Bounds for Learning the Kernel: Rademacher Chaos Complexity

Yiming Ying enxyy@bris.ac.uk
Colin Campbell C.Campbell@bris.ac.uk
Department of Engineering Mathematics
University of Bristol
Queen's Building, Bristol, BS8 1TR, UK

Abstract

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

1. Introduction

Kernel methods (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004) such as Support Vector Machines (SVMs) have been extensively used for supervised learning tasks such as classification and regression. 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. This topic has recently received increasing attention which is often termed multi-kernel learning (MKL) in Machine Learning and nonparametric group lasso in Statistics. Lanckriet et al. (2004) proposed a semi-definite programming (SDP) approach to automatically learn a linear combination of candidate kernels for the case of SVMs. This approach was improved by Bach et al. (2004) who used sequential minimization optimization (SMO) and by Sonnenburg et al. (2006) who reformulated it as a semi-infinite linear programming (SILP) task. Other approaches include the so-called COSSO estimate for additive models (Lin and Zhang, 2006), hyperkernels (Ong and Smola, 2005), Bayesian probabilistic kernel learning models (Girolami and Rogers, 2005), and kernel discriminant analysis (Ye et al., 2008). Such MKL formulations have been successfully demonstrated in combining multiple heterogeneous data sources to enhance biological inference (Lanckriet et al., 2004).

The above mentioned MKL algorithms (Lanckriet et al., 2004; Bach et al., 2004; Sonnenburg et al., 2006) are based on the dual formulation of binary SVM to learn the linear combination of a finite set of candidate kernels. Departing from the primal problem, a general regularization framework for the kernel learning problem is formulated in Micchelli and Pontil (2005); Wu et al. (2006) with a potentially *infinite* number of candidate kernels. A difference of convex (DC) programming approach was proposed in Argyriou et al. (2006) for this framework. Specifically, let $\mathbb{N}_n = \{1, 2, \cdots, n\}$ for any $n \in \mathbb{N}$. 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 reflected 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 (basis) 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 regularization scheme (Micchelli and Pontil, 2005; Wu et al., 2006) for MKL can be cast as a two-layer minimization problem:

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

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 algorithm (1) with loss function ϕ.

The objective of statistical generalization analysis of MKL algorithms is to study its properties such as generalization and the characterization of the complexity of MKL system which are essential for building its theoretical foundations. Theoretical work towards this direction was pursued by Bousquet and Herrmann (2003); Lanckriet et al. (2004); Micchelli et al. (2005); Srebro and Ben-David (2006); Ying and Zhou (2007). In this paper we adopt the spirit of Rademacher complexity bounds for empirical risk minimization (ERM) and SVM with a single kernel (Bartlett et al., 2006; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002) and develop an appealing generalization bound for general MKL algorithm (1). Our novel approach is based on the principal theory of U-processes (e.g. Arcones and Giné (1993); De La Peña and Giné (1999)) which can yield tighter generalization bounds than previous approaches.

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 algorithm (1) reduces to investigation of the suprema of a homogeneous Rademacher chaos process 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 give a 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 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* (Devroye et al., 1997) 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 $\text{sign}(f_{\mathbf{z}}^{\phi})$ with the regularization scheme (1). For brevity, throughout this note we restrict our interest to a class of loss functions used in Wu et al. (2006), see also a general definition of classification loss functions in Bartlett et al. (2006).

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 $\text{sgn}(f(x))$ 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 $\text{sign}(f_{\mathbf{z}}^{\phi})$ approximates the Bayes rule f_c. More specifically, we aim to estimate the *excess misclassification error*

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

for the multi-kernel regularized classification algorithm (1). As shown in Zhang (2004); Bartlett et al. (2006), the excess misclassification error can usually 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 Zhang (2004) that $f_{\phi} = f_c$ and the

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

One can refer to Zhang (2004); Bartlett et al. (2006) 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* defined by

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

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

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

Then, we know from Ying and Zhou (2007) 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 Ying and Zhou (2007).

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 ball of candidate RKHSs be denoted by $\mathcal{B}_{\mathcal{K}} := \left\{ f : \quad f \in \mathcal{H}_K \text{ and } \|f\|_K \leq 1, \ K \in \mathcal{K} \right\}$. By the definition of $f_{\mathbf{z}}^{\phi}$, we get, for some RKHS $\mathcal{H}_K$, that $\frac{1}{n}\sum_{i=1}^{n} \phi(y_i f_{\mathbf{z}}^{\phi}(x_i)) + \lambda \|f_{\mathbf{z}}^{\phi}\|_K^2 \leq \frac{1}{n}\sum_{i=1}^{n} \phi(0) + \lambda \|0\|_K^2 = 1$. 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

$$f_{\mathbf{z}}^{\phi}, \ f_{\lambda}^{\phi} \in \mathcal{B}_{\lambda} := \frac{1}{\sqrt{\lambda}} \mathcal{B}_{\mathcal{K}} := \left\{ \frac{f}{\sqrt{\lambda}} : f \in \mathcal{B}_{\mathcal{K}} \right\}. \tag{6}$$

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

$$M_{\lambda}^{\psi} = \sup \left\{ |\psi(t)| : \ \forall |t| \leq \kappa\sqrt{1/\lambda} \right\}, \tag{7}$$

and denote its local Lipschitz constant by

$$C_{\lambda}^{\psi} = \sup \left\{ \frac{|\psi(x) - \psi(x')|}{|x - x'|} : \ \forall |x|, |x'| \leq \kappa\sqrt{\frac{1}{\lambda}} \right\}. \tag{8}$$

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

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 De La Peña and Giné (1999) 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 $\{\epsilon_i : i \in \mathbb{N}_n\}$ are independent Rademacher random variables. Also, $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$ are 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*

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

We refer to the expectation of its suprema

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

as the empirical Rademacher chaos complexity over F.

