Learning Theory Estimates via Integral Operators and Their Approximations†

Steve Smale
Toyota Technological Institute at Chicago
1427 East 60th Street, Chicago, IL 60637, USA
E-mail: smale@math.berkeley.edu

Ding-Xuan Zhou
Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, CHINA
E-mail: mazhou@math.cityu.edu.hk

April 13, 2005

§1. Introduction

This report on learning theory is written in the spirit of:

The best understanding of what one can see comes from theories of what one can't see.

This thought has been expressed in a number of ways by different scientists, and is supported everywhere. Obvious choices vary from gravity to economic equilibrium. For learning theory we see its expression in the focus on the regression function defined by an unknown measure and through data independent estimates.

This perspective on learning theory is hardly novel with us. Already in the last century, Niyogi and Girosi [6] wrote in this style.

A basic model we shall take throughout the paper is to assume that samples are drawn from a (joint) probability measure ρ on $Z = X \times Y$ with a compact metric space X and $Y = \mathbb{R}$. Our primary objective is the **regression function** of ρ defined as

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X.$$

Here $\rho(y|x)$ is the conditional distribution at x induced by ρ.

† The first author is partially supported by NSF grant 0325113. The second author is supported by the Research Grants Council of Hong Kong [Project No. CityU 103704].

The regression problem in learning theory (see [2, 6] and the references therein) aims at good approximations $f_{\mathbf{z}}$ of the regression function, constructed by learning algorithms from a set of random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn independently according to ρ. To understand the approximation, we estimate the error $\|f_{\mathbf{z}} - f_\rho\|_\infty$ or $\|f_{\mathbf{z}} - f_\rho\|_{C^s}$ or $\|f_{\mathbf{z}} - f_\rho\|_\rho$, where $\|f\|_\rho = \|f\|_{L^2_{\rho_X}} = \{\int_X |f(x)|^2 d\rho_X\}^{1/2}$ denotes the L^2 norm in the space $L^2_{\rho_X}$ and ρ_X the marginal distribution of ρ on X.

The learning algorithm we investigate in this paper is a Tikhonov regularization scheme associated with Mercer kernels.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, *i.e.*, for any finite set of distinct points $\{x_1, \cdots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is positive semidefinite. Such a kernel is called a **Mercer kernel**.

The **Reproducing Kernel Hilbert Space** (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined to be the closure [1] of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product[1] denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$.

The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K. \tag{1.1}$$

Denote $\kappa = \sqrt{\sup_{x \in X} K(x, x)}$. Then (1.1) implies that $\mathcal{H}_K \subset C(X)$ and

$$\|f\|_\infty \leq \kappa \|f\|_K, \qquad \forall f \in \mathcal{H}_K. \tag{1.2}$$

[1] Notice that the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is only positive semidefinite, it is possible that for a nonzero vector $(c_i)_{i=1}^\ell$ there holds $\sum_{i,j=1}^\ell c_i K(x_i, x_j) c_j = 0$. However, as a function on X, $\sum_{i=1}^\ell c_i K_{x_i} \equiv 0$. To show this [1], take an arbitrary point $x_{\ell+1} \in X$. By the definition of the Mercer kernel, the $(\ell+1) \times (\ell+1)$ matrix $(K(x_i, x_j))_{i,j=1}^{\ell+1}$ is still positive semidefinite. It follows that the quadratic function of the real variable $t = c_{\ell+1}$

$$\sum_{i,j=1}^{\ell+1} c_i K(x_i, x_j) c_j = 0 + 2 \sum_{i=1}^{\ell} c_i K(x_i, x_{\ell+1}) t + K(x_{\ell+1}, x_{\ell+1}) t^2$$

is nonnegative everywhere. By letting $t \to \pm 0$, we see that $\sum_{i=1}^\ell c_i K(x_i, x_{\ell+1}) = 0$, that is, the function $\sum_{i=1}^\ell c_i K_{x_i}$ vanishes on the arbitrary point $x_{\ell+1}$, hence is zero identically on X. This shows that $\|\cdot\|_K$ is not only a seminorm, but a norm of the Hilbert space $\mathcal{H}_K$.

