Generalization Bounds for Learning the Kernel

Yiming Ying and Colin Campbell
Department of Engineering Mathematics
University of Bristol, Bristol
BS8 1TR, United Kingdom
{enxyy, C.Campbell}@bris.ac.uk

Abstract

In this paper we develop a novel probabilistic generalization bound for learning the kernel problem. First, we show that the generalization analysis of the kernel learning algorithms reduces to investigation of the suprema of the Rademacher chaos process of order two over candidate kernels, which we refer to as *Rademacher chaos complexity*. Next, we show how to estimate the empirical Rademacher chaos complexity by well-established metric entropy integrals and pseudo-dimension of the set of candidate kernels. Our new methodology mainly depends on the principal theory of U-processes. Finally, we establish satisfactory excess generalization bounds and misclassification error rates for learning Gaussian kernels and general radial basis kernels.

1 Introduction

Kernel methods such as Support Vector Machines (SVM) have been extensively applied to supervised learning tasks such as classification and regression, see e.g. [9, 23, 24, 28]. The performance of a kernel machine largely depends on the data representation via the choice of kernel function. Hence, one central issue in kernel methods is the problem of kernel selection. To automate kernel learning algorithms, it is desirable to integrate the process of selecting kernels into the learning algorithms.

Kernel learning can range from the width parameter selection of Gaussian kernels to obtaining an optimal linear combination from a set of finite candidate kernels. The latter is often referred to as multiple kernel learning (MKL) in Machine Learning and nonparametric Group Lasso [3, 32] in Statistics. Lanckriet et al. [17] pioneered work on MKL and proposed a semi-definite programming (SDP) approach to automatically learn a linear combination of candidate kernels for the case of SVMs. Similar problems studied recently include the so-called COSSO estimate for additive model [18], hyperkernels [22], Bayesian probabilistic kernel learning models [14], and kernel discriminant analysis [30] etc. Such MKL formulations have been successfully demonstrated in combining multiple heterogeneous data sources to enhance biological inference [17].

The above mentioned MKL algorithms learn the linear combination of a finite set of candidate kernels. A general regularization framework including kernel hyper-parameter learning and MKL was formulated in [20, 29] with a potentially *infinite* number of candidate kernels which is generally referred to as the *learning the kernel problem*. Specifically, let $\mathbb{N}_n = \{1, 2, \cdots, n\}$ for any $n \in \mathbb{N}$ and we are interested in the classification problem on the input space $X \subseteq \mathbb{R}^d$ and output space $Y = \{\pm 1\}$. The relation between input X and output Y is specified by a set of training samples $\mathbf{z} = \{z_i = (x_i, y_i) : x_i \in X,\ y_i \in Y,\ i \in \mathbb{N}_n\}$ which are identically and independently distributed (i.i.d.) according to an unknown distribution ρ on $Z = X \times Y$. Let $\mathcal{K}$ be a prescribed (possible infinite) set of candidate (base) kernels and denote the candidate reproducing kernel Hilbert space (RKHS) with kernel K by $\mathcal{H}_K$ with norm $\|\cdot\|_K$. In addition, we always assume that the quantity $\kappa := \sup_{K\in\mathcal{K}, x\in X} \sqrt{K(x,x)}$ is finite. Then the general kernel learning scheme in [20, 29] can be cast as a two-layer minimization problem:

$$f_{\mathbf{z}}^{\phi} = \arg \min_{\substack{f\in\mathcal{H}_K \\ K\in\mathcal{K}}} \left\{ \frac{1}{n} \sum_{i\in\mathbb{N}_n} \phi\big(y_i f(x_i)\big) + \lambda \|f\|_K^2 \right\}. \qquad (1)$$

where $\phi : \mathbb{R} \to [0, \infty)$ is a prescribed loss function and λ is a positive regularization parameter. We emphasize that the superscript ϕ means that the solution $f_{\mathbf{z}}^{\phi}$ is produced by (1) with loss function ϕ. When the loss function ϕ is chosen to the hinge loss for SVM and $\mathcal{K}$ is the linear combination of the set of finite base kernels $\{K_\ell : \ell \in \mathbb{N}_m\}$, i.e. $\mathcal{K} := \{\sum_{\ell\in\mathbb{N}_m} \lambda_\ell \mathcal{K}_\ell : \sum_{\ell\in\mathbb{N}_m} \lambda_\ell = 1, \lambda_\ell \geq 0,\ \forall \ell \in \mathbb{N}_m\}$, then the above kernel learning framework (1) is reduced to the SVM-based MKL formulation [17]. If we choose the set of base kernels as $\mathcal{K} = \left\{e^{-\sigma\|x-t\|^2} : \sigma > 0\right\}$, the above formulation (1) is generally reduced to the problem of Gaussian kernel hyper-parameter learning.

Statistical generalization analysis of learning the kernel system (1) was pursued by [7, 17, 31, 21, 25]. In this paper we leverage Rademacher complexity bounds for empirical risk minimization (ERM) and for SVM with a single kernel [4, 5, 16] and develop a novel generalization bound for kernel learning system (1). Our new approach is based on the principal theory of U-processes (e.g. [11]) which can yield tight generalization bounds.

This paper is organized as follows. In Section 2 we review necessary background for generalization analysis and illustrate our main results. Section 3 discusses related work

and compares our results with those in the literature. Our main idea is developed in Section 4. There we show the generalization analysis of the kernel learning problem (1) reduces to investigation of the suprema of the homogeneous Rademacher chaos of order two over candidate kernels, which we refer to as *Rademacher chaos complexity*. In Section 5 we show how to estimate the Rademacher chaos complexity using metric entropy integrals and the pseudo-dimension of the set of candidate kernels. Examples for learning Gaussian kernels and radial basis kernels are given in Section 6 to illustrate our proposed generalization analysis. In Section 7 we present the conclusion and discussion of possible extensions.

2 Main Results

In this section we outline our main contributions. Before we do this, let us review the objective of generalization analysis for multiple kernel learning focusing on classification problems.

2.1 Target of Analysis

A classifier $\mathcal{C}$ assigns, for each point x, a prediction $\mathcal{C}(x) \in Y$. The prediction power of classifiers is measured by the *misclassification error* which is defined, for a classifier $\mathcal{C} : X \to Y$, by

$$\mathcal{R}(\mathcal{C}) := \int_{X\times Y} P(y \neq \mathcal{C}(x)|x)d\rho(x,y). \tag{2}$$

The best classifier is called the *Bayes rule* [13] which minimizes the misclassification error over all classifiers: $f_c = \arg\inf \mathcal{R}(\mathcal{C})$.

We are interested in the statistical behavior of the *multi-kernel regularized classifier* given by $\mathrm{sign}(f_{\mathbf{z}}^{\phi})$ with the regularization scheme (1). For brevity, throughout this note we restrict our interest to a class of loss functions used in [29], see also a general definition of classification loss functions in [4].

Definition 1 *A function $\phi : \mathbb{R} \to [0,\infty)$ is called a normalized classifying loss if it is convex, $\phi'(0) < 0$, $\inf_{t\in\mathbb{R}} \phi(t) = 0$, and $\phi(0) = 1$.*

The convexity and the condition $\phi'(0) < 0$ in the definition of the normalized classifying loss implies that $\phi(yf(x)) > \phi(0) > 0$ whenever $yf(x) < 0$ (i.e. when $\mathrm{sgn}(f(x))$ misclassifies the true label y). The true error or *generalization error* is defined as

$$\mathcal{E}^{\phi}(f) = \int_{X\times Y} \phi(yf(x))d\rho(x,y),$$

and the target function f_{ρ}^{ϕ} is defined by $f_{\rho}^{\phi} = \arg\min_f \mathcal{E}^{\phi}(f)$. Examples of normalized classifying losses include the hinge loss $\phi(t) = (1-t)_+$ for soft margin SVM, general q-norm soft margin SVM loss $\phi(t) = (1-t)_+^q$ with $q > 1$, and the least square loss $\phi(t) = (1-t)^2$.

