will bound $\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_if\left(\boldsymbol{x}_i\right)\right|\right]$.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{f\in\mathcal{B}_\lambda}\frac{1}{n}\left|\sum_{i=1}^n\epsilon_if\left(\boldsymbol{x}_i\right)\right|\right]\\
=&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\frac{1}{\lambda}\sup_{K\in\mathcal{K}_{mul}}\sup_{\|f\|_K\leq1}\frac{1}{n}\left|\left\langle\sum_{i=1}^n\epsilon_iK(\cdot,\boldsymbol{x}_i),f\right\rangle_K\right|\right]\\
=&\frac{1}{\lambda}\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{K\in\mathcal{K}_{mul}}\frac{1}{n}\left|\sum_{i,j=1}^n\epsilon_i\epsilon_jK\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\right|^{\frac{1}{2}}\right]\\
\leq&\frac{1}{\lambda}\sqrt{\frac{2U_n(\mathcal{K}_{mul})}{n}}+\frac{1}{\lambda}\sup_{K\in\mathcal{K}_{mul}}\frac{\sqrt{\mathrm{trace}(\mathbf{K})}}{n}\\
\leq&\frac{1}{\lambda}\sqrt{\frac{2U_n(\mathcal{K}_{mul})}{n}}+\frac{A}{\lambda\sqrt{n}}.
\end{aligned}\tag{9}$$

Substituting inequalities (8) (9) into (7), we have

$$\mathbb{E}_D[\Phi(D)] \leq \frac{4}{\gamma\lambda}\sqrt{\frac{2U_n(\mathcal{K}_{mul})}{n}} + \frac{4A}{\gamma\lambda\sqrt{n}} + \frac{2}{\sqrt{n}}. \tag{10}$$

Combining with (6) and (10), then

$$\mathbb{E}(f) \leq \hat{E}_D(f) + \frac{4}{\gamma\lambda}\sqrt{\frac{2U_n(\mathcal{K}_{mul})}{n}} + \frac{4A}{\gamma\lambda\sqrt{n}} + \frac{2}{\sqrt{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

□

C.3. Proof of Lemma 5.3

Proof. Let $\mathcal{K}_{mul}$ be a domain of l-layer deep kernels. The empirical Rademacher chaos complexity $\hat{U}_D(\mathcal{K}_{mul})$ can be bounded by the metric entropy integral as follows:

Lemma C.3 (Theorem 2 in (Ying & Campbell, 2010))**.** *For any $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, there holds*

$$\hat{U}_D(\mathcal{K}) \leq r^2 + 24e\int_0^{r^2} \log\left[1 + \mathcal{N}\left(\varepsilon, \mathcal{K}, d_2\right)\right] d\varepsilon.$$

We assume that $0 \in \mathcal{K}$, then we have that

$$\hat{U}_D(\mathcal{K}) \leq 24e\int_0^{r^2} \log\left[\mathcal{N}_2(\varepsilon, \mathcal{K}, n) + 1\right] d\varepsilon.$$

Since $\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n) \leq \left(\frac{8l^2eA^4}{\varepsilon^2}\right)^{ld^{\mathrm{k}}_{max}}$, we have that

$$\begin{aligned}
U_n(\mathcal{K}_{mul}) &= \mathbb{E}_D\left[\hat{U}_D(\mathcal{K}_{mul})\right] \\
&\leq 24e\,\mathbb{E}_D\left[\int_0^{r^2} \log\left[\mathcal{N}_2(\varepsilon, \mathcal{K}_{mul}, n) + 1\right] d\varepsilon\right] \\
&\leq 24e\,\mathbb{E}_D\left[\int_0^{A^2} \log\left[\left(\frac{8l^2eA^4}{\varepsilon^2}\right)^{ld^{\mathrm{k}}_{max}} + 1\right] d\varepsilon\right] \\
&\leq 24e\int_0^{A^2} \ln\left(\frac{8.4l^2eA^4}{\varepsilon^2}\right)^{ld^{\mathrm{k}}_{max}} d\varepsilon \\
&= 24eld^{\mathrm{k}}_{max}\left[A^2\ln(8.4l^2e) + \int_0^{A^2} \ln\left(\frac{A^4}{\varepsilon^2}\right) d\varepsilon\right] \\
&= 24eld^{\mathrm{k}}_{max}\left[A^2\ln(8.4l^2e) + 4A^2\right] \\
&\leq 172eA^2ld^{\mathrm{k}}_{max} + 48eA^2ld^{\mathrm{k}}_{max}\ln l \\
&\leq 220eA^2ld^{\mathrm{k}}_{max}\ln l \quad (l \geq 3).
\end{aligned}$$

□

D. Estimating the Uniform Covering Number and the Rademacher Chaos Complexity

D.1. Proof of Theorem 6.1

Proof. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d^{\mathrm{p}}}\}$ be the set that can be pseudo-shattered by $\mathcal{F}$, then according to the definition of pseudo-dimension, there are real numbers $\{r_i \in \mathbb{R} : i \in [d^{\mathrm{p}}]\}$ such that for any $\boldsymbol{y} \in \{-1, 1\}^{d^{\mathrm{p}}}$ there is a function $f \in \mathcal{F}$ with property $\operatorname{sgn}\left(f(\boldsymbol{x}_i) - r_i\right) = y_i$ for any $i \in [d^{\mathrm{p}}]$.

