Error Bounds for Learning the Kernel

Charles A. Micchelli
Department of Mathematics and Statistics
State University of New York
The University at Albany
1400 Washington Avenue, Albany, NY 12222, USA

Massimiliano Pontil
Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genoa, Italy
and
Department of Computer Sciences
University College London
Gower Street, London WC1E, England, UK

Qiang Wu
Department of Mathematical Sciences
Middle Tennessee State University
1301 E Main Street, Murfreesboro, TN 37132, USA
E-mail: *qwu@mtsu.edu*

Ding-Xuan Zhou
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong, P.R. China
E-mail: *mazhou@cityu.edu.hk*

Abstract

The problem of learning the kernel function has received considerable attention in machine learning. Much of the work has focused on kernel selection criteria, particularly on minimizing a regularized error functional over a prescribed set of kernels. Empirical studies indicate that this approach can enhance statistical performance and is computationally feasible. In this paper, we present a theoretical analysis of its generalization error. We establish for a wide variety of classes of kernels, such as the set of *all* multivariate Gaussian kernels, that this learning method generalizes well and, when the regularization parameter is appropriately chosen, it is consistent. A central role in our analysis is played by the interaction between the sample error and the approximation error.

Keywords: Kernel, regularization, error decomposition, least square, support vector machine, Gaussians

Mathematics Subject Classification 2010: 68T05, 68Q32

1 Introduction

A widely used approach for learning a function from empirical data consists in minimizing a regularization functional which models a trade-off between an error term, measuring the fit to the data, and a smoothness term, measuring the function complexity. Specifically, in this paper we focus on learning methods which, given a set of examples $\mathbf{z} = \{(x_j, y_j) : j \in \mathbb{N}_m\} \subseteq Z := X \times Y$, sampled *i.i.d.* according to an unknown distribution ρ supported on $X \times Y$, where $Y \subseteq \mathbb{R}$, estimates a real-valued function by solving the variational problem

$$\min_{f \in \mathcal{H}_K} \mathcal{E}_\lambda(f, K) \tag{1.1}$$

where $\mathcal{E}_\lambda(f, K) := \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2$, $\mathcal{E}_{\mathbf{z}}(f)$ is the *empirical error* of the function f on the data $\mathbf{z}$, namely,

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{j \in \mathbb{N}_m} \ell(y_j, f(x_j))$$

as measured by a prescribed nonnegative *loss function* $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, λ is a positive parameter, $\mathbb{N}_m := \{1, \ldots, m\}$ and $\mathbb{R}_+ = \{t : t \geq 0\}$. The minimum in (1.1) is taken over all functions $f \in \mathcal{H}_K$, a *reproducing kernel Hilbert space* (RKHS) with reproducing kernel K [5]. This approach has a long history. It has been studied, from different perspectives, in statistics, in optimal estimation, and more recently, has been a focus of attention in machine learning theory, see, for example, [13, 25] and the references therein for a discussion. The choice of the loss function ℓ leads to different learning methods among which the prominent ones are square loss regularization and support vector machines.

When the kernel K is fixed, the algorithm (1.1) is well understood, see, for example, [10, 12, 32, 31, 36, 43] and the references therein. The choice of the parameter λ plays a central role in the method as it allows one to control the smoothness of the function f, thereby avoiding overfitting. Theoretically, it is chosen by a trade-off between the estimates for the sample error and the approximation error, see, for example, [12, 10, 32, 31].

A more challenging task is the choice of the kernel. This has motivated various studies addressing the problem of minimizing functional (1.1) not only over $f \in \mathcal{H}_K$ but also over K in some prescribed class $\mathcal{K}$ of kernels [8, 18, 22, 24, 27, 30]. That is, we consider the variational problem

$$(K_{\mathbf{z}}, f_{\mathbf{z}}) := \operatorname{argmin}\Big\{\mathcal{E}_\lambda(f, K) : K \in \mathcal{K},\ f \in \mathcal{H}_K\Big\}. \tag{1.2}$$

When the set $\mathcal{K}$ is a convex and bounded subset of the set of positive definite kernels, this problem can be reformulated as a regularized empirical error minimization problem, in which the regularizer is a Banach space norm induced by the class $\mathcal{K}$. In particular if $\mathcal{K}$ is the convex hull of finitely many basic kernels, then that norm is a mixed norm involving the reproducing kernel norms of the basic kernels [28]. This point of view provides a useful interpretation for the problem of learning the kernel, however it is not central in the present paper and will be addressed further.

Motivated by the need to improve the approximation error, this scheme was also studied in [36]. Practical experience with this method [2, 3, 6, 8, 22, 23, 24] indicates that it can enhance the performance of the learning algorithm and is computationally efficient to solve.

For a discussion of the hypotheses on $\mathcal{K}$ which ensure that the minimum above exists see [27, 36, 28].

In this paper we focus on the problem of bounding the *generalization error* of $f_{\mathbf{z}}$, namely, $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\ell^*)$, where $\mathcal{E}(f)$ is the *expected error* of f, $\mathcal{E}(f) := \mathbb{E}\,\ell(y, f(x))$, the expectation $\mathbb{E}$ being over the probability measure ρ, and f_ℓ^* is the *target function* defined as $f_\ell^* := \arg\min \mathcal{E}(f)$, where the minimum is taken over all measurable functions. Our analysis holds for a wide class of kernels $\mathcal{K}$ with two basic assumptions. First, we require that the class $\mathcal{K}$ is uniformly bounded, that is,

$$\kappa = \sup_{K\in\mathcal{K}} \sup_{x\in X} \sqrt{K(x,x)} < \infty.$$

Second, we demand that all kernels in $\mathcal{K}$ are continuous. Therefore, it follows by the reproducing kernel property of $\mathcal{H}_K$ [5], for all $K \in \mathcal{K}$ and $f \in \mathcal{H}_K$, that

$$\|f\|_\infty := \max_{x\in X} |f(x)| \le \kappa \|f\|_K, \tag{1.3}$$

an inequality which we will use repeatedly in our subsequent analysis.

In our analysis a probabilistic upper bound on the sample error is achieved by estimating the Rademacher complexity of the set

$$\mathcal{K}_0 = \{K(x,\cdot) : K \in \mathcal{K}, x \in X\}. \tag{1.4}$$

The results are presented in Section 3 and proved in Section 4. Earlier analysis similar to that which appears here may be found in [8, 22], however, as far as we know we achieve greater generality than is available so far. In particular our work applies to continuously parameterized kernel classes, such as those investigated in [3, 42]. Recent papers have addressed the problem of computing the Rademancher average of function classes given by the union of finitely many reproducing kernels or their convex hull [11, 25, 21]. A more general approach is presented in [40, 41], in which Rademacher chaos complexity is employed to get faster rates under certain conditions. In Section 5 we apply these bounds to the important case of Gaussian kernels with *arbitrary* variance and illustrate our results in the case of support vector machines and regularized least squares. Section 6 presents final remarks and future direction of research.

2 Preventing overfitting?

We proceed our presentation of the probabilistic analysis for the generalization error by proving a positive lower bound for the regularization functional $\mathcal{E}_\lambda$ in (1.1) which is valid for any set of uniformly bounded kernels $\mathcal{K}$. This observation is not relevant for our subsequent error bounds, however it suggests that overfitting would not occur provided the regularization parameter is appropriately chosen.

Below, if $K \in \mathcal{K}$ we denote by $K(\mathbf{x})$ the $m \times m$ Gram matrix $(K(x_i, x_j) : i, j \in \mathbb{N}_m)$ where $\mathbf{x} = (x_i : i \in \mathbb{N}_m) \in X^m$. We also define the vector $\mathbf{y} := (y_i : i \in \mathbb{N}_m) \in \mathbb{R}^m$.