The target of error analysis is to understand how $\mathrm{sign}(f_{\mathbf{z}}^{\phi})$ approximates the Bayes rule f_c. More specifically, we aim to estimate the *excess misclassification error*

$$\mathcal{R}(\mathrm{sign}(f_{\mathbf{z}}^{\phi})) - \mathcal{R}(f_c)$$

for the multi-kernel regularized classification algorithm (1). As shown in [33, 4], the excess misclassification error usually can be bounded by the *excess generalization error*:

$$\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_{\phi}), \tag{3}$$

and we refer to the relation between these two excess errors as the *comparison inequality*. For example, for a SVM hinge loss we know [33] that $f_{\phi} = f_c$ and

$$\mathcal{R}(\mathrm{sign}(f_{\mathbf{z}}^{\phi})) - \mathcal{R}(f_c) \leq \mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_c). \tag{4}$$

One can refer to [4, 33] for more comparison inequalities for general loss functions.

Consequently, it suffices to bound the excess generalization error (3). To this end, we introduce the error decomposition of algorithm (1). Let the empirical error $\mathcal{E}_{\mathbf{z}}$ be defined, for any f, by $\mathcal{E}_{\mathbf{z}}^{\phi}(f) = \frac{1}{n}\sum_{j\in\mathbb{N}_n} \phi(y_j f(x_j))$. We also introduce the *regularization error* $\mathcal{D}(\lambda)$ defined by

$$\mathcal{D}(\lambda) = \inf_{K\in\mathcal{K}} \inf_{f\in\mathcal{H}_K} \Big\{\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda\|f\|_K^2\Big\},$$

and also call the minimizer f_{λ}^{ϕ} of the regularization error the *regularization function*. Also, we define the *sample error* $\mathcal{S}_{\mathbf{z},\lambda}$ by

$$\mathcal{S}_{\mathbf{z},\lambda} = \Big\{\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\mathbf{z}}^{\phi})\Big\} + \Big\{\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\lambda}^{\phi}) - \mathcal{E}^{\phi}(f_{\lambda}^{\phi})\Big\}.$$

Then, we know from [31] that the *error decomposition* holds true:

$$\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \leq \mathcal{D}(\lambda) + \mathcal{S}_{\mathbf{z},\lambda}. \tag{5}$$

Throughout this paper, for simplicity we always assume the existence of the empirical solution $f_{\mathbf{z}}^{\phi}$ and the regularization function f_{λ}^{ϕ}, see discussions in Appendix B of [31].

To estimate the sample error $\mathcal{S}_{\mathbf{z},\lambda}$, we need to find the hypothesis space of $f_{\mathbf{z}}^{\phi}$ and f_{λ}^{ϕ}. Let the union of the unit balls of candidate RKHSs be denoted by

$$\mathcal{B}_{\mathcal{K}} := \Big\{f :\ f \in \mathcal{H}_K \text{ and } \|f\|_K \leq 1,\ K \in \mathcal{K}\Big\}. \tag{6}$$

By the definition of $f_{\mathbf{z}}^{\phi}$, we get, for some RKHS $\mathcal{H}_K$, that $\frac{1}{n}\sum_{i=1}^n \phi\big(y_i f_{\mathbf{z}}^{\phi}(x_i)\big) + \lambda\|f_{\mathbf{z}}^{\phi}\|_K^2 \leq \frac{1}{n}\sum_{i=1}^n \phi\big(0\big) + \lambda\|0\|_K^2 = 1$. Hence, $\|f_{\mathbf{z}}^{\phi}\|_K \leq \sqrt{1/\lambda}$. Likewise, for some kernel $K \in \mathcal{K}$, $\|f_{\lambda}^{\phi}\|_K \leq \sqrt{1/\lambda}$. This implies, for any samples $\mathbf{z}$, that

$$f_{\mathbf{z}}^{\phi},\ f_{\lambda}^{\phi} \in \mathcal{B}_{\lambda} := \frac{1}{\sqrt{\lambda}}\,\mathcal{B}_{\mathcal{K}} := \Big\{\frac{f}{\sqrt{\lambda}} : f \in \mathcal{B}_{\mathcal{K}}\Big\}. \tag{7}$$

Hence, $\|f_{\mathbf{z}}^{\phi}\|_{\infty} < \kappa\sqrt{1/\lambda}$ and $\|f_{\lambda}^{\phi}\|_{\infty} < \kappa\sqrt{1/\lambda}$. Finally, for a locally Lipschitz continuous function $\psi : \mathbb{R} \to [0,\infty)$ we need the constant defined by

$$M_{\lambda}^{\psi} = \sup\Big\{|\psi(t)| :\ |t| \leq \kappa\sqrt{1/\lambda}\Big\}, \tag{8}$$

and denote the local Lipschitz constant by

$$C_{\lambda}^{\psi} = \sup\Big\{\frac{|\psi(x) - \psi(x')|}{|x - x'|} :\ \forall |x|, |x'| \leq \kappa\sqrt{\frac{1}{\lambda}}\Big\}. \tag{9}$$

If $\psi = \phi$ is convex, then ϕ's left derivative ϕ'_- and right one ϕ'_+ are well defined and C_{λ}^{ϕ} is identical to

$$C_{\lambda}^{\phi} = \sup\big\{\max\big(|\phi'_-(t)|, |\phi'_+(t)|\big) :\ |t| \leq \kappa\sqrt{1/\lambda}\big\}.$$

2.2 Main Theorems

Our generalization analysis depends on the suprema of the *homogeneous Rademacher chaos of order two* over a class of functions defined as follows, see Chapter 3.2 of [11] for a general definition of Rademacher chaos of order m for any $m \in \mathbb{N}$.

Definition 2 *Let F be a class of functions on $X \times X$ and let $\{\epsilon_i : i \in \mathbb{N}_n\}$ be independent Rademacher random variables. Also, let $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$ be independent random variables distributed according to a distribution μ on X. The homogeneous Rademacher chaos process of order two, with respect to the Rademacher variable ε, is a random variable system defined by*

$$\big\{\hat{U}_f(\varepsilon) = \frac{1}{n} \sum_{i,j \in \mathbb{N}_n, i<j} \varepsilon_i \varepsilon_j f(x_i, x_j) : f \in F\big\},$$

and we refer to the expectation of its superma

$$\hat{\mathscr{U}}_n(F) = \mathbb{E}_\varepsilon[\sup_{f \in F} |\hat{U}_f(\varepsilon)|]$$

as the empirical Rademacher chaos complexity over F.

It is worth mentioning that the Rademacher process

$$\big\{\frac{1}{\sqrt{n}} \sum_{i \in \mathbb{N}_n} \varepsilon_i f(x_i) : f \in F\big\}$$

for Rademacher averages can be regarded as a *homogeneous Rademacher chaos process of order one*. The nice application of U-processes to the generalization analysis of the ranking and scoring problem is recently developed in [10].

Our first main result shows that the excess generalization error of MKL algorithms can be bounded by the empirical Rademacher chaos complexity over the set of candidate kernels.