It is worth mentioning that the Rademacher process $\{\frac{1}{\sqrt{n}} \sum_{i \in \mathbb{N}_n} \varepsilon_i f(x_i) : f \in F\}$ 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 ranking and scoring problem is recently developed in Clémencon et al. (2008).

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 $\delta \in (0,1)$, with probability at least $1-\delta$, there holds*

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \leq 4C_\lambda^\phi \left(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n}\right)^{\frac{1}{2}} + 4\kappa C_\lambda^\phi \Big(\frac{1}{n\lambda}\Big)^{\frac{1}{2}} + 3M_\lambda^\phi \Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}, \quad (9)$$

and

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}(f_\rho^\phi) \leq 8C_\lambda^\phi \left(\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n}\right)^{\frac{1}{2}} + 8C_\lambda^\phi \kappa \Big(\frac{1}{n\lambda}\Big)^{\frac{1}{2}} + 6M_\lambda^\phi \Big(\frac{\ln(\frac{2}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{4}{\sqrt{n}} + \mathcal{D}(\lambda). \quad (10)$$

In practice, the empirical complexity $\hat{\mathcal{U}}_n(\mathcal{K})$ can be estimated from finite samples. In analogy to the data-dependent risk bounds of Rademacher averages (Bartlett et al., 2006), we can get margin bounds for learning the kernel problems using Rademacher chaos complexities.

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

$$\psi(t) = \begin{cases} 1, & t \le 0 \\ 1 - \frac{t}{\gamma}, & 0 < t \le \gamma \\ 0, & t > \gamma \end{cases} \tag{11}$$

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

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

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

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

where the last equality follows from the positive semi-definiteness of kernel K. Hence, the Rademacher chaos complexity can be estimated by

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

Consequently, Corollary 4 implies that

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

This coincides with the bound in Bartlett and Mendelson (2002) for the single kernel case with solutions $f_{\mathbf{z}}^{\phi}$ in the function space

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

Now we apply the well-established theory of U processes to estimate Rademacher chaos complexity by pseudo-dimension of candidate kernels. For this purpose, we recall the definition of kernel pseudo-dimension of a class of kernel functions on the product space $X \times X$, see Anthony and Bartlett (1999).

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*

$$\hat{\mathcal{U}}_n(\mathcal{K}) \leq C(1+\kappa)^2 d_{\mathcal{K}} \ln(2en^2). \tag{12}$$

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

$$\mathcal{K}_{\text{sc}} = \{e^{-\sigma\|x-t\|^2} : \quad \sigma \in [0, \infty)\}. \tag{13}$$

The second class of candidate kernels is more general as considered in Micchelli et al. (2005): 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

$$\mathcal{K}_{\text{rbf}} = \Big\{ \int_0^\infty e^{-\sigma\|x-t\|^2} dp(\sigma) : p \in \mathcal{M}(\mathbb{R}^+) \Big\} \tag{14}$$

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

Corollary 7 *Let candidate kernels be given by equation (13) and (14). Then, there exists a universal constant C, such that, for $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, there holds*

$$\hat{\mathcal{U}}_n(\mathcal{K}_{rbf}) \leq \hat{\mathcal{U}}_n(\mathcal{K}_{sc}) \leq C\ (1+\kappa)^2 \ln(2en^2).$$

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

$$\text{conv}(\mathcal{K}) := \left\{ \sum_{j \in \mathbb{N}_m} \lambda_\ell K_\ell : K_\ell \in \mathcal{K}, \lambda_\ell \geq 0, \sum_{\ell \in \mathbb{N}_m} \lambda_\ell = 1, m \in \mathbb{N} \right\}.$$

Then, it is easy to check, by the definition of the Rademacher chaos complexity, that

$$\hat{\mathcal{U}}_n(\text{conv}(\mathcal{K}_{\text{rbf}})) \leq \hat{\mathcal{U}}_n(\mathcal{K}_{\text{rbf}}), \quad \hat{\mathcal{U}}_n(\text{conv}(\mathcal{K}_{\text{sc}})) \leq \hat{\mathcal{U}}_n(\mathcal{K}_{\text{sc}}).$$

One can also refer to Srebro and Ben-David (2006) for more examples of Gaussian kernels with low rank covariance matrices.

Combining Theorems 3 with 6, 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

$$\mathcal{E}^{\phi}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \leq C\left(C_{\lambda}^{\phi}\left(\frac{d_{\mathcal{K}}\ln n}{n\lambda}\right)^{\frac{1}{2}} + M_{\lambda}^{\phi}(\frac{\ln\frac{2}{\delta}}{n})^{\frac{1}{2}}\right) + \mathcal{D}(\lambda). \tag{15}$$

From the above equation, by choosing λ appropriately we can derive meaningful excess generalization error rates with respect to sample number n, and hence excess misclassification error rates by the comparison inequalities such as equation (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.

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 $yf_{sp}(x) > 0$ 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*

$$\rho_X\left\{x \in X : |f_{sp}(x)| < \Delta t\right\} \leq c_\theta t^\theta, \qquad \forall t > 0. \tag{16}$$

Observe that condition (16) with $\theta = \infty$ is equivalent to

$$\rho_X\{x \in X : |f_{\mathrm{sp}}(x)| < \gamma t\} = 0, \qquad \forall\ 0 < t < 1.$$

That is, $|f_{\mathrm{sp}}(x)| \geq \gamma$ 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 Steinwart and Scovel (2005) 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 $a_n \leq cb_n$ for all $n \in \mathbb{N}$.