Proposition 2.1. *Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a nonnegative loss function with the property that for any $c > 0$ there exists $\delta > 0$ such that*

$$\ell(u, v) \geq \delta |u - v|^2 \text{ for all } u, v \in \mathbb{R} \text{ satisfying } |u - v| \leq c.$$

If $\mathbf{y} \neq 0$ then for every $\lambda > 0$ and $\mathbf{z} \in Z^m$ then there exists $\mu > 0$ such that $\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \mu$.

Proof. We note that

$$\lambda \|f_\mathbf{z}\|^2_{K_\mathbf{z}} \leq \mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \leq \mathcal{E}_\lambda(0, K_\mathbf{z}) = \bar{\ell} := \frac{1}{m} \sum_{i \in \mathbb{N}_m} \ell(y_i, 0).$$

Hence, using inequality (1.3), we have, for every $i \in \mathbb{N}_m$, that

$$|f_\mathbf{z}(x_i)| \leq \kappa \|f_\mathbf{z}\|_{K_\mathbf{z}} \leq \kappa \sqrt{\bar{\ell}/\lambda}.$$

We define $\|\mathbf{y}\|_\infty := \max_{i \in \mathbb{N}_m} |y_i|$ and observe that the choice $c = \|\mathbf{y}\|_\infty + \kappa \sqrt{\bar{\ell}/\lambda}$ ensures that $|y_i - f_\mathbf{z}(x_i)| \leq c$ and, so, by hypothesis there is a corresponding $\delta > 0$ such that

$$\ell(y_i, f_\mathbf{z}(x_i)) \geq \delta (y_i - f_\mathbf{z}(x_i))^2.$$

Consequently, we obtain that

$$\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \delta \mathcal{Q}_\mathbf{z}(f_\mathbf{z}) + \lambda \|f_\mathbf{z}\|^2_{K_\mathbf{z}} \geq \delta \mathcal{Q}_{\hat{\lambda}}(K_\mathbf{z})$$

where $\hat{\lambda} := \frac{\lambda}{\delta}$,

$$\mathcal{Q}_\mathbf{z}(f_\mathbf{z}) := \frac{1}{m} \sum_{i \in \mathbb{N}_m} (y_i - f_\mathbf{z}(x_i))^2$$

and

$$\mathcal{Q}_\lambda(K_\mathbf{z}) := \min_{f \in \mathcal{H}_{K_\mathbf{z}}} \{\mathcal{Q}_\mathbf{z}(f) + \lambda \|f\|^2_{K_\mathbf{z}}\}.$$

According to [27, Lemma 3.1] we have that

$$\mathcal{Q}_{\hat{\lambda}}(f_\mathbf{z}) = \hat{\lambda}(\mathbf{y}, (K_\mathbf{z}(\mathbf{x}) + m\hat{\lambda} I)^{-1} \mathbf{y}) \geq \frac{\hat{\lambda} \|\mathbf{y}\|^2}{m(\kappa^2 + m\hat{\lambda})}$$

and the result follows by noting that

$$\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \delta \frac{\hat{\lambda} \|\mathbf{y}\|^2}{m(\kappa^2 + m\hat{\lambda})} = \frac{\delta \lambda \|\mathbf{y}\|^2}{m(\delta \kappa^2 + m\lambda)}.$$

■

Proposition 2.1 applies to the square loss but not to the hinge loss used in support vector machines. To deal with the hinge loss we modify the proof above and obtain the following result.

Proposition 2.2. *Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a nonnegative loss function with the property that for some $c > 0$ there exists a $\delta > 0$ such that*

$$\ell(u, v) \geq \delta |u - v|^2 \text{ for all } u \in Y \text{ and } |v| \leq c.$$

If $\mathbf{y} \neq 0$ for every $\lambda > 0$ and $\mathbf{z} \in Z^m$ then there exists $\mu > 0$ such that $\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \mu$.

Proof. First note if $\|f_\mathbf{z}\|_{K_\mathbf{z}} \geq \frac{c}{\kappa}$ then we have that

$$\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \lambda \|f_\mathbf{z}\|^2_{K_\mathbf{z}} \geq \frac{\lambda c^2}{\kappa^2}.$$

Otherwise, there holds the inequality $\|f_\mathbf{z}\|_{K_\mathbf{z}} \leq \frac{c}{\kappa}$ which implies by (1.3), for every $i \in \mathbb{N}_m$, that

$$|f_\mathbf{z}(x_i)| \leq \kappa \|f_\mathbf{z}\|_{K_\mathbf{z}} \leq c.$$

Therefore, the assumption on the loss function ℓ tells us that there is a $\delta > 0$ such that $\ell(y_i, f_\mathbf{z}(x_i)) \geq \delta(y_i - f_\mathbf{z}(x_i))^2$. We now use the same argument as in the proof of Proposition 2.1 to obtain that

$$\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \tilde{\mu} := \frac{\delta\lambda\|\mathbf{y}\|^2}{m(\delta\kappa^2 + m\lambda)}$$

and our conclusion follows by taking $\mu = \min\left\{\frac{\lambda c^2}{\kappa^2}, \tilde{\mu}\right\}$. ■

The assumption on the loss function in Proposition 2.2 covers both the square loss and the hinge loss given by the formula $\ell(u, v) = (1 - uv)_+ = \max\{0, 1 - uv\}$ which is used in support vector machine classification, see for example [16]. To see this, take $c = \frac{1}{2}$ and recall that for binary classification $Y = \{-1, 1\}$. Consequently, if $u \in Y$ and $|v| \leq \frac{1}{2}$, then $|uv| \leq \frac{1}{2}$ and $\frac{1}{2} \leq |u - v| \leq \frac{3}{2}$ from which it follows that $(1 - uv)_+ \geq (\frac{1}{2}) \geq \frac{2}{9}|u - v|^2$.

The lower bound above says that $\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z})$ is bounded away from zero when the set $\mathcal{K}$ is uniformly bounded. This suggests that, with additional information on the target function and with an appropriate choice of λ our approach may be free of overfitting, a phenomenon which occurs when the empirical error is zero but the expected error in far from zero. We shall confirm this fact by our analysis below.

3 Error bound

In this section, we present our results of generalization error analysis. For this purpose, we require some notation. First, we follow [10, 37, 36] and introduce the projection operator, defined for every measurable function $f : X \to \mathbb{R}$ and positive constant T, as $\pi_T(f)(x) = \text{sgn}(f)(x)\min(|f(x)|, T)$, where $\text{sgn}(f)(x) = 1$, if $f(x) \geq 0$ and -1 otherwise, namely,

$$\pi_T(f)(x) = \begin{cases} -T, & \text{if } f(x) < -T \\ f(x), & \text{if } f(x) \in [-T, T] \\ T, & \text{if } f(x) > T. \end{cases}$$

The constant T is called the projection level and is useful for providing a better estimate for classification but it is not needed for regression problems. Next, we define the *truncated sample error* as

$$\mathcal{S}_{\mathbf{z}}(m,\lambda,f,T)=\big\{\mathcal{E}\left(\pi_T(f_{\mathbf{z}})\right)-\mathcal{E}_{\mathbf{z}}(\pi_T(f_{\mathbf{z}}))\big\}+\big\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(f)\big\}$$

and the corresponding *sample error* as

$$\mathcal{S}_{\mathbf{z}}(m,\lambda,f)=\mathcal{S}_{\mathbf{z}}(m,\lambda,f,\infty)=\big\{\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\big\}+\big\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(f)\big\}.$$