Theorem 3 *Let ϕ be a normalized classifying loss. Then, for any $\delta \in (0,1)$ we have, with probability at least $1-\delta$, that*

$$\begin{aligned}\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \quad &\le 4C_\lambda^\phi \left(\frac{2\hat{\mathscr{U}}_n(\mathcal{K})}{\lambda n}\right)^{\frac{1}{2}} + 4\kappa C_\lambda^\phi \Big(\frac{1}{n\lambda}\Big)^{\frac{1}{2}} \\ &\quad +3M_\lambda^\phi \Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}},\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}(f_\rho^\phi) \quad &\le 8C_\lambda^\phi \left(\frac{2\hat{\mathscr{U}}_n(\mathcal{K})}{\lambda n}\right)^{\frac{1}{2}} + 8C_\lambda^\phi \kappa \Big(\frac{1}{n\lambda}\Big)^{\frac{1}{2}} \\ &\quad +3M_\lambda^\phi \Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{4}{\sqrt{n}} + \mathcal{D}(\lambda).\end{aligned}$$

In practice, the empirical complexity $\hat{\mathscr{U}}_n(\mathcal{K})$ can be estimated from finite samples. In analogy to the data-dependent risk bounds of Rademacher averages [4], we can get margin bounds of Rademacher chaos complexities for learning the kernel problems.

Corollary 4 *Let $\gamma > 0,\ 0 < \delta < 1$ and define the* margin cost function *by*

$$\psi(t) = \begin{cases} 1, & t \le 0 \\ 1 - \frac{t}{\gamma}, & 0 < t \le \gamma \\ 0, & t > \gamma \end{cases} \tag{10}$$

Then, with probability at least $1-\delta$, there holds

$$\begin{aligned}\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) \quad &\le \mathcal{E}_{\mathbf{z}}^\psi(f_{\mathbf{z}}^\phi) + 4\Big(\frac{2\hat{\mathscr{U}}_n(\mathcal{K})}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 4\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} \\ &\quad +3\Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.\end{aligned}$$

Theorem 3 and Corollary 4 will be proved in Section 4. When $\mathcal{K}$ only has a single kernel K, we have

$$\begin{aligned}\hat{\mathscr{U}}_n(K) \quad &\le \mathbb{E}_\varepsilon \Big|\tfrac{1}{n}\textstyle\sum_{i,j \in \mathbb{N}_n} \varepsilon_i \varepsilon_j K(x_i, x_j)\Big| \\ &\quad + \Big|\tfrac{1}{n}\textstyle\sum_{i \in \mathbb{N}_n} K(x_i, x_i)\Big| \\ &= \mathbb{E}_\varepsilon \tfrac{1}{n}\textstyle\sum_{i,j \in \mathbb{N}_n} \varepsilon_i \varepsilon_j K(x_i, x_j) \\ &\quad + \tfrac{1}{n}\textstyle\sum_{i \in \mathbb{N}_n} K(x_i, x_i)\end{aligned}$$

where the last equality follows from the positive semi-definiteness of kernel K. Hence, denote by $\mathbf{K}$ the matrix $(K(x_i, x_j))_{ij \in \mathbb{N}_n}$, the Rademacher chaos complexity can be estimated as follows:

$$\hat{\mathscr{U}}_n(K) \le \frac{2}{n} \sum_{i \in \mathbb{N}_n} K(x_i, x_i) = \frac{2}{n} \mathrm{trace}(\mathbf{K}).$$

Consequently, Corollary 4 implies that

$$\begin{aligned}\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}}^\phi)) \quad &\le \mathcal{E}_{\mathbf{z}}^\psi(f_{\mathbf{z}}^\phi) + \frac{8}{\gamma} \frac{\sqrt{\mathrm{trace}(\mathbf{K})}}{n\sqrt{\lambda}} \\ &\quad +4\kappa\Big(\frac{1}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} + 3\Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}},\end{aligned}$$

which coincides with the bound in [5] for the single kernel case with solutions $f_{\mathbf{z}}^\phi$ in the function space

$$\{f = \sum_{i \in \mathbb{N}_n} \alpha_i K(x_i, \cdot) : \sum_{i,j \in \mathbb{N}_n} \alpha_i \alpha_j K(x_i, x_j) \le \frac{1}{\lambda}\}.$$

Now we apply the well-established theory of U processes to estimate Rademacher chaos complexity by the pseudo-dimension of the set of candidate kernels. For this purpose, we recall the definition of the kernel pseudo-dimension of a class of kernel functions on the product space $X \times X$, see [2].

Definition 5 *Let $\mathcal{K}$ be a set of reproducing kernel functions mapping from $X \times X$ to $\mathbb{R}$. We say that $S_m = \{(x_i, t_i) \in X \times X : i \in \mathbb{N}_m\}$ is pseudo-shattering by $\mathcal{K}$ if there are real numbers $\{r_i \in \mathbb{R} : i \in \mathbb{N}_m\}$ such that for any $b \in \{-1,1\}^m$ there is a function $K \in \mathcal{K}$ with property $sgn(K(x_i, t_i) - r_i) = b_i$ for any $i \in \mathbb{N}_m$. Then, we define a pseudo-dimension $d_{\mathcal{K}}$ of $\mathcal{K}$ to be the maximum cardinality of S_m that is pseudo-shattered by $\mathcal{K}$.*

The Rademacher chaos complexity can be bounded using pseudo-dimensions.

Theorem 6 *Denote the pseudo-dimension of $\mathcal{K}$ by $d_{\mathcal{K}}$. Then, there exists a universal constant C such that, for any $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, there holds*

$$\hat{\mathscr{U}}_n(\mathcal{K}) \le C(1+\kappa)^2 d_{\mathcal{K}} \ln(2en^2). \tag{11}$$

For Gaussian-type kernels, we can explicitly bound the empirical Rademacher chaos complexities. First, consider the set of scalar candidate kernels given by

$$\mathcal{K}_{\text{gau}} = \big\{e^{-\sigma\|x-t\|^2} : \ \sigma \in [0,\infty)\big\}. \tag{12}$$

The second class of candidate kernels is more general as considered in [21]: the whole class of *radial basis kernels*. Let $\mathcal{M}(\mathbb{R}^+)$ be the class of probabilities on $\mathbb{R}^+$. We consider the candidate kernel defined by

$$\mathcal{K}_{\text{rbf}} = \Big\{ \int_0^\infty e^{-\sigma\|x-t\|^2} dp(\sigma) : p \in \mathcal{M}(\mathbb{R}^+)\Big\} \tag{13}$$

For the above specific sets of base kernels, we can have the following result by estimating the pseudo-dimension of $\mathcal{K}_{\text{gau}}$.

Corollary 7 *Let candidate kernels be given by equation (12) and (13). Then, there exists a universal constant C, such that, for $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, there holds*

$$\hat{\mathscr{U}}_n(\mathcal{K}_{rbf}) \le \hat{\mathscr{U}}_n(\mathcal{K}_{gau}) \le C\,(1+\kappa)^2 \ln(2en^2).$$

Theorem 6 and Corollary 7 will be proved in Section 5. Define the convex hull of $\mathcal{K}$ by

$$\begin{aligned}\text{conv}\big(\mathcal{K}\big) \ := \big\{&\textstyle\sum_{j\in\mathbb{N}_m} \lambda_\ell K_\ell : K_\ell \in \mathcal{K}, \lambda_\ell \ge 0,\\ &\textstyle\sum_{\ell\in\mathbb{N}_m} \lambda_\ell = 1, m \in \mathbb{N}\big\}.\end{aligned}$$

Then, it is easy to check, by the definition of the Rademacher chaos complexity, that $\hat{\mathscr{U}}_n\big(\text{conv}\big(\mathcal{K}_{\text{rbf}}\big)\big) \le \hat{\mathscr{U}}_n(\mathcal{K}_{\text{rbf}})$ and $\hat{\mathscr{U}}_n\big(\text{conv}\big(\mathcal{K}_{\text{gau}}\big)\big) \le \hat{\mathscr{U}}_n(\mathcal{K}_{\text{gau}})$. One can also see [25] for more examples of Gaussian kernels with low rank covariance matrices.