Example 1 *Let $\phi(t) = (1-t)_+$ be the hinge loss and consider the MKL algorithm (1) with $\mathcal{K}$ given by either $\mathcal{K}_{sc}$ 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*

$$\mathcal{R}(sgn(f_{\mathbf{z}}^{\phi})) - \mathcal{R}(f_c) \leq \mathcal{O}\Big([\ln n + \ln(2/\delta)]^{\frac{1}{2}}\Big(\frac{1}{n}\Big)^{\frac{\theta}{3\theta+2}}\Big).$$

The proof of this example is postponed to Section 6. Other examples such as least square loss regression can be found in Section 6. In this case we need to consider the function approximation (De Vito et al., 2006; Smale and Zhou, 2004) on a domain of $\mathbb{R}^d$.

3. Related Work

Statistical bounds with Rademacher complexities were first pursued by Lanckriet et al. (2004) and Bousquet and Herrmann (2003) for kernel learning from a linear combination of finite candidate kernels. The Rademacher complexities are estimated by the eigenvalues of the candidate kernel matrix over the inputs. Ying and Zhou (2007) first showed that the statistical generalization performance of MKL algorithms essentially relied on V_γ-dimension (see e.g. Alon et al., 1997; Anthony and Bartlett, 1999) of

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

There, the empirical covering number of $\mathcal{K}_X$ was also estimated. Based on these main results, the following generalization bounds of Rademacher averages were established in Ying and Zhou (2007); Micchelli et al. (2005)[1]:

$$\mathcal{E}(f_{\mathbf{z}}^{\phi}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^{\phi}) \leq 4C_{\lambda}^{\phi}\left(\frac{2R_n(\mathcal{K}_X)}{\sqrt{n}\lambda}\right)^{\frac{1}{2}} + 4\kappa C_{\lambda}^{\phi}\Big(\frac{1}{\sqrt{n}\lambda}\Big)^{\frac{1}{2}} + M_{\lambda}^{\phi}\Big(\frac{\ln(\frac{1}{\delta})}{n}\Big)^{\frac{1}{2}} + \frac{1}{\sqrt{n}}.$$

Here, the Rademacher complexity $R_n(\mathcal{K}_X)$ is defined by $\sup_{f\in\mathcal{K}_X}\frac{1}{\sqrt{n}}|\sum_{i\in\mathbb{N}_n}\varepsilon_i f(x_i)|$ which is often bounded by $\mathcal{O}(d_{\mathcal{K}}\ln n)$ by using metric entropy integrals, see Theorem 20 in Ying and Zhou (2007). Hence, the resultant rates are quite loose with dependence on the sample number of order $n^{-\frac{1}{4}}$. In contrast, by combining Theorem 3 with Theorem 6 our new bound is of order $n^{-\frac{1}{2}}$. Specifically, for the hinge loss for soft margin SVM classification, under the same conditions of Example 1 with $\mathcal{K} = \mathcal{K}_{\text{SC}}$, the following rates was obtained there

$$\mathbb{E}[\mathcal{R}(\text{sgn}(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)] = \mathcal{O}((\log n)^{\frac{1}{2}} n^{-\frac{\theta}{2(2+3\theta)}}).$$

In contrast, we can get $\mathcal{O}((\log n)^{\frac{1}{2}} n^{-\frac{\theta}{2+3\theta}})$ as stated in Example 1, and hence our Rademacher chaos complexity approach greatly improves the results in Ying and Zhou (2007).

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 $M_{\lambda}^{\phi}\Big(\ln(\frac{1}{\delta})/n\Big)^{\frac{1}{2}}$ appears.

Subsequently, Srebro and Ben-David (2006) employed a different approach by directly estimating 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, for any sample $\mathbf{z}$, that

$$f^\phi_\mathbf{z} \in \frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$$

where $\mathcal{B}_\mathcal{K}$ is the same as the notation $\mathcal{F}_\mathcal{K}$ used in Srebro and Ben-David (2006). A simple modification of Theorem 2 in Srebro and Ben-David (2006) to the function class $\frac{1}{\sqrt{\lambda}}\mathcal{B}_\mathcal{K}$ yields

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

Since $\mathcal{R}^\gamma_\mathbf{z}(f^\phi_\mathbf{z}) \ge \mathcal{E}^\psi_\mathbf{z}(f^\phi_\mathbf{z})$ with margin cost function ψ defined by equation (11), Corollary 4 implies

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

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 Srebro and Ben-David (2006) is that the dependence on the pseudo-dimension and margin is in a additive form, i.e. $d_\mathcal{K} + \ln\frac{1}{\gamma^2}$. The Rademacher approach is of multiplicative form $d_\mathcal{K}\ln\frac{1}{\gamma^2}$ 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 (Srebro and Ben-David, 2006) would result in quite loose generalization bounds. To see this, we focus on the hinge loss and recall the scaling version of Theorem 1 there:

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

There are two ways to get generalization bounds from this 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 Mendelson (2003) (or Lemma 3.4 of Alon et al. (1997)) to the function class $\phi \circ \mathcal{B}_\lambda$ implies that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \leq \sup_{f \in \mathcal{B}_\lambda} |\mathcal{E}^\phi(f) - \mathcal{E}_{\mathbf{z}}^\phi(f)| \leq 8\mathbb{E}[\mathcal{N}_\infty(\varepsilon, \phi \circ \mathcal{B}_\lambda, \mathbf{z})] e^{-\frac{n\varepsilon^2\lambda}{128\kappa^2}},$$

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 $d_{\mathbf{z}}(f, g) = \sup_{i \in \mathbb{N}_n} |f(z_i) - g(z_i)|$. Note for the hinge loss, $\mathcal{N}_\infty(\varepsilon, \phi \circ \mathcal{B}_\lambda, \mathbf{z}) \leq \mathcal{N}_\infty(\varepsilon, \mathcal{B}_\lambda, \mathbf{x}) = \mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda})$. Hence, with probability at least $1 - \delta$, there holds

