Research Note RN/05/09

Department of Computer Science, University College London

Error bounds for learning the kernel[1]

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

Department of Computer Sciences
University College London
Gower Street, London WC1E, England, UK
E-mail: *m.pontil@cs.ucl.ac.uk*

Qiang Wu, Ding-Xuan Zhou

Dept of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong, P.R. China
E-mail: {*wuqiang*, *mazhou*}@*math.cityu.edu.hk*

8 June, 2005

Abstract

The problem of learning the kernel function has recently received considerable attention in machine learning. Much of the work so far 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.

[1]This work was partially supported by NSF Grant No. ITR-0312113, EPSRC Grant No. GR/T18707/01, IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and the University Grants Council of Hong Kong (Project No. CityU 103704).

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 and $\mathbb{N}_m := \{1, \ldots, m\}$. 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, see [2]. This approach has a long history. It has been studied, from different perspectives, in statistics, in optimal recovery, and more recently, has been a focus of attention in machine learning theory, see, for example, [14] and references therein for a discussion. The choice of the loss function ℓ leads to different learning methods among which the most prominent have been square loss regularization and support vector machines.

When the kernel K is fixed, the algorithm (1.1) is well understood, see, for example, [6, 7, 16, 17, 21, 23] 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 approximation error, see, for example, [7, 6, 16, 17].

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 [5, 9, 10, 13, 14, 15]. That is, we consider the variational problem

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

This scheme was also introduced in [21], motivated by improving the approximation error. Practical experience with this method [1, 3, 5, 10, 12, 13] 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 [14, 21].

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 to the measure ρ, and $f_\ell^* = \arg\min \mathcal{E}(f)$, the *target function* with the minimum 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, so that it follows by the reproducing kernel property of $\mathcal{H}_K$ [2], for all $K \in \mathcal{K}$ and $f \in \mathcal{H}_K$, that

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

an inequality which we will use repeatedly in the paper.

A key step in our analysis is a probabilistic upper bound on the sample error which is achieved by estimating the Rademacher complexity of the $\mathcal{K}_0 = \{K(x, \cdot) : K \in \mathcal{K}, x \in X\}$.

Our main results are presented in Section 3 and proved in Section 4. Analysis similar to that which appear here may be found in [5, 10], however, as far as we know we achieve greater generality than is available so far. 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 regularization networks.

2 Preventing overfitting?

We precede 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 $\mathcal{K}$. 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 $\ell(u, v) \geq \delta |u - v|^2$ for all $u, v \in \mathbb{R}$ satisfying $|u - v| \leq c$. If $\mathbf{y} \neq 0$ for every $\lambda > 0$ and $\mathbf{z} \in Z^m$ then there exists $\rho > 0$ such that $\mathcal{E}_\lambda(f_\mathbf{z}, K_\mathbf{z}) \geq \rho$.*

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 [14, 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

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

■

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, if λ is appropriately chosen our approach may be free of over-fitting, 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 Main results

In this section, we present our main results. For this purpose, we require some notation. First, we follow [6, 20, 21] and introduce the projection operator defined, for every measurable function $f : X \to \mathbb{R}$ as $\pi_T(f)(x) = \operatorname{sgn}(f)(x) \min(|f(x)|, T)$, where $\operatorname{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and -1 otherwise, namely,

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

$T > 0$ is called the projection level. The projection operator is useful for providing a better estimate for classification. However, it is not needed for regression problems. Next we define the *truncated sample error* as

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

and the *sample error* as

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

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

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

The function f in the above equation can be arbitrarily chosen, however, only proper choices lead to good estimates of the approximation error. A good choice is $f = f_\lambda^*$ where

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

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

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

It measures the approximation ability of the hypothesis space, $\{f : f \in \mathcal{H}_K, K \in \mathcal{K}\}$, to the target function f_ℓ^* and is determined by the structure of the loss function, the underlying distribution and the hypothesis space. If some prior knowledge 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 $\Psi(\infty) = 0$.

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

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

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

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

To bound the second term in the above equation we observe that

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

The result follows by combining these two inequalities. ■

The sample error consists of two terms. The second term $\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)$ 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 $z \in Z$. It requires more effort to bound 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 $\mathbf{P}(\varepsilon = -1) = \mathbf{P}(\varepsilon = 1) = \frac{1}{2}$. 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

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

If $\mathcal{F}$ is a real-valued function class, $c \in \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$ and h a bounded function, we define the i 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, see, for example, [4].

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* $c \in \mathbb{R}$, *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* $R(\mathcal{F} + h, m) \leq R(\mathcal{F}, m) + \frac{1}{\sqrt{m}}\|f\|_\infty$.