The third quantity we need is the *regularization error* of a function $f\in\mathcal{H}_K$ which is defined as

$$\mathcal{A}(f)=\mathcal{E}(f)-\mathcal{E}(f^*_\ell)+\lambda\|f\|_K^2$$

where, recall $f^*_\ell:=\arg\min\mathcal{E}(f)$. The regularization error of f is a regularized version of the approximation error $\mathcal{E}(f)-\mathcal{E}(f^*_\ell)$. The function f in the above equation can be arbitrarily chosen, however, only proper choices lead to good estimates of the regularization error. A good choice is $f=f^*_\lambda$ where

$$(K^*_\lambda,f^*_\lambda)=\arg\min_{K\in\mathcal{K}}\min_{f\in\mathcal{H}_K}\big\{\mathcal{E}(f)+\lambda\|f\|_K^2\big\}.$$

The regularization error of f^*_λ will be denoted by $\mathcal{A}^*(\lambda)$, that is, we have that

$$\mathcal{A}^*(\lambda)=\mathcal{A}(f^*_\lambda)=\min_{K\in\mathcal{K}}\min_{f\in\mathcal{H}_K}\big\{\mathcal{E}(f)-\mathcal{E}(f^*_\ell)+\lambda\|f\|_K^2\big\}.$$

This quantity measures the approximation ability of the hypothesis space, $\{f:f\in\mathcal{H}_K,K\in\mathcal{K}\}$, to represent the target function f^*_ℓ and is determined by the structure of the loss function, the distribution ρ underlying the data and the hypothesis space. If some prior knowledge on the target function is available, we can estimate the decay rate of the regularization error, as we shall do in Section 5. Finally, we need the *residual loss* defined, for every $T>0$, as

$$\Psi(T)=\sup_{(x,y)\in Z}\sup_{f:X\to\mathbb{R}}\big\{\ell(y,\pi_T(f)(x))-\ell(y,f(x))\big\}.$$

Note that $\Psi(T)\geq 0$ for all $T>0$ and $\Psi(\infty)=0$. The use of residual error for a unified error decomposition framework was introduced in [35] and applied to the analysis of logistic classification in [39].

Proposition 3.1. *For every $K\in\mathcal{K}$, $f\in\mathcal{H}_K$ and any $T>0$, there holds the inequality*

$$\mathcal{E}\left(\pi_T(f_{\mathbf{z}})\right)-\mathcal{E}(f^*_\ell)\leq\mathcal{S}_{\mathbf{z}}(m,\lambda,f,T)+\Psi(T)+\mathcal{A}(f).$$

In particular, taking $T=\infty$, we have that $\mathcal{E}\left(f_{\mathbf{z}}\right)-\mathcal{E}(f^*_\ell)\leq\mathcal{S}_{\mathbf{z}}(m,\lambda,f)+\mathcal{A}(f)$.

Proof. We write $\mathcal{E}(\pi(f_{\mathbf{z}}))-\mathcal{E}(f^*_\ell)$ as

$$\begin{aligned}&\big\{\mathcal{E}(\pi_T(f_{\mathbf{z}}))-\mathcal{E}_{\mathbf{z}}(\pi_T(f_{\mathbf{z}}))\big\}+\big\{(\mathcal{E}_{\mathbf{z}}(\pi_T(f_{\mathbf{z}}))+\lambda\|f_{\mathbf{z}}\|^2_{K_{\mathbf{z}}})-(\mathcal{E}_{\mathbf{z}}(f)+\lambda\|f\|^2_K)\big\}\\&\quad+\big\{\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(f)\big\}+\big\{\mathcal{E}(f)-\mathcal{E}(f^*_\ell)+\lambda\|f\|^2_K\big\}-\lambda\|f_{\mathbf{z}}\|^2_{K_{\mathbf{z}}}.\end{aligned}$$

To bound the second term in the above equation we observe, for every $K \in \mathcal{K}$ and $f \in \mathcal{H}_K$, that

$$\mathcal{E}_{\mathbf{z}}(\pi_T(f_{\mathbf{z}})) + \lambda\|f_{\mathbf{z}}\|^2_{K_{\mathbf{z}}} \leq \Psi(T) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|^2_{K_{\mathbf{z}}} \leq \Psi(T) + \mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|^2_K.$$

The result follows by combining this inequality with the above definitions. ■

We note that the error decomposition in Proposition 3.1 is different from the traditional technique that bounds the excess error by sample error and approximation error. It contains an additional residual loss term. Similar non-traditional error decompositions have also appeared in the analysis of learning with sample dependent hypothesis space [38] and the analysis of minimum error entropy algorithm [19, 20, 17].

The sample error consists of two terms. The second term $\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)$ is the deviation between the empirical mean and the expectation of $\ell(y, f(x))$ respectively. This is a fixed random variable on Z which is easy to bound. The first term, $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$, is the deviation between the expected value and the empirical mean of $\ell(y, f_{\mathbf{z}}(x))$ with respect to $z = (x, y) \in Z$. More effort is required to bound it because the function $f_{\mathbf{z}}$ varies with $\mathbf{z}$ and, so, we need to deal with a set of random variables. For this purpose, we use the notion of Rademacher complexity.

Definition 3.2. We say that the random variable ε is a Rademacher variable if $\text{Prob}(\varepsilon = -1) = \text{Prob}(\varepsilon = 1) = \frac{1}{2}$. Let $\mathcal{F}$ be a function class on Z and ε_i, $i \in \mathbb{N}_m$ be a set of Rademacher variables. The Rademacher complexity on $\mathcal{F}$ is defined as

$$R(\mathcal{F}, m) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i \in \mathbb{N}_m} \varepsilon_i f(z_i) \right| \right]$$

where the expectation is over the i.i.d. Rademacher variables ϵ_i and the i.i.d. variables z_i.

If $\mathcal{F}$ is a class of real-valued functions on Z, $c \in \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$ and h a bounded function, we define the sets $c\mathcal{F} := \{cf : f \in \mathcal{F}\}$, $\phi \circ \mathcal{F} := \{\phi(f) : f \in \mathcal{F}\}$ and $\mathcal{F} + h = \{f + h : f \in \mathcal{F}\}$. We recall some simple properties of the Rademacher complexity for $\mathcal{F}$, see, for example, [7].

Lemma 3.3. *If $\mathcal{F}$ is a function class on Z then we have that*

(i) $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{m} \sum_{i \in \mathbb{N}_m} f(z_i) - \mathbb{E} f\right|\right] \leq 2R(\mathcal{F}, m)$;

(ii) *For every $c \in \mathbb{R}$, there holds $R(c\mathcal{F}, m) = |c|R(\mathcal{F}, m)$;*

(iii) *If ϕ is a Lipschitz function with Lipschitz constant L and $\phi(0) = 0$, then we have that $R(\phi \circ \mathcal{F}, m) \leq 2LR(\mathcal{F}, m)$;*

(iv) *If h is a bounded function, then $R(\mathcal{F} + h, m) \leq R(\mathcal{F}, m) + \frac{1}{\sqrt{m}}\|h\|_\infty$.*

We also note that it is straightforward to see that $R(co\mathcal{F}, m) = R(\mathcal{F}, m)$ where $co\mathcal{F}$ is the convex hull of $\mathcal{F}$. Moreover, if we let $\overline{\mathcal{F}}$ be the closure of $\mathcal{F}$, that is, the set of functions on Z with the property that there is a sequence $\{f_n\}$ of functions on Z such that, for any $z \in Z$, we have that $\lim_{n\to\infty} f_n(z) = f(z)$ then we also have $R(\overline{\mathcal{F}}, m) = R(\mathcal{F}, m)$. Consequently, any upper bound for the class $\mathcal{K}$ extends to the larger class $\overline{co\mathcal{K}}$.