Combining Theorems 3, 6 with Corollary 7, for learning the kernel problem (1) with the set of candidate kernels $\mathcal{K} = \mathcal{K}_{\text{rbf}}$ or $\mathcal{K} = \mathcal{K}_{\text{gau}}$ the excess generalization bound can be summarized as follows: there exists a universal constant C such that, with probability at least $1-\delta$ there holds

$$\begin{aligned}\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_\rho^\phi) \le C\Big(&C_\lambda^\phi\big(\tfrac{\ln n}{n\lambda}\big)^{\frac12}\\ &+M_\lambda^\phi\big(\tfrac{\ln\frac{2}{\delta}}{n}\big)^{\frac12}\Big) + \mathcal{D}(\lambda).\end{aligned} \tag{14}$$

From the above equation, by choosing λ appropriately we can derive meaningful excess generalization error rates with respect to the sample number n, and hence excess misclassification error rates by the comparison inequalities such as inequality (4). To this end, we usually assume conditions on the distribution ρ or some regularity condition on the target function f_ρ^ϕ under which the regularization error $\mathcal{D}(\lambda)$ decays polynomially. For instance, we can employ the following condition introduced in [8].

Definition 8 *We say that ρ is separable by $\{\mathcal{H}_K : K \in \mathcal{K}\}$ if there is some $f_{sp} \in \mathcal{H}_{\bar K}$ with some $\bar K \in \mathcal{K}$ such that $y f_{sp}(x) > 0$ almost surely. It has separation exponent $\theta \in (0,\infty]$ if we can choose f_{sp} and positive constants Δ, c_θ such that $\|f_{sp}\|_{\bar K} = 1$ and*

$$\rho_X\big\{x \in X : |f_{sp}(x)| < \Delta t\big\} \le c_\theta t^\theta, \qquad \forall t > 0. \tag{15}$$

Observe that condition (15) with $\theta = \infty$ is equivalent to $\rho_X\big\{x \in X : |f_{\text{sp}}(x)| < \gamma t\big\} = 0, \quad \forall\ 0 < t < 1$. That is, $|f_{\text{sp}}(x)| \ge \gamma$ almost everywhere. Thus, separable distributions with separation exponent $\theta = \infty$ correspond to strictly separable distributions. Other assumptions on the distribution ρ such as the geometric noise condition introduced in [26, 27] are possible to achieve polynomial decays of the regularization error.

We are now ready to state misclassification error rates. Hereafter, the expression $a_n = \mathcal{O}(b_n)$ means that there exists an absolute constant c such that $a_n \le c b_n$ for all $n \in \mathbb{N}$.

Example 1 *Let $\phi(t) = (1-t)_+$ be the hinge loss and consider the kernel learning formulation (1) with $\mathcal{K}$ given by either $\mathcal{K}_{gau}$ or $\mathcal{K}_{rbf}$. Suppose that the separation condition holds true with exponent $\theta > 0$. Then, by choosing $\lambda = n^{-\frac{2+\theta}{(2+3\theta)}}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ there holds*

$$\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \le \mathcal{O}\Big(\big[\ln n + \ln(2/\delta)\big]^{\frac12}\Big(\frac1n\Big)^{\frac{\theta}{3\theta+2}}\Big).$$

The proof of this example is postponed to Section 6. Other examples such as least square loss regression can be established. In this case we need to consider the function approximation [12, 31] on a domain of $\mathbb{R}^d$.

3 Related Work

Statistical bounds with Rademacher complexities were first pursued by [17, 7] for learning the kernel from a linear combination of finite candidate kernels. The Rademacher complexities are estimated by the eigenvalues of the candidate kernel matrix over the inputs.

It was established by Ying and Zhou [31] that the union space $\mathcal{B}_\mathcal{K}$ is a uniform Glivenko-Cantelli (uGC) class (see definition in [1]) if and only if, for any $\gamma > 0$, the V_γ-dimension of

$$\mathcal{K}_X = \{K(\cdot,x) : x \in X, K \in \mathcal{K}\}$$

is finite. There, the empirical covering number of $\mathcal{K}_X$ was also estimated. Based on these main results, the Rademacher bounds were established in [31, 21][1]:

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^\phi) \ &\le 4C_\lambda^\phi\Big(\tfrac{2R_n(\mathcal{K}_X)}{\sqrt{n}\lambda}\Big)^{\frac12} + 4\kappa C_\lambda^\phi\Big(\tfrac{1}{\sqrt{n}\lambda}\Big)^{\frac12}\\ &\quad + M_\lambda^\phi\Big(\tfrac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac12} + \tfrac{2}{\sqrt{n}}.\end{aligned}$$

Here, the Rademacher complexity $R_n(\mathcal{K}_X)$ is defined by $\mathbb{E}\sup_{f\in\mathcal{K}_X}\frac{1}{\sqrt{n}}\big|\sum_{i\in\mathbb{N}_n}\varepsilon_i f(x_i)\big|$ which is often bounded by $\mathcal{O}\big(d_\mathcal{K}\ln n\big)$ by using metric entropy integrals, see Theorem 20 in [31]. Hence, the resultant rates are quite loose whose dependence on the sample number is of order $n^{-\frac14}$ in comparison with our new bound of order $n^{-\frac12}$ summarized in

[1] This bound is originally given in the form of expectation. However, it is easy to convert it to the current probabilistic form by the bounded difference inequality from which the extra term $M_\lambda^\phi\big(\ln(\frac{1}{\delta})/n\big)^{\frac12}$ appears.

equation (14). Specifically, for the hinge loss, as stated in Example 1 we can get a better rate $\mathcal{O}\big((\log n)^{\frac{1}{2}} n^{-\frac{\theta}{2+3\theta}}\big)$ in comparison with the rate $\mathcal{O}\big((\log n)^{\frac{1}{2}} n^{-\frac{\theta}{2(2+3\theta)}}\big)$ given in [31].

Subsequently, Srebro and Ben-David [25] employed matrix analysis techniques to directly estimate the empirical covering number of $\mathcal{B}_\mathcal{K}$ with the pseudo-dimension of the candidate kernels. Margin bounds were established for SVM. Specifically, let

$$\mathcal{R}^\gamma_\mathbf{z}(f) = \frac{|\{i : y_i f(x_i) < \gamma\}|}{n}.$$

Note $f^\phi_\mathbf{z} \in \frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$ where $\mathcal{B}_\mathcal{K}$ is the same as the notation $\mathcal{F}_\mathcal{K}$ used in [25]. A simple modification of Theorem 2 in [25] to the function class $\frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$, for any margin cost function ψ defined by equation (10), there holds

$$\begin{aligned}\mathcal{R}(\mathrm{sgn}(f^\phi_\mathbf{z})) \quad &\le \mathcal{R}^\gamma_\mathbf{z}(f^\phi_\mathbf{z}) + \Big(8(2+d_\mathcal{K}) \ln \tfrac{128en^3\kappa^2}{\gamma^2\lambda d_\mathcal{K}} \\ &\quad +256\tfrac{\kappa^2}{\gamma^2\lambda} \ln \tfrac{128n\kappa^2}{\gamma^2\lambda} + \ln \tfrac{1}{\delta}\Big)^{\frac{1}{2}}/\sqrt{n}.\end{aligned}$$

Since

$$\mathcal{R}^\gamma_\mathbf{z}(f^\phi_\mathbf{z}) \ge \mathcal{E}^\psi_\mathbf{z}(f^\phi_\mathbf{z})$$

Theorem 4 implies

$$\begin{aligned}\mathcal{R}(\mathrm{sgn}(f^\phi_\mathbf{z})) \quad &\le \mathcal{R}^\gamma_\mathbf{z}(f^\phi_\mathbf{z}) + 8C\Big(\tfrac{2(1+\kappa)^2 d_\mathcal{K} \ln(2en^2)}{n\lambda\gamma^2}\Big)^{\frac{1}{2}} \\ &\quad +4\Big(\tfrac{\ln \frac{2}{\delta}}{n}\Big)^{\frac{1}{2}}.\end{aligned}$$

Comparing the above two margin bounds, there is no logarithmic margin term, i.e. $\ln \frac{1}{\gamma^2}$, in our bound. One possible advantage of the direct empirical covering approach [25] is that the dependence on the pseudo-dimension and margin is roughly in a additive form, i.e. $d_\mathcal{K} \ln \frac{1}{\gamma^2} + \frac{1}{\gamma^2} \ln \frac{1}{\gamma^2}$. The Rademacher approach is of multiplicative form $\sqrt{\frac{d_\mathcal{K}}{\gamma^2}}$ due to the contraction inequality of Rademacher averages for the margin cost function.