The learning algorithm we study here is a Tikhonov regularized one as in [5] with $\lambda > 0$:

$$\textbf{Learning Scheme} \qquad f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m}\big(f(x_i)-y_i\big)^2+\lambda\|f\|_K^2\right\}. \tag{1.3}$$

To understand (1.3), following our previous studies on Shannon sampling [10, 11], we define the **sampling operator** $S_{\mathbf{x}} : \mathcal{H}_K \to \mathrm{I\!R}^m$ associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of X by

$$S_{\mathbf{x}}(f) = \big(f(x_i)\big)_{i=1}^m.$$

The adjoint of the sampling operator, $S_{\mathbf{x}}^T : \mathrm{I\!R}^m \to \mathcal{H}_K$, is given by

$$S_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K_{x_i}, \qquad c \in \mathrm{I\!R}^m.$$

We know from [2, 11] that a solution $f_{\mathbf{z},\lambda}$ of (1.3) exists, is unique and given by

$$f_{\mathbf{z},\lambda} = \big(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1}\frac{1}{m}S_{\mathbf{x}}^T y. \tag{1.4}$$

Our goal is to understand how $f_{\mathbf{z},\lambda}$ approximates f_ρ and how the decay of the regularization parameter $\lambda = \lambda(m)$ leads to convergence rates. The rates for this approximation in $L^2_{\rho_X}$ have been considered in [3, 4, 16, 11, 14], while the approximation in the space $\mathcal{H}_K$ (hence in $L^\infty_{\rho_X}$ by (1.2) and in C^s by [17]) has been shown in [11]. (An early version of Theorem 1 below appeared in a late version of [11], and was subsequently removed.) In this paper we provide a simpler approach with stronger convergence rates.

§2. Main Results on the Errors in $\mathcal{H}_K$

A data-free limit of (1.3) is

$$f_\lambda := \arg\min_{f\in\mathcal{H}_K}\big\{\|f-f_\rho\|_\rho^2+\lambda\|f\|_K^2\big\}. \tag{2.1}$$

Since $\lambda > 0$, a solution of (2.1) exists, is unique and given by [3]

$$f_\lambda = \big(L_K+\lambda I\big)^{-1}L_K f_\rho, \tag{2.2}$$

where $L_K : L^2_{\rho_X} \to \mathcal{H}_K$ is an integral operator defined by

$$L_K(f)(x) := \int_X K(x,y)f(y)d\rho_X(y), \qquad x \in X.$$

The operator L_K can also be defined as a self-adjoint operator on $\mathcal{H}_K$ or on $L^2_{\rho_X}$. We shall use the same notion L_K for these operators defined on different domains.

Towards estimating $f_{\mathbf{z},\lambda} - f_\rho$ in various norms, compare (1.4) with (2.2). First consider the random variable $\xi := yK_x$ on (Z,ρ) with values in the Hilbert space $\mathcal{H}_K$. We see that

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) = \frac{1}{m}\sum_{i=1}^m y_i K_{x_i} = \frac{1}{m}S_{\mathbf{x}}^T y, \qquad E(\xi) = \int_X K_x \int_Y y d\rho(y|x) d\rho_X(x) = L_K f_\rho$$

which shows that $\frac{1}{m}S_{\mathbf{x}}^T y$ is a good approximation of $L_K f_\rho$. Second with a function $f \in \mathcal{H}_K$, look at the random variable $\xi := f(x)K_x$ on (X,ρ_X) with values in $\mathcal{H}_K$. Again we have

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) = \frac{1}{m}\sum_{i=1}^m f(x_i) K_{x_i} = \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} f, \qquad E(\xi) = \int_X K_x f(x) d\rho_X(x) = L_K f$$

meaning that $\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}}$ is a good approximation of L_K. Thus $\big(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1}$ should approximate $\big(L_K + \lambda I\big)^{-1}$ well, and one would expect from (1.4) and (2.2) good error analysis of $f_{\mathbf{z},\lambda} - f_\lambda$ in the space $\mathcal{H}_K$. Such a result following this idea is stated in the following Theorem 1. The proof will be carried out in detail in Section 3 by applying a Bennett inequality to the random variable $(y - f_\lambda(x))K_x$ with values in the Hilbert space $\mathcal{H}_K$.

We assume that for some $M \geq 0$, $|y| \leq M$ almost surely, that is, $\rho(y|x)$ is supported on $[-M, M]$ for almost every $x \in X$. Then $\|f_\rho\|_\rho \leq \|f_\rho\|_\infty \leq M$.

Theorem 1. *Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \leq M$ almost surely. Then for any $0 < \delta < 1$, with confidence $1 - \delta$ there holds*

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{6\kappa M \log\big(2/\delta\big)}{\sqrt{m}\lambda}.$$

Using Theorem 1, we will prove our total error estimates in the $\|\cdot\|_K$ norm.