We are now in a position to state our main results for the estimation of the sample error. To this end, we need, for $t \in \mathrm{I\!R}_+$, two quantities,

$$\Phi(t) := \sup_{y \in Y} \sup_{|s| \le t} \ell(y, s)$$

and

$$L(t) = \sup_{y \in Y} \sup_{|s_1|,|s_2| \le t} \frac{|\ell(y, s_1) - \ell(y, s_2)|}{|s_1 - s_2|}.$$

We also introduce the constants $\gamma = \sqrt{\Phi(0)/\lambda}$ and $\tau := \min\{T, \kappa\gamma\}$ as they appear often in our subsequent analysis. Note that we suppress the dependency of these constants on λ as it is only later that we shall adjust λ to obtain our estimates for the approximation error. Recall, in the theorem stated next, the proof of which is given in Section 4, that the set $\mathcal{K}_0$ was defined by equation (1.4).

Theorem 3.4. *If $f \in \mathcal{H}_K$, $\delta \in (0, 1)$ then with confidence $1 - \delta$ there holds*

$$\mathcal{S}_{\mathbf{z}}(m, \lambda, f, T) \le 4L(\tau)\gamma\sqrt{R(\mathcal{K}_0, m)} + \frac{2\Phi(0)}{\sqrt{m}} + \left(\frac{1}{2}\Phi(\tau) + \Phi(\|f\|_\infty)\right)\sqrt{\frac{2\log\frac{2}{\delta}}{m}}.$$

When projection is not involved, the sample error is bounded in the corollary below.

Corollary 3.5. *For $\delta \in (0, 1)$, with confidence $1 - \delta$ there holds*

$$\mathcal{S}_{\mathbf{z}}(m, \lambda, f^*_\lambda) \le 4L(\kappa\gamma)\gamma\sqrt{R(\mathcal{K}_0, m)} + \frac{1}{\sqrt{m}}\left(2\Phi(0) + \frac{3}{2}\Phi(\kappa\gamma)\sqrt{2\log\frac{2}{\delta}}\right).$$

Note that if $\lim_{m\to\infty} R(\mathcal{K}_0, m) = 0$ and $\lim_{\lambda\to 0} \mathcal{A}^*(\lambda) = 0$, we can choose $\lambda = \lambda(m)$ in such a way that $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f^*_\ell)$ tends to zero in probability as m tends to infinity. In other words, under the above hypotheses, our results imply the consistency of algorithm (1.2). Moreover, the convergence rates can be derived when quantitative estimates of $R(\mathcal{K}_0, m)$ and $\mathcal{A}^*(\lambda)$ are available, see our examples in Section 5.

One may find in the literature similar bounds for transductive learning where the error on test data is bounded in terms of the error on training data and the empirical Rademacher complexity of the kernel matrices; see [8, 22]. However, these results do not imply our results for inductive learning. More recent results in [11, 25, 21] have the advantage of computability of the empirical Rademacher complexity in terms of the trace or eigenvalues of the kernel matrix, they may not provide useful bounds for RBF kernel classes such as the Gaussians with arbitrary variances. As we shall see in Section 5 our results overcome this difficulty.

4 Estimating the sample error

In this section, we provide the proofs for the estimate of the sample error described earlier. They are based on the lemmas below. The first lemma bounds the second term in the sample error.

Lemma 4.1. *Let f be a bounded function. For every $\delta \in (0,1)$, with confidence $1-\delta$ there holds*

$$\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \leq \Phi\left(\|f\|_\infty\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{m}}.$$

Proof. Consider the random variable $\xi = \ell(y, f(x))$. Notice that $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i \in \mathbb{N}_m} \xi(z_i)$ and $\mathcal{E}(f) = \mathbb{E}\,\xi$. By our assumption, $0 < \xi \leq \Phi(\|f\|_\infty)$ which implies that $|\xi - \mathbb{E}\,\xi| \leq \Phi(\|f\|_\infty)$. Therefore, the conclusion follows by applying the one-sided Hoeffding inequality to the random variable ξ, see, for example, [14]. ■

The second lemma concerns some ancillary inequalities for $f \in \mathcal{H}_K$.

Lemma 4.2. *If $f \in \mathcal{H}_K$ and $K \in \mathcal{K}$ then*

(i) $\|f\|_\infty \leq \kappa \sqrt{\mathcal{A}(f)/\lambda}$

(ii) $\|f^*_\lambda\|_\infty \leq \kappa\gamma$

(iii) $\|f_{\mathbf{z}}\|_{K_{\mathbf{z}}} \leq \gamma$

(iv) $\|\pi_T(f)\|_\infty \leq \min\{T, \kappa\|f\|_K\}$.

Proof. The first claim follows directly from the fact that

$$\lambda\|f\|_K^2 \leq \mathcal{E}(f) - \mathcal{E}(f^*_\ell) + \lambda\|f\|_K^2 = \mathcal{A}(f)$$

and (1.3). Note that

$$\mathcal{A}^*(\lambda) \leq \mathcal{E}(0) - \mathcal{E}(f^*_\ell) \leq \Phi(0).$$

Hence, the second claim is a consequence of the first one. The third claim follows in a manner identical to the second one and the last claim follows from the definition of π_T and inequality (1.3). ■

By inequalities (iii) and (iv) above it follows that

$$\mathcal{E}(\pi_T(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi_T(f_{\mathbf{z}})) \leq g(\mathbf{z}) := \sup_{f \in \gamma\mathcal{B}_{\mathcal{K}}} \big(\mathcal{E}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}}(\pi_T(f))\big) \tag{4.1}$$

where $\mathcal{B}_{\mathcal{K}}$ is the union of the unit balls in $\mathcal{H}_K$ over $K \in \mathcal{K}$, that is,

$$\mathcal{B}_{\mathcal{K}} = \bigcup_{K \in \mathcal{K}} \Big\{ f \in \mathcal{H}_K : \|f\|_K \leq 1 \Big\}.$$

The third lemma applies the McDiarmid's inequality, see [26], to the random variable $g(\mathbf{z})$ to measure the difference between $g(\mathbf{z})$ and $\mathbb{E}\,g(\mathbf{z})$.

Lemma 4.3. *Let $g(\mathbf{z})$ be defined as above. For every $\delta \in (0,1)$, with confidence $1-\delta$ there holds*

$$g(\mathbf{z}) \leq \mathbb{E}\,g(\mathbf{z}) + \Phi\left(\tau\right) \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

Proof. Denote by $\mathbf{z}_i'$ the sample which coincides with $\mathbf{z}$ except for the i-th pair $z_i = (x_i, y_i)$ replaced by $z_i' = (x_i', y_i')$. Consequently, we observe that

$$\begin{aligned}
g(\mathbf{z}) - g(\mathbf{z}_i') &= \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} (\mathcal{E}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}}(\pi_T(f))) - \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} (\mathcal{E}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}_i'}(\pi_T(f))) \\
&\leq \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} (\mathcal{E}_{\mathbf{z}_i'}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}}(\pi_T(f))) \\
&= \frac{1}{m} \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} (\ell(y_i', \pi_T(f(x_i'))) - \ell(y_i, \pi_T(f(x_i)))) \\
&\leq \frac{1}{m}\Phi(\tau)
\end{aligned}$$

where the last inequality follows from inequality (iv) of Lemma 4.2. Interchanging the roles of $\mathbf{z}$ and $\mathbf{z}_i'$, in the above computation, gives us the inequality

$$|g(\mathbf{z}) - g(\mathbf{z}_i')| \leq \frac{1}{m}\Phi(\tau)$$

and, so, by McDiarmid's inequality we have that

$$\text{Prob}\{g(\mathbf{z}) - \mathbb{E}\, g(\mathbf{z}) > \varepsilon\} \leq \exp\left(-\frac{2m\varepsilon^2}{\Phi^2(\tau)}\right)$$

and the desired result follows. ■

The last lemma estimates $\mathbb{E}\, g(\mathbf{z})$ in terms of the Rademacher complexity of the set $\mathcal{K}_0$.

