Journal of Machine Learning Research 8 (2007) 249-276 Submitted 6/06; Revised 10/06; Published 2/07

Learnability of Gaussians with Flexible Variances

Yiming Ying Y.YING@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK

Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: Peter L. Bartlett

Abstract

Gaussian kernels with flexible variances provide a rich family of Mercer kernels for learning algorithms. We show that the union of the unit balls of reproducing kernel Hilbert spaces generated by Gaussian kernels with flexible variances is a uniform Glivenko-Cantelli (uGC) class. This result confirms a conjecture concerning learnability of Gaussian kernels and verifies the uniform convergence of many learning algorithms involving Gaussians with changing variances. Rademacher averages and empirical covering numbers are used to estimate sample errors of multi-kernel regularization schemes associated with general loss functions. It is then shown that the regularization error associated with the least square loss and the Gaussian kernels can be greatly improved when flexible variances are allowed. Finally, for regularization schemes generated by Gaussian kernels with flexible variances we present explicit learning rates for regression with least square loss and classification with hinge loss.

Keywords: Gaussian kernel, flexible variances, learning theory, Glivenko-Cantelli class, regularization scheme, empirical covering number

1. Introduction

Let X be a metric space in $\mathbb{R}^n$ and $Y \subseteq \mathbb{R}$. In learning theory, we are interested in the approximation of functions or function relations from samples. The functions are from an input space X to an output space Y, and samples are drawn according to a Borel probability measure ρ on the space $Z := X \times Y$. The target function f_ρ^V that we want to learn or approximate is a minimizer (may not be unique) of some error functional $\mathcal{E}(f) = \int_Z V(y, f(x)) d\rho$ induced by a loss function $V : Y \times Y \to \mathbb{R}_+$, that is,

$$f_\rho^V = \arg\min \Big\{ \mathcal{E}(f) : \ f \text{ is a measurable function from } X \text{ to } Y \Big\}. \tag{1}$$

To define f_ρ^V, we denote ρ_X as the marginal distribution of ρ on X and $d\rho(\cdot|x)$ the conditional distribution. Then the error can be written as $\mathcal{E}(f) = \int_X \int_Y V(y, f(x)) d\rho(y|x) d\rho_X(x)$, and we choose f_ρ^V to be a minimizer of the pointwise error: for almost every $x \in X$,

$$f_\rho^V(x) = \arg\min_{t \in Y} \int_Y V(y,t) d\rho(y|x).$$

The idea of many learning algorithms is to approximate f_ρ^V or its generalizing ability (measured by the error in terms of the loss V) by minimizing the empirical error $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i))$, or a penalized version, where $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a set of samples drawn independently according to ρ. The law of large numbers tells us that $\mathcal{E}_{\mathbf{z}}(f) \to \mathcal{E}(f)$ in probability for a fixed function f. This leads to the natural expectation that a minimizer $f_{\mathbf{z}}$ of $\mathcal{E}_{\mathbf{z}}$ over a set of functions $\mathcal{G}$, called the hypothesis space, would approximate a minimizer $f_{\mathcal{G}}$ of $\mathcal{E}$ in $\mathcal{G}$ (whose error is close to that of f_ρ^V when $\mathcal{G}$ is large): $\mathcal{E}(f_{\mathbf{z}}) \to \mathcal{E}(f_{\mathcal{G}})$ as $m \to \infty$. This approximation behaves well when the hypothesis space $\mathcal{G}$ enjoys the uniform convergence property (see Vapnik, 1998; Alon et al., 1997).

Throughout the paper we choose Y to be a subset of $\mathbb{R}$, the loss function V has an extended domain $V : Y \times \mathbb{R} \to \mathbb{R}_+$, and $f_\rho^V(x)$ is defined to be a minimizer of $\min_{t\in\mathbb{R}} \int_Y V(y,t)d\rho(y|x)$. In particular, for the binary classification problem (Devroye et al., 1997), we take $Y = \{1, -1\}$ and $V(y,t) = \phi(yt)$ with $\phi : \mathbb{R} \to \mathbb{R}_+$. For the hinge loss $\phi(yt) = \max\{1 - yt, 0\}$ used in the support vector machine for classification (Cortes and Vapnik, 1995), the target function $f_\rho^V(x) = f_c(x)$ is called the *Bayes rule* defined as $f_c(x) = 1$ for $\rho(y = 1|x) \geq \rho(y = -1|x)$, and $f_c(x) = -1$ otherwise. In this case, the uniform convergence can be characterized by the finiteness of the VC-dimension of $\mathcal{G}$ (see, e.g., Vapnik, 1998).

For the regression problem, we choose $Y = \mathbb{R}$ and $V(y,t) = \psi(y-t)$ with $\psi : \mathbb{R} \to \mathbb{R}_+$. In particular, for the least square loss $\psi(y-t) = (y-t)^2$, $f_\rho^V(x) = f_\rho(x) = \int_Y y d\rho(y|x)$ is the regression function (induced by conditional means). When Y is a closed interval on $\mathbb{R}$, the uniform convergence of real-valued function space $\mathcal{G}$ can be characterized by the property: for every $\varepsilon > 0$, there holds

$$\lim_{\ell\to+\infty} \sup_{\mu} \Pr\left\{ \sup_{m\geq\ell} \sup_{f\in\mathcal{G}} \left| \frac{1}{m}\sum_{i=1}^m f(x_i) - \int_X f(x)d\mu \right| > \varepsilon \right\} = 0. \tag{2}$$

Here Pr denotes the probability with respect to the samples $x_1, x_2, \ldots$, independently drawn according to a Borel probability distribution μ on X. The supremum is taken with respect to all such probability distributions. Following Dudley et al. (1991), we say that $\mathcal{G}$ is *uniform Glivenko-Cantelli* (uGC) if it satisfies the equality (2) for any $\varepsilon > 0$, which is equivalent to the finiteness of V_γ-dimension for any $\gamma > 0$ (Alon et al., 1997), see Section 4.

In this paper, we restrict our attention to the uniform convergence of kernel-based learning algorithms. A function $K : X \times X \to \mathbb{R}$ is called a *reproducing kernel* if it is symmetric and positive semidefinite, that is, for any finite set of distinct points $\{x_1, \cdots, x_\ell\} \subset \Omega$, the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is positive semidefinite. If moreover, K is continuous, then we call such a reproducing kernel a *Mercer kernel*. The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_x = K(x,\cdot) : x \in X\}$ with the reproducing property (Aronszajn, 1950)

$$f(x) = \langle f, K_x\rangle_K, \qquad \forall x \in X, f \in \mathcal{H}_K. \tag{3}$$

Learning algorithms considered here can be stated as minimization problems in $\mathcal{H}_K$. The reproducing property (3) makes the minimization over $\mathcal{H}_K$ be realized by an optimization procedure in $\mathbb{R}^m$. Consider the Tikhonov regularization scheme (Evgeniou et al., 2000) associated with the kernel K, loss V and a regularization parameter $\lambda > 0$ defined as

$$\widetilde{f}_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)) + \lambda \|f\|_K^2 \right\}.$$

The Representer Theorem (Schölkopf et al., 2001; Wahba, 1990) tells us $\widetilde{f}_{\mathbf{z},\lambda} = \sum_{i=1}^m c_i K_{x_i}$ with $(c_i)_{i=1}^m \in \mathbb{R}^m$. When V is convex with respect to the second variable, this leads to a convex optimization problem in $\mathbb{R}^m$. In particular, for the least square loss or the hinge loss, it is a convex quadratic programming optimization problem which can handle large data settings. For discussions on the error rates of these schemes, see, for example, Vapnik (1998); Zhang (2004); Steinwart and Scovel (2005). This paper aims at Tikhonov regularization schemes associated with a set of kernels.

1.1 Multi-kernel Regularization Schemes

Multi-kernel regularization schemes have attracted attention recently due to applications in multi-task learning (Evgeniou and Pontil, 2004; Lanckriet et al., 2004), mixture density estimation (Li and Barron, 1999; Rakhlin et al., 2005), multi-kernel regularized classifiers (Chapelle et al., 2002; Cristianini et al., 1998; Wu et al., 2007), and many others. These schemes involve a set of Mercer kernels $\{K^\sigma\}_{\sigma\in\Sigma}$ with an index set Σ and take the following form with the regularization parameter $\lambda > 0$

$$f_{\mathbf{z},\lambda} := \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \left\{ \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)) + \lambda \|f\|_{K^\sigma}^2 \right\}, \tag{4}$$

where the loss function $V(y,\cdot)$ is usually convex for any $y \in Y$.

Throughout the paper we assume the existence of a solution to the optimization problem (4). This assumption is satisfied when Σ is a compact metric space and $K^\sigma(x,y)$ is continuous with respect to $\sigma \in \Sigma$ for each fixed pair $(x,y) \in X \times X$. See Wu et al. (2007). When $\Sigma = (0,b]$ (a noncompact index set) with $0 < b < \infty$, each K^σ is a Gaussian kernel and V is the least-square loss, the existence will be verified in Appendix B under some conditions on the sample $\mathbf{z}$.

When we consider the convergence of $\mathcal{E}(f_{\mathbf{z},\lambda})$ to $\mathcal{E}(f_\rho^V)$, we need to find the hypothesis space for the uniform convergence here. Since $f_{\mathbf{z},\lambda}$ is the minimizer in (4), with a special choice $f = 0$ we see that whenever $f_{\mathbf{z},\lambda} \in \mathcal{H}_{K^\sigma}$, the quantity $\lambda \|f_{\mathbf{z},\lambda}\|_{K^\sigma}^2$ is bounded by

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda \|f_{\mathbf{z},\lambda}\|_{K^\sigma}^2 \le \mathcal{E}_{\mathbf{z}}(0) + 0 \le \frac{1}{m}\sum_{i=1}^m V(y_i,0) \le \|V(y,0)\|_{L_\rho^\infty(Z)}.$$

When ρ satisfies $\|V(y,0)\|_{L_\rho^\infty(Z)} \le M < \infty$ (a strong assumption for regression problems, excluding Gaussian noise), we have $\|f_{\mathbf{z},\lambda}\|_{K^\sigma} \le \sqrt{M/\lambda}$ for all $\mathbf{z} \in Z^m$. This leads to the following hypothesis set.

Definition 1 *The normalized hypothesis set $\mathcal{H}$ associated with Mercer kernels $\{K^\sigma\}_{\sigma\in\Sigma}$ is defined as*

$$\mathcal{H} = \cup_{\sigma\in\Sigma}\{f \in \mathcal{H}_{K^\sigma} : \|f\|_{K^\sigma} \le 1\}. \tag{5}$$

The above analysis tells us that if $\|V(y,0)\|_{L_\rho^\infty(Z)} \le M$, then $f_{\mathbf{z},\lambda} \in \sqrt{\frac{M}{\lambda}}\,\mathcal{H} = \left\{\sqrt{\frac{M}{\lambda}} f : \ f \in \mathcal{H}\right\}$ for almost every $\mathbf{z} \in Z^m$. For the study of uniform convergence and error analysis on the multi-kernel scheme (4), a basic question is whether $\mathcal{H}$ is uGC. This is the main question investigated in this paper. The reproducing property of RKHS plays an essential role in our subsequent investigations.

The first purpose of this paper is to show that the uGC property of $\mathcal{H}$ is equivalent to that of a smaller set consisting of fundamental functions from the RKHS.

Theorem 2 *Let $\{K^\sigma\}_{\sigma\in\Sigma}$ be a set of Mercer kernels on X with*

$$\kappa := \sup_{\sigma\in\Sigma}\sup_{x\in X}\sqrt{K^\sigma(x,x)} < \infty. \tag{6}$$

Then the set $\mathcal{H}$ defined by (5) is uGC if and only if the normalized fundamental set

$$\mathcal{F} = \mathcal{F}_\Sigma = \{K_x^\sigma : \sigma \in \Sigma, x \in X\} \tag{7}$$

is uGC.

Theorem 2 will be proved in Section 3. The reproducing property (3) tells us that $\|K_x^\sigma\|_{K^\sigma} = \sqrt{K^\sigma(x,x)} \le \kappa$ for any $x \in X$ and $\sigma \in \Sigma$. Therefore, the normalized condition (6) on the kernels $\{K_\sigma : \sigma \in \Sigma\}$ is essential for $\frac{1}{\kappa}\mathcal{F} \subseteq \mathcal{H}$ (a scaling). Since $\mathcal{F}$ contains much less functions than $\mathcal{H}$, checking the uGC property for $\mathcal{F}$ is potentially much simpler than that for $\mathcal{H}$. We shall use this idea to establish the learnability of Gaussian kernels with flexible variances.

1.2 uGC Property for Gaussians with Flexible Variances

The second purpose of this paper is to verify the learnability of Gaussian kernels with flexible variances stated in the form of the following uGC property. The theorem will be proved in Section 4.

Theorem 3 *Let $n \in \mathbb{N}$ and X be any subset of $\mathbb{R}^n$. Let*

$$K^\sigma(x,y) = \exp\left\{-\sum_{i=1}^n \frac{(x_i-y_i)^2}{\sigma_i^2}\right\} \quad \text{for } \sigma = (\sigma_1,\ldots,\sigma_n) \in (0,+\infty)^n. \tag{8}$$

Define $\mathcal{H}$ by (5) with $\Sigma = (0,+\infty)^n$. Then $\mathcal{H}$ is uGC.

Note that each kernel in (8) is, in a way, normalized: its $C(X)$ norm is 1, ensuring the kernel to be uniformly bounded by 1 and (6) to be satisfied with $\kappa = 1$.