However, considering that our main target is to estimate generalization bounds and excess misclassification errors, the direct approach [25] would result in quite loose generalization bounds. To see this, we focus on the hinge loss and recall the scaling version of Theorem 1 in [25]:

$$\mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda}) \le 2\big(\frac{en^2\kappa^2}{\varepsilon\sqrt{\lambda}}\big)^{d_\mathcal{K}} \big(\frac{16n\kappa^2}{\varepsilon^2\lambda}\big)^{\frac{64\kappa^2}{\varepsilon^2\lambda} \ln\left(\frac{\varepsilon\sqrt{\lambda}en}{8\kappa}\right)}.$$

There are two ways to get generalization bounds from the above covering number: the Rademacher approach with entropy integrals and the classical method. We point out that the first approach does not work since the entropy

$$\ln \mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda}) = \mathcal{O}(\varepsilon^{-2})$$

which tells us the entropy integral

$$\int_0^\infty \sqrt{\ln \mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon)} d\varepsilon = \infty.$$

The second approach is a classical method. For example, applying Theorem 2.3 of [19] (or Lemma 3.4 of [1]) to the function class $\phi \circ \mathcal{B}_\lambda$ implies that

$$\begin{aligned}\mathcal{E}^\phi(f^\phi_\mathbf{z}) - \mathcal{E}^\phi_\mathbf{z}(f^\phi_\mathbf{z}) \quad &\le \sup_{f\in\mathcal{B}_\lambda} \big|\mathcal{E}^\phi(f) - \mathcal{E}^\phi_\mathbf{z}(f)\big| \\ &\le 8\mathbb{E}\big[\mathcal{N}_\infty(\varepsilon, \phi\circ\mathcal{B}_\lambda, \mathbf{z})\big] e^{-\frac{n\varepsilon^2\lambda}{128\kappa^2}},\end{aligned}$$

where $\phi \circ \mathcal{B}_\lambda = \{\phi(yf(x)) : f \in \mathcal{B}_\lambda\}$ and $\mathcal{N}_\infty(\varepsilon, T, \mathbf{z})$ is the empirical covering number defined, for any $f, g \in T$, by the pseudo-metric $d_\mathbf{z}(f,g) = \sup_{i\in\mathbb{N}_n} \big|f(z_i) - g(z_i)\big|$. Note for the hinge loss, $\mathcal{N}_\infty(\varepsilon, \phi\circ\mathcal{B}_\lambda, \mathbf{z}) \le \mathcal{N}_\infty(\varepsilon, \mathcal{B}_\lambda, \mathbf{x}) = \mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda})$. Hence, with probability at least $1-\delta$, there holds

$$\mathcal{E}^\phi(f^\phi_\mathbf{z}) - \mathcal{E}^\phi_\mathbf{z}(f^\phi_\mathbf{z}) \le \varepsilon$$

where ε satisfies the equation

$$\frac{n\varepsilon^2\lambda}{128\kappa^2} \ge \ln \mathbb{E}\big[\mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda})\big] + \ln \frac{8}{\delta}. \tag{16}$$

Consequently, from equation (16) we have, at least for $\varepsilon \le 1$, that $\varepsilon \ge 64\kappa\Big(\frac{\ln\left(16n\kappa^2/\lambda\right)}{n\lambda^2}\Big)^{\frac{1}{4}}$, which makes the generalization bound unacceptably loose, and hence leads to loose excess misclassification error bounds.

Moreover, Rademacher approaches are usually more flexible. For instance, it is unknown how to directly estimate the pseudo-dimension of RBF kernels $\mathcal{K}_{\mathrm{rbf}}$ and hence it could be a problem to directly apply the approach of [25]. The Rademacher approaches can handle this general case using the Rademacher chaos complexity of $\mathcal{K}_{\mathrm{gau}}$ instead of directly using that of $\mathcal{K}_{\mathrm{rbf}}$ as stated in Corollary 7 in Section 2.

4 Generalization Bounds by Rademacher Chaos

In this section we show that the excess generalization bound for the kernel learning formulation (1) can be bounded by well-established Rademacher chaos of order two as stated in Theorem 3.

To prove this theorem, we recall the definition of the ordinary *Rademacher complexity*, see e.g. [5, 15, 16]. For any class F of bounded functions and any $n \in \mathbb{N}$, the empirical Rademacher complexity is defined by

$$\hat{R}_n(F) := \frac{1}{\sqrt{n}} \sup_{f\in F} \Big| \sum_{i\in\mathbb{N}_n} \epsilon_i f(z_i)\Big|$$

where $\{z_i : i \in \mathbb{N}_n\}$ are independent random variables distributed according to μ. Its useful properties can be found in, e.g. [5, 16, 19].

Now we assemble the necessary materials to obtain the main technical lemma.

Lemma 9 *Suppose the cost function $\psi : \mathbb{R} \to [0,\infty)$ is locally Lipschitz continuous with $\psi(0) = 1$. Let $\mathcal{B}_\lambda$ be defined by equation (7) and M^ψ_λ, $\mathcal{C}^\psi_\lambda$ be respectively defined by (8) and (9). Then, with probability at least $1-\delta$, there holds*

$$\begin{aligned}\sup_{f\in\mathcal{B}_\lambda} \big|\mathcal{E}^\psi(f) - \mathcal{E}^\psi_\mathbf{z}(f)\big| &\le 4C^\psi_\lambda \Big(\tfrac{2\hat{\mathscr{U}}_n(\mathcal{K})}{\lambda n}\Big)^{\frac{1}{2}} \\ &\quad +4\kappa C^\psi_\lambda \Big(\tfrac{1}{n\lambda}\Big)^{\frac{1}{2}} + 3M^\psi_\lambda \Big(\tfrac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \tfrac{2}{\sqrt{n}}.\end{aligned}$$

Proof: For any $\mathbf{z} = \{(x_i, y_i) : i \in \mathbb{N}_n\}$, let $\mathbf{z}' = \{(x_i, y_i) : i \in \mathbb{N}_n\}$ be the same copy of $\mathbf{z}$ with k-th sample

replaced by sample (x_k', y_k'). Since ψ is nonnegative, the bounded difference coefficient is given by

$$\begin{aligned}&\Big|\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}}(f)\big|-\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}'}(f)\big|\Big|\\&\le\sup_{f\in\mathcal{B}_\lambda}\Big|\mathcal{E}^\psi_{\mathbf{z}}(f)-\mathcal{E}^\psi_{\mathbf{z}'}(f)\Big|\\&=\tfrac{1}{n}\sup_{f\in\mathcal{B}_\lambda}\Big|\psi(y_kf(x_k))-\psi(y_k'f(x_k'))\Big|\le M^\psi_\lambda/n.\end{aligned}$$