Lemma 4.4. *We have that*

$$\mathbb{E}\, g(\mathbf{z}) \leq 4L(\tau)\gamma\sqrt{R(\mathcal{K}_0, m)} + \frac{2\Phi(0)}{\sqrt{m}}.$$

Proof. We use Lemma 3.3 repeatedly and verify that

$$\begin{aligned}
\mathbb{E}\, g(\mathbf{z}) &\leq 2\,\mathbb{E} \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \ell(y_i, f(x_i))\right| \\
&\leq 2\,\mathbb{E} \sup_{f\in\gamma\mathcal{B}_{\mathcal{K}}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i (\ell(y_i, f(x_i)) - \ell(y_i, 0))\right| + \frac{2\Phi(0)}{\sqrt{m}} \\
&\leq 4L(\tau)\, R(\gamma\mathcal{B}_{\mathcal{K}}, m) + \frac{2\Phi(0)}{\sqrt{m}} \\
&= 4L(\tau)\,\gamma\, R(\mathcal{B}_{\mathcal{K}}, m) + \frac{2\Phi(0)}{\sqrt{m}}.
\end{aligned}$$

By the definition of $\mathcal{B}_{\mathcal{K}}$ and the reproducing kernel property we have that

$$\begin{aligned}\sup_{f\in\mathcal{B}_{\mathcal{K}}}\left|\frac{1}{m}\sum_{i\in\mathbb{N}_m}\varepsilon_i f(x_i)\right| &= \frac{1}{m}\sup_{K\in\mathcal{K}}\sup_{\|f\|_K\le 1}\left|\left\langle\sum_{i\in\mathbb{N}_m}\varepsilon_i K_{x_i}, f\right\rangle_K\right| \\ &= \frac{1}{m}\sup_{K\in\mathcal{K}}\left\|\sum_{i\in\mathbb{N}_m}\varepsilon_i K_{x_i}\right\|_K \\ &= \frac{1}{m}\sup_{K\in\mathcal{K}}\left(\sum_{i,j\in\mathbb{N}_m}\varepsilon_i\varepsilon_j K(x_i,x_j)\right)^{1/2} \\ &\le \frac{1}{\sqrt{m}}\sup_{K\in\mathcal{K}}\left(\sup_{t\in X}\left|\sum_{i\in\mathbb{N}_m}\varepsilon_i K(x_i,t)\right|\right)^{1/2},\end{aligned}$$

where the last inequality follows from the Hölder inequality and the fact $|\varepsilon_j| = 1$. Hence, by Jensen's inequality we conclude that

$$R(\mathcal{B}_{\mathcal{K}}, m) \le \frac{1}{\sqrt{m}}\left(\mathbb{E}\sup_{K\in\mathcal{K}}\sup_{t\in X}\left|\sum_{i\in\mathbb{N}_m}\varepsilon_i K(x_i,t)\right|\right)^{1/2} = \sqrt{R(\mathcal{K}_0, m)}.$$

This finishes the proof. ■

We note that the proof of Theorem 3.4 follows by combing inequality (4.1), Lemmas 4.3, 4.4 and 4.1. As for Corollary 3.5, we choose $T = \infty$ and $f = f_\lambda^*$ in Theorem 3.4 and use inequality (ii) of Lemma 4.2.

5 Learning with Gaussians

In this section we specify our results to the family of Gaussian kernels, that is, we assume that $X \subset \mathbb{R}^n$ and consider the family of kernels

$$\mathcal{G} := \{G_\sigma : \sigma \in (0,\infty)\},$$

where $G_\sigma(x,y) = \exp(-\sigma\|x-y\|^2)$, $x, y \in \mathbb{R}^n$ and $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. We shall prove the following result about the Gaussian family $\mathcal{G}$.

Proposition 5.1. *For each $n \in \mathbb{N}$, there exists a constant d_n such that, for all $m \in \mathbb{N}$ $R(\mathcal{G}_0, m) \le d_n \frac{\log m}{\sqrt{m}}$.*

To prove the proposition we use some tools from empirical processes, see for example [33]. To this end, we recall the concept of covering numbers.

Definition 5.2. Let $(\mathscr{M}, d)$ be a pseudo-metric space and S a subset of $\mathscr{M}$. For every $\varepsilon > 0$, the *covering number* of S by balls of radius ε with respect to d, denoted by $\mathscr{N}(S,\varepsilon,d)$, is defined as the minimal number of balls of radius ε whose union covers S, namely,

$$\mathscr{N}(S,\varepsilon,d) = \min\left\{n \in \mathbb{N} : \text{ there exist } \{s_j\}_{j=1}^n \subset \mathscr{M} \text{ such that } S \subseteq \bigcup_{j=1}^n B(s_j,\varepsilon)\right\}$$

where $B(s_j,\varepsilon) := \{s \in \mathscr{M} : d(s,s_j) < \varepsilon\}$.

Next, we introduce the p-norm empirical covering number. Let d_p denote the normalized ℓ^p–metric on the Euclidean space $\mathbb{R}^m$ defined, for all $a = (a_i : i \in \mathbb{N}_m), b = (b_i : i \in \mathbb{N}_m) \in \mathbb{R}^m$, as $d_p(a,b) = \left(\frac{1}{m}\sum_{i\in\mathbb{N}_m} |a_i - b_i|^p\right)^{1/p}$.

Definition 5.3. Let $\mathcal{F}$ be a class of bounded functions defined on X, $\mathbf{x} = (x_i : i \in \mathbb{N}_m) \in X^m$ and $\mathcal{F}|_{\mathbf{x}} = \{(f(x_i) : i \in \mathbb{N}_m) : f \in \mathcal{F}\} \subseteq \mathbb{R}^m$. For $1 \leq p \leq \infty$, we define the *p-norm empirical covering number* of $\mathcal{F}$ associated to $\mathbf{x}$ as $\mathscr{N}_{p,\mathbf{x}}(\mathcal{F},\varepsilon) = \mathscr{N}(\mathcal{F}|_{\mathbf{x}}, \varepsilon, d_p)$. Moreover, we let

$$\mathscr{N}_p(\mathcal{F},\varepsilon,m) := \sup_{\mathbf{x}\in X^m} \mathscr{N}_{p,\mathbf{x}}(\mathcal{F},\varepsilon).$$

Proof of Proposition 5.1. By [42, Lemma 12 and Lemma 13], there exists some constant c_n depending only on n such that $\log \mathscr{N}_\infty(\mathcal{G}_0,\varepsilon,m) \leq \frac{c_n}{\varepsilon}\left(\log\frac{m}{\varepsilon}\right)^2$. Since $\mathscr{N}_2(\mathcal{F},\varepsilon,m) \leq \mathscr{N}_\infty(\mathcal{F},\varepsilon,m)$ we also have that

$$\log \mathscr{N}_2(\mathcal{G}_0,\varepsilon,m) \leq \frac{c_n}{\varepsilon}\left(\log\frac{m}{\varepsilon}\right)^2. \tag{5.1}$$

Next, we recall that if $\mathcal{F}$ is a bounded function class whose 2-norm empirical covering number $\mathscr{N}_2(\mathcal{F},\varepsilon,m)$ is finite for all $\varepsilon > 0$ then the Rademacher average can be bounded in the following manner [33]

$$R(\mathcal{F},m) \leq \frac{1}{\sqrt{m}}\int_0^U \sqrt{\log \mathscr{N}_2(\mathcal{F},\varepsilon,m)}\, d\varepsilon, \tag{5.2}$$

where $U = \sup_{f\in\mathcal{F}} \mathbb{E} f^2$. But, for each $f \in \mathcal{G}$ we have that $\mathbb{E} f^2 \leq 1$. Hence, by combining inequalities (5.1) and (5.2) we conclude there is a constant d_n such that for each m

$$R(\mathcal{G}_0,m) \leq \frac{c_n}{\sqrt{m}}\int_0^1 \frac{1}{\sqrt{\varepsilon}}\log\frac{m}{\varepsilon}\, d\varepsilon \leq d_n\frac{\log m}{\sqrt{m}}.$$

■

Using this estimate in Theorem 3.4 and Corollary 3.5, provides an estimate for the sample error. These estimates suggest a way to choose the regularization parameter and compute the learning rate. If some prior knowledge is available on the target function we can estimate the decay of the regularization error. To further illustrate our results, we consider two classical learning algorithms: regularized least squares (RLS) and support vector machine (SVM) classification.