When X is compact and the index set is restricted to be $[a,+\infty)^n$ with $a > 0$, we know from Theorem 3 in Zhou (2003) that for any $s \in \mathbb{N}$, the set $\mathcal{H}$ defined by (5) is included in a ball of $C^s(X)$ with a finite radius. Hence, the closure of $\mathcal{H}$ in $C(X)$ is compact. This belongs to the well-studied case (Cucker and Smale, 2001; Cucker and Zhou, 2007) that the closure of the hypothesis space $\mathcal{H}$ is compact, and thereby satisfies the uniform convergence condition.

When $a = 0$ and the index set becomes $(0,\infty)^n$, the closure is not compact any more. This observation led the second author to raise the following open problem in Zhou (2003): if we denote

$$K_\sigma(x,y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}, \qquad \text{with } \sigma > 0, \tag{9}$$

is the function set

$$\mathcal{H}_0 = \cup_{\sigma>0}\left\{ f(x) = \sum_{i=1}^\ell c_i K_\sigma(x_i,x) : \ \sum_{i,j=1}^\ell c_i K_\sigma(x_i,x_j)c_j \le 1, \ x_i \in [0,1]^n, \text{ and } \ell \in \mathbb{N} \right\} \tag{10}$$

involving the Gaussian kernels (9) with isotropic variances uGC? Using Theorem 3, we know that the answer to the question is positive.

Theorem 4 *Let $X = [0,1]^n$ and K_σ be given by (9). Define $\mathcal{H}_0$ by (10). Then $\mathcal{H}_0$ is uGC. It is contained in the unit ball of $C(X)$, but its closure in $C(X)$ is not compact.*

Proof The first statement follows from Theorem 3 and the fact that any subset of a uGC set is uGC.

Since each function f from $\mathcal{H}_0$ satisfies $\|f\|_{C(X)} \leq 1$, this set is contained in the unit ball of $C(X)$.

To see the last statement, we apply the Arzelá-Ascoli Theorem (see Yosida, 1980, P.85) which asserts that a subset of $C(X)$ has compact closure if and only if it is bounded and equicontinuous. Take $f_\sigma(x) = K_\sigma(x,0) = \exp\{-\frac{|x|^2}{\sigma^2}\} \in \mathcal{H}_0$ with $\sigma \in (0,\infty)$. For the neighborhood $[0,\varepsilon)^n$ of 0 with $\varepsilon > 0$, we see that

$$\sup_{\sigma\in(0,\infty)} \sup_{t\in[0,\varepsilon)^n} |f_\sigma(t) - f_\sigma(0)| = \sup_{\sigma\in(0,\infty)} \left|\exp\{-\frac{n\varepsilon^2}{\sigma^2}\} - 1\right| = 1 \not\to 0.$$

Therefore, the set of functions $\mathcal{H}_0$ is not equicontinuous. By the Arzelá-Ascoli Theorem, the closure of $\mathcal{H}_0$ is not compact. ■

The rest of the paper is organized as follows. In the next section we show the implications of the above main theorems to the error analysis for the regularization scheme (4). In particular, we present error rates of two examples involving Gaussian kernels with flexible variances. Sections 3 and 4 are distributed to prove Theorem 2 and 3 respectively. In Sections 5 and 6 we develop error bounds for the multi-kernel regularization scheme (4), especially for Gassian kernels with flexible variances. This is the last purpose of this paper. We postpone the derivation of error rates for regression with least square loss and classification with hinge loss to the end of this paper.

2. Applications to Error Analysis: Two Examples

The theorems in Section 1 give us qualitative results on the learnability of multi-kernel scheme (4). They can be deepened, which yields quantitative error rates for $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^V)$. In particular, we expect to derive explicit rates for multi-kernel regularized learning algorithms associated with Gaussian kernels with flexible variances.

To see how $\mathcal{E}(f_{\mathbf{z},\lambda})$ approximate $\mathcal{E}(f_\rho^V)$, assume $f_{\mathbf{z},\lambda} \in \mathcal{H}_{K^\sigma}$ for some $\sigma \in \Sigma$ and choose some $\sigma' \in \Sigma$ and $f_\lambda \in \mathcal{H}_{K^{\sigma'}}$ called a *regularizing function* (Smale and Zhou, 2004). Write

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^V) &= \left\{\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\}\right\} - \lambda\|f_{\mathbf{z},\lambda}\|_{K^\sigma}^2 \\ &+ \left\{(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_{K^\sigma}^2) - (\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_{K^{\sigma'}}^2)\right\} \\ &+ \left\{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^V) + \lambda\|f_\lambda\|_{K^{\sigma'}}^2\right\}.\end{aligned}$$

The definition (4) tells us that the middle term above is at most 0. The first term $\left\{\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\}\right\}$ is called the *sample error*. Its second part $\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda) = \frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)$ involving a single random variable $\xi = V(y, f(x))$ on Z can be easily estimated. The last term called the *regularization error* is independent of the samples (Niyogi and Girosi,

1996; Cucker and Smale, 2001; Smale and Zhou, 2003) and measures the approximation ability of the multi-kernel space $\cup_{\sigma\in\Sigma}\mathcal{H}_{K^\sigma}$.

Definition 5 *The regularization error associated with the regularizing function $f_\lambda \in \mathcal{H}_{K^\sigma}$ with $\sigma \in \Sigma$ is defined as*

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho^V) + \lambda \|f_\lambda\|_{K^\sigma}^2.$$

The regularization error of the system (4) is

$$\widetilde{\mathcal{D}}(\lambda) = \min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \{\mathcal{E}(f) - \mathcal{E}(f_\rho^V) + \lambda \|f\|_{K^\sigma}^2\}, \tag{11}$$

where f_λ takes the special form

$$f_\lambda = f_\lambda^V := \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \{\mathcal{E}(f) + \lambda \|f\|_{K^\sigma}^2\}. \tag{12}$$

Thus we have the *error decomposition*:

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^V) \le \left\{ \{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} \right\} + \mathcal{D}(\lambda). \tag{13}$$

What is left for the error analysis is the term $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})$. It involves a set of random variables $\xi_{\mathbf{z}} = V(y, f_{\mathbf{z},\lambda}(x))$ with $\mathbf{z} \in Z^m$. That is also the motivation to study the uniform convergence of $\mathcal{H}$ (Vapnik, 1998).

By the error decomposition (13), the learning rate depends on trading off the sample error and the regularization error. The decay of regularization error $\widetilde{\mathcal{D}}(\lambda)$ relies on the regularity (smoothness) of the target function f_ρ^V for most commonly used loss functions which will be discussed in Section 6. Sample error estimates depend on the capacity of the hypothesis space $\sqrt{M/\lambda}\,\mathcal{H}$ which can be studied (see, e.g., Koltchinskii and Panchenko, 2002) by means of its covering numbers and Rademacher complexity (see Section 5). However, it is not easy to estimate the Rademacher complexity of $\sqrt{M/\lambda}\,\mathcal{H}$. In Section 5 we provide an alternative way to estimate the sample error by computing the Rademacher complexity of the fundamental set $\mathcal{F}$.

Let us give two examples both involving the Gaussian kernels (8) with flexible variances to illustrate the learning rates whose proofs will be given in Section 6.

The first example is regularized regression with the least square loss $V(y,t) = (y-t)^2$. Here $Y = \mathbb{R}$. Then the *multi-kernel least square regularized regression algorithm* (4) associated with Gaussian kernels (8) can be written as

$$f_{\mathbf{z},\lambda} = \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \left\{ \frac{1}{m}\sum_{i=1}^m (y_i - f(x_i))^2 + \lambda \|f\|_{K^\sigma}^2 \right\}. \tag{14}$$

By the special feature of the least square loss, the distance between $f_{\mathbf{z},\lambda}$ and the target function f_ρ is often measured by the weighted L^2 metric in $L^2_{\rho_X}$ defined as $\|f\|_{L^2_{\rho_X}} = \left(\int_X |f(x)|^2 d\rho_X\right)^{1/2}$. When ρ_X is the Lebesgue measure, we denote the metric as $\|f\|_{L^2(X)}$ in Example 1. Also, denote $H^s(X)$ to be the Sobolev space (e.g., Stein, 1970) with index $s > 0$ on X.

For this multi-kernel algorithm, we have the following learning rates achieved by special choices of the regularization parameter $\lambda = \lambda(m)$.

Example 1 *Let $X \subseteq \mathbb{R}^n$ be a domain with Lipschitz boundary. Define $f_{\mathbf{z},\lambda}$ by (14) with the Gaussians (8). Assume $f_\rho \in H^s(X)$ for $s > 0$ and $|y| \le M_0$ almost surely.*

(1) If $n/2 < s \le n/2+2$ then for any $0 < \varepsilon < 2s-n$, we have

$$\mathbb{E}\Big[\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}(X)}\Big] = O\Big((\log m)^{\frac{1}{2}} m^{-\frac{2s-n-\varepsilon}{4(4s-n-2\varepsilon)}}\Big) \quad \text{by taking } \lambda = m^{-\frac{2s-\varepsilon}{4(4s-n-2\varepsilon)}}.$$

(2) If X is bounded, ρ_X is the Lebesgue measure, and $s \le 2$ then by choosing $\lambda = m^{-\frac{2s+n}{4(4s+n)}}$, we get

$$\mathbb{E}\Big[\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2(X)}\Big] = O\Big((\log m)^{\frac{1}{2}} m^{-\frac{s}{2(4s+n)}}\Big).$$

In the above example we are considering the function approximation (De Vito et al., 2005; Smale and Zhou, 2005) on a domain of $\mathbb{R}^n$, so the learning rate is poor if the dimension n is large. However, in many situations, the input space X is a low-dimensional manifold embedded in the large-dimensional space $\mathbb{R}^n$. In such a situation, the learning rates may be greatly improved. This will not be discussed in this paper because the discussion involves the function approximation on Riemannian manifolds (see, e.g., Ye and Zhou, 2007), which is out of our scope here.

The second example is regularized classification with the hinge loss $V(y,t) = (1-yt)_+ := \max\{1-yt, 0\}$. Here $Y = \{1,-1\}$ and we are interested in functions $\mathcal{C} : X \to Y$ called binary classifiers which divide X into two classes. The target function is the Bayes rule f_c.

The *multi-kernel SVM regularized classification algorithm* (4) associated with Gaussian kernels (8) is defined to be a minimizer of the following optimization problem

$$f_{\mathbf{z},\lambda} = \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \Big\{\frac{1}{m}\sum_{i=1}^m (1-y_i f(x_i))_+ + \lambda\|f\|^2_{K^\sigma}\Big\}. \tag{15}$$

Then the sign function $\operatorname{sgn}(f_{\mathbf{z},\lambda})$ is used as a classifier where $\operatorname{sgn}(f)(x) = 1$ for $f(x) \ge 0$ and $\operatorname{sgn}(f)(x) = -1$ otherwise.

The prediction power of classifiers is measured by the misclassification error. The *misclassification error* for a classifier $\mathcal{C} : X \to Y$ is defined to be

$$\mathcal{R}(\mathcal{C}) := \Pr\{\mathcal{C}(x) \ne y\} = \int_X P(y \ne \mathcal{C}(x)|x) d\rho_X. \tag{16}$$

The Bayes rule is the classifier which minimizes the misclassification error.

The error analysis of classification algorithms often aims at understanding the approximating behaviors of the *excess misclassification error*

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)$$

as the sample size m becomes large. Our learning rate for the hinge loss assumes a separable condition which was introduced by Chen et al. (2004) as follows.

Definition 6 *We say that ρ is separable by $\mathcal{H}_\Sigma$ if there is some $f_{sp} \in \mathcal{H}_{K^\sigma}$ with some $\sigma \in \Sigma$ 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 $\|f_{sp}\|_{K^\sigma} = 1$ and*

$$\rho_X\big\{x \in X : |f_{sp}(x)| < \Delta t\big\} \le c_\theta t^\theta, \qquad \forall t > 0. \tag{17}$$

Observe that condition (17) with $\theta = \infty$ is equivalent to

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

That is, $|f_{sp}(x)| \geq \gamma$ almost everywhere. Thus, separable distributions with separation exponent $\theta = \infty$ are exactly strictly separable distributions. Recall (Shawe-Taylor et al., 1998; Vapnik, 1998) that ρ is said to be *strictly separable* with margin $\gamma > 0$ if ρ is separable together with the requirement $y f_{sp}(x) \geq \gamma$ almost everywhere. The separation condition is different from the Tsybakov's noise condition (Tsybakov, 2004): the former involves a function f_{sp} from $\mathcal{H}_{K^\sigma}$ and describes the approximation of the RKHS (Chen et al., 2004), while the latter is about the distribution ρ only.

The learning rates for the multi-kernel SVM algorithm (15) can be stated as follows.

Example 2 *Let $f_{\mathbf{z},\lambda}$ be defined by (15) with $\{K^\sigma\}$ given by (8). If ρ is separable by $\mathcal{H}_\Sigma$ with some separation exponent $\theta > 0$, then by choosing $\lambda = m^{-\frac{2+\theta}{2(2+3\theta)}}$, we have*

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

It is observed in applications that allowing flexible variances improves the learnability of Gaussian kernels. As we shall see in Section 6 for the least square loss, when the regression function f_ρ has Sobolev smoothness, the regularization error (11) associated with Gaussians with flexible variances decays as $O(\lambda^s)$ for some $s > 0$. This has also been confirmed theoretically by Steinwart and Scovel (2005) for classification with the hinge loss, under some geometric noise condition for the distribution. Such a decay is impossible for the regularization error associated with a single Gaussian kernel, at least when ρ_X is the Lebesgue measure on X, as shown by Smale and Zhou (2003). This demonstrates that learning algorithms using Gaussian kernels with flexible variances have advantages for many applications.