By McDiarmid's bounded difference inequality (e.g. [13]), with probability $1-\frac{\delta}{2}$ there holds that

$$\begin{aligned}&\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}}(f)\big|\\&\le\mathbb{E}\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}}(f)\big|+M^\psi_\lambda\Big(\frac{\ln\frac{2}{\delta}}{2n}\Big)^{\frac12}.\end{aligned}\tag{17}$$

Consequently, the first term on the righthand side of the above inequality can be estimated by the standard symmetrization arguments. Indeed, with probability at least $1-\frac{\delta}{2}$, there holds

$$\begin{aligned}&\mathbb{E}\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}}(f)\big|\Big]\\&\le2\mathbb{E}\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{B}_\lambda}\tfrac{1}{n}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i\psi(y_if(x_i))\Big|\Big]\\&\le2\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{B}_\lambda}\tfrac{1}{n}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i\psi(y_if(x_i))\Big|\\&\quad+2M^\psi_\lambda\big(\tfrac{\ln\frac{2}{\delta}}{2n}\big)^{\frac12},\end{aligned}\tag{18}$$

where the last inequality used again the McDiarmid's bounded difference inequality. Note that $\|f\|_\infty\le\kappa\sqrt{1/\lambda}$ for all $f\in\mathcal{B}_\lambda$. Then, from the definition of C^ψ_λ given by equation (9), $\bar\psi=\psi-\psi(0):\mathbb{R}\to\mathbb{R}$ has the Lipschitz constant C^ψ_λ and $\bar\psi(0)=0$. Applying the contraction property of Rademacher averages (e.g. Property (4) of Theorem 12 in [5] or Lemma 16 in [31]) implies that, with probability $1-\frac{\delta}{2}$

$$\begin{aligned}&\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i\psi(y_if(x_i))\Big|\Big]\\&\le\mathbb{E}_\varepsilon\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i\bar\psi(y_if(x_i))\Big|+\mathbb{E}_\varepsilon\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i\Big|\\&\le2C^\psi_\lambda\,\mathbb{E}_\varepsilon\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_if(x_i)\Big|+\Big(\mathbb{E}_\varepsilon\sum_{i,j\in\mathbb{N}_n}\epsilon_i\epsilon_j\Big)^{1/2}\\&\le2C^\psi_\lambda\,\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_if(x_i)\Big|\Big]+\sqrt{n},\end{aligned}$$

where, in the second inequality, we used the assumption that $\psi(0)=1$. Finally, we can rewrite $\mathbb{E}_\varepsilon\sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_if(x_i)\Big|$ as

$$\begin{aligned}&\mathbb{E}_\varepsilon\quad\sqrt{\tfrac{1}{\lambda}}\sup_{K\in\mathcal{K}}\sup_{\|f\|_K\le1}\Big|\langle\sum_{i\in\mathbb{N}_n}\epsilon_iK_{x_i},f\rangle_K\Big|\\&\quad=\sqrt{\tfrac{1}{\lambda}}\mathbb{E}_\varepsilon\sup_{K\in\mathcal{K}}\Big|\textstyle\sum_{i,j\in\mathbb{N}_n}\varepsilon_i\varepsilon_jK(x_i,x_j)\Big|^{\frac12}\\&\quad\le\sqrt{\tfrac{2n}{\lambda}}\sqrt{\hat{\mathscr{U}}_n(\mathcal{K})}+\sqrt{\tfrac{1}{\lambda}}\sup_{K\in\mathcal{K}}\sqrt{\mathrm{trace}(\mathbf{K})},\end{aligned}$$

where $\mathbf{K}=(K(x_i,x_j))_{i,j\in\mathbb{N}_n}$. Putting all the above inequalities back into (18) yields that

$$\begin{aligned}&\mathbb{E}\Big[\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\psi(f)-\mathcal{E}^\psi_{\mathbf{z}}(f)\big|\Big]\\&\le4C^\psi_\lambda\sqrt{\tfrac{2\hat{\mathscr{U}}_n(\mathcal{K})}{\lambda n}}+4C^\psi_\lambda\kappa\Big(\tfrac{1}{\lambda n}\Big)^{\frac12}+\tfrac{2}{\sqrt{n}}+2M^\psi_\lambda\big(\tfrac{\ln\frac{2}{\delta}}{2n}\big)^{\frac12}\end{aligned}$$

where the last inequality used the fact that $\mathrm{trace}(\mathbf{K})\le\kappa^2n$. This combining with inequality (17) yields the desired result. □

We are ready to prove Theorem 3 and Corollary 4.

Proof of Theorem 3: Recall that $f_{\mathbf{z}},f_\lambda\in\mathcal{B}_\lambda$, hence $\mathcal{S}_{\mathbf{z},\lambda}\le2\sup_{f\in\mathcal{B}_\lambda}\big|\mathcal{E}^\phi(f)-\mathcal{E}^\phi_{\mathbf{z}}(f)\big|$. Note that ϕ is a normalized classifying loss. Then, putting Lemma 9 with $\psi=\phi$ and the error decomposition (5) together yielding the desired theorem. □

Proof of Corollary 4: The margin-based cost function ψ obviously satisfies the conditions in Lemma 9 with $C^\psi_\lambda=\frac{1}{\gamma}$ and $M^\psi_\lambda=1$. Since $\chi_{y\neq\mathrm{sgn}(f(x))}\le\psi(yf(x))$, there holds that $\mathcal{R}(\mathrm{sgn}(f^\phi_{\mathbf{z}}))\le\mathcal{E}^\psi(f^\phi_{\mathbf{z}})$ which, combining with Theorem 3, yields the desired assertion. □

5 Estimating the Rademacher Chaos Complexity

In this section we further estimate the Rademacher chaos complexity $\hat{\mathscr{U}}_n(\mathcal{K})$ by the metric entropy integral, and then prove Theorem 6 and Corollary 7 as stated in Section 2.

Now let $\mathcal{G}$ be a set of functions on $X\times X$ and $\mathbf{x}=\{x_i\in X:i\in\mathbb{N}_n\}$, define the l^2 empirical metric of two functions $f,g\in\mathcal{G}$ by

$$d_{\mathbf{x}}(f,g)=\Big(\frac{1}{n^2}\sum_{i,j\in\mathbb{N}_n,i<j}|f(x_i,x_j)-g(x_i,x_j)|^2\Big)^{\frac12}.$$

The *empirical covering number* $\mathcal{N}_2(\mathcal{G},\mathbf{x},\eta)$ is the smallest number of balls with pseudo-metric $d_{\mathbf{x}}$ required to cover $\mathcal{G}$.

The empirical Rademacher chaos complexity $\hat{\mathscr{U}}_n(\mathcal{K})$ can be bounded by the metric entropy integral as follows.