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z}}^\phi) \leq \varepsilon,$$

with ε satisfying the equation

$$\frac{n\varepsilon^2\lambda}{128\kappa^2} \geq \ln \mathbb{E}[\mathcal{N}_n(\mathcal{F}_\mathcal{K}, \varepsilon\sqrt{\lambda})] + \ln \frac{8}{\delta}. \tag{17}$$

Hence, from equation (17) we have, at least for $\varepsilon \leq 1$, that $\varepsilon \geq 64\kappa \Big(\frac{\ln\big(16n\kappa^2/\lambda\big)}{n\lambda^2} \Big)^{\frac{1}{4}}$. This tells us that the sample complexity is of the form of $\mathcal{O}(n^{-\frac{1}{4}})$ which makes the generalization bound unacceptably loose, and hence leads to loose misclassification error bounds.

Moreover, Rademacher approaches are usually more flexible. For instance, it is unknown how to directly estimate the pseudo-dimension of the RBF kernels $\mathcal{K}_{\mathrm{rbf}}$ and hence it could be a problem to directly apply the approach of Srebro and Ben-David (2006). The Rademacher approaches can handle this general case using the Rademacher chaos complexity of $\mathcal{K}_{\mathrm{sc}}$ 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 of MKL algorithm (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 averages*, see e.g. Bartlett and Mendelson (2002); Bartlett et al. (2005); Koltchinskii (2001); Koltchinskii and Panchenko (2002).

Definition 9 *Let μ be a probability measure on Ω and F be a class of uniformly bounded and measurable functions on Ω. For any $n \in \mathbb{N}$, define the random variable 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 *and* $\{\epsilon_i : i = 1, \ldots, n\}$ *are independent Rademacher random variables, that is,* $P(\epsilon_i = +1) = P(\epsilon_i = -1) = 1/2$. *Also, we often call* $R_n(F) := \mathbb{E}[\hat{R}_n(F)] = \mathbb{E}_\mu \mathbb{E}_\varepsilon[R_n(F)]$ *the Rademacher averages (complexity) over the class* F.

Hence, $\hat{R}_n(F)$ is the suprema of the Rademacher process $\{\frac{1}{\sqrt{n}} \sum_{i \in \mathbb{N}_n} \epsilon_i f(z_i) : f \in F\}$ indexed by F which can also be regarded as the homogenous Rademacher chaos process of order one. Some useful properties of Rademacher averages are summarized in the following proposition, see e.g. Bartlett and Mendelson (2002); Ledoux and Talagrand (1991).

Proposition 10 *Let* F *be a class of uniformly bounded, real-valued, and measurable functions on* (Ω, μ). *Then, the following properties hold true.*

(a) For every $c \in \mathbb{R}$, $\mathbb{E}_\varepsilon R_n(cF) = |c| \mathbb{E}_\varepsilon R_n(F)$, *where* $cF = \{cf : f \in F\}$.

(b) If for each $i \in \mathbb{N}_n$, $\phi_i : \mathbb{R} \to \mathbb{R}$ *is a function with* $\phi_i(0) = 0$ *having a Lipschitz constant* c_i, *then for any* $\{x_i \in X : i \in \mathbb{N}_n\}$,

$$\mathbb{E}_\varepsilon \Big[\sup_{f \in F} | \sum_{i \in \mathbb{N}_n} \epsilon_i \phi_i(f(x_i))| \Big] \leq 2 \mathbb{E}_\varepsilon \Big[\sup_{f \in F} \Big| \sum_{i \in \mathbb{N}_n} c_i \epsilon_i f(x_i)| \Big].$$

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

Lemma 11 *Suppose the cost function* ψ *is Lipschitz continuous with* $\psi(0) = 1$. *Let* $\mathcal{B}_\lambda$ *be defined by equation (6) and* M_λ^ψ, $\mathcal{C}_\lambda^\psi$ *be respectively defined by (7) and (8). Then, with probability at least* $1 - \delta$, *there holds*

$$\sup_{f \in \mathcal{B}_\lambda} |\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f)| \leq 4 C_\lambda^\psi \Big(\frac{2 \hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n} \Big)^{\frac{1}{2}} + 4 \kappa C_\lambda^\psi \Big(\frac{1}{n\lambda} \Big)^{\frac{1}{2}} + 3 M_\lambda^\psi \Big(\frac{\ln(\frac{2}{\delta})}{n} \Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}}.$$

Proof By McDiarmid's bounded difference inequality (McDiarmid, 1989), with probability $1 - \frac{\delta}{2}$ there holds that

$$\sup_{f \in \mathcal{B}_\lambda} |\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f)| \leq \mathbb{E} \sup_{f \in \mathcal{B}_\lambda} |\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f)| + M_\lambda^\psi \Big(\frac{\ln \frac{2}{\delta}}{2n} \Big)^{\frac{1}{2}}. \tag{18}$$