Another way to obtain improved error rates is to take kernels changing with the sample size. Kernels of this kind include polynomial kernels with changing degrees (Zhou and Jetter, 2006) and Gaussian kernels with changing variances (Steinwart, 2001; Steinwart and Scovel, 2005). Though the learning rates given by Steinwart and Scovel (2005) is comparable to those in Example 2, the main difficulty there is a requirement of some information about the distribution ρ for the choice of the kernel (similar to the choice of the regularization parameter according to some regularity properties of ρ). Compared to that, when no information about ρ is available, the algorithm (4) still produces the empirically optimal kernel from the kernel set, as part of the optimization problem.

3. Reducing the Hypothesis Set

In this section we show how the uGC property of the normalized hypothesis set can be reduced to that of the normalized fundamental set, hence establish Theorem 2. For $\sigma \in \Sigma$, denote $\mathcal{H}_\sigma = \mathcal{H}_{K^\sigma}, \langle\cdot,\cdot\rangle_\sigma = \langle\cdot,\cdot\rangle_{K^\sigma}$, and $\kappa^\sigma = \sup_{x\in X}\sqrt{K^\sigma(x,x)}$. Assume $\kappa^\sigma < \infty$.

Let μ be a Borel probability distribution on X. For $\{x_i\}_{i=1}^m$ drawn according to μ, we denote $\mathbb{E}_m(f) = \frac{1}{m}\sum_{i=1}^m f(x_i)$ and $\mathbb{E}(f) = \int_X f(x)d\mu$.

To prove Theorem 1, we need the following proposition about the function

$$F^\sigma(x) := \int_X K^\sigma(x,y)d\mu(y), \qquad x \in X. \tag{18}$$

Proposition 7 *Let K^σ be a Mercer kernel on X and μ be a Borel probability distribution. Define the function F^σ by (18). Then we have*

(a) $F^\sigma \in \mathcal{H}_\sigma$;

(b) $\langle f, F^\sigma\rangle_\sigma = \int_X f(y)d\mu(y)$ for every $f \in \mathcal{H}_\sigma$.

Proof Define a linear functional T_σ on $\mathcal{H}_\sigma$ as

$$T_\sigma(f) = \int_X f(y)d\mu(y).$$

Since μ is a Borel probability measure on X, we know by the reproducing property (3)

$$|T_\sigma(f)| \le \int_X |\langle f, K_y^\sigma\rangle_\sigma| d\mu(y) \le \kappa^\sigma \|f\|_\sigma, \qquad \forall f \in \mathcal{H}_\sigma.$$

This means T_σ is a bounded linear functional on the Hilbert space $\mathcal{H}_\sigma$. By the Riesz Representation Theorem of Hilbert spaces, we know that there exists a function $g^\sigma \in \mathcal{H}_\sigma$ such that

$$T_\sigma(f) = \langle f, g^\sigma\rangle_\sigma, \qquad \forall f \in \mathcal{H}_\sigma. \tag{19}$$

In particular, for the function $f = K_x^\sigma$ lying in $\mathcal{H}_\sigma$ with x being an arbitrarily fixed point in X, there holds

$$T_\sigma(K_x^\sigma) = \int_X K_x^\sigma(y)d\mu(y) = \int_X K^\sigma(x,y)d\mu(y) = \langle K_x^\sigma, g^\sigma\rangle_\sigma = g^\sigma(x).$$

This equals to $F^\sigma(x)$ according to the definition (18). Hence $F^\sigma(x) = g^\sigma(x)$ for every $x \in X$. Therefore $F^\sigma = g^\sigma \in \mathcal{H}_\sigma$ which proves property (a). Property (b) is an immediate consequence of equality (19) since $g^\sigma = F^\sigma$. ■

Property (a) of Proposition 7 means that the integral and the inner product with K_y^σ can be interchanged:

$$\langle f, \int_X K^\sigma(\cdot,y)d\mu(y)\rangle_\sigma = \int_X f(y)d\mu(y) = \int_X \langle f, K_y^\sigma\rangle_\sigma d\mu(y).$$

Lemma 8 *Let K^σ be a Mercer kernel on X and μ be a Borel probability distribution. Denote $G^\sigma(x) = \frac{1}{m}\sum_{i=1}^m K^\sigma(x_i,x) - F^\sigma(x)$ for $m \in \mathbb{N}$ and $\mathbf{x} = (x_i)_{i=1}^m \in X^m$. Then, the following statements are true.*

(a) $G^\sigma \in \mathcal{H}_\sigma$ and $\sup_{f\in\mathcal{H}_\sigma, \|f\|_\sigma\le 1} |\mathbb{E}_m(f) - \mathbb{E}(f)| = \|G^\sigma\|_\sigma$.

(b) $\|G^\sigma\|_\sigma \le \sqrt{2\|G^\sigma\|_{C(X)}}$ and $\|G^\sigma\|_{C(X)} \le \kappa^\sigma \|G^\sigma\|_\sigma$.

Proof By property (a) of Proposition 7, $G^\sigma = \frac{1}{m}\sum_{i=1}^m K_{x_i}^\sigma - F^\sigma \in \mathcal{H}_\sigma$. This in connection with (3) and property (b) of Proposition 7 tells us that for any $f \in \mathcal{H}_\sigma$,

$$\mathbb{E}_m(f) - \mathbb{E}(f) = \langle f, \frac{1}{m}\sum_{i=1}^m K_{x_i}^\sigma - F^\sigma\rangle_\sigma = \langle f, G^\sigma\rangle_\sigma.$$

Then

$$\sup_{f\in\mathcal{H}_\sigma, \|f\|_\sigma\le 1} |\mathbb{E}_m(f) - \mathbb{E}(f)| = \sup_{f\in\mathcal{H}_\sigma, \|f\|_\sigma\le 1} |\langle f, G^\sigma\rangle_\sigma| = \|G^\sigma\|_\sigma.$$

This proves the first statement.

To verify the second statement, we compute the norm $\|G^\sigma\|_\sigma$ by the definition and property (b) of Proposition 7. It yields

$$\|G^\sigma\|_\sigma^2 = \frac{1}{m^2}\sum_{i,j=1}^m K^\sigma(x_i,x_j) - \frac{2}{m}\sum_{i=1}^m \int_X K_{x_i}^\sigma(y)d\mu(y) + \int_X F^\sigma(y)d\mu(y).$$

But $\int_X K_{x_i}^\sigma(y)d\mu(y) = \int_X K^\sigma(x_i,y)d\mu(y) = F^\sigma(x_i)$. So we have

$$\begin{aligned}\|G^\sigma\|_\sigma^2 &= \tfrac{1}{m}\sum_{i=1}^m \left\{\frac{1}{m}\sum_{j=1}^m K^\sigma(x_i,x_j) - F^\sigma(x_i)\right\} \\ &\quad - \int_X \left\{\frac{1}{m}\sum_{i=1}^m K^\sigma(x_i,y) - F^\sigma(y)\right\}d\mu(y) \\ &= \frac{1}{m}\sum_{i=1}^m G^\sigma(x_i) - \int_X G^\sigma(y)d\mu(y) \\ &\le 2\sup_{y\in X}|G^\sigma(y)| = 2\|G^\sigma\|_{C(X)}.\end{aligned}$$

This gives the first inequality of the second statement.

The second inequality of the second statement follows directly from the reproducing property:

$$|G^\sigma(y)| = |\langle G^\sigma, K_y^\sigma\rangle_\sigma| \le \|G^\sigma\|_\sigma \|K_y^\sigma\|_\sigma \le \kappa^\sigma \|G^\sigma\|_\sigma, \qquad \forall y \in X.$$

This completes the proof of the lemma. ■

Now we are ready to prove Theorem 2 stated in Section 1.

Proof of Theorem 2. Recall $\kappa = \sup_{\sigma\in\Sigma}\kappa^\sigma$ and the definition (5) of the set $\mathcal{H}$. By Part (a) of Lemma 8,

$$\sup_{f\in\mathcal{H}}|\mathbb{E}_m(f) - \mathbb{E}(f)| = \sup_{\sigma\in\Sigma}\sup_{f\in\mathcal{H}_\sigma,\|f\|_\sigma\le 1}|\mathbb{E}_m(f) - \mathbb{E}(f)| = \sup_{\sigma\in\Sigma}\|G^\sigma\|_\sigma.$$

On the other hand, since $\{K_y^\sigma : \sigma\in\Sigma, y\in X\}$ is exactly the set $\mathcal{F}$ according to its definition (7), we see that

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_m(f) - \mathbb{E}(f)| = \sup_{\sigma\in\Sigma}\sup_{y\in X}\left|\frac{1}{m}\sum_{i=1}^m K_y^\sigma(x_i) - F^\sigma(y)\right| = \sup_{\sigma\in\Sigma}\|G^\sigma\|_{C(X)}.$$

This in connection with Lemma 8 implies that for any $\varepsilon > 0$ and $\ell \in \mathbb{N}$, there holds

$$\begin{aligned}\left\{\sup_{m\ge\ell}\sup_{f\in\mathcal{F}}|\mathbb{E}_m(f) - \mathbb{E}(f)| > \kappa\varepsilon\right\} &\subseteq \left\{\sup_{m\ge\ell}\sup_{f\in\mathcal{H}}|\mathbb{E}_m(f) - \mathbb{E}(f)| > \varepsilon\right\} \\ &\subseteq \left\{\sup_{m\ge\ell}\sup_{f\in\mathcal{F}}|\mathbb{E}_m(f) - \mathbb{E}(f)| > \frac{\varepsilon^2}{2}\right\}.\end{aligned}$$

Therefore, $\mathcal{H}$ is uGC if and only if $\mathcal{F}$ is. This proves Theorem 2. ■

4. Gaussian Kernels Provide uGC Hypothesis Sets

In this section we establish the uGC property of Gaussians with flexible variances stated in Theorem 3. This will be done by verifying a criterion for uGC sets in terms of empirical covering numbers given by Dudley et al. (1991). For $1 \le p < +\infty$ and $\mathbf{x} = (x_i)_{i=1}^m \in X^m$, we denote the l^p empirical metric of two functions f, g as

$$d_{\mathbf{x},p}(f,g) = \Big(\frac{1}{m}\sum_{i=1}^{m} |f(x_i) - g(x_i)|^p\Big)^{1/p}$$

and for $p = \infty$

$$d_{\mathbf{x},\infty}(f,g) = \max_{1\le i\le m} |f(x_i) - g(x_i)|.$$

Definition 9 *Let $\mathcal{G}$ be a set of functions on X, $1 \le p \le +\infty$ and $\mathbf{x} = (x_i)_{i=1}^m \in X^m$. The empirical covering number $\mathcal{N}_p(\mathcal{G}, \mathbf{x}, \eta)$ is defined to be the minimal integer N such that there are N functions $\{g^j\}_{j=1}^N \subset \mathcal{G}$ satisfying*

$$\min_{1\le j\le N} d_{\mathbf{x},p}(g, g^j) \le \eta, \qquad \forall g \in \mathcal{G}.$$

The metric entropy of $\mathcal{G}$ is defined as

$$H_{m,p}(\mathcal{G}, \eta) = \sup_{\mathbf{x}\in X^m} \log \mathcal{N}_p(\mathcal{G}, \mathbf{x}, \eta), \qquad m \in \mathbb{N}, \eta > 0.$$

The following criterion was established by Dudley et al. (1991).

Lemma 10 *A set $\mathcal{G}$ of functions from X to $[0,1]$ is uGC if and only if for some $1 \le p \le +\infty$, there holds*

$$\lim_{m\to\infty} \frac{H_{m,p}(\mathcal{G}, \eta)}{m} = 0, \qquad \forall \eta > 0.$$

To continue, we need the following combinatorial dimensions.

Definition 11 *Let $\mathcal{G}$ be a set of functions from X to $[0,1]$. We say that $A \subset X$ is V_γ shattered (P_γ shattered) by $\mathcal{G}$ if there is a number $\alpha \in \mathbb{R}$ (a function $s : A \to [0,1]$) with the following property: For every subset E of A there exists some function $f_E \in \mathcal{G}$ such that $f_E(x) \le \alpha - \gamma$ ($f_E(x) \le s(x) - \gamma$) for every $x \in A \setminus E$, and $f_E(x) \ge \alpha + \gamma$ ($f_E(x) \ge s(x) + \gamma$) for every $x \in E$. The V_γ dimension of $\mathcal{G}$, $V_\gamma(\mathcal{G})$, (The P_γ dimension of $\mathcal{G}$, $P_\gamma(\mathcal{G})$,) is the maximal cardinality of a set $A \subset X$ that is V_γ shattered (P_γ shattered) by $\mathcal{G}$.*

Based on Lemma 10, Alon et al. (1997) showed that $\mathcal{G}$ is uGC if and only if the V_γ dimension or P_γ dimension of $\mathcal{G}$ is finite for every $\gamma > 0$.

In addition, we need the following relation between the dimensions and the empirical covering numbers essentially proved by Alon et al. (1997). A complete proof is given in Appendix A.

Lemma 12 *Let $\mathcal{G}$ be a set of functions from X to $[0,1]$.*

(a) $V_\gamma(\mathcal{G}) \le P_\gamma(\mathcal{G}) \le (\frac{2}{\gamma} + 1) V_{\gamma/2}(\mathcal{G})$ for any $\gamma > 0$.