Lemma 10 *There exists a universal constant C such that, for any $\mathbf{x}=\{x_i:i\in\mathbb{N}_n\}$, there holds*

$$\hat{\mathscr{U}}_n(\mathcal{K})\le C\int_0^\infty\log\mathcal{N}_2(\mathcal{K}\cup\{0\},\mathbf{x},\varepsilon)d\varepsilon.$$

Proof: We rely on [11] to prove this result. Let $\mathbf{X}_T=\{X_s:s\in T\}$ be a real-valued homogeneous Rademacher chaos process of order two and the pseudo-distance defined by $\rho_{\mathbf{x}}(s,t)=(\mathbb{E}|X_s-X_t|^2)^{\frac12}$. Then, by Corollary 5.1.8 in [11] we know that there exists a universal constant C such that

$$\mathbb{E}_\varepsilon\sup_{s,t\in T}|X_s-X_t|\le C\int_0^\infty\big[\log\mathcal{N}(\mathbf{X}_T,\rho_{\mathbf{x}},\varepsilon)\big]d\varepsilon\tag{19}$$

In our context, for $\mathbf{x}=\{x_i\in X:i\in\mathbb{N}_n\}$, define the Rademacher chaos process of order two indexed by

$$\Big\{X_K=\frac{1}{n}\sum_{i,j\in\mathbb{N}_n,i<j}\varepsilon_i\varepsilon_jK(x_i,x_j):K\in\mathcal{K}\cup\{0\}\Big\}.$$

Observe that

$$\begin{aligned}\rho_{\mathbf{x}}(K,K')^2&=\mathbb{E}|X_K-X'_K|^2\\&=\tfrac{1}{n^2}\sum_{i<j,i'<j'}\mathbb{E}\big[\varepsilon_i\varepsilon_j\varepsilon_{i'}\varepsilon_{j'}(K(x_i,x_j)-K'(x_i,x_j))\\&\qquad\times(K(x_{i'},x_{j'})-K'(x_{i'},x_{j'}))\big]\\&=\sum_{i<j}\big|K(x_i,x_j)-K'(x_i,x_j)\big|^2/n^2=d_{\mathbf{x}}(K,K')^2.\end{aligned}$$

Hence, $\mathcal{N}(\mathbf{X}_T, d_{\mathbf{x}}, \varepsilon) = \mathcal{N}_2(\mathcal{K} \cup \{0\}, d_{\mathbf{x}}, \varepsilon)$ for any $\varepsilon > 0$. Consequently, applying equation (19) yields the desired assertion. □

It is worth mentioning that the standard entropy integral for bounding the superma of Rademacher chaos processes of order one (Rademacher averages) is of the form

$$\int_0^{\infty} \sqrt{\log \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon)} d\varepsilon.$$

One can see [11] for general entropy integrals to bound the superma of Rademacher chaos processes of order m for any $m \in \mathbb{N}$. Also, it is worth noting that

$$\int_0^{\infty} \log \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) d\varepsilon = \int_0^{\kappa^2} \log \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) d\varepsilon$$

since $\mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) = 1$ whenever ε is larger than κ^2.

The empirical covering number can further be bounded by the shattering dimension of the set of candidate kernels.

Lemma 11 *If the pseudo-dimension $d_{\mathcal{K}}$ of the set of basis kernels is finite, then we have that*

$$\mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) \leq 1 + \Big(\frac{en^2\kappa^2}{\varepsilon d_{\mathcal{K}}}\Big)^{d_{\mathcal{K}}}.$$

Proof: Note, for any $K', K \in \mathcal{K}$, that $d_{\mathbf{x}}(K', K) \leq D_{\infty}^{\mathbf{x}}(K', K) := \sup_{i \in \mathbb{N}_n} \big|K(x_i, x_j) - K'(x_i, x_j)\big|$ and denote by $\mathcal{N}_{\infty}(\mathcal{K}_X \cup \{0\}, \mathbf{x}, \varepsilon)$ the empirical covering number with pseudo-metric $D_{\infty}^{\mathbf{x}}$. Hence, $\mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) \leq \mathcal{N}_{\infty}(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon)$. Now, applying that the relation (see Chapter 11 of [2] and also Lemma 3 of [25]) between the covering number and pseudo-dimension implies that

$$\mathcal{N}_{\infty}(\mathcal{K}, \mathbf{x}, \varepsilon) \leq \Big(\frac{en^2\kappa^2}{\varepsilon d_{\mathcal{K}}}\Big)^{d_{\mathcal{K}}}.$$

Consequently,

$$\begin{aligned} \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) &\leq \mathcal{N}_{\infty}(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) \\ &\leq 1 + \mathcal{N}_{\infty}(\mathcal{K}, \mathbf{x}, \varepsilon) \leq 1 + \Big(\tfrac{en^2\kappa^2}{\varepsilon d_{\mathcal{K}}}\Big)^{d_{\mathcal{K}}} \end{aligned}$$

which completes the assertion. □

We are in a position to apply Lemma 11 and Lemma 10 to prove Theorem 6.

Proof of Theorem 6: Since $d_{\mathcal{K}} \geq 1$ and $\frac{en^2\kappa^2}{\varepsilon} \geq 1$ for any $0 < \varepsilon \leq \kappa^2$, we have that

$$1 + \big(\frac{en^2\kappa^2}{d_{\mathcal{K}}\varepsilon}\big)^{d_{\mathcal{K}}} \leq \big(\frac{2en^2\kappa^2}{\varepsilon}\big)^{d_{\mathcal{K}}}.$$

Combinig this fact with Lemma 11 and Lemma 10, we have that

$$\begin{aligned} \hat{\mathscr{U}}_n(\mathcal{K}) &\leq C \int_0^{\kappa^2} \log \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon) d\varepsilon \\ &\leq C \int_0^{\kappa^2} \ln\big(\frac{2en^2\kappa^2}{\varepsilon}\big)^{d_{\mathcal{K}}} d\varepsilon \\ &\leq C d_{\mathcal{K}} \Big[\int_0^{\kappa^2} 2 \ln \sqrt{\frac{\kappa^2}{\varepsilon}} d\varepsilon + \kappa^2 \ln(2en^2)\Big] \\ &\leq C d_{\mathcal{K}} \Big[\int_0^{\kappa^2} 2\sqrt{\frac{\kappa^2}{\varepsilon}} d\varepsilon + \kappa^2 \ln(2en^2)\Big] \\ &\leq 5C(1+\kappa)^2 d_{\mathcal{K}} \ln(2en^2). \end{aligned}$$

This finishes the assertion. □

For the set of scalar Gaussian kernels given by equation (12), we have the following estimation.

Lemma 12 *Consider the set of basis kernels $\mathcal{K}_{gau}$ given by equation (12), then $d_{\mathcal{K}_{gau}} = 1$.*

Proof: It is obvious that there exists at least one pair of points $(x, t) \in X \times X$ such that it is pseudo-shattering by $\mathcal{K}$. Now assume that two pairs of points (x_1, t_1) and (x_2, t_2) are shattering by $\mathcal{K}$. By Definition 5, that means there exists $r_1, r_2 \in \mathbb{R}$ and $\sigma, \sigma' \in [0, \infty)$ such that

$$e^{-\sigma\|x_1 - t_1\|^2} > r_1, \;\; e^{-\sigma\|x_2 - t_2\|^2} < r_2,$$

and

$$e^{-\sigma'\|x_1 - t_1\|^2} < r_1, \;\; e^{-\sigma'\|x_2 - t_2\|^2} > r_2.$$

Hence,

$$e^{-\sigma\|x_1 - t_1\|^2} > e^{-\sigma'\|x_1 - t_1\|^2},$$

and

$$e^{-\sigma\|x_2 - t_2\|^2} < e^{-\sigma'\|x_2 - t_2\|^2}.$$

Equivalently,

$$\sigma < \sigma', \;\text{ and } \sigma > \sigma',$$

which is obviously a contradiction. Consequently, the pseudo-dimension of $\mathcal{K}_{\text{gau}}$ is identical to one. □

We are ready to prove Corollary 7 on the estimation of the Rademacher chaos complexities of $\mathcal{K}_{\text{gau}}$ and $\mathcal{K}_{\text{rbf}}$.