5.1 Regularized least squares

In the sequel, we denote by $\rho(y|x)$ the conditional probability of y for a given point $x \in X$ and by ρ_X the marginal distribution of ρ on X. Recall that $\kappa = 1$ for the class $\mathcal{G}$. In regression problems we assume that $|y| \leq M$ almost surely. In RLS, the loss function takes the form $\ell(y,f(x)) = (y - f(x))^2$. A standard argument shows that the target function is given by the regression function, that is,

$$f_\ell^*(x) = \int_Y y d\rho(y|x)$$

and, for every function $f \in L^2(\rho_X)$, there holds

$$\mathcal{E}(f) - \mathcal{E}(f_\ell^*) = \|f - f_\ell^*\|^2_{L^2(\rho_X)}.$$

Moreover, it is easy to verify, for $t \in \mathbb{R}_+$, that $\Phi(t) \leq (M+t)^2$ and $L(t) \leq 2(M+t)$. Putting these estimates into Corollary 3.5, when $\lambda \leq 1$, we obtain, with confidence $1-\delta$, that

$$\mathcal{S}_{\mathbf{z}}(m, \lambda, f_\lambda^*) \leq \left(16\sqrt{d_n} + 2 + 6\sqrt{2\log\tfrac{2}{\delta}}\right) M^2 \frac{\sqrt{\log m}}{\lambda m^{1/4}}.$$

Corollary 5.4. *If there are constants $c > 0$ and $\beta \in (0,1]$ such that for all $\lambda > 0$ $\mathcal{A}^*(\lambda) \leq c\lambda^\beta$, then there is a constant c' such that for any m there exists a λ such that, with confidence $1-\delta$,*

$$\|f_{\mathbf{z}} - f_\ell^*\|^2_{L^2(\rho_X)} \leq c'\left(16M^2\sqrt{d_n} + 2M^2 + 6M^2\sqrt{2\log\tfrac{2}{\delta}}\right)^{\frac{\beta}{1+\beta}} \left(\frac{\sqrt{\log m}}{m^{1/4}}\right)^{\frac{\beta}{1+\beta}}. \qquad (5.3)$$

In proving the corollary we have used the fact that the function $h(\lambda) = a/\lambda + c\lambda^\beta$, $\lambda > 0$ achieves its minimum at $\lambda = \hat{\lambda} = (a/c\beta)^{\frac{1}{1+\beta}}$. A direct computation gives, for some constact c', that

$$h(\hat{\lambda}) = (c\beta)^{\frac{1}{1+\beta}}(1 + 1/\beta)a^{\frac{\beta}{\beta+1}} = c'a^{\frac{\beta}{\beta+1}}.$$

The result follows by setting

$$a = \left(16\sqrt{d_n} + 2 + 6\sqrt{2\log(2/\delta)}\right) M^2\sqrt{\log m}/m^{\frac{1}{4}}$$

and a direct computation.

Corollary 5.4 tells us that the learning rate can be computed once the regularization error is estimated. Let us illustrate this by an example.

Example 1. *If $d\rho_X$ is the Lebesgue measure on X and f_ℓ^* is a restriction to X of a function in*

$$H^s(\mathbb{R}^n) = \left\{ f \in L^2(\mathbb{R}^n) : \|f\|_{H^s} = \left((2\pi)^{-n}\int_{\mathbb{R}^n}(1+|\xi|^2)^s|\hat{f}(\xi)|^2 d\xi\right)^{1/2} < \infty \right\},$$

where $\hat{f}$ denotes the Fourier transform of f, then

$$\mathcal{A}^*(\lambda) \leq c\lambda^{\frac{4s}{2n+4s+ns}}$$

with

$$c = (\pi^2 + \pi^{-n/2})\|f_\ell^*\|^2_{L^2} + \|f_\ell^*\|^2_{H^s}.$$

Hence, by Corollary 5.4 there exist constants c' and λ such that with confidence $1-\delta$

$$\|f_{\mathbf{z}} - f_\ell^*\|^2_{L^2} \leq c'\left(16M^2\sqrt{d_n} + 2 + 6M^2\sqrt{2\log\tfrac{2}{\delta}}\right)^{\frac{4s}{2n+8s+ns}} \left(\frac{\log m}{\sqrt{m}}\right)^{\frac{2s}{2n+8s+ns}}.$$

Proof. For every $\sigma \in (0, \infty)$, we define functions $f_{\ell,\sigma}$, for every $x \in \mathbb{R}^n$

$$f_{\ell,\sigma}(x) = \left(\frac{\sigma}{\pi}\right)^{n/2} \int_X G_\sigma(x, y) f_\ell^*(y) dy.$$

By [32, Lemma 8.1] we have that $f_{\ell,\sigma} \in \mathcal{H}_{G_\sigma}$ and $\|f_{\ell,\sigma}\|_{G_\sigma} \leq \pi^{-\frac{n}{4}} \sigma^{\frac{n}{4}} \|f_\ell^*\|_{L^2}$. Moreover, we have that

$$\begin{aligned}
\|f_{\ell,\sigma} - f_\ell^*\|_{L^2}^2 &= \left\| \left(e^{-\frac{\pi\|\cdot\|^2}{\sigma}} - 1 \right) \hat{f}_\ell^* \right\|_{L^2}^2 \\
&= \int_{\|t\| \leq \sigma^{\frac{1}{2+s}}} \left(e^{-\frac{\pi\|t\|^2}{\sigma}} - 1 \right)^2 |\hat{f}_\ell^*(t)|^2 dt + \int_{\|t\| > \sigma^{\frac{1}{2+s}}} \left(e^{-\frac{\pi\|t\|^2}{\sigma}} - 1 \right)^2 |\hat{f}_\ell^*(t)|^2 dt \\
&\leq \int_{\|t\| \leq \sigma^{\frac{1}{2+s}}} \left(\frac{\pi\|t\|^2}{\sigma} \right)^2 |\hat{f}_\ell^*(t)|^2 dt + \int_{\|t\| > \sigma^{\frac{1}{2+s}}} |\hat{f}_\ell^*(t)|^2 dt \\
&\leq \int_{\|t\| \leq \sigma^{\frac{1}{2+s}}} \pi^2 \sigma^{-\frac{2s}{2+s}} |\hat{f}_\ell^*(t)|^2 dt + \int_{\|t\| > \sigma^{\frac{1}{2+s}}} \sigma^{-\frac{2s}{2+s}} |\hat{f}_\ell^*(t)|^2 \|t\|^{2s} dt \\
&\leq \left(\pi^2 \|f_\ell^*\|_{L^2}^2 + \|f_\ell^*\|_{H^s}^2 \right) \sigma^{-\frac{2s}{2+s}}.
\end{aligned}$$