(b) For every $m \in \mathbb{N}$ and $0 < \varepsilon < 1$, there holds

$$\sup_{\mathbf{x}\in X^m} \mathcal{N}_\infty(\mathcal{G},\mathbf{x},\varepsilon) \le 2\left(\frac{4m}{\varepsilon^2}\right)^{1+d\log\left(\frac{2em}{\varepsilon}\right)}, \qquad d := P_{\varepsilon/4}(\mathcal{G}).$$

With the above preparations, we can prove Theorem 3. First, let us begin with the univariate case.

Lemma 13 *Let X be a subset of $\mathbb{R}$ and $\mathcal{F}$ be given by (7) with $K^\sigma(x,y) = \exp\{-\frac{(x-y)^2}{\sigma^2}\}$ and $\Sigma = (0,+\infty)$. Then $V_\gamma(\mathcal{F}) \le 2$ for every $\gamma > 0$.*

Proof Suppose to the contrary that $V_\gamma(\mathcal{F}) \ge 3$ for some $\gamma > 0$. It means that there is a set $A = \{x_1,x_2,x_3\} \subset X$ with $x_1 < x_2 < x_3$ which is V_γ-shattered by $\mathcal{F}$. That is, there is some $\alpha \in \mathbb{R}$ such that for every $E \subseteq A$ there exists some function $f_E \in \mathcal{F}$ satisfying

$$\begin{cases} f_E(x) \le \alpha - \gamma, & \text{for } x \in A \setminus E, \\ f_E(x) \ge \alpha + \gamma, & \text{for } x \in E. \end{cases}$$

Take $E = \{x_1,x_3\} \subset A$. Then there is a function $f_E \in \mathcal{F}$ such that $f_E(x_2) \le \alpha - \gamma$ and $f_E(x_1)$, $f_E(x_3) \ge \alpha + \gamma$. The function f_E can be represented as $f_E(x) = \exp\{-\frac{(x-y)^2}{\sigma^2}\}$ for some $y \in X$ and $\sigma \in (0,\infty)$. It can be extended to the whole real line with the same expression and we denote this extended function as $\widetilde{f}_E$. The function $\widetilde{f}_E$ has no local minimum on $\mathbb{R}$. However, it is continuous and satisfies

$$\widetilde{f}_E(x_1) = f_E(x_1) > f_E(x_2) = \widetilde{f}_E(x_2), \qquad \widetilde{f}_E(x_3) = f_E(x_3) > f_E(x_2) = \widetilde{f}_E(x_2).$$

So the minimum value of $\widetilde{f}_E$ on the closed interval $[x_1,x_3]$ is achieved on the open interval (x_1,x_3). Consequently, $\widetilde{f}_E$ has a local minimum on $\mathbb{R}$, which is a contradiction. ■

Next, we prove Theorem 3 by Lemmas 10 and 13. Here the tensor product form of the functions in the set $\mathcal{F}$ plays an essential role.

Proof of Theorem 3. By Theorem 2, we need to show that $\mathcal{F}$ is uGC. Consider another set of functions on $\mathbb{R}^n$ defined as

$$\widetilde{\mathcal{F}} = \{K_x^\sigma = K^\sigma(x,\cdot) : \sigma \in \Sigma, x \in \mathbb{R}^n\}.$$

If $\widetilde{\mathcal{F}}$ is uGC as a set of functions from $\mathbb{R}^n$ to $[0,1]$, then its restriction onto X is also uGC, which would imply the uGC property of $\mathcal{F}$. Therefore, it suffices to prove the uGC of $\widetilde{\mathcal{F}}$.

Define sets of univariate functions as

$$\mathcal{F}^j = \left\{ e^{-(t-s_j)^2/\sigma_j^2} : \sigma_j \in (0,\infty), s_j \in \mathbb{R} \right\}, \qquad j = 1,2,\ldots,n.$$

Fix $j \in \{1,2,\ldots,n\}$. Consider the set $\mathcal{F}^j$ of functions from $\mathbb{R}$ to $[0,1]$. Lemma 13 tells us that $V_\gamma(\mathcal{F}^j) \le 2$ for every $\gamma > 0$. Applying Lemma 12 to $\mathcal{G} = \mathcal{F}^j$ of functions on $\mathbb{R}$, we find that for every $0 < \varepsilon < 1$, $P_{\varepsilon/4}(\mathcal{F}^j) \le \frac{16}{\varepsilon} + 2$ and

$$\sup_{\mathbf{t}\in\mathbb{R}^m} \mathcal{N}_\infty(\mathcal{F}^j,\mathbf{t},\varepsilon) \le \mathcal{N}_0 := 2\left(\frac{4m}{\varepsilon^2}\right)^{1+\left(\frac{16}{\varepsilon}+2\right)\log\left(\frac{2em}{\varepsilon}\right)}. \tag{20}$$

Now we apply (20) to estimate the empirical covering number of the set $\widetilde{\mathcal{F}}$. To this end, let $0<\varepsilon<1$ and $\mathbf{x}=(x_i)_{i=1}^m$ where each point $x_i\in\mathbb{R}^n$ can be expressed as a vector $x_i=(x_{i,1},\dots,x_{i,n})$. Let $j\in\{1,2,\dots,n\}$. Consider $\mathbf{t}:=(x_{i,j})_{i=1}^m\in\mathbb{R}^m$. The bound (20) tells us that there are $\mathcal{N}_0$ pairs $\{(\sigma_{j,\ell},s_{j,\ell})\}_{\ell=1}^{\mathcal{N}_0}$ with $\sigma_{j,\ell}\in(0,\infty)$ and $s_{j,\ell}\in\mathbb{R}$ representing $\mathcal{N}_0$ functions $\left\{g^\ell(t)=e^{-(t-s_{j,\ell})^2/\sigma_{j,\ell}^2}\right\}_{\ell=1}^{\mathcal{N}_0}\subset\mathcal{F}^j$ such that

$$\min_{1\le\ell\le\mathcal{N}_0} d_{\mathbf{t}}(g,g^\ell)=\min_{1\le\ell\le\mathcal{N}_0}\left\{\max_{1\le i\le m}|g(x_{i,j})-g^\ell(x_{i,j})|\right\}\le\varepsilon,\qquad\forall g\in\mathcal{F}^j.$$

That is, for any $\sigma_j\in(0,\infty)$ and $s_j\in\mathbb{R}$, we can find some $\ell\in\{1,\dots,\mathcal{N}_0\}$ satisfying

$$\left|\exp\left\{-\frac{(x_{i,j}-s_j)^2}{\sigma_j^2}\right\}-\exp\left\{-\frac{(x_{i,j}-s_{j,\ell})^2}{\sigma_{j,\ell}^2}\right\}\right|\le\varepsilon,\qquad\forall i=1,\dots,m. \tag{21}$$

Now we choose a set $\mathcal{F}_\varepsilon$ of functions on $\mathbb{R}^n$ consisting of $\mathcal{N}_0^n$ functions

$$f_{\ell_1,\ell_2,\dots,\ell_n}(\cdot)=\exp\left\{-\sum_{j=1}^n\frac{(\cdot_j-s_{j,\ell_j})^2}{\sigma_{j,\ell_j}^2}\right\},\qquad\ell_1,\ell_2,\dots,\ell_n\in\{1,2,\dots,\mathcal{N}_0\}.$$

Each function $f\in\widetilde{\mathcal{F}}$ can be expressed as

$$f(x_i)=\exp\left\{-\sum_{j=1}^n\frac{(x_{i,j}-s_j)^2}{\sigma_j^2}\right\}$$

with $\{\sigma_j\}_{j=1}^n\subset(0,\infty)^n$ and $\{s_j\}_{j=1}^n\subset\mathbb{R}^n$. We can choose some $\{\ell_j\}_{\ell=1}^n\in\{1,\dots,\mathcal{N}_0\}^n$ satisfying (21). Then

$$|f(x_i)-f_{\ell_1,\ell_2,\dots,\ell_n}(x_i)|\le\sum_{p=1}^n\exp\left\{-\sum_{j=p+1}^n\frac{(x_{i,j}-s_j)^2}{\sigma_j^2}\right\}\varepsilon\le n\varepsilon.$$

Thus, we have

$$d_{\mathbf{x},\infty}(f,f_{\ell_1,\ell_2,\dots,\ell_n})=\max_{1\le i\le m}|f(x_i)-f_{\ell_1,\ell_2,\dots,\ell_n}(x_i)|\le n\varepsilon.$$

By the definition of empirical covering numbers, we have

$$\mathcal{N}_\infty(\widetilde{\mathcal{F}},\mathbf{x},n\varepsilon)\le\mathcal{N}_0^n=2^n\left(\frac{4m}{\varepsilon^2}\right)^{n+\left(\frac{16}{\varepsilon}+2\right)n\log\left(\frac{2em}{\varepsilon}\right)}.$$

Therefore, for any $0<\varepsilon<1$,

$$H_{m,\infty}(\widetilde{\mathcal{F}},n\varepsilon)\le n\log 2+\left(n+\left(\frac{16}{\varepsilon}+2\right)n\log\left(\frac{2em}{\varepsilon}\right)\right)\log\frac{4m}{\varepsilon^2}. \tag{22}$$

Observe that $\mathcal{N}_\infty(\widetilde{\mathcal{F}},\mathbf{x},\varepsilon)=1$ for any $\varepsilon\ge1$. Hence $H_{m,\infty}(\widetilde{\mathcal{F}},n\varepsilon)=0$ for any $\varepsilon\ge1$. Combining this observation with (22) implies that

$$\lim_{m\to\infty}\frac{H_{m,\infty}(\widetilde{\mathcal{F}},n\varepsilon)}{m}=0,\quad\text{for any }\varepsilon>0.$$

Hence $\widetilde{\mathcal{F}}$ is uGC by Lemma 10. This completes the proof of Theorem 3. ■

The proof of Theorem 3 actually gives estimates for the empirical covering number which can be used to bound the sample error in (13) as we shall do in the following section.

5. Error Bound by Rademacher Averages

In order to bound the error $\mathcal{E}(f_{\mathbf{z},\sigma}) - \mathcal{E}(f_\rho^V)$, we know from the error decomposition (13) that it is sufficient to estimate the sample error and the regularization error. In particular, we will provide the error bounds for the multi-kernel scheme (4) generated by Gaussians with flexible variances.

5.1 Sample Error Estimate

In this subsection we are mainly concerned about the sample error. The regularization error will be discussed in the next subsection. The estimate of the sample error for the regularized multi-kernel scheme (4) involves the hypothesis space

$$\mathcal{H}_\lambda = \bigcup_{\sigma\in\Sigma} \Big\{ f \in \mathcal{H}_{K_\sigma} : \|f\|_{K_\sigma} \le \sqrt{\frac{M}{\lambda}} \Big\}. \tag{23}$$

Below, we show how to get sample error estimates by Rademacher complexities (Bartlett and Mendelson, 2002; Koltchinskii, 2001; Koltchinskii and Panchenko, 2002) of the reduced hypothesis space $\mathcal{F}$ defined by (7), which is potentially easier to compute than that of $\mathcal{H}_\lambda$. Let's first introduce Rademacher average over a set of functions F on Ω.

Definition 14 *Let μ be a probability measure on Ω and F be a class of uniformly bounded functions. For every integer m, let*

$$R_m(F) := \mathbb{E}_\mu \mathbb{E}_\varepsilon \Big[\frac{1}{m} \sup_{f\in F} \Big| \sum_{i=1}^m \varepsilon_i f(z_i) \Big| \Big]$$

where $\{z_i\}_{i=1}^m$ are independent random variables distributed according to μ and $\{\varepsilon_i\}_{i=1}^m$ are independent Rademacher random variables, that is, $P(\varepsilon_i = +1) = P(\varepsilon_i = -1) = 1/2$.

Turn to the multi-kernel regularization scheme (4). If the loss function $V(y,t)$ is convex with respect to t, its left and right partial derivatives with respect to the second variable, denoted as $V'_-(y,t), V'_+(y,t)$ respectively for simplicity, exist for every $y \in Y$. Throughout this section, we assume the loss function is admissible in the following sense.

Definition 15 *We say that the loss function $V : Y \times \mathbb{R} \to \mathbb{R}_+$ is admissible (with respect to ρ) if it is convex with respect to the second variable, $M = \|V(y,0)\|_{L^\infty_\rho(Z)} < +\infty$ and, for any $\lambda > 0$ there holds*

$$C_\lambda = \sup \Big\{ \max(|V'_-(y,t)|, |V'_+(y,t)|) : y \in Y,\ |t| \le \kappa \sqrt{\frac{M}{\lambda}} \Big\} < +\infty. \tag{24}$$

We also need the following lemma (Bartlett and Mendelson, 2002; Ledoux and Talagrand, 1991) summarizing some of the properties of Rademacher averages. A complete proof is given in Appendix A.

Lemma 16 *Let F be a class of uniformly bounded real-valued functions on (Ω, μ) and $m \in \mathbb{N}$.*

(a) For every $c \in \mathbb{R}$, $R_m(cF) = |c| R_m(F)$, where $cF = \{cf : f \in F\}$.