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 $z_j \in Z$, $j \in \mathbb{N}_m$, we have that $\lim_{n\to\infty} f_n(z_i) = f(z_i)$ 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 on the estimation of the sample error. To this end, we need two quantities,

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

and

$$L(t) = \sup_{y \in Y} \sup_{|f_1|, |f_2| \leq t} \frac{|\ell(y, f_1) - \ell(y, f_2)|}{|f_1 - f_2|}.$$

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.

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

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

When the 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*

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

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.

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*

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

Proof. Consider the random variable $\xi = \ell(y, f(x))$. It is easy to see 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, $0 < \xi \le \Phi(\|f\|_\infty)$ which implies that $|\xi - \mathbb{E}\,\xi| \le \Phi(\|f\|_\infty)$. The conclusion follows by applying the one-sided Hoeffding inequality to the random variable ξ, see, for example, [8]. ■

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

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

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

(ii) $\|f_\lambda^*\|_\infty \le \kappa\gamma$

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

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

Proof. The first claim follows directly from the fact that

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

and (1.3). Note that $\mathcal{A}^*(\lambda) \le \mathcal{E}(0) - \mathcal{E}(f_\ell^*) + \lambda\cdot 0 \le \Phi(0)$. 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

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

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

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

The third lemma applies the McDiarmid's inequality ([11]) 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*

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

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 $z_i' = (x_i', y_i')$. One can easily check that

$$g(\mathbf{z}) - g(\mathbf{z}_i') = \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left(\mathcal{E}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}}(\pi_T(f))\right) - \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left(\mathcal{E}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}_i'}(\pi_T(f))\right)$$

$$\leq \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left(\mathcal{E}_{\mathbf{z}_i'}(\pi_T(f)) - \mathcal{E}_{\mathbf{z}}(\pi_T(f))\right)$$

$$= \frac{1}{m} \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left(\ell(y_i', \pi_T(f(x_i'))) - \ell(y_i, \pi_T(f(x_i)))\right) \leq \frac{1}{m} \Phi(\tau)$$

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

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

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

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

and the result follows. ∎

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

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

$$\mathbb{E}\, g(\mathbf{z}) \leq 2\, \mathbb{E} \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \ell(y_i, f(x_i)) \right|$$

$$\leq 2\, \mathbb{E} \sup_{f \in \gamma \mathcal{B}_{\mathcal{K}}} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i (\ell(y_i, f(x_i)) - \ell(y_i, 0)) \right| + \frac{2\Phi(0)}{\sqrt{m}}$$

$$\leq 4L(\tau)\, R(\gamma \mathcal{B}_{\mathcal{K}}, m) + \frac{2\Phi(0)}{\sqrt{m}}$$

$$= 4L(\tau)\, \gamma\, R(\mathcal{B}_{\mathcal{K}}, m) + \frac{2\Phi(0)}{\sqrt{m}}$$

By the definition of $\mathcal{B}_{\mathcal{K}}$ and the reproducing kernel property [2], we have that

$$\sup_{f\in\mathcal{B}_{\mathcal{K}}}\left|\frac{1}{m}\sum_{i\in\mathbb{N}_m}\varepsilon_i f(x_i)\right| = \frac{1}{m}\sup_{K\in\mathcal{K}}\sup_{\|f\|_K\leq 1}\left|\left\langle\sum_{i\in\mathbb{N}_m}\varepsilon_i K_{x_i}, f\right\rangle_K\right|$$

$$= \frac{1}{m}\sup_{K\in\mathcal{K}}\left\|\sum_{i\in\mathbb{N}_m}\varepsilon_i K_{x_i}\right\|_K$$

$$= \frac{1}{m}\sup_{K\in\mathcal{K}}\left(\sum_{i,j\in\mathbb{N}_m}\varepsilon_i\varepsilon_j K(x_i,x_j)\right)^{1/2}$$

$$\leq \frac{1}{\sqrt{m}}\sup_{K\in\mathcal{K}}\left(\sup_{t\in X}\left|\sum_{i\in\mathbb{N}_m}\varepsilon_i K(x_i,t)\right|\right)^{1/2}$$

Hence, by Jensen's inequality we conclude that

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

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. Corollary 3.5, instead, follows by taking $T=\infty$ and $f=f^*_\lambda$ in Theorem 3.4 and using 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 $\mathcal{K}\equiv\mathcal{G}:=\{G_\sigma:\sigma\in(0,\infty)\}$, 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:

Proposition 5.1. *There exists a constant C_n such that $R(\mathcal{K}_0,m)\leq C_n\frac{\log m}{\sqrt{m}}$.*

To prove the proposition we use the notion of Dudley's entropy integral. To this end, we need the concept of covering numbers.