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

$$\begin{aligned} &\mathbb{E} \Big[\sup\nolimits_{f \in \mathcal{B}_\lambda} |\mathcal{E}^\psi(f) - \mathcal{E}_{\mathbf{z}}^\psi(f)| \Big] \leq 2 \mathbb{E} \mathbb{E}_\varepsilon \Big[\sup\nolimits_{f \in \mathcal{B}_\lambda} \tfrac{1}{n} | \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i))| \Big] \\ &\leq 2 \mathbb{E}_\varepsilon \Big[\sup\nolimits_{f \in \mathcal{B}_\lambda} \tfrac{1}{n} | \sum_{i \in \mathbb{N}_n} \epsilon_i \psi(y_i f(x_i))| \Big] + 2 M_\lambda^\psi (\frac{\ln \frac{2}{\delta}}{2n})^{\frac{1}{2}}, \end{aligned} \tag{19}$$

where the last inequality used again the McDiarmid's bounded difference inequality. Note that $\|f\|_\infty \leq \kappa\sqrt{1/\lambda}$ for all $f \in \mathcal{B}_\lambda$. Then, from the definition of C_λ^ψ given by equation (8), $\bar{\psi} = \psi - \psi(0) : \mathbb{R} \to \mathbb{R}$ has the Lipschitz constant C_λ^ψ and $\bar{\psi}(0) = 0$. Applying the contraction property of Rademacher averages (e.g. Property (b) of Proposition 10) implies that, with probability $1 - \frac{\delta}{2}$

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

Finally, we know that

$$\begin{aligned}
\mathbb{E}_\varepsilon \sup_{f\in\mathcal{B}_\lambda}\Big|\sum_{i\in\mathbb{N}_n}\epsilon_i f(x_i)\Big| = \mathbb{E}_\varepsilon\ & \sqrt{\tfrac{1}{\lambda}}\sup_{K\in\mathcal{K}}\sup_{\|f\|_K\leq 1}\Big|\langle\sum_{i\in\mathbb{N}_n}\epsilon_i K_{x_i}, f\rangle_K)\Big| \\
&= \sqrt{\tfrac{1}{\lambda}}\mathbb{E}_\varepsilon \sup_{K\in\mathcal{K}}\Big|\textstyle\sum_{i,j\in\mathbb{N}_n}\varepsilon_i\varepsilon_j K(x_i,x_j)\Big|^{\frac{1}{2}} \\
&\leq \sqrt{\tfrac{2n}{\lambda}}\sqrt{\hat{\mathcal{U}}_n(\mathcal{K})} + \sqrt{\tfrac{1}{\lambda}}\sup_{K\in\mathcal{K}}\sqrt{trace(\mathbf{K})},
\end{aligned}$$

where $\mathbf{K} = (K(x_i, x_j))_{i,j\in\mathbb{N}_n}$. Putting all the above inequalities back into (19) yields that

$$\mathbb{E}\Big[\sup_{f\in\mathcal{B}_\lambda}|\mathcal{E}^\psi(f) - \mathcal{E}_\mathbf{z}^\psi(f)|\Big] \leq 4C_\lambda^\psi\sqrt{\frac{2\hat{\mathcal{U}}_n(\mathcal{K})}{\lambda n}} + 4C_\lambda^\psi\kappa\Big(\frac{1}{\lambda n}\Big)^{\frac{1}{2}} + \frac{2}{\sqrt{n}} + 2M_\lambda^\psi(\frac{\ln\frac{2}{\delta}}{2n})^{\frac{1}{2}},$$

where we used the fact that trace$(\mathbf{K}) \leq \kappa^2 n$. Putting this back into inequality (18) yields the desired result. ■

We are ready to prove Theorem 3 and Corollary 4.

Proof of Theorem 3: Recall that $f_\mathbf{z}, f_\lambda \in \mathcal{B}_\lambda$ and note that ϕ is a normalized classifying loss. Then, applying Lemma 11 with $\psi = \phi$ implies inequality (9). For the second assertion, observe that $\mathcal{S}_{\mathbf{z},\lambda} \leq 2\sup_{f\in\mathcal{B}_\lambda}|\mathcal{E}^\phi(f) - \mathcal{E}_\mathbf{z}^\phi(f)|$. Hence, putting Lemma 11 with $\psi = \phi$ and the error decomposition (5) together yields the desired inequality (10). ■

Proof of Corollary 4: The margin-based cost function ψ obviously satisfies the conditions in Lemma 11 with $C_\lambda^\psi = \frac{1}{\gamma}$ and $M_\lambda^\psi = 1$. Since $\chi_{y\neq\text{sgn}(f(x))} \leq \psi(yf(x))$, there holds that $\mathcal{R}(\text{sgn}(f_\mathbf{z}^\phi)) \leq \mathcal{E}^\psi(f_\mathbf{z}^\phi)$ which, combining with inequality (9) in Theorem 3, yields the desired assertion. ■

5. Estimating the Rademacher Chaos Complexity

It is well-known that Rademacher complexity can be estimated by the metric entropy integral involving covering numbers and nice applications to Statistical Learning Theory can be found in Bartlett et al. (2006); Bartlett and Mendelson (2002); Bartlett et al. (2005); Koltchinskii and Panchenko (2002); Mendelson (2003) and references therein. In this section we discuss how to estimate the Rademacher chaos complexity $\hat{\mathcal{U}}_n(\mathcal{K})$ by the metric entropy integral, and then prove Theorem 6 and Corollary 7 as stated in Section 2.

First, parallel to the properties of Rademacher averages, it is useful to outline some properties of the Rademacher chaos complexity some of which may be interesting in its own right.

Proposition 12 *Let $F_1, \cdots, F_k$ and H be classes of real functions on $X \times X$. Then the following holds true.*

(a) If $F \subseteq H$ then $\hat{\mathcal{U}}_n(F) \leq \hat{\mathcal{U}}_n(H)$.

(b) $\hat{\mathcal{U}}_n(conv(F)) = \hat{\mathcal{U}}_n(F)$.

(c) For any $c \in \mathbb{R}$, $\hat{\mathcal{U}}_n(cF) = |c|\hat{\mathcal{U}}_n(F)$.

(d) $\hat{\mathcal{U}}_n(\sum_{i\in\mathbb{N}_k} F_i) \leq \sum_{i\in\mathbb{N}_k} \hat{\mathcal{U}}_n(F_i)$.