(b) If for each $i \in \{1,\dots,m\}$, $\phi_i : \mathbb{R} \to \mathbb{R}$ is a function with $\phi_i(0) = 0$ having a Lipschitz constant c_i, then for any $\{x_i\}_{i=1}^m$,

$$\mathbb{E}_\varepsilon\Big[\sup_{f\in F}|\sum_{i=1}^m \varepsilon_i\phi_i(f(x_i))|\Big] \le 2\mathbb{E}_\varepsilon\Big[\sup_{f\in F}\Big|\sum_{i=1}^m c_i\varepsilon_i f(x_i)|\Big].$$

The sample error analysis of multi-kernel regularization scheme (4) involves the Rademacher complexity of the fundamental space $\mathcal{F}$ denoted by

$$R_m(\mathcal{F}) = \mathbb{E}_{\rho_X}\mathbb{E}_\varepsilon\Big[\frac{1}{m}\sup_{f\in\mathcal{F}}|\sum_{i=1}^m \varepsilon_i f(x_i)|\Big].$$

Lemma 17 *Let $\mathcal{H}_\lambda$ be defined by (23), then*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{H}_\lambda}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|\Big] \le 4C_\lambda\sqrt{\frac{M}{\lambda}}\left(R_m(\mathcal{F})\right)^{1/2} + \frac{2M}{\sqrt{m}}.$$

Proof Let $\mathcal{V}_\lambda'$ be a set of functions on Z defined as

$$\mathcal{V}_\lambda' = \Big\{V(y,f(x)) :\ f\in\mathcal{H}_\lambda\Big\}.$$

Using standard symmetrization arguments (e.g., Van der Vaart and Weller, 1996, Lemma 2.3.1), one can see that

$$\mathbb{E}_\rho\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|\Big] \le 2\mathbb{E}_\rho\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}\frac{1}{m}|\sum_{i=1}^m\varepsilon_i V(y_i,f(x_i))|\Big] := 2R_m(\mathcal{V}_\lambda').$$

To handle $R_m(\mathcal{V}_\lambda')$, we apply Lemma 16. To this end, note that $\|f\|_\infty \le \kappa\sqrt{M/\lambda}$ for all $f\in\mathcal{H}_\lambda$, for fixed $\{y_i\}_{i=1}^m \in Y^m$. If we define functions

$$\phi_i(t) = \begin{cases} V(y_i,t)-V(y_i,0) & \text{when } |t|\le\kappa\sqrt{M/\lambda} \\ V(y_i,\kappa\sqrt{M/\lambda})-V(y_i,0) & \text{when } t\ge\kappa\sqrt{M/\lambda} \\ V(y_i,-\kappa\sqrt{M/\lambda})-V(y_i,0) & \text{when } t\le-\kappa\sqrt{M/\lambda}, \end{cases}$$

then $\phi_i(t):\mathbb{R}\to\mathbb{R}$ has the Lipschitz constant $c_i = C_\lambda$ and $\phi_i(0)=0$ for any i. Applying Lemma 16 to the space $\mathcal{H}_\lambda$, then $\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}\Big|\sum_{i=1}^m\varepsilon_i V(y_i,f(x_i))\Big|\Big]$ is bounded by

$$\begin{aligned}
\mathbb{E}_\varepsilon\ &\Big[\sup_{f\in\mathcal{H}_\lambda}\Big|\sum_{i=1}^m\varepsilon_i\phi_i(f(x_i))\Big|\Big] + \mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}\Big|\sum_{i=1}^m\varepsilon_i V(y_i,0)\Big|\Big] \\
&\le 2\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}|\sum_{i=1}^m C_\lambda\varepsilon_i f(x_i)\Big|\Big] + \Big[\mathbb{E}_\varepsilon\Big|\sum_{i=1}^m\varepsilon_i V(y_i,0)\Big|^2\Big]^{1/2} \\
&= 2C_\lambda\sqrt{\tfrac{M}{\lambda}}\,\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}}|\sum_{i=1}^m\varepsilon_i f(x_i)|\Big] + \left(\mathbb{E}_\varepsilon\Big[\sum_{i,j=1}^m\varepsilon_i V(y_i,0)V(y_j,0)\varepsilon_j\Big]\right)^{1/2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}_\lambda}\Big|\sum_{i=1}^m\varepsilon_i V(y_i,f(x_i))\Big|\Big] \ &\le 2C_\lambda\sqrt{\tfrac{M}{\lambda}}\,\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{H}}|\sum_{i=1}^m\varepsilon_i f(x_i)|\Big] \\
&\quad + M\sqrt{m}.
\end{aligned} \tag{25}$$

For the first term on the right hand side of the above inequality, we use the reproducing property (3) and obtain

$$\begin{aligned}\sup_{f\in\mathcal{H}}|\sum_{i=1}^m \varepsilon_i f(x_i)| &= \sup_{\sigma\in\Sigma}\sup_{\substack{f\in\mathcal{H}_{K^\sigma}\\ \|f\|_{K^\sigma}\le 1}}|\langle\sum_{i=1}^m \varepsilon_i K_{x_i}^\sigma, f\rangle_{K^\sigma}| = \sup_{\sigma\in\Sigma}\left\|\sum_{i=1}^m \varepsilon_i K_{x_i}^\sigma\right\|_{K^\sigma}\\ &= \left[\sup_{\sigma\in\Sigma}\sum_{i,j=1}^m \varepsilon_i K^\sigma(x_i,x_j)\varepsilon_j\right]^{1/2}.\end{aligned} \tag{26}$$

But by the definition of the fundamental set $\mathcal{F}$, we know that

$$\begin{aligned}\sup_{\sigma\in\Sigma}\sum_{i,j=1}^m \varepsilon_i K^\sigma(x_i,x_j)\varepsilon_j \le\ & m\sup_{\sigma\in\Sigma}\sup_{s\in X}|\sum_{i=1}^m \varepsilon_i K^\sigma(x_i,s)|\\ &= m\sup_{f\in\mathcal{F}}|\sum_{i=1}^m \varepsilon_i f(x_i)|.\end{aligned} \tag{27}$$

Combining the estimates (25), (26), and (27) implies that

$$\begin{aligned}R_m(\mathcal{V}_\lambda) &\le 2C_\lambda\sqrt{\tfrac{M}{\lambda}}\,\mathbb{E}_{\rho_X}\mathbb{E}_\varepsilon\left[\tfrac{1}{m}\sup_{f\in\mathcal{F}}|\sum_{i=1}^m \varepsilon_i f(x_i)|\right]^{1/2}+\frac{M}{\sqrt{m}}\\ &\le 2C_\lambda\sqrt{\tfrac{M}{\lambda}}\,\Big(R_m(\mathcal{F})\Big)^{1/2}+\tfrac{M}{\sqrt{m}}.\end{aligned}$$

This finishes Lemma 17. ■

The following error bound for the multi-kernel scheme (4) is a straightforward consequence of the sample error estimate and the error decomposition (13).

Theorem 18 *Let V be admissible with C_λ given by (24). Define $f_{\mathbf{z},\lambda}$ by (4). Then we have*

$$\mathbb{E}\Big[\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho^V)\Big]\le 4C_\lambda\sqrt{\frac{M}{\lambda}}\Big(R_m(\mathcal{F})\Big)^{1/2}+\frac{2M}{\sqrt{m}}+\widetilde{\mathcal{D}}(\lambda).$$

Proof By the error decomposition (13), a special choice $f_\lambda^V\in\mathcal{H}_{K^\sigma}$ with some $\sigma\in\Sigma$ defined by (12) gives us that

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho^V)\le\left\{\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\mathcal{E}_{\mathbf{z}}(f_\lambda^V)-\mathcal{E}(f_\lambda^V)\right\}+\widetilde{\mathcal{D}}(\lambda).$$

Together with the fact $\mathbb{E}(\mathcal{E}_{\mathbf{z}}(f_\lambda^V))=\mathcal{E}(f_\lambda^V)$ and $f_{\mathbf{z},\lambda}\in\mathcal{H}_\lambda$ defined in (23), we know that

$$\mathbb{E}\left[\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})+\mathcal{E}_{\mathbf{z}}(f_\lambda^V)-\mathcal{E}(f_\lambda^V)\right]\le\mathbb{E}\Big[\sup_{f\in\mathcal{H}_\lambda}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|\Big]$$

which in connection with Lemma 17 yields the desired estimate. ■

To estimate the Radmacher average $R_m(\mathcal{F})$ in Theorem 18, one can resort to the following bound using the empirical l^2 covering number, which is due to Dudley (1999); Van der Vaart and Weller (1996).

Lemma 19 *Let F be a class of uniformly bounded functions. Then there exists an absolute constant C such that, for any sample $\{x_i\}_{i=1}^m$, there holds*

$$\frac{1}{\sqrt{m}}\mathbb{E}_\varepsilon\Big[\sup_{f\in F}|\sum_{i=1}^m \varepsilon_i f(x_i)|\Big] \le C\int_0^{+\infty}\sqrt{\log \mathcal{N}_2(F,\mathbf{x},\varepsilon)}d\varepsilon.$$

Applying the estimate in the proof for Theorem 3, we can compute the Radmacher average $R_m(\mathcal{F})$ for Gaussian kernels with flexible variances, and hence yields the subsequent error bound.

Theorem 20 *Let X be a subset of $\mathbb{R}^n$ and V be admissible. Define $f_{\mathbf{z},\lambda}$ by (4) and the Gaussian kernels by (8). Then we have*

$$\mathbb{E}\Big[\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho^V)\Big] \le C_\lambda\sqrt{\frac{C'M}{\lambda}}\Big(\frac{\log^2 m}{m}\Big)^{1/4}+\frac{2M}{\sqrt{m}}+\widetilde{\mathcal{D}}(\lambda), \tag{28}$$

where C' is a constant independent of m or λ.

Proof Recall the fundamental function set $\mathcal{F}$ defined by (7). Consider the larger set

$$\widetilde{\mathcal{F}} = \{K_x^\sigma = K^\sigma(x,\cdot) : \sigma\in\Sigma, x\in\mathbb{R}^n\}.$$

The estimate (22) tells us that for $0<\varepsilon<1$,

$$\begin{aligned}\log\mathcal{N}_2(\widetilde{\mathcal{F}},\mathbf{x},\varepsilon) &\le \log\mathcal{N}_\infty(\widetilde{\mathcal{F}},\mathbf{x},\varepsilon)\\ &\le n\log 2+\Big(n+\Big(\tfrac{16n}{\varepsilon}+2\Big)n\log\Big(\tfrac{2enm}{\varepsilon}\Big)\Big)\log\tfrac{4n^2m}{\varepsilon^2}.\end{aligned}$$

Since $K^\sigma\le 1$ for each $\sigma\in\Sigma$, we see that $\log\mathcal{N}_2(\widetilde{\mathcal{F}},\mathbf{x},\varepsilon)=0$ for any $\varepsilon\ge 1$. Applying Lemma 19, we have

$$R_m(\widetilde{\mathcal{F}})\le C''\frac{\log^2 m}{\sqrt{m}},$$

where C'' is an absolute constant independent of m. The trivial fact $R_m(\mathcal{F})\le R_m(\widetilde{\mathcal{F}})$ together with Theorem 18 gives us the desired result. ■

In order to get explicit error rates, we see from Theorems 18 and 20 that what left is to estimate the regularization error.

5.2 Regularization Error with Gaussians

In this subsection we exclusively focus on the multi-kernel regularization error (11) associated with the least square loss and Gaussian kernels. We show that how the Fourier analysis (Stein, 1970) can be applied to get the polynomial decay of the regularization error under Sobolev smoothness condition on the regression function.

Before we go to the main point, it is worth briefly mentioning why the multi-kernel regularization error can improve the error rates. To this end, note that, for the regularization error of a single Gaussian kernel, it was proved by Smale and Zhou (2003) that the polynomial decay $O(\lambda^s)$ for some $s>0$ is impossible under the Sobolev smoothness hypothesis on the regression function f_ρ. Actually, it only decays logarithmically $O((\log(1/\lambda))^{-s})$ for some $s>0$. Putting this logarithmical

decay back into (28) and trading off λ and m, we notice that the error rate is unacceptably slow. Below we show that the multi-kernel scheme (4) associated with the least square loss and Gaussian kernels (9) with flexible isotropic variances can give regularization errors $\mathcal{D}(\lambda)$ of polynomial decays $O(\lambda^\beta)$ for some $\beta > 0$, under the assumption of Sobolev smoothness on f_ρ. Combining this polynomial decay with (28) can give rise to tighter bounds compared to the single kernel case.

To estimate the multi-kernel regularization error (11), we introduce some basic facts about RKHS (Aronszajn, 1950). Let Ω be a domain (bounded or not) in $\mathbb{R}^n$. Let $\sigma > 0$. Consider the RKHS induced by the Gaussian kernel given in (9) as $K_\sigma(x,y) = \exp\{-\frac{|x-y|^2}{\sigma^2}\} : \Omega \times \Omega \to \mathbb{R}$. We denote it as $\mathcal{H}_{K_\sigma}(\Omega)$ with norm $\|\cdot\|_{\mathcal{H}_{K_\sigma}(\Omega)}$ for simplicity. Let $\widetilde{\Omega} \subset \Omega$. By restricting the kernel K_σ to $\tilde{\Omega} \times \tilde{\Omega}$, it also induces an RKHS in $\tilde{\Omega}$ denoted as $\mathcal{H}_{K_\sigma}(\tilde{\Omega})$. Then we know (Aronszajn, 1950) that

$$\mathcal{H}_{K_\sigma}(\tilde{\Omega}) = \left\{ g = f|_{\tilde{\Omega}} : \ f \in \mathcal{H}_{K_\sigma}(\Omega) \right\} \tag{29}$$

with norm

$$\|g\|_{\mathcal{H}_{K_\sigma}(\tilde{\Omega})} = \inf\{\|f\|_{\mathcal{H}_{K_\sigma}(\Omega)} : \ f|_{\tilde{\Omega}} = g\}. \tag{30}$$

Define the integral operator L_K associated with a Mercer kernel K and a Borel measure μ on Ω as

$$L_K f(x) := \int_\Omega K(x,t) f(t) d\mu(t), \quad x \in \Omega, \ f \in L^2_\mu(\Omega).$$

If Ω is a compact domain in $\mathbb{R}^n$, then L_K is a positive, self-adjoint, compact operator and its range lies in $C(\Omega)$. Take the square root $L_K^{1/2}$ of L_K, then

$$\|L_K^{1/2} f\|_K = \|f\|_{L^2_\mu(\Omega)} \qquad \forall f \in L^2_\mu(\Omega). \tag{31}$$

When $\Omega = \mathbb{R}^n$ and μ is the Lebesgue measure, we define for $\sigma > 0$

$$f^\sigma(x) = L_{K_\sigma} f(x) = \int_{\mathbb{R}^n} K_\sigma(x,y) f(y) dy, \quad x \in \mathbb{R}^n, \ f \in L^2(\mathbb{R}^n).$$

As in Steinwart and Scovel (2005), we shall use these functions as approximations of f_ρ to estimate the regularization error.