Definition 5.2. Let $(\mathscr{M},d)$ be a pseudo-metric space and $S\subset\mathscr{M}$ a subset. 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 $1 \le p \le \infty$, 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. A recent result of Ying and Zhou [22] asserts that there exists some constant c_n depending only on n such that $\log \mathscr{N}_\infty(\mathcal{K}_0,\varepsilon,m) \le \frac{c_n}{\varepsilon}\left(\log\frac{m}{\varepsilon}\right)^2$. Since $\mathscr{N}_2(\mathcal{F},\varepsilon,m) \le \mathscr{N}_\infty(\mathcal{F},\varepsilon,m)$ we also have that

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

Next, we recall a result from [19] which states 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

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

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

$$R(\mathcal{K}_0,m) \le \frac{c_n}{\sqrt{m}}\int_0^1 \frac{1}{\sqrt{\varepsilon}}\log\frac{m}{\varepsilon}\, d\varepsilon \le C_n\frac{\log m}{\sqrt{m}}$$

for some constant C_n depending only on n. ■

Putting this estimate into Theorem 3.4 or Corollary 3.5, we get the estimate for the sample error. This suggests a way to choose the regularization parameter and compute the learning rate if some prior knowledge is available on the target function so that we can estimate the decay of the regularization error. To further illustrate our results, we consider two classical learning algorithms: regularization networks (RN) and support vector machine (SVM) classification.

5.1 Regularization networks

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. Note that $\kappa = 1$ for the class $\mathcal{G}$ under consideration. In regression problems we assume that $|y| \le M$ almost surely. In RN, 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,

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

Moreover, 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 that $\Phi(t) \le (M+t)^2$ and $L(t) \le 2(M+t)$. Putting these quantities into Corollary 3.5. If $\lambda \le 1$ we obtain, with confidence $1-\delta$, that

$$\mathcal{S}_{\mathbf{z}}(m, \lambda, f_\lambda^*) \le \left(18 + 6\sqrt{2\log\frac{2}{\delta}}\right) M^2 \frac{\sqrt{\log m}}{\lambda m^{1/4}}.$$

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

$$\|f_{\mathbf{z}} - f_\ell^*\|^2_{L^2(\rho_X)} \le C' \left(18M^2 + 6M^2\sqrt{2\log\frac{2}{\delta}}\right)^{\frac{\beta}{1+\beta}} \left(\frac{\sqrt{\log m}}{m^{1/4}}\right)^{\frac{\beta}{1+\beta}}. \tag{5.3}$$

In proving the corollary we have use 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

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

The result follows by setting $A = \left(18 + 6\sqrt{2\log(2/\delta)}\right) M^2\sqrt{\log m}/m^{\frac{1}{4}}$ 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^* extends to a function in*

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

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

$$\mathcal{A}^*(\lambda) \le C\lambda^{\frac{4s}{4n+4s+ns}}$$

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 $1-\delta$

$$\|f_{\mathbf{z}} - f_\ell^*\|^2_{L^2} \le C' \left(18M^2 + 6M^2\sqrt{2\log\frac{2}{\delta}}\right)^{\frac{s}{4n+8s+ns}} \left(\frac{\sqrt{\log m}}{m^{1/4}}\right)^{\frac{4s}{4n+8s+ns}}.$$

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

$$\begin{aligned}
\|f_{\ell,\sigma} - f_\ell^*\|_{L^2}^2 &= \left\| \left(e^{-\frac{\pi\|\cdot\|^2}{\sigma}} - 1 \right) \hat{f}_\ell^* \right\|_{L^2}^2 \\
&= \int_{\|t\| \le \sigma^{\frac{2}{4+s}}} \left(e^{-\frac{\pi\|t\|^2}{\sigma}} - 1 \right)^2 |\hat{f}_\ell^*(t)|^2 dt + \int_{\|t\| > \sigma^{\frac{2}{4+s}}} \left(e^{-\frac{\pi\|t\|^2}{\sigma}} - 1 \right)^2 |\hat{f}_\ell^*(t)|^2 dt \\
&\le \int_{\|t\| \le \sigma^{\frac{2}{4+s}}} \pi^2 \sigma^{-\frac{2s}{4+s}} |\hat{f}_\ell^*(t)|^2 dt + \int_{\|t\| > \sigma^{\frac{2}{4+s}}} \sigma^{-\frac{2s}{4+s}} |\hat{f}_\ell^*(t)|^2 \|t\|^s dt \\
&\le \left(\pi^2 \|f_\ell^*\|_{L^2}^2 + \|f_\ell^*\|_{H^s}^2 \right) \sigma^{-\frac{2s}{4+s}}.
\end{aligned}$$

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

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

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

By the analysis in [17], for any fixed $\beta, \sigma > 0$, $\inf_{f \in \mathcal{H}_{G_\sigma}} \{\mathcal{E}(f) - \mathcal{E}(f_\ell^*) + \lambda \|f\|_K\}$ 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 impossible. Thus, Example 1 ensures that algorithm (1.2) significantly improves the approximation power and learning ability.