Proof of Corollary 7: For the RBF kernels set $\mathcal{K}_{\text{rbf}}$, note, for any $\{x_i : i \in \mathbb{N}_n\}$, that $\hat{\mathscr{U}}_n(\mathcal{K}_{\text{rbf}})$ is bounded by

$$\begin{aligned} &\mathbb{E}_{\varepsilon} \sup_{p \in \mathcal{M}(\mathbb{R}^+)} \Big| \int_0^{\infty} \sum_{i<j} \varepsilon_i \varepsilon_j e^{-\sigma\|x_i - x_j\|^2} dp(\sigma) \Big| / n \\ &\leq \mathbb{E}_{\varepsilon} \sup_{\sigma \in \mathbb{R}^+} \big|\sum_{i<j} \varepsilon_i \varepsilon_j e^{-\sigma\|x_i - x_j\|^2}\big| / n \leq \hat{\mathscr{U}}_n(\mathcal{K}_{\text{gau}}). \end{aligned}$$

Then, the assertion follows immediately by combining Theorem 6 with Lemma 12. □

More examples such as Gaussian kernels with covariance matrices are illustrated in [25] whose pseudo-dimensions can directly be estimated using the techniques developed in Chapter 11 of [2].

6 Error rates

We are in a position to derive explicit error rates by trading off the sample error estimated by Rademacher chaos complexity and the regularization error.

Proof of Example 1: First note, for the hinge loss, that $C_{\lambda}^{\phi} = 1$ and $M_{\lambda}^{\phi} \leq 1 + \frac{\kappa}{\sqrt{\lambda}}$. Then, putting Theorems 3, 6 and Corollary 7 together, with probability at least $1 - \delta$ there holds that

$$\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_c) \leq \mathcal{O}\Big(\Big(\tfrac{\ln n}{n\lambda}\Big)^{\frac{1}{2}} + \Big(\tfrac{\ln \frac{2}{\delta}}{n\lambda}\Big)^{\frac{1}{2}}\Big) + \mathcal{D}(\lambda).$$

In addition, we know from [31] that if the distribution enjoys the weakly separation condition with exponent θ then

the regularization error decays as $\mathcal{D}(\lambda) = \mathcal{O}\Big(\lambda^{\frac{\theta}{\theta+2}}\Big)$. Letting $\lambda = n^{-\frac{\theta+2}{3\theta+2}}$ and noting the comparison inequality (4) yields the desired result. □

The last example is for the least square loss for classification. In this case, the target function f_ρ^ϕ is referred to as the *regression function* defined, for any $x \in X$, by $f_\rho(x) = P(Y = 1|x) - P(Y = -1|x)$. Similar error rates could be derived using some ideas in [8] for q-norm soft margin SVM loss and logistic regression loss.

Example 2 *Let X be a domain in $\mathbb{R}^d$ with Lipschitz boundary. Assume the regression function f_ρ belongs to the Sobolev space $H^s(X)$ with some $0 < s \leq 2$. If, moreover, the marginal distribution ρ_X is the Lebesgue measure then, by choosing $\lambda = n^{-\frac{2s+d}{2(4s+d)}}$, with probability at least $1-\delta$ there holds*

$$\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \mathcal{O}\Big(\Big[\ln n + \ln\frac{2}{\delta}\Big]^{\frac{1}{4}} n^{-\frac{s}{2(4s+d)}}\Big).$$

The proof is the same as the argument for Example 1 in [31]. However, we replace the rough bounds given there by our new tight generalization bound (e.g. equation (14)). Ignoring the difference of the forms to express error rates using expectations and probabilistic inequalities, Example 2 yields that $\mathcal{O}\Big((\ln n)^{\frac{1}{4}}\ n^{-\frac{2s-d-\varepsilon}{4(4s-d-2\varepsilon)}}\Big)$. Likewise, for the case $0 < s \leq 2$ and ρ_X is the Lebesgue measure, we got improved rates $\mathcal{O}\Big(\big(\ln n\big)^{\frac{1}{4}} n^{-\frac{s}{2(4s+d)}}\Big)$ in comparison with $\mathcal{O}\Big(\big(\ln n\big)^{\frac{1}{4}} n^{-\frac{s}{4(4s+d)}}\Big)$ obtained previously. Hence, our new error rates substantially improve those in [31].

7 Conclusion

In this paper we provided a novel statistical generalization bound for kernel learning algorithms which extends and improves the previous work in the literature [17, 7, 31, 21, 25]. The main tools are based on the theory of U-processes such as the so-called homogeneous Rademacher chaos of order two and metric entropy integrals involving empirical covering numbers.

There are several questions remaining to be further studied. Firstly, it would be interesting to get fast error rates with respect to the sample number as those in [4, 26, 27, 29]. For this purpose, the extension of localized Rademacher averages [6] to the scenario of multiple kernel learning would be useful. Secondly, it would be possible to give generalization bounds based on decoupling Gaussian chaos of order two. Thirdly, as mentioned in Section 3, it remains unknown how to get additive margin bounds using Rademacher approaches. Finally, the empirical Rademacher chaos complexity can be estimated from finite samples, and hence another direction for further investigation is to apply it to practical kernel learning problems.

Acknowledgements

We wish to thank the anonymous referees for their careful reading and valuable comments. This work is supported by EPSRC grant EP/E027296/1.

References

[1] N. Alon, S. Ben-David, S. N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, **44**: 615–631, 1997.

[2] M. Anthony and P.L. Bartlett. *Neural Networks Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] F. Bach. Consistency of the group Lasso and multiple kernel learning, *J. of Machine Learning Research*, **9**: 1179-1225, 2008.

[4] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. of the American Statistical Association*, **473**: 138-156, 2006.

[5] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. of Machine Learning Research*, **3**: 463–482, 2002.

[6] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, **33**: 1497–1537, 2005.

[7] O. Bousquet and D.J.L. Herrmann. On the complexity of learning the kernel matrix. *NIPS*, 2003.

[8] D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou. Support vector machine soft margin classifiers: error analysis. *J. of Machine Learning Research* **5**: 1143–1175, 2004.

[9] F. Cucker and D.X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.

[10] S. Clémencon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, **36**: 844–874, 2008.

[11] V.H. De La Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, New York, 1999.

[12] E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, **5**: 59–85, 2006.

[13] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1997.

[14] M. Girolami and S. Rogers. Hierarchic Bayesian models for kernel learning. *ICML*, 2005.

[15] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47: 1902–1914, 2001.

[16] V. Koltchinskii and V. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, **30**: 1–50, 2002.

[17] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. of Machine Learning Research*, **5**: 27–72, 2004.

[18] Y. Lin and H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, **34**: 2272–2297, 2006.

[19] S. Mendelson. A few notes on Statistical Learning Theory. In *Advanced Lectures in Machine Learning*, (S. Mendelson, A.J. Smola Eds), LNCS 2600: 1-40, Springer 2003.

[20] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization, *J. of Machine Learning Research*, **6**: 1099–1125, 2005.

[21] C. A. Micchelli, M. Pontil, Q. Wu, and D. X. Zhou. Error bounds for learning the kernel. Technical Report, City University of Hong Kong, 2005.

[22] C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *J. of Machine Learning Research* **6** 1043–1071, 2005.

[23] B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, Cambridge, MA, USA, 2002.

[24] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

[25] N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. *COLT*, 2006.

[26] I. Steinwart and C. Scovel. Fast rates for support vector machines. *COLT*, 2005.

[27] I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, **35**: 575-607, 2007.

[28] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New york, 2008.

[29] Q. Wu, Y. Ying, and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, **23**: 108-13, 2007.

[30] J. Ye, S. Ji, and J. Chen. Multi-class discriminant kernel learning via convex programming, *J. of Machine Learning Research*, **9** 719–758, 2008.

[31] Y. Ying and D. X. Zhou. Learnability of Gaussians with flexible variances. *J. of Machine Learning Research*, **8**: 249-276, 2007.

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, Series B, **68**: 49-67, 2006.

[33] Tong Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**: 56-85, 2004.