For any $\boldsymbol{y} = (y_1, \ldots, y_{d^{\mathrm{P}}}) \in \{-1, 1\}^{d^{\mathrm{P}}}$, we can take the real number r_i as $r_i = \mathbf{w'}^{\top}\Phi(\boldsymbol{x}_i)$ for any $i \in [d^{\mathrm{P}}]$, then $\exists\ \mathbf{w}, \mathbf{w'}$ such that

$$y_i\left(\mathbf{w}^{\top}\Phi(\boldsymbol{x}_i) - r_i\right) \geq 1 \ \ (\forall i \in [d^{\mathrm{P}}]),$$

i.e.,

$$y_i\left(\mathbf{w}^{\top}\Phi(\boldsymbol{x}_i) - \mathbf{w'}^{\top}\Phi(\boldsymbol{x}_i)\right) \geq 1 \ \ (\forall i \in [d^{\mathrm{P}}]),$$

let $\tilde{\mathbf{w}} = \mathbf{w} - \mathbf{w'}$, then we have

$$y_i\left(\tilde{\mathbf{w}}^{\top}\Phi(\boldsymbol{x}_i)\right) \geq 1 \ \ (\forall i \in [d^{\mathrm{P}}]).$$

Summing these d^{P} inequalities, we have

$$\begin{aligned} d^{\mathrm{P}} &\leq \tilde{\mathbf{w}}^{\top}\sum_{i=1}^{d^{\mathrm{P}}} y_i\Phi(\boldsymbol{x}_i) \\ &\leq \|\tilde{\mathbf{w}}\|\left\|\sum_{i=1}^{d^{\mathrm{P}}} y_i\Phi(\boldsymbol{x}_i)\right\| \leq \Lambda\left\|\sum_{i=1}^{d^{\mathrm{P}}} y_i\Phi(\boldsymbol{x}_i)\right\|. \end{aligned}$$

Since the above inequality holds for any $\boldsymbol{y} \in \{-1, 1\}^{d^{\mathrm{P}}}$, taking the expectation of $y_1, \ldots, y_{d^{\mathrm{P}}}$ on both sides, where $y_1, \ldots, y_{d^{\mathrm{P}}}$ obey independent and uniform distribution. From the independence assumption, we have $\mathbb{E}[y_i y_j] = \mathbb{E}[y_i]\,\mathbb{E}[y_j], i \neq j$. From the uniform distribution, we have $\mathbb{E}[y_i y_j] = 0\ (i \neq j), \mathbb{E}[y_i y_j] = 1\ (i = j)$. Then,

$$\begin{aligned} d^{\mathrm{P}} &\leq \Lambda\mathbb{E}_{\boldsymbol{y}}\left[\left\|\sum_{i=1}^{d^{\mathrm{P}}} y_i\Phi(\boldsymbol{x}_i)\right\|\right] \\ &\leq \Lambda\left[\mathbb{E}_{\boldsymbol{y}}\left[\left\|\sum_{i=1}^{d^{\mathrm{P}}} y_i\Phi(\boldsymbol{x}_i)\right\|^2\right]\right]^{1/2} \\ &= \Lambda\left[\sum_{i,j=1}^{d^{\mathrm{P}}} \mathbb{E}_{\boldsymbol{y}}[y_i y_j]\left(\Phi(\boldsymbol{x}_i)^{\top}\Phi(\boldsymbol{x}_j)\right)\right]^{1/2} \\ &= \Lambda\left[\sum_{i=1}^{d^{\mathrm{P}}}\left(\Phi(\boldsymbol{x}_i)^{\top}\Phi(\boldsymbol{x}_i)\right)\right]^{1/2} \\ &= \Lambda\left[\sum_{i=1}^{d^{\mathrm{P}}}\left(K(\boldsymbol{x}_i, \boldsymbol{x}_i)\right)\right]^{1/2} \\ &\leq \Lambda r\sqrt{d^{\mathrm{P}}}, \end{aligned}$$

where the second inequality uses the Jensen's inequality and the property of convex functions. Hence, we have $d^{\mathrm{P}} \leq r^2\Lambda^2$. □

D.2. Proof of Theorem 6.2

Proof. It can be known that $\mathcal{K}_{\text{conv}}(S) \subseteq \mathcal{K}_{\text{line}}(S) \subseteq \mathcal{K}_{\text{span}}(S)$, and $\mathcal{K}_{\text{span}}(S)$ is a vector space with dimension $d(\mathcal{K}_{\text{span}}(S)) \leq k$. Since the fact that the pseudo-dimension of a k-dimensional vector space of real-valued functions is k, the following relationship of the kernel pseudo-dimension holds:

$$d^{\mathrm{k}}\left(\mathcal{K}_{\text{conv}}(S)\right) \leq d^{\mathrm{k}}\left(\mathcal{K}_{\text{line}}(S)\right) \leq d^{\mathrm{k}}\left(\mathcal{K}_{\text{span}}(S)\right) = d\left(\mathcal{K}_{\text{span}}(S)\right) \leq k. \tag{11}$$