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 misclassification error $\mathcal{R}(f) = \mathbf{P}\{f(X) \ne Y\}$. The optimal classifier, which yields the minimal misclassification 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 $X_y := \left\{x \in X : \mathbf{P}(y|x) > \frac{1}{2}\right\}$, and $X_0 := \left\{x \in X : \mathbf{P}(1|x) = \frac{1}{2}\right\}$ then f^* 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 misclassification error.

SVM classification uses the loss function $\ell(y, f(x)) = \max\{1 - yf(x), 0\}$ and the target function is $f_\ell^* = f^*$, [18]. It computes a real-valued function $f_{\mathbf{z}}$ on X and gives the classifier $\operatorname{sgn}(f_{\mathbf{z}})$.

In classification problems, we need only the sign of the functions to predict the label. Hence, the projection will be used.Observe that $\Psi(T) = 0$ if $T \ge 1$ and, for all t, notice that $\Phi(t) = (1+t)$ and $L(t) = 1$. Combining Proposition 3.1, Theorem 3.4 with $T = 1$ and $f = f_\lambda^*$, Proposition 5.1 we obtain

$$\mathcal{E}(\pi_1(f_{\mathbf{z}})) - \mathcal{E}(f^*) \le \left(6 + 3\sqrt{2\log\tfrac{2}{\delta}}\right) \frac{\sqrt{\log m}}{\lambda^{1/2} m^{1/4}} + \mathcal{A}^*(\lambda).$$

recalling, for all real-valued functions $f : X \to Y$, that

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

(see, for example, [23, 20]) 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, with confidence $1 - \delta$*

$$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f^*) \leq C' \left(6 + 3\sqrt{2 \log \tfrac{2}{\delta}}\right)^{\frac{2\beta}{1+2\beta}} \left(\frac{\log m}{\sqrt{m}}\right)^{\frac{\beta}{1+2\beta}}.$$

As an example of this observation we consider distributions satisfying the geometric noise condition, [16]. 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 $t > 0$,

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

Thus, applying [16, 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 $1 - \delta$*

$$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f^*) \leq D' \left(6 + 3\sqrt{2 \log \tfrac{2}{\delta}}\right)^{\frac{2\alpha}{3\alpha+1}} \left(\frac{\log m}{\sqrt{m}}\right)^{\frac{\alpha}{3\alpha+1}}.$$

6 Conclusion

We have provided an analysis of the generalization error for learning a kernel function within a broader class $\mathcal{K}$ than previously considered in the literature. 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 regularization networks and support vector machines. In the future we plan to apply this analysis to other useful kernel classes $\mathcal{K}$. In particular, [14] considers the convex set of kernels parameterized by a locally compact set Σ, namely $\mathcal{K} = \left\{\int_\Sigma G(\sigma)dp(\sigma) : p \in \mathcal{P}(\Sigma)\right\}$, 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. 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 $\mathcal{K}_0$ and $\mathcal{G}_0$ are the same. This study may reveal good kernel classes $\mathcal{K}$ which have with faster learning rates than the one described here.

References

[1] A. Argyriou, C.A. Micchelli and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. *Proc. 18-th Annual Conf. on Learning Theory (COLT'05)*, 2005.

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

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

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

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

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

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

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

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

[10] 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.

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

[12] Y. Lee, Y. Kim, S. Lee and J.Y. Koo. Structured multicategory support vector machine with ANOVA decomposition. Tech. Rep. 743, Dept of Statistics, The Ohio State University, 2004.

[13] Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models – COSSO. Institute of Statistics Mimeo Series 2556, NCSU, January 2003.

[14] C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. Preprint, 2004.

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

[16] C. Scovel and I. Steinwart. Fast rates for support vector machines. Preprint, 2004.

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

[18] 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.

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

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

[21] Q. Wu, Y.M. Ying and D.X. Zhou. Multi-kernel regularized classifiers. Preprint, 2004.

[22] Y.M. Ying and D.X. Zhou. Learnability of Gaussians with flexible variances, Preprint, 2004.

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