Since $G_\sigma \in \mathcal{G}$ and $f_{\ell,\sigma} \in \mathcal{H}_{G_\sigma}$ for all $\sigma \in (0, \infty)$, we have that

$$\mathcal{A}^*(\lambda) \leq \inf_{\sigma \in (0,\infty)} \left\{ \|f_{\ell,\sigma} - f_\ell^*\|_{L^2}^2 + \lambda \|f_{\ell,\sigma}\|_{G_\sigma}^2 \right\}.$$

By taking $\sigma = \lambda^{-\frac{2(2+s)}{2n+4s+ns}}$ we obtain the desired estimate for $\mathcal{A}^*(\lambda)$. ■

By the analysis in [31], for any fixed $\beta, \sigma > 0$, $\inf_{f \in \mathcal{H}_{G_\sigma}} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\ell^*) + \lambda \|f\|_{G_s}^2 \right\}$ cannot decay with rate $\mathcal{O}(\lambda^\beta)$ as $\lambda \to 0^+$. Hence, a polynomial decay of $\|f_{\mathbf{z}} - f_\ell^*\|_{L^2(\rho_X)}^2$ is generally impossible. Thus, Example 1 ensures that the multiple kernel learning algorithm (1.2) significantly improves the approximation power and learning ability over the algorithm with a single kernel of fixed bandwidth. Similar improvement and fast rate can also be achieved by varying and validating the bandwidth parameter [32, 15].

5.2 Support vector machine classification

In binary classification we choose $Y = \{1, -1\}$ and wish to find a classifier $f : X \to Y$. The prediction power of f is measured by the classification error

$$\mathcal{R}(f) = \text{Prob}\left\{ f(X) \neq Y \right\}.$$

The optimal classifier, which yields the minimal classification error is called the *Bayes rule*: $f^* = \arg\min \mathcal{R}(f)$ with the minimum taken over all classifiers $f : X \to Y$. If, for every $y \in Y$, we let

$$X_y := \left\{ x \in X : \text{Prob}(y|x) > \frac{1}{2} \right\}$$

and

$$X_0 := \left\{ x \in X : \text{Prob}(1|x) = \frac{1}{2} \right\},$$

Then the Bayes rule takes the form $f^*(x) = y$ if $x \in X_y$ for $y \in Y$. We note, in passing, that, unless X_0 is empty, the Bayes rule is not unique. The performance of a classification algorithm is measured by the approximation ability of the output classifier to the Bayes rule with respect to the classification error.

SVM classification uses the loss function $\ell(y, f(x)) = \max\{1 - yf(x), 0\}$ and the target function is $f^*_\ell = f^*$ [34]. It computes the real-valued function $f_{\mathbf{z}}$ which solves problem (1.2) and gives the classifier $\operatorname{sgn}(f_{\mathbf{z}})$. In order to bound the excess classification error of $f_{\mathbf{z}}$ we use the projection operator and recall, for all real-valued functions $f : X \to Y$, that

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f^*) \leq \mathcal{E}(\pi_1(f)) - \mathcal{E}(f^*) \tag{5.4}$$

see, for example, [43, 37]. We observe that $\Psi(T) = 0$ if $T \geq 1$ and, for all $t \in \mathbb{R}_+$, that $\Phi(t) = (1+t)$ and $L(t) = 1$. Combining Proposition 3.1, Theorem 3.4 with $T = 1$ and $f = f^*_\lambda$ and Proposition 5.1 we obtain that

$$\mathcal{E}(\pi_1(f_{\mathbf{z}})) - \mathcal{E}(f^*) \leq \left(4\sqrt{d_n} + 2 + 3\sqrt{2\log\tfrac{2}{\delta}}\right) \frac{\sqrt{\log m}}{\lambda^{1/2} m^{1/4}} + \mathcal{A}^*(\lambda).$$

Using the above equation and equation (5.4) we we obtain the following corollary.

Corollary 5.5. *If there exist constants $c > 0$ and $\beta \in (0, 1]$ such that, for all $\lambda > 0$, $\mathcal{A}^*(\lambda) \leq c\lambda^\beta$ then, for any m there is a choice of λ such that, for some constant c', with confidence $1 - \delta$*

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f^*) \leq c' \left(4\sqrt{d_n} + 2 + 3\sqrt{2\log\tfrac{2}{\delta}}\right)^{\frac{2\beta}{1+2\beta}} \left(\frac{\log m}{\sqrt{m}}\right)^{\frac{\beta}{1+2\beta}}.$$

As an example of this observation we consider distributions satisfying the geometric noise condition [32]. To this end, let $\tau(x) = d(x, X_0 \bigcup X_{-i})$ for $x \in X_i$, $i = 1, -1, 0$. We say ρ has geometric noise exponent $\alpha > 0$ if there exists $c > 0$ such that, for all $t > 0$,

$$\int_X \left|\text{Prob}(1|x) - \text{Prob}(-1|x)\right| e^{-\tau^2(x)/t} d\rho_X(x) \leq ct^{-\alpha n/2}.$$

Thus, applying [32, Theorem 2.14] and Corollary 5.5, we are led to the following example.

Example 2. *If X is a subset of the unit ball in $\mathbb{R}^n$ and ρ has geometric noise exponent $\alpha > 0$ with constant c, then there is a constant $d > 0$ such that, for all $\lambda > 0$, $\mathcal{A}^*(\lambda) \leq d\lambda^{\alpha/(\alpha+1)}$. Hence, there exist a constant d' and a choice of λ such that with confidence $1 - \delta$*

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f^*) \leq d' \left(4\sqrt{d_n} + 2 + 3\sqrt{2\log\tfrac{2}{\delta}}\right)^{\frac{2\alpha}{3\alpha+1}} \left(\frac{\log m}{\sqrt{m}}\right)^{\frac{\alpha}{3\alpha+1}}.$$

6 Discussion

We have provided an analysis of the generalization error for a general kernel learning method based on a regularization scheme within a class $\mathcal{K}$ of uniformly bounded reproducing kernels.

When $\mathcal{K}$ is the family of Gaussian kernels with arbitrary variance, our analysis guarantees the consistency of the learning algorithm and provides good error rates for the case of regularized least squares and support vector machines.

We note that an earlier version of this paper appeared in [29]. The sample error bound was motivated by an earlier version of [42] released in 2004, which established the necessary and sufficient condition for the learnability and consistency of learning the kernel problem and derived the covering number for the union of the unit balls in RKHSs with Gaussians. In the late published version, the work [42] also derived error bounds which in certain instances improve upon those given in [29].

A number of research questions can be studied starting from the framework presented in this paper. We close with highlighting some useful kernel classes $\mathcal{K}$ and learning schemes which stem out of the main theme of this paper and which would be valuable subject of future work.

- In [27] a convex set of kernels parameterized by a locally compact set Σ is considered, namely
$$\mathcal{K} = \left\{ \int_{\Sigma} G(\sigma) dp(\sigma) : p \in \mathcal{P}(\Sigma) \right\}, \tag{6.1}$$
where for each $\sigma \in \Sigma$, $G(\sigma) : X \times X \to \mathbb{R}$ is a prescribed kernel which depends continuously on σ and $\mathcal{P}(\Sigma)$ is the set of all probability measures on Σ. This study reveals good kernel classes $\mathcal{K}$ which have faster learning rates than the one obtained for single kernel. For example, when $\Sigma \subseteq \mathbb{R}_+$ and the function $G(\sigma)$ is a multivariate Gaussian kernel with variance σ then $\mathcal{K}$ equals the closed convex hull of $\mathcal{G}$, that is, the class of radial kernels, and the Rademacher complexities of $\overline{co}\mathcal{G}_0$ and $\mathcal{G}_0$ are the same.