We then have the following lemma for the pseudo-dimension and compositions with non-decreasing functions:

Lemma D.1 (Theorem 11.3 in (Anthony & Bartlett, 1999)). *Suppose $\mathcal{F}$ is a class of real-valued functions and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-decreasing function. Let $\sigma(\mathcal{F})$ denote the class $\{\sigma \circ f : f \in \mathcal{F}\}$. Then $d^{\mathrm{k}}(\sigma(\mathcal{F})) \leq d^{\mathrm{k}}(\mathcal{F})$.*

Since the g^i operations produced by the compositions of common-used kernels are monotonous (non-decreasing) with respect to their input kernels, therefore, if the kernel domain of their input kernels is denoted by $\mathcal{K}^1_{mul}$, then $d^{\mathrm{k}}(\mathcal{K}^1_{mul}) = d(\mathcal{K}^1_{mul}) = m$. For any $i \in \{2, \ldots, l\}$, we have

$$d^{\mathrm{k}}(\mathcal{K}^i_{mul}) \leq d^{\mathrm{k}}(g^i \circ h^{i-1}(\mathcal{K}^{i-1}_{mul})) \leq d^{\mathrm{k}}(h^{i-1}(\mathcal{K}^{i-1}_{mul})) \leq d^{\mathrm{k}}(\mathcal{K}^{i-1}_{mul}). \tag{12}$$

Hence, $d^{\mathrm{k}}(\mathcal{K}^i_{mul}) \leq m$, and the bound of the Rademacher chaos complexity can be derived immediately. □

E. A Lower Bound Based on the Rademacher Complexity

E.1. Proof of Theorem 7.1

Proof. We first define a function class of DKMs as follows:

$$\mathcal{G} = \{\boldsymbol{x} \mapsto \sum_{i=1}^{n} \alpha_i K^l(\boldsymbol{x}, \boldsymbol{x}_i) : \sum_{i,j} \alpha_i \alpha_j K^p(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 1, K^p(\boldsymbol{x}_i, \boldsymbol{x}_i) \leq A^2, K^p(\cdot,\cdot) \in \mathcal{K}^p_*, \boldsymbol{x}_i \in \mathcal{X}, \forall p \in [l]\},$$

where the class of q-th layer deep kernels for some $q \in \{2, \cdots, l\}$ is defined as follows:

$$\mathcal{K}^q_* = \left\{K^q(\cdot,\cdot) = g^q\left(\left[K^{q-1}(\cdot,\cdot)\right]\right)\right\},$$

where $K^{q-1}(\cdot,\cdot) \in \mathcal{K}^{q-1}_*$, g^q is the nonlinear function produced by $K^q(\cdot,\cdot)$ which composites $K^{q-1}(\cdot,\cdot)$, and the classes of all other layers deep kernels are linear kernel classes for any $p \in [l]$ and $p \neq q$, i.e., the g^p operation is the identity transformation. Therefore, the l-layer deep kernel can be denoted as $K^l(\boldsymbol{x}, \boldsymbol{x}') = K^q(\boldsymbol{x}, \boldsymbol{x}')$. Note that the linear kernel does not produce nonlinear mapping, and $K^q(\cdot,\cdot)$ represents a single-layer nonlinear kernel. It can be shown that $\mathcal{G} \subseteq \mathcal{F}$.

$$\begin{aligned}
&\hat{R}_D(\mathcal{G}) \\
=&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right|\right] \\
=&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{W}^l:\|\boldsymbol{W}^l\|\leq 1}\sup_{\Phi^l:\|\Phi^l\|\leq A}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i {\boldsymbol{W}^l}^\top \Phi^l(\boldsymbol{x}_i)\right|\right] \\
=&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{W}^q:\|\boldsymbol{W}^q\|\leq 1}\sup_{\phi^q:\|\phi^q\|\leq A}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i {\boldsymbol{W}^q}^\top \phi^q(\boldsymbol{x}_i)\right|\right] \\
=&\sup_{\phi^q:\|\phi^q\|\leq A}\frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^{n}\epsilon_i\phi^q(\boldsymbol{x}_i)\right\| \\
\geq&\sup_{\phi^q:\|\phi^q\|\leq A}\frac{1}{n}c\sqrt{\sum_{i=1}^{n}\|\phi^q(\boldsymbol{x}_i)\|^2} \quad \text{(Use Khintchine-Kahane inequality, } c>0\text{)} \\
=&\sup_{\phi^q:\|\phi^q\|\leq A}\frac{1}{n}c\sqrt{\sum_{i=1}^{n}K^q(\boldsymbol{x}_i,\boldsymbol{x}_i)} \\
=&\frac{cA}{\sqrt{n}}.
\end{aligned}$$

Since $\mathcal{G}$ is a subset of $\mathcal{F}$, so $\hat{R}_D(\mathcal{F}) \geq \hat{R}_D(\mathcal{G}) \geq \frac{cA}{\sqrt{n}}$. □