(e) For any $1 < q < p < \infty$, the Khinchin-type inequality holds true

$$\left(\mathbb{E}_\varepsilon \sup_{f\in F} |\hat{U}_f(\varepsilon)|^q\right)^{\frac{1}{q}} \leq \left(\mathbb{E}_\varepsilon \sup_{f\in F} |\hat{U}_f(\varepsilon)|^p\right)^{\frac{1}{p}} \leq \left(\frac{p-1}{q-1}\right) \left(\mathbb{E}_\varepsilon \sup_{f\in F} |\hat{U}_f(\varepsilon)|^q\right)^{\frac{1}{q}}$$

Proof Properties (a), (c), and (d) are directly from Definition 2 of the Rademacher chaos complexity. To prove Property (b), we note, for any $m \in \mathbb{N}$, $f_k \in F$, and $\{\lambda_k : k \in \mathbb{N}_m\}$ satisfying $\sum_k \lambda_k = 1$ and $\lambda_k \geq 0$, that

$$\begin{aligned}\left|\sum_{i,j,i<j} \varepsilon_i\varepsilon_j \sum_{k\in\mathbb{N}_m} \lambda_k f_k(x_i,x_j)\right| &\leq \sum_{k\in\mathbb{N}_m} \lambda_k \left|\sum_{i<j} \varepsilon_i\varepsilon_j f_k(x_i,x_j)\right| \\ &\leq \sup_{f\in F} \left|\sum_{i<j} \varepsilon_i\varepsilon_j f(x_i,x_j)\right|.\end{aligned}$$

Since $\lambda_k, f_k \in F$ are arbitrary, $\hat{\mathcal{U}}_n(\mathrm{conv(F)}) \leq \hat{\mathcal{U}}_n(F)$. The reverse inequality is obvious which completes the proof of the desired Property (b). The last property is from Theorem 3.2.2 of De La Peña and Giné (1999). ■

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

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

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 $\hat{\mathcal{U}}_n(\mathcal{K})$ can be bounded by the metric entropy integral as follows.

Proposition 13 *There exists a universal constant C such that, for any $\mathbf{x} = \{x_i : i \in \mathbb{N}_n\}$, there holds*

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

Proof We rely on Arcones and Giné (1993); De La Peña and Giné (1999) to prove this result. Let $\mathbf{X}_T = \{X_s : s \in T\}$ be a real-valued homogeneous Rademacher chaos process and the pseudo-distance defined by

$$\rho_{\mathbf{x}}(s,t) = (\mathbb{E}_\varepsilon |X_s - X_t|^2)^{\frac{1}{2}}.$$

Then, by Corollary 5.1.8 in De La Peña and Giné (1999) or Proposition 2.6 of Arcones and Giné (1993) we know that there exists a universal constant C such that

$$\mathbb{E}_\varepsilon \sup_{s,t \in T} |X_s - X_t| \leq C \int_0^\infty \log \mathcal{N}(\mathbf{X}_T, \rho_{\mathbf{x}}, \varepsilon) d\varepsilon \tag{20}$$

Applying this result to our context, for $\mathbf{x} = \{x_i \in X : i \in \mathbb{N}_n\}$, let the Rademacher chaos process of order two be defined by

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

Moreover, for any $K, \tilde{K} \in \mathcal{K}$ there holds

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

Hence, for any $\varepsilon > 0$ we have that

$$\mathcal{N}(\mathbf{X}_T, \rho_{\mathbf{x}}, \varepsilon) = \mathcal{N}_2(\mathcal{K} \cup \{0\}, \mathbf{x}, \varepsilon).$$

Consequently, applying equation (20) yields the desired assertion. ■

It is worth mentioning that the standard entropy integral for bounding the suprema 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 refer to Arcones and Giné (1993); De La Peña and Giné (1999) for more general entropy integrals to bound the suprema 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 be further estimated by the shattering dimension of the set of candidate kernels.

Lemma 14 *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

$$d_{\mathbf{x}}(K', K) \leq D_\infty^{\mathbf{x}}(K', K) := \sup_{i,j \in \mathbb{N}_n} |K(x_i, x_j) - K'(x_i, x_j)|.$$

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 the relation (see Chapter 11 of Anthony and Bartlett (1999) and also Lemma 3 of Srebro and Ben-David (2006)) between 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,

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

This completes the proof of the desired assertion. ■

We are in a position to apply Lemma 14 and Proposition 13 to prove Theorem 6.

Proof of Theorem 6: From Lemma 14 and Proposition 13, we have that

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

This completes the assertion. ■

For the set of scalar Gaussian kernels given by equation (13), we have the following estimation.

Lemma 15 *Consider the set of basis kernels $\mathcal{K}_{sc}$ given by equation (13), then $d_{\mathcal{K}_{sc}}=1$.*

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 of pseudo-dimension, there exists $r_1,r_2\in\mathbb{R}$ and $\sigma,\sigma'\in[0,\infty)$ such that

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

and

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

Hence,

$$e^{-\sigma\|x_1-t_1\|^2}>e^{-\sigma'\|x_1-t_1\|^2},\text{ and }e^{-\sigma\|x_2-t_2\|^2}<e^{-\sigma'\|x_2-t_2\|^2}.$$

Equivalently, $\sigma<\sigma'$, and $\sigma>\sigma'$, which is obviously a contradiction. Consequently, the pseudo-dimension of $\mathcal{K}_{\rm SC}$ is identical to one. □

We are ready to prove Corollary 7 with estimation of the Rademacher chaos complexities of $\mathcal{K}_{\rm sc}$ and $\mathcal{K}_{\rm rbf}$.