- Problems addressed in this paper naturally extend to the context of operator valued kernels, as considered, for example, in [9]. Among the several classes of operator valued kernels which would be valuable to analyse from a statistical learning theory point of view, we mention the general class
$$\mathcal{K}_{\text{operator}} = \{AK : A \in \mathcal{A},\ K \in \mathcal{K}\}, \tag{6.2}$$
where $\mathcal{K}$ is a class of scalar kernels, e.g. the class (6.1) above, and $\mathcal{A}$ is a subset of the set of bounded positive operators on a Hilbert space $\mathcal{Y}$. Operator valued kernels arise in various application areas, in particular they are instrumental in multitask learning. In this setting the kernels are matrix valued (hence sometimes called multitask kernels) and the set $\mathcal{A}$ in equation (6.2) is a subset of the set of positive definite matrices. It is interesting to note that if $\mathcal{A}$ is the set of all positive definite matrices with trace bounded by one, and the set $\mathcal{K}$ is a singleton then problem (1.1) is equivalent to trace norm regularization in a reproducing kernel Hilbert space, as considered [1, 4]. Therefore, if we further enlarge the class $\mathcal{K}$, the approach considered in this paper gives rise to the problem of learning the kernel for trace norm regularization, which could be a valuable direction of future study.

- The learning scheme in equation (1.2) involves the minimization of the sum of the regularized empirical error over all functions $f \in \mathcal{H}_K$ and over all kernels $K \in \mathcal{K}$.

The regularization over f is necessary in order to avoid overfitting, however the regularization over K is implicit in the choice of the class $\mathcal{K}$. Therefore, it is natural to consider an alternative two stage optimization approach, in which we first minimize the regularized empirical error over $f \in \mathcal{H}_K$ and subsequently the empirical error of the minimizer $f_{\mathbf{z}}$ is minimized error over $K \in \mathcal{K}$. For example in the case of the square loss function a direct computation gives that

$$\frac{1}{m} \sum_{i \in \mathbb{N}_m} (y_i - f_{\mathbf{z}}(x_i))^2 = \lambda(\mathbf{y}, (K_{\mathbf{z}}(\mathbf{x}) + m\lambda I)^{-2}\mathbf{y}) \tag{6.3}$$

whereas, as noted in Section 2,

$$\frac{1}{m} \sum_{i \in \mathbb{N}_m} (y_i - f_{\mathbf{z}}(x_i))^2 + \lambda \|f_{\mathbf{z}}\|_K^2 = \lambda(\mathbf{y}, (K_{\mathbf{z}}(\mathbf{x}) + m\lambda I)^{-1}\mathbf{y}). \tag{6.4}$$

Note that these two expressions differ only in the exponents that appear in the right hand side of equations (6.3) and (6.4). In the future, it would be interesting to investigate how these different exponents in the objective function affects the learning rate. This observation may also be relevant for binary classification with the hinge loss function.

Acknowledgments

This work was supported in part by the EPSRC, American National Science Foundation, Research Grants Council of Hong Kong [Project No. CityU 104113] and National Natural Science Foundation of China [No.s 11471292, 11461161006, and 11671171].

References

[1] A. Argyriou, T. Evgeniou, M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] A. Argyriou, R. Hauser, C. A. Micchelli and M. Pontil. A DC-programming algorithm for kernel selection. In *Proceedings of the 23rd international conference on Machine learning*, pp. 41–48, 2006, ACM.

[3] A. Argyriou, C. A. Micchelli and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. In *Proc. 18th Annual Conf. on Learning Theory*, 2005.

[4] A. Argyriou, C. A. Micchelli, M. Pontil. On spectral learning. *Journal of Machine Learning Research*, 11:935-953, 2010.

[5] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**: 337–404, 1950.

[6] F. R. Bach, G. R. G Lanckriet and M. I. Jordan. Multiple kernels learning, conic duality, and the SMO algorithm. In *Proc. of the Int. Conf. on Machine Learning (ICML'04)*, 2004.

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

[8] O. Bousquet and D. J. L. Herrmann. On the complexity of learning the kernel matrix. In *Advances in Neural Information Processing Systems*, **15**, 2003.

[9] A. Caponnetto, C. A., Micchelli, M. Pontil, Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008

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

[11] C. Cortes, M. Mohri, and A. Rostamizadeh, Generalization bounds for learning kernels, In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 247–254, 2010.

[12] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory. *Found. Comput. Math.*, **2**: 413–428, 2002.

[13] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*, vol. 24 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, New York, NY, USA, 2007.

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

[15] M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels, *Electron. J. Stat.* 7: 1-42, 2013.

[16] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.

[17] J. Fan, T. Hu, Q. Wu, and D. X. Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis.* 41:164-189, 2016.

[18] M. Herbster. Relative loss bounds and polynomial-time predictions for the K-LMS-NET algorithm. In *Proc. 15th Int. Conf. Algorithmic Learning Theory*, October 2004.

[19] T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Learning theory approach to a minimum error entropy criterion. *Journal of Machine Learning Research.* 14:377-397, 2013.

[20] T. Hu, J. Fan, Q. Wu, and D. X. Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications.* 13(4):437-455, 2015.

[21] Z. Hussain and J. Shawe-Taylor, Improved loss bounds for multiple kernel learning, In *JMLR Workshop and Conference Proceedings Volume 15: AISTATS 2011*, pp. 370–377, 2011.

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

[23] Y. Lee, Y. Kim, S. Lee and J. Y. Koo. Structured multicategory support vector machine with analysis of variance decomposition. *Biometrika*, 93(3): 555–571, 2006.

[24] Y. Lin and H. H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5): 2272–2297, 2006.

[25] A. Maurer and M. Pontil. Structured sparsity and generalization, *Journal of Machine Learning Research*, 13: 671–690, 2012.

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

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

[28] C. A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, 66:297–319, 2007.

[29] C. A. Micchelli, M. Pontil, Q. Wu, D. X. Zhou. Error bounds for learning the kernel. Research Note RN/05/09 Department of Computer Science, UCL, 2005.

[30] C. S. Ong, A. J. Smola, and R. C. Williamson. Hyperkernels. In *Advances in Neural Information Processing Systems*, **15**, S. Becker et. al (Eds.), MIT Press, Cambridge, MA, 2003.

[31] S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, **1**: 17–41, 2003.

[32] I. Steinwart and C. Scovel. Fast rates for support vector machines using Guassian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

[33] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes.* Springer-Verlag, New York, 1996.

[34] G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the Randomized GACV, in *Advances in Kernel Methods - Support Vector Learning*, Schölkopf, Burges and Smola, eds., MIT Press, pages 69–88, 1999.

[35] Q. Wu. *Classification and Regularization in Learning Theory*, VDM Verlag, 2009.

[36] Q. Wu, Y. Ying and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

[37] Q. Wu and D. X. Zhou. Support vector machines: linear programming versus quadratic programming. *Neural computation*, **17**: 1160–1187, 2005.

[38] Q. Wu and D. X. Zhou. Learning with sample dependent hypothesis spaces. *Computers and Mathematics with Applications*, 56:2896–2907, 2008.

[39] D. H. Xiang. Logistic classification with varying Gaussian, *Computers & Mathematics with Applications*, 61(2): 397-407, 2011

[40] Y. Ying and C. Campbell. Generalization bounds for learning the kernel problem. In *Proceedings of the 23rd Conference on Learning Theory (COLT 2009)*, pages 407–416, 2009.

[41] Y. Ying and C. Campbell. Rademacher chaos complexity for learning the kernel. *Neural Computation*, 22(11), 2010

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

[43] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.*, **32**: 56–85, 2004.