Lemma 21 *Let f^σ be defined for $f \in L^2(\mathbb{R}^n)$ as above. Then $f^\sigma \in \mathcal{H}_{K_\sigma}(\mathbb{R}^n)$ and*

$$\|f^\sigma\|_{\mathcal{H}_{K_\sigma}(\mathbb{R}^n)} \le (\sqrt{\pi}\sigma)^{n/2} \|f\|_{L^2(\mathbb{R}^n)}. \tag{32}$$

Proof We shall use notations and results on limits of reproducing kernels (see Theorem I in Section 9 of Aronszajn (1950)).

Denote E_j as the closed ball of $\mathbb{R}^n$ with radius j centered at zero. Then $\mathbb{R}^n = \bigcup_{j\in\mathbb{N}} E_j$. For $j \le j'$ and $f_{j'} \in \mathcal{H}_{K_\sigma}(E_{j'})$, the properties (29) and (30) of the restriction of RKHS tell us that

$$f_{j'}|_{E_j} \in \mathcal{H}_{K_\sigma}(E_j) \ \text{ and } \ \|f_{j'}|_{E_j}\|_{\mathcal{H}_{K_\sigma(E_j)}} \le \|f_{j'}\|_{\mathcal{H}_{K_\sigma}(E_{j'})}.$$

In order to show that $f^\sigma \in \mathcal{H}_{K_\sigma}(\mathbb{R}^n)$ and (32) holds, by Theorem I in Section 9 of Aronszajn (1950) it is sufficient to prove for $f_{\sigma,j} := f^\sigma|_{E_j}$ that

$$f_{\sigma,j} \in \mathcal{H}_{K_\sigma}(E_j) \quad \text{and} \quad \liminf_{j\to\infty} \|f_{\sigma,j}\|_{\mathcal{H}_{K_\sigma}(E_j)} \le (\sqrt{\pi}\sigma)^{n/2} \|f\|_{L^2(\mathbb{R}^n)}. \tag{33}$$

To this end, for $j \le j'$, define

$$f_{\sigma,j,j'}(x) := \int_{E_{j'}} K_\sigma(x,y) f(y) dy \to f_{\sigma,j}(x) \quad \text{uniformly in} \quad C(E_j). \tag{34}$$

Then $f_{\sigma,j,j'} \in \mathcal{H}_{K_\sigma}(E_{j'})$ by (31) since $f \in L^2(E_{j'})$. Using (29) and (30) with $K_{\sigma,j'} := K_\sigma|_{E_{j'} \times E_{j'}}$, it yields

$$\| f_{\sigma,j,j'}|_{E_j} \|^2_{\mathcal{H}_{K_\sigma}(E_j)} \le \| L_{K_{\sigma,j'}}(f|_{E_{j'}}) \|^2_{\mathcal{H}_{K_\sigma}(E_{j'})}. \tag{35}$$

By (31), we have

$$\begin{aligned} \| L_{K_{\sigma,j'}}(f|_{E_{j'}}) \|^2_{\mathcal{H}_{K_\sigma}(E_{j'})} &= \| L^{1/2}_{K_{\sigma,j'}}(f|_{E_{j'}}) \|_{L^2(E_{j'})} = \langle L_{K_{\sigma,j'}}(f|_{E_{j'}}), f|_{E_{j'}} \rangle_{L^2(E_{j'})} \\ &\le \| L_{K_{\sigma,j'}}(f|_{E_{j'}}) \|_{L^2(E_{j'})} \| f|_{E_{j'}} \|_{L^2(E_{j'})}. \end{aligned} \tag{36}$$

Note that for each $x \in E_{j'}$, there holds $\int_{E_{j'}} K_\sigma(x,t)dt \le \int_{\mathbb{R}^n} K_\sigma(x,t)dt = (\sqrt{\pi}\sigma)^n$. We get by the Schwarz inequality

$$\begin{aligned} \| L_{K_{\sigma,j'}} f \|^2_{L^2(E_{j'})} &= \int_{E_{j'}} \left| \int_{E_{j'}} K_\sigma(x,t) f(t) dt \right|^2 dx \\ &\le \int_{E_{j'}} \Big\{ \int_{E_{j'}} K_\sigma(x,t) dt \Big\} \Big\{ \int_{E_{j'}} K_\sigma(x,t) |f(t)|^2 dt \Big\} dx \\ &\le (\sqrt{\pi}\sigma)^n \int_{E_{j'}} |f(t)|^2 \Big\{ \int_{E_{j'}} K_\sigma(x,t) dx \Big\} dt \\ &\le (\sqrt{\pi}\sigma)^{2n} \| f \|^2_{L^2(\mathbb{R}^n)}. \end{aligned}$$

Putting this estimate into (36), it follows from (35) that

$$\| f_{\sigma,j,j'}|_{E_j} \|_{\mathcal{H}_{K_\sigma}(E_j)} \le \| L_{K_{\sigma,j'}}(f|_{E_{j'}}) \|_{\mathcal{H}_{K_\sigma}(E_{j'})} \le (\sqrt{\pi}\sigma)^{n/2} \| f \|_{L^2(\mathbb{R}^n)}.$$

Since the fixed ball of $\mathcal{H}_{K_\sigma}(E_j)$ with radius $(\sqrt{\pi}\sigma)^{n/2}\| f \|_{L^2(\mathbb{R}^n)}$ centered at zero is weakly compact, there exists a subsequence $\{j'_\ell\}_{\ell \in \mathbb{N}}$ of $\{j'\}$ such that

$$f_{\sigma,j,j'_\ell}|_{E_j} \rightharpoonup f^* \qquad \text{in } \mathcal{H}_{K_\sigma}(E_j) \quad \text{as} \quad \ell \to \infty.$$

Therefore

$$\begin{aligned} \| f^* \|^2_{\mathcal{H}_{K_\sigma}(E_j)} &= \lim_{\ell \to \infty} \langle f^*, f_{\sigma,j,j'_\ell}|_{E_j} \rangle_{\mathcal{H}_{K_\sigma}(E_j)} \\ &\le \| f^* \|_{\mathcal{H}_{K_\sigma}(E_j)} \liminf_{\ell \to \infty} \| f_{\sigma,j,j'_\ell}|_{E_j} \|_{\mathcal{H}_{K_\sigma}(E_j)} \end{aligned}$$

which tells us that

$$\| f^* \|_{\mathcal{H}_{K_\sigma}(E_j)} \le \liminf_{\ell \to \infty} \| f_{\sigma,j,j'_\ell}|_{E_j} \|_{\mathcal{H}_{K_\sigma}(E_j)} \le (\sqrt{\pi}\sigma)^{n/2} \| f \|_{L^2(\mathbb{R}^n)}. \tag{37}$$

By the reproducing property (3), we also have, for each $x \in E_j$

$$f_{\sigma,j,j'_\ell}|_{E_j}(x) = \langle f_{\sigma,j,j'_\ell}|_{E_j}, K_{\sigma,j}(x,\cdot) \rangle_{\mathcal{H}_{K_\sigma}(E_j)} \to \langle f^*, K_{\sigma,j}(x,\cdot) \rangle_{\mathcal{H}_{K_\sigma}(E_j)} = f^*(x)$$

which in connection with (34) gives us that

$$f_{\sigma,j} = f^* \in \mathcal{H}_{K_\sigma}(E_j).$$

Together with (37) and (33), we know that Lemma 21 holds true. ■

Proposition 22 *Let X be a domain in $\mathbb{R}^n$ with Lipschitz boundary. Suppose $f_\rho \in H^s(X)$ for some $s > 0$. Consider the multi-kernel scheme (4) with the least square loss $V(y,t) = (y-t)^2$ and the Gaussian kernels (9) with $\Sigma = (0,\infty)$. Then, the following statements hold true.*

(1) If $n/2 < s \le n/2+2$, then there holds

$$\widetilde{\mathcal{D}}(\lambda) \le \inf_{\sigma\in(0,\infty)} \inf_{f\in\mathcal{H}_{K_\sigma}} \Big\{ \|f - f_\rho\|^2_{C(X)} + \lambda \|f\|^2_{K_\sigma} \Big\}.$$

For any $0 < \varepsilon < 2s-n$, there exists a constant $C_{\varepsilon,s,X}$ such that

$$\widetilde{\mathcal{D}}(\lambda) \le C_{\varepsilon,s,X} \|f_\rho\|^2_{H^s(X)} \, \lambda^{\frac{2s-\varepsilon-n}{2s-\varepsilon}}.$$

(2) If X is bounded, ρ_X is the Lebesgue measure on X and $s \le 2$, then there exists a constant $C_{s,n,X}$ independent of λ such that

$$\widetilde{\mathcal{D}}(\lambda) \le C_{s,n,X} \|f_\rho\|^2_{H^s(X)} \, \lambda^{\frac{2s}{2s+n}}. \tag{38}$$

Proof For the least square loss, we have

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}(X)}. \tag{39}$$

Since X has a Lipschitz boundary, we know (Stein, 1970) that there exists an extension function $\widetilde{f}_\rho \in H^s(\mathbb{R}^n)$ and an absolute constant $C_{s,X}$ such that

$$\widetilde{f}_\rho|_X = f_\rho \quad \text{and} \quad \|\widetilde{f}_\rho\|_{H^s(\mathbb{R}^n)} \le C_{s,X} \|f_\rho\|_{H^s(X)}. \tag{40}$$

Define the normalized kernel $\widetilde{K}_\sigma = (\sqrt{\pi}\sigma)^{-n} K_\sigma$. Let

$$f^\sigma_\rho(x) = (\sqrt{\pi}\sigma)^{-n} L_{K_\sigma}(\widetilde{f}_\rho)(x) = \int_{\mathbb{R}^n} \widetilde{K}_\sigma(x,y)\, \widetilde{f}_\rho(y) dy, \qquad x \in \mathbb{R}^n.$$

Then we know that f^σ_ρ belongs to $\mathcal{H}_{K_\sigma}(\mathbb{R}^n)$ by Lemma 21. Combined with the fact (29), it follows that $g^\sigma_\rho := f^\sigma_\rho|_X \in \mathcal{H}_{K_\sigma}(X)$. Take $f = g^\sigma_\rho$ in the definition of the regularization error $\widetilde{\mathcal{D}}(\lambda)$. We see by (39) that

$$\widetilde{\mathcal{D}}(\lambda) \le \inf_{\sigma\in(0,\infty)} \{ \|g^\sigma_\rho - f_\rho\|^2_{L^2_{\rho_X}(X)} + \lambda \|g^\sigma_\rho\|^2_{K_\sigma} \}. \tag{41}$$

By the relations (29) and (30) on the restriction of RKHS and Lemma 21, we see that

$$\|g^\sigma_\rho\|_{K_\sigma} \le \|f^\sigma_\rho\|_{\mathcal{H}_{K_\sigma}(\mathbb{R}^n)} = (\sqrt{\pi}\sigma)^{-n} \|L_{K_\sigma} \widetilde{f}_\rho\|_{\mathcal{H}_{K_\sigma}(\mathbb{R}^n)} \le (\sqrt{\pi}\sigma)^{-n/2} \|\widetilde{f}_\rho\|_{L^2(\mathbb{R}^n)}.$$

Together with (40), it yields

$$\|g^\sigma_\rho\|_{K_\sigma} \le C_{s,X} (\sqrt{\pi}\sigma)^{-n/2} \|f_\rho\|_{H^s(X)}. \tag{42}$$

To bound the right hand side of (41), we need the Fourier transform defined for $f \in L^1(\mathbb{R}^n)$ as

$$\widehat{f}(\xi) = \int_{\mathbb{R}^n} f(x) e^{-ix\cdot\xi} dx, \qquad \xi \in \mathbb{R}^n.$$

It has a natural extension to $L^2(\mathbb{R}^n)$ satisfying $\|\widehat{f}\|_{L^2(\mathbb{R}^n)} = (2\pi)^{n/2}\|f\|_{L^2(\mathbb{R}^n)}$ (the Plancherel formula). The norm for functions in the Sobolev space $H^s(\mathbb{R}^n)$ can be expressed as $\|f\|_{H^s(\mathbb{R}^n)} = (2\pi)^{-n/2}(\int_{\mathbb{R}^n}(1+|\xi|^2)^s|\widehat{f}(\xi)|^2 d\xi)^{1/2}$. One nice property of the Fourier transform says that the Fourier transform of the convolution $f*g(x) = \int_{\mathbb{R}^n} f(x-y)g(y)dy$ equals $\widehat{f}(\xi)\widehat{g}(\xi)$. It implies that $\widehat{f_\rho^\sigma}(\xi) = e^{-\sigma^2|\xi|^2/4}\widehat{\widetilde{f}_\rho}(\xi)$ since the Fourier transform of the function $(\sqrt{\pi}\sigma)^{-n}e^{-|x|^2/\sigma^2}$ is $e^{-\sigma^2|\xi|^2/4}$.