Proof of Corollary 7: The estimation of $\hat{\mathcal{U}}_n(\mathcal{K}_{\rm sc})$ follows immediately by combining Theorem 6 with Lemma 15. For the RBF kernels set $\mathcal{K}_{\rm rbf}$, note, for any $\{x_i : i \in \mathbb{N}_n\}$, that

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

This completes the assertion. ■

More examples such as Gaussian kernels with covariance matrices are illustrated in Srebro and Ben-David (2006) where these pseudo-dimensions can be directly estimated using the techniques developed in Chapter 11 of Anthony and Bartlett (1999).

6. Error Rates

We are now in a position to derive explicit error rates by trading off the sample error of Rademacher chaos complexity and the regularization error in the error decomposition inequality (5). In subsequent examples we emphasize that the set of basis kernels are given by either equation (13) or the RBF kernels defined by equation (14).

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

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

In addition, we know from Theorem 10 of Chen et al. (2004) 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. ■

Now we turn our attention to general q-norm soft margin SVM losses $\phi(t) = (1-t)_+^q$ for $q \in (1,\infty)$ for classification. In this case the target function f_ρ^ϕ becomes

$$f_\rho^\phi(x) = f_q(x) = \frac{(1+f_\rho(x))^{\frac{1}{q-1}} - (1-f_\rho(x))^{\frac{1}{q-1}}}{(1+f_\rho(x))^{\frac{1}{q-1}} + (1-f_\rho(x))^{\frac{1}{q-1}}},$$

where $f_\rho(x) := P(Y=1|x) - P(Y=-1|x)$.

Example 2 *Let $\phi(t) = (1-t)_+^q$ for some $q \in (1,\infty)$ and suppose that the separation condition holds true with exponent $\theta > 0$. Then, choosing $\lambda = n^{-\frac{q\theta}{4+2(2q+1)\theta}}$ with probability at least $1-\delta$ there holds*

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

Proof First observe that $C_\lambda^\phi \leq (1+\frac{1}{\sqrt{\lambda}})^{q-1}$ and $M_\lambda^\phi \leq (1+\frac{\kappa}{\sqrt{\lambda}})^q$. Hence, from Theorems 3, 6 and Corollary 7 we know, for any $\lambda \in (0,1)$, that

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

Also, we know from Theorem 10 of Chen et al. (2004) 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 $\lambda = n^{-\frac{q(\theta+2)}{2+(2q+1)\theta}}$ yields that

$$\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_q) \leq \mathcal{O}\Big(\Big[\ln n + \ln\frac{2}{\delta}\Big]^{\frac{1}{2}} n^{-\frac{q\theta}{2+(2q+1)\theta}}\Big).$$

Recall the comparison inequality (Theorem 14 of Chen et al. (2004)) for q-norm SVM:

$$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \sqrt{2\Big(\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_q)\Big)}.$$

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

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

which completes the proof of the example. ■

Our last example is the least square loss for classification which is extensively studied in the single kernel case (Caponnetto and De Vito, 2007; De Vito et al., 2006; Smale and Zhou, 2004; Zhang, 2004). In this case, in order to get meaningful rates of the regularization error $\mathcal{D}(\lambda)$ we can assume the target function enjoys some Sobolev smoothness, see e.g Stein (1970) for its precise definition. Recall in the regression case, the target function $f_\rho^\phi = f_\rho(x)$ for any $x \in X$ usually referred to as the *regression function* and the nature of least square loss implies that

$$\mathcal{E}(f_{\mathbf{z}}^\phi) - \mathcal{E}(f_\rho) = \int_X |f_{\mathbf{z}}^\phi(x) - f_\rho(x)|^2 d\rho_X(x).$$

Example 3 *Let X be a domain in $\mathbb{R}^d$ with Lipschitz boundary. Assume the regression function $f_\rho \in H^s(X)$ with some $s > 0$. Then the following holds true.*

1. *If $d/2 < s \leq d/2 + 2$ then for any $0 < \varepsilon < 2s - d$, by taking $\lambda = n^{-\frac{2s-\varepsilon}{2(4s-d-2\varepsilon)}}$, with probability at least $1 - \delta$ there holds*

$$\begin{aligned}\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) &\leq \Big(\int_X |f_{\mathbf{z}}^\phi(x) - f_\rho(x)|^2 d\rho_X(x)\Big)^{\frac{1}{2}} \\ &\leq \mathcal{O}\Big(\Big[\ln n + \ln\tfrac{2}{\delta}\Big]^{\frac{1}{4}} n^{-\frac{2s-d-\varepsilon}{4(4s-d-2\varepsilon)}}\Big).\end{aligned}$$

2. *If X is bounded, ρ_X is the Lebesgue measure, and $0 < s \leq 2$ then by choosing $\lambda = n^{-\frac{2s+d}{2(4s+d)}}$, with probability at least $1 - \delta$, there holds*

$$\begin{aligned}\mathcal{R}(sgn(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) &\leq \Big(\int_X |f_{\mathbf{z}}^\phi - f_\rho|^2 d\rho_X(x)\Big)^{\frac{1}{2}} \\ &\leq \mathcal{O}\Big(\Big[\ln n + \ln\tfrac{2}{\delta}\Big]^{\frac{1}{4}} n^{-\frac{s}{2(4s+d)}}\Big).\end{aligned}$$

Proof For the least square loss, we observe that $C_\lambda^\phi = 2(1 + \frac{1}{\sqrt{\lambda}})$ and $M_\lambda^\phi \leq (1 + \frac{\kappa}{\sqrt{\lambda}})^2$. Then, we know from Theorem 3, Theorem 6 and Corollary 7 that

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}^\phi) - \mathcal{E}(f_\rho) &= \int_X |f_{\mathbf{z}}^\phi(x) - f_\rho(x)|^2 d\rho_X(x) \\ &\leq \mathcal{O}\Big(\Big(\tfrac{\ln n}{n\lambda^2}\Big)^{\frac{1}{2}} + \Big(\tfrac{\ln\frac{2}{\delta}}{n\lambda^2}\Big)^{\frac{1}{2}} + \tfrac{1}{\sqrt{n}}\Big) + \mathcal{D}(\lambda).\end{aligned} \tag{21}$$

Then, for the first assertion we know from Proposition 22 of Ying and Zhou (2007) that

$$\mathcal{D}(\lambda) \leq \mathcal{O}\Big(\lambda^{\frac{2s-\varepsilon-d}{2s-\varepsilon}}\Big).$$

Putting the above two equations together and letting $\lambda = n^{-\frac{2s-\varepsilon}{2(4s-2\varepsilon-d)}}$ implies that

$$\int_X |f_{\mathbf{z}}^\phi(x) - f_\rho(x)|^2 d\rho_X(x) \leq \mathcal{O}\Big(\Big[\ln n + \ln\frac{2}{\delta}\Big]^{\frac{1}{2}} n^{-\frac{2s-d-\varepsilon}{2(4s-d-2\varepsilon)}}\Big).$$

Hence, the desired result follows from the comparison inequality (Chen et al., 2004; Bartlett et al., 2006; Zhang, 2004) for the least square loss:

$$\mathcal{R}(\mathrm{sign}(f_{\mathbf{z}}^\phi)) - \mathcal{R}(f_c) \leq \sqrt{2\Big(\mathcal{E}^\phi(f_{\mathbf{z}}^\phi) - \mathcal{E}^\phi(f_\rho)\Big)}. \tag{22}$$

The proof of the second assertion is similar as above. Recall that Proposition 22 of Ying and Zhou (2007) implies that the regularization error is estimated as follows:

$$\mathcal{D}(\lambda) \leq \mathcal{O}\Big(\lambda^{\frac{2s}{2s+d}}\Big).$$

Combining this with inequality (21) and the comparison inequality (22), with choice $\lambda = n^{-\frac{2s+d}{2(4s+d)}}$ we get the desired second assertion. ■

We end this section with a comparison with error rates in Ying and Zhou (2007) on the least square loss for classification. In Example 1 there, it was proven that: if $d/2 < s \leq d/2 + 2$ then for any $0 < \varepsilon < 2s - d$, we have that

$$\begin{aligned}\mathbb{E}\Big[\int_X |f_{\mathbf{z}}^{\phi}(x) - f_\rho(x)|^2 d\rho_X(x)\Big]^{\frac{1}{2}} &\leq \Big(\mathbb{E}\Big[\int_X |f_{\mathbf{z}}^{\phi}(x) - f_\rho(x)|^2 d\rho_X(x)\Big]\Big)^{\frac{1}{2}} \\ &\leq \mathcal{O}\Big((\ln n)^{\frac{1}{4}}\ n^{-\frac{2s-d-\varepsilon}{8(4s-d-2\varepsilon)}}\Big).\end{aligned}$$

Ignoring the difference of the forms to express error rates using expectations and probabilistic inequalities, Example 3 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 $\mathcal{O}\Big((\ln n)^{\frac{1}{4}} n^{-\frac{s}{2(4s+d)}}\Big)$ in comparison with $\mathcal{O}\Big((\ln n)^{\frac{1}{4}} n^{-\frac{s}{4(4s+d)}}\Big)$ obtained previously. Hence, our new error rates substantially improve those in Ying and Zhou (2007).

7. Conclusion

In this paper we provided a novel statistical generalization bound for kernel learning algorithms which extends and improves previous work in the literature (Lanckriet et al., 2004; Micchelli et al., 2005; Srebro and Ben-David, 2006; Wu et al., 2006; Ying and Zhou, 2007). 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 Bartlett et al. (2006); Steinwart and Scovel (2005); Wu et al. (2006). For this purpose, the extension of localized Rademacher averages (Bartlett et al., 2005) to the scenario of multiple kernel learning would be useful.

- Secondly, it would be interesting to investigate generalization bounds based on decoupling Gaussian chaos of order two, see its definition in De La Peña and Giné (1999).

- 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.

Acknowledgments

This work is supported by EPSRC grant EP/E027296/1.

References

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.

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

M. A. Arcones and E. Giné. Limit theorems for U-processes, *The Annals of Probability*, **21**: 1494–1542, 1993.

A. Argyriou, R. Hauser, C. Micchelli, and M. Pontil. A DC-programming algorithm for kernel selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

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

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.

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

O. Bousquet and D.J.L. Herrmann. On the complexity of learning the kernel matrix. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

A. Caponnetto and E. De Vito. Optimal Rates for Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, **7**: 331-368, 2007.

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.

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

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

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.

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

M. Girolami and S. Rogers. Hierarchic Bayesian models for kernel learning, In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

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

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

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.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Press, New York, 1991.

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

C. McDiarmid. On the method of bounded differences. In Surveys in Combinatorics 1989, pages 148–188. Cambridge University Press, 1989.

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.

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

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.

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

I.J. Schoenberg. Metric spaces and completely monotone functions, *Ann. of Math.* 39: 811-841, 1938.

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

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, **41**: 279–305, 2004.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of 19th Annual Conference on Learning Theory (COLT)*, 2006.

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

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *J. of Machine Learning Research*, **7**: 1531–1565, 2006.

I. Steinwart and C. Scovel. Fast rates for support vector machines. In *Proceedings of 18th Annual Conference on Learning Theory (COLT)*, 2005.

E. M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, Princeton, New Jersey, 1970.

V. Vapnik. *Statistical Learning Theory*, John Wiley & Sons, 1998.

Q. Wu, Y. Ying, and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 2006.

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

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

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