(1) For any marginal distribution ρ_X, there holds $\|f\|_{L^2_{\rho_X}} \le \|f\|_{C(X)}$. Then the first inequality follows from (39).

Since $X \subseteq \mathbb{R}^n$ and $(n+\varepsilon)/2 > n/2$, we know that the Sobolev space $H^{(n+\varepsilon)/2}(\mathbb{R}^n)$ can be embedded into $C(\mathbb{R}^n)\cap L^\infty(\mathbb{R}^n)$, there exists a constant $C'_{\varepsilon,X}$ such that $\|f\|_{L^\infty(\mathbb{R}^n)} \le C'_{\varepsilon,X}\|f\|_{H^{(n+\varepsilon)/2}(\mathbb{R}^n)}$. It follows that

$$\|g_\rho^\sigma - f_\rho\|_{C(X)} \le \|f_\rho^\sigma - \widetilde{f}_\rho\|_{L^\infty(\mathbb{R}^n)} \le C'_{\varepsilon,X}\|f_\rho^\sigma - \widetilde{f}_\rho\|_{H^{(n+\varepsilon)/2}(\mathbb{R}^n)}.$$

Write

$$\|f_\rho^\sigma - \widetilde{f}_\rho\|^2_{H^{(n+\varepsilon)/2}(\mathbb{R}^n)} = (2\pi)^{-n}\int_{\mathbb{R}^n}(1+|\xi|^2)^{(n+\varepsilon)/2}\left|(e^{-\sigma^2|\xi|^2/4}-1)\widehat{\widetilde{f}}_\rho(\xi)\right|^2 d\xi.$$

Since $\frac{s}{2} - \frac{n+\varepsilon}{4} < 1$, we have $|e^{-\sigma^2|\xi|^2/4}-1| \le (\sigma^2|\xi|^2/4)^{\frac{s}{2}-\frac{n+\varepsilon}{4}}$. Hence

$$\begin{aligned}\|f_\rho^\sigma - \widetilde{f}_\rho\|^2_{H^{(n+\varepsilon)/2}(\mathbb{R}^n)} &\le \tfrac{\sigma^{2s-(n+\varepsilon)}}{(2\pi)^n}\int_{\mathbb{R}^n}(1+|\xi|^2)^{(n+\varepsilon)/2}(|\xi|^2)^{s-(n+\varepsilon)/2}|\widehat{\widetilde{f}}_\rho(\xi)|^2 d\xi \\ &\le \tfrac{\sigma^{2s-(n+\varepsilon)}}{(2\pi)^n}\int_{\mathbb{R}^n}(1+|\xi|^2)^s|\widehat{\widetilde{f}}_\rho(\xi)|^2 d\xi = \sigma^{2s-(n+\varepsilon)}\|\widetilde{f}_\rho\|^2_{H^s(\mathbb{R}^n)}.\end{aligned}$$

In connection with (40), this implies that

$$\|g_\rho^\sigma - f_\rho\|_{C(X)} \le C'_{\varepsilon,X}\sigma^{s-(n+\varepsilon)/2}C_{s,X}\|f_\rho\|_{H^s(X)}.$$

Combining with (42) and choosing $\sigma = \lambda^{1/(2s-\varepsilon)}$ this proves the first statement of the proposition.

(2) If X is bounded and ρ_X is the Lebesgue measure on X, then $\|g_\rho^\sigma - f_\rho\|^2_{L^2_{\rho_X}} = \|g_\rho^\sigma - f_\rho\|^2_{L^2(X)}$ and by the Plancherel formula,

$$\|g_\rho^\sigma - f_\rho\|^2_{L^2(X)} \le \|f_\rho^\sigma - \widetilde{f}_\rho\|^2_{L^2(\mathbb{R}^n)} = (2\pi)^{-n}\int_{\mathbb{R}^n}\left|(e^{-\sigma^2|\xi|^2/4}-1)\widehat{\widetilde{f}}_\rho(\xi)\right|^2 d\xi.$$

Observe from the restriction $s \le 2$ that $1-e^{-t} \le (1-e^{-t})^{s/2} \le t^{s/2}$ for $t > 0$. Applying this to $t = \sigma^2|\xi|^2/4$ we obtain

$$\begin{aligned}\|g_\rho^\sigma - f_\rho\|^2_{L^2(X)} &\le \tfrac{\sigma^{2s}}{(2\pi)^n 4^s}\int_{\mathbb{R}^n}|\xi|^{2s}|\widehat{\widetilde{f}}_\rho(\xi)|^2 d\xi \\ &\le \sigma^{2s}\|\widetilde{f}_\rho\|^2_{H^s(\mathbb{R}^n)} \le C^2_{s,X}\sigma^{2s}\|f_\rho\|^2_{H^s(X)},\end{aligned}$$

where we have used (40) in the last inequality. Putting this estimate and (42) into (41) gives us the second statement of the proposition corresponding to the choice $\sigma = \lambda^{1/(2s+n)}$. ■

Finally, we are able to derive error rates for the multi-kernel regularization scheme (4) associated with Gaussian kernels (8) with flexible variances. This is done by putting the improved regularization error bound in Proposition 22 into the total error bound in Theorem 20. We demonstrate the approach for regression with least square loss and for classification with hinge loss.

6. Error Rates with Gaussians

In this section we prove Examples 1 and 2. First, let us consider Example 1, that is, the case of the least square regularized regression: $Y = \mathbb{R}$. If $|y| \le M_0$ almost surely, the least square loss $V(y,s) = (y-s)^2$ is admissible with respect to ρ and $M = \|V(y,0)\|_{L^\infty_\rho(Z)} \le M_0^2$, $C_\lambda \le 2M_0(1+\kappa/\sqrt{\lambda})$. By Theorem 20 and the special property (39) of the loss function, we immediately get the following result.

Proposition 23 *Let $V(y,t) = (y-t)^2$, $X \subseteq \mathbb{R}^n$ and $\{K^\sigma\}$ be given by (8). Define $f_{\mathbf{z},\lambda}$ by (14). If $0 < \lambda \le 1$, then there exists a constant $\widetilde{C}$ independent of m, λ such that*

$$\mathbb{E}\Big[\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho^2\Big] \le \widetilde{C}\Big(\frac{\log^2 m}{m\lambda^4}\Big)^{1/4} + \widetilde{\mathcal{D}}(\lambda). \tag{43}$$

Using this proposition, we can provide the proof of Example 1.

Proof of Example 1. Since the set (8) of Gaussian kernels with general variances $\sigma = (\sigma_1, \ldots, \sigma_n) \in (0,\infty)^n$ contains the set (9) of Gaussian kernels with isotropic variances, we see that

$$\widetilde{\mathcal{D}}(\lambda) \le \inf_{\sigma\in(0,\infty)} \inf_{f\in\mathcal{H}_{K_\sigma}} \Big\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|^2_{K_\sigma}\Big\}.$$

(1) When $n/2 < s \le n/2+2$ and $0 < \varepsilon < 2s-n$, we know from Proposition 22 that

$$\widetilde{\mathcal{D}}(\lambda) \le C_{\varepsilon,s,X}\|f_\rho\|^2_{L^2_{\rho_X}(X)}\lambda^{\frac{2s-\varepsilon-n}{2s-\varepsilon}}.$$

Putting this into (43) and choosing $\lambda = m^{-\frac{2s-\varepsilon}{4(4s-n-2\varepsilon)}}$ verifies the first error estimate in Example 1.

(2) If X is bounded, ρ_X is the Lebesgue measure on X, and $s \le 2$, we apply the bound (38) in Proposition 22. Together with the above inequality, we know that

$$\widetilde{\mathcal{D}}(\lambda) \le C_{s,n,X}\|f_\rho\|^2_{L^2_{\rho_X}(X)}\lambda^{\frac{2s}{2s+n}}.$$

In connection with the error bound (43) we see that when $\lambda = m^{-\frac{2s+n}{4(4s+n)}}$, the error estimate in Part (2) holds true. ■

Now we move on to establish Example 2 for the regularized classification with the hinge loss $V(y,s) = (1-ys)_+$. In this case we take $Y = \{1,-1\}$. It is easy to see that V is admissible with $M = 1$ and $C_\lambda = 1$. An important relation between the excess misclassification error and the excess error was given by Zhang (2004) as

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_c), \qquad \forall f : X \to \mathbb{R}.$$

Then the following result is an easy consequence of Theorem 20.

Proposition 24 *Let $Y = \{1,-1\}$, $V(y,t) = (1-yt)_+$, $X \subseteq \mathbb{R}^n$ and $\{K^\sigma\}$ be given by (8). Define $f_{\mathbf{z},\lambda}$ by (15). If $0 < \lambda \le 1$, then there exists a constant C' independent of m, λ such that*

$$\mathbb{E}\Big[\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)\Big] \le \sqrt{\frac{C'}{\lambda}}\Big(\frac{\log^2 m}{m}\Big)^{1/4} + \widetilde{\mathcal{D}}(\lambda).$$

We are in a position to prove Example 2 by Proposition 24.

Proof of Example 2. It was shown by Chen et al. (2004) that if ρ is separable with some exponent $\theta > 0$ then there exists some constant c' such that $\widetilde{\mathcal{D}}(\lambda) \le c'\lambda^{\frac{\theta}{2+\theta}}$. Choosing $\lambda = m^{-\frac{2+\theta}{2(2+3\theta)}}$ gives us the desired result. ■

Acknowledgments

We would like to thank the referees for their constructive suggestions and comments. This work was supported by the Research Grants Council of the Hong Kong [Project No. CityU 103704], City University of Hong Kong [Project No. 7001983], and the National Science Fund for Distinguished Young Scholars of China [Project No. 10529101]. The corresponding author is Ding-Xuan Zhou. Yiming Ying is on leave from Institute of Mathematics, Chinese Academy of Science, Beijing 100080, P. R. China.

Appendix A.

This appendix includes complete proofs of two lemmas which are essentially proved in Alon et al. (1997); Ledoux and Talagrand (1991) with slightly different forms.

Proof of Lemma 12. Part (a) is an easy consequence of Lemma 2.4 in Alon et al. (1997).

Set $d_0 = \min\{m, d\}$. As in the proof of Lemma 3.5 of Alon et al. (1997), one can bound the empirical covering number by packing numbers which can then be estimated by Lemma 3.3 in Alon et al. (1997) as

$$\sup_{\mathbf{x}\in X^m} \mathcal{N}_\infty(\mathcal{G}, \mathbf{x}, \varepsilon) \le 2\left(m\left(\frac{2}{\varepsilon}\right)^2\right)^{\log y + 1},$$

where

$$y = \sum_{i=1}^{d_0} \binom{m}{i} \left(\frac{2}{\varepsilon}\right)^i \le \left(\frac{2}{\varepsilon}\right)^{d_0} \left(\frac{em}{d_0}\right)^{d_0} \le \left(\frac{2em}{\varepsilon}\right)^{d_0} \le \left(\frac{2em}{\varepsilon}\right)^{d}.$$

This verifies Part (b). ■

Proof of Lemma 16. The first statement is immediate from the definition of Rademacher averages.

For the second statement, we use Theorem 4.12 in Ledoux and Talagrand (1991). It tells us the following result: If T is a bounded subset of $\mathbb{R}^m$, each function ϕ_i with the Lipschitz constant not more than 1 satisfies $\phi_i(0) = 0$, and a function $G : \mathbb{R}_+ \to \mathbb{R}_+$ is convex and nondecreasing, then there holds

$$\mathbb{E}_\varepsilon G\Big(\frac{1}{2}\sup_{t\in T}|\sum_{i=1}^m \varepsilon_i \phi_i(t_i)|\Big) \le \mathbb{E}_\varepsilon G\Big(\sup_{t\in T}|\sum_{i=1}^m \varepsilon_i t_i|\Big). \tag{44}$$

Fixed $\{x_i\}_{i=1}^m$. Then $T := \{(c_1 f(x_1), \cdots, c_m f(x_m)) : f \in F\}$ is a bounded subset of $\mathbb{R}^m$ since F is uniformly bounded. Applying (44) with $G(u) = u$ and $\widetilde{\phi}_i(x) = \phi_i(x/c_i)$, we have

$$\begin{aligned}\mathbb{E}_\varepsilon\Big[\sup_{f\in F}|\sum_{i=1}^m \varepsilon_i \phi_i(f(x_i))|\Big] &= \mathbb{E}_\varepsilon\Big[\sup_{t\in T}|\sum_{i=1}^m \varepsilon_i \widetilde{\phi}_i(t_i)|\Big] \le 2\mathbb{E}_\varepsilon\Big[\sup_{t\in T}|\sum_{i=1}^m \varepsilon_i t_i|\Big] \\ &= 2\mathbb{E}_\varepsilon\Big[\sup_{f\in F}\Big|\sum_{i=1}^m c_i \varepsilon_i f(x_i)|\Big]\end{aligned}$$

which proves our second statement. ■

Appendix B.

The arguments in this paper do not change much if we replace a minimizer of the optimization problem (4) by an ε-*minimizer* $f^{\varepsilon}_{\mathbf{z},\lambda} \in \mathcal{H}_{K^\sigma}$ (with some $\sigma \in \Sigma$) satisfying

$$\frac{1}{m}\sum_{i=1}^{m} V(y_i, f^{\varepsilon}_{\mathbf{z},\lambda}(x_i)) + \lambda \| f^{\varepsilon}_{\mathbf{z},\lambda}\|^2_{K^\sigma} \le \min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K^\sigma}} \left\{ \frac{1}{m}\sum_{i=1}^{m} V(y_i, f(x_i)) + \lambda \| f\|^2_{K^\sigma} \right\} + \varepsilon.$$

The existence of an ε-minimizer for any given $0 < \varepsilon \le 1$ can be seen in Wu et al. (2007) where it was shown that $\min_{f\in\mathcal{H}_{K^\sigma}} \{\frac{1}{m}\sum_{i=1}^{m} V(y_i, f(x_i)) + \lambda \| f\|^2_{K^\sigma}\}$ is continuous as a function of $\sigma \in \Sigma$ if $K^\sigma(x,x')$ is continuous with respect to $\sigma \in \Sigma$ for each fixed pair $(x,x') \in X \times X$.

In this appendix, we verify the existence for the scheme (4) involving the least-square loss and Gaussians (9) with flexible variances under some mild conditions on the sample $\mathbf{z}$. Note that the shortest distance between pairs of distinct points from $\{x_1,\ldots,x_m\}$ is achieved by some pair (x_{i_1}, x_{i_2}) with $i_1 \ne i_2$. The condition for our existence result is that such a minimizing pair is unique and the sample values y_{i_1}, y_{i_2} have the same sign. Since x_{i_1} and x_{i_2} are close, this assumption of having the same sign for the sample values is reasonable.

Proposition 25 *Let $\lambda > 0$, X be a compact subset of $\mathbb{R}^n$ and $K_\sigma(x,y) = \exp\{-\frac{|x-y|^2}{\sigma^2}\}$ for $x,y \in X$. Consider the scheme with $0 < b < \infty$*

$$f_{\mathbf{z},\lambda} := \arg \min_{\sigma\in(0,b]} \min_{f\in\mathcal{H}_{K^\sigma}} \left\{ \frac{1}{m}\sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \| f\|^2_{K^\sigma} \right\}. \tag{45}$$

If we can find $\{i_1 \ne i_2\} \subset \{1,\ldots,m\}$ such that

$$y_{i_1} y_{i_2} > 0 \quad \textit{and} \quad |x_{i_1} - x_{i_2}| \le |x_i - x_j| \textit{ for any } \{i \ne j\} \subset \{1,\ldots,m\}$$

with equality valid only for $(i,j) = (i_1,i_2)$ or (i_2,i_1), then the existence of a solution to (45) holds true.

Proof For $\sigma > 0$ we denote

$$e_{\mathbf{z},\lambda}(\sigma) = \min_{f\in\mathcal{H}_{K^\sigma}} \left\{ \frac{1}{m}\sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \| f\|^2_{K^\sigma} \right\}.$$

The minimizer of this one-layer optimization problem exists, is unique, and can be expressed as $f^\sigma_{\mathbf{z},\lambda} = \sum_{i=1}^m c_i^\sigma K^\sigma_{x_i}$. The coefficient vector $c^\sigma := (c_i^\sigma)_{i=1}^m$ is the solution of the linear system

$$([K^\sigma]_{\mathbf{x}} + m\lambda I_m)\, c = \mathbf{y}$$

where $\mathbf{y}$ is the vector $(y_i)_{i=1}^m$ and $[K^\sigma]_{\mathbf{x}}$ is the $m \times m$ matrix $(K^\sigma(x_i,x_j))_{i,j=1}^m$. Each main diagonal entry of $[K^\sigma]_{\mathbf{x}}$ is 1. A simple computation yields $\frac{1}{m}\sum_{i=1}^m (f^\sigma_{\mathbf{z},\lambda}(x_i) - y_i)^2 = m\lambda^2 \|c^\sigma\|^2, \| f^\sigma_{\mathbf{z},\lambda}\|^2_{K^\sigma} = (c^\sigma)^T [K^\sigma]_{\mathbf{x}} c^\sigma$ and

$$e_{\mathbf{z},\lambda}(\sigma) = \lambda \mathbf{y}^T \left([K^\sigma]_{\mathbf{x}} + m\lambda I_m\right)^{-1} \mathbf{y}.$$

In the following, we will show that $\lim_{\sigma\to 0} e_{\mathbf{z},\lambda}(\sigma)$ is strictly larger than $e_{\mathbf{z},\lambda}(\sigma)$ for any $\sigma \in (0,\sigma_2]$ with some $\sigma_2 > 0$. Thereby, the minimizer of $e_{\mathbf{z},\lambda}$ should be achieved at a positive number in $(0,b]$ since $e_{\mathbf{z},\lambda}(\cdot)$ is a continuous function.

Without loss of generality, we assume that $i_1 = 1, i_2 = 2$. Denote $a_\sigma = \exp\{-\frac{|x_1-x_2|^2}{\sigma^2}\}$. We see that $a_\sigma > 0$ and $\lim_{\sigma\to 0} a_\sigma = 0$. Our assumption on the distances $|x_i - x_j| \geq |x_1 - x_2|$ $(i \neq j)$ tells us that the off-diagonal entries of $[K^\sigma]_{\mathbf{x}}$ decays to 0 faster than a_σ except for the two entries at $(1,2)$ and $(2,1)$. That is,

$$[K^\sigma]_{\mathbf{x}} + m\lambda I_m = A_\sigma + a_\sigma B_\sigma, \quad A_\sigma := \begin{bmatrix} m\lambda+1 & a_\sigma & 0 \\ a_\sigma & m\lambda+1 & 0 \\ 0 & 0 & (m\lambda+1)I_{m-2} \end{bmatrix}$$

and $\lim_{\sigma\to 0} B_\sigma = 0$. The inverse of A_σ has a nice form

$$A_\sigma^{-1} = \begin{bmatrix} \frac{m\lambda+1}{(m\lambda+1)^2-a_\sigma^2} & -\frac{a_\sigma}{(m\lambda+1)^2-a_\sigma^2} & 0 \\ -\frac{a_\sigma}{(m\lambda+1)^2-a_\sigma^2} & \frac{m\lambda+1}{(m\lambda+1)^2-a_\sigma^2} & 0 \\ 0 & 0 & \frac{1}{m\lambda+1} I_{m-2} \end{bmatrix}.$$

Recall that the norm of an $m \times m$ matrix A is defined by $\|A\| = \sup\{\|Ax\| : \|x\| = 1\}$. If in addition, A is symmetric, then $\|A\| = \max\{|\lambda_i| : i = 1,2,\ldots,m\}$ where $\{\lambda_i : i = 1,2,\cdots,m\}$ are the eigenvalues of A. Hence,

$$\|A_\sigma^{-1}\| \leq \frac{1}{m\lambda + 1 - a_\sigma} \leq \frac{1}{m\lambda}.$$

Moreover, since $\lim_{\sigma\to 0} B_\sigma = 0$, there exists $0 < \sigma_1 \leq b$ such that

$$\|A_\sigma^{-1/2} B_\sigma A_\sigma^{-1/2}\| \leq 1 \quad \forall 0 < \sigma \leq \sigma_1.$$

Therefore, for any $0 < \sigma \leq \sigma_1$ there holds

$$\|(I + a_\sigma A_\sigma^{-1/2} B_\sigma A_\sigma^{-1/2})^{-1}\| \leq \frac{1}{1-a_\sigma}.$$

If we write $([K^\sigma]_{\mathbf{x}} + m\lambda I_m)^{-1}$ as $A_\sigma^{-1/2}(I + a_\sigma A_\sigma^{-1/2} B_\sigma A_\sigma^{-1/2})^{-1} A_\sigma^{-1/2}$, then

$$([K^\sigma]_{\mathbf{x}} + m\lambda I_m)^{-1} - A_\sigma^{-1} = -a_\sigma A_\sigma^{-1/2}(I + a_\sigma A_\sigma^{-1/2} B_\sigma A_\sigma^{-1/2})^{-1} A_\sigma^{-1/2} B_\sigma A_\sigma^{-1}.$$

Consequently,

$$\|\,([K^\sigma]_{\mathbf{x}} + m\lambda I_m)^{-1} - A_\sigma^{-1}\| \leq \frac{a_\sigma}{1-a_\sigma}\Big(\frac{1}{m\lambda}\Big)^2 \|B_\sigma\|.$$

This implies that

$$e_{\mathbf{z},\lambda}(\sigma) \leq \lambda \mathbf{y}^T A_\sigma^{-1} \mathbf{y} + \lambda \frac{a_\sigma \|B_\sigma\|}{1-a_\sigma}\Big(\frac{1}{m\lambda}\Big)^2.$$

But

$$\lambda \mathbf{y}^T A_\sigma^{-1} \mathbf{y} = \frac{\lambda}{m\lambda+1}\|\mathbf{y}\|^2 + \frac{\lambda a_\sigma^2 (y_1^2 + y_2^2)}{(m\lambda+1)((m\lambda+1)^2 - a_\sigma^2)} - \frac{2\lambda a_\sigma y_1 y_2}{(m\lambda+1)^2 - a_\sigma^2},$$

which means that $e_{\mathbf{z},\lambda}(\sigma) - \frac{\lambda}{m\lambda+1}\|\mathbf{y}\|^2$ is bounded by

$$\lambda a_\sigma \Big[\frac{\|B_\sigma\|}{1-a_\sigma}\Big(\frac{1}{m\lambda}\Big)^2 + \frac{a_\sigma(y_1^2+y_2^2)}{(m\lambda+1)((m\lambda+1)^2-a_\sigma^2)} - \frac{2y_1y_2}{(m\lambda+1)^2-a_\sigma^2}\Big].$$

Using the properties that $y_1 y_2 > 0$, $\lim_{\sigma\to 0} a_\sigma = 0$ and $\lim_{\sigma\to 0} \|B_\sigma\| = 0$ again, we know there exists $0 < \sigma_2 \le \sigma_1$ such that

$$e_{\mathbf{z},\lambda}(\sigma) < \frac{\lambda}{m\lambda+1}\|\mathbf{y}\|^2, \qquad \forall 0 < \sigma \le \sigma_2.$$

We also observe that $\lim_{\sigma\to 0}[K^\sigma]_{\mathbf{x}} + m\lambda I_m = (m\lambda+1)I_m$. Hence

$$\lim_{\sigma\to 0} e_{\mathbf{z},\lambda}(\sigma) = \frac{\lambda}{m\lambda+1}\|\mathbf{y}\|^2 > e_{\mathbf{z},\lambda}(\sigma) \qquad \forall 0 < \sigma \le \sigma_2.$$

It means that the infimum in (45) cannot be achieved as $\sigma \to 0$. By the continuity of $e_{\mathbf{z},\lambda}(\cdot)$, the existence of a solution to (45) follows from that of the optimization problem for σ lying in a compact subset of $(0, b]$ proved in Wu et al. (2007). ■

The above existence result largely depends on the least square loss and the assumption on the data. It remains an open problem on how to prove the existence of the minimizer of the multi-kernel scheme (4) associated with general loss functions and data.

References

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

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.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

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

N. Cristianini, J. Shawe-Taylor, and C. Campbell. Dynamically adapting kernels in support vector machines. In *Advances in Neural Information Processing Systems* 11 (M. S. Kearns, S. A. Solla, and D. A. Cohn, eds), MIT Press, 1999.

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

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002.

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

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, to appear, 2007.

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.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics 63, Cambridge University Press, 1999.

R. M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli Classes. *Journal of Theoretical Probability*, 4:485–510, 1991.

T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proceedings of 10th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004. http://www.cs.ucl.ac.uk/staff/M.Pontil/reading/mt-kdd.pdf

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

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. L. Bartlett, P. L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal 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.

J. Li, and A. Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems* 12 (S. A. Solla, K. L. Todd, K.-R. Müller eds.), MIT Press, 2000.

P. Niyogi and F. Girosi. On the relationships between generalization error, hypothesis complexity and sample complexity for radial basis functions. *Neural Computation*, 8:819–842, 1996.

V. S. Pugachev, and I. N. Sinitsyn. *Lectures on Functional Analysis and Applications*. World Scientific, Singapore, 1999.

A. Rakhlin, D. Panchenko, and S. Mukherjee. Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229, 2005.

B. Schölkopf, B. R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory, Lecture Notes in Artificial Intelligence*, 2111: 416–426, 2001.

J. Shawe-Taylor, P. L. Bartlett, S. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.

S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1:17–41, 2003.

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

S. Smale and D. X. Zhou. Shannon sampling II. Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19:285–302, 2005.

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

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

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

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

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

A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, 1996.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Q. Wu, Y. Ying, and D. X. Zhou. Multi-kernel Regularized Classifiers. *Journal of Complexity*, forthcoming.

G. B. Ye and D. X. Zhou. Learning and approximation by Gaussians on Riemannian manifolds. *Advances in Computational Mathematics*, forthcoming.

K. Yosida. *Functional Analysis*. 6th edition, Springer-Verlag, 1980.

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

D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18:739–767, 2002.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49:1743–1752, 2003.

D. X. Zhou and K. Jetter. Approximation with polynomial kernels and SVM classifiers. *Advances in Computational Mathematics*, 25:323–344, 2006.