Theorem 2. *Let* $\mathbf{z}$ *be randomly drawn according to* ρ *satisfying* $|y| \leq M$ *almost surely. Assume that* f_ρ *is in the range of* L_K^r *for some* $\frac{1}{2} < r \leq 1$. *Take the regularization parameter as* $\lambda = \left(3\kappa M / \|L_K^{-r} f_\rho\|_\rho\right)^{\frac{2}{1+2r}} m^{-\frac{1}{1+2r}}$. *For any* $0 < \delta < 1$, *with confidence* $1-\delta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq 4 \log\big(2/\delta\big) \left(3\kappa M\right)^{\frac{2r-1}{2r+1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{2}{1+2r}} \left(\frac{1}{m}\right)^{\frac{2r-1}{4r+2}}. \tag{2.3}$$

In the estimate (2.3), $\|L_K^{-r} f_\rho\|_\rho$ is a key factor, but also perhaps the most elusive factor. It is finite by the hypothesis that f_ρ lies in the range of L_K^r. Here L_K^r makes sense as the rth power of L_K since $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ is self-adjoint and non-negative. In fact, the image of L_K^r is contained in $\mathcal{H}_K$ if $r \geq 1/2$. Then $\|L_K^{-r} f_\rho\|_\rho$ measures a complexity of the regression function. Think of f_ρ with many oscillations having this measure large.

The convergence in $\mathcal{H}_K$ implies the convergence in $C^s(X)$ under some conditions on K. Here $C^s(X)$ is the space of all functions on $X \subset \mathrm{I\!R}^n$ whose partial derivatives up to order s are continuous with $\|f\|_{C^s(X)} = \sum_{|\alpha| \leq s} \|D^\alpha f\|_\infty$, and $C^{s+\epsilon}(X)$ denotes the subspace (of $C^s(X)$) of functions with these partial derivatives to be Hölder ϵ on X.

It was proved in [17] that when $K \in C^{2s+\epsilon}(X \times X)$ with $0 < \epsilon < 2$ and X is the closure of a domain in $\mathrm{I\!R}^n$, the inclusion $\mathcal{H}_K \subset C^{s+\epsilon/2}(X)$ is well defined and bounded. But the norm of the inclusion, depending on X, was not explicitly given in [17]. Here we find the norm of the well defined inclusion $\mathcal{H}_K \subset C^s(X)$ as

$$\|f\|_{C^s(X)} \leq 4^s \|K\|_{C^{2s}}^{1/2} \|f\|_K, \qquad \forall f \in \mathcal{H}_K. \tag{2.4}$$

To see this, let $x \in X$ and $h \in \mathrm{I\!R}^n$ such that $x+h, \ldots, x+sh \in X$. Then the reproducing property (1.1) tells us that

$$\left| |h|^{-s} \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} f(x+jh) \right| = \left| \langle f, |h|^{-s} \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} K_{x+jh} \rangle_K \right|$$
$$\leq \|f\|_K \left| |h|^{-s} \sum_{i=0}^{s} \binom{s}{i} (-1)^{s-i} |h|^{-s} \sum_{j=0}^{s} \binom{s}{j} (-1)^{s-j} K(x+ih, x+jh) \right|^{1/2}.$$

Taking h to be vectors along an axis with $|h| \to 0$ gives bounds for the partial derivatives. For $\alpha \in \mathbb{Z}_+^n$ with $|\alpha| \leq s$, we have $\|D^\alpha f\|_\infty \leq \|K\|_{C^{2s}}^{1/2} \|f\|_K$. This proves (2.4). Then Theorem 2 in connection with (2.4) implies the following convergence rate in $C^s(X)$.

Corollary 1. *Under the assumption and the choice of λ in Theorem 2, if X is the closure of a domain in $\mathbb{R}^n$ and K is $C^{2s+\epsilon}$ for some $s \in \mathbb{N}$ and $\epsilon > 0$, then with confidence $1-\delta$,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{C^s(X)} \le 4^{1+s} \log\bigl(2/\delta\bigr) \|K\|_{C^{2s}}^{\frac{6r-1}{4r+2}} (3M)^{\frac{2r-1}{2r+1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{2}{1+2r}} \left(\frac{1}{m}\right)^{\frac{2r-1}{4r+2}}. \tag{2.5}$$

The extreme situation is when $r=1$. In this case, we have

Corollary 2. *Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \le M$ almost surely. If $\|L_K^{-1} f_\rho\|_\rho < \infty$ and $\lambda = \bigl(3\kappa M / \|L_K^{-1} f_\rho\|_\rho\bigr)^{2/3} m^{-1/3}$, with confidence $1-\delta$ we have*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \le 4 \log\bigl(2/\delta\bigr) (3\kappa M)^{1/3} \|L_K^{-1} f_\rho\|_\rho^{2/3} \left(\frac{1}{m}\right)^{1/6}.$$

If moreover, X is the closure of a domain in $\mathbb{R}^n$ and $K \in C^{2s+\epsilon}(X \times X)$, then

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{C^s(X)} \le 4^{1+s} \log\bigl(2/\delta\bigr) \|K\|_{C^{2s}}^{\frac{5}{6}} (3M)^{1/3} \|L_K^{-1} f_\rho\|_\rho^{2/3} \left(\frac{1}{m}\right)^{1/6}.$$

Remark. *The other extreme is when $r \to 1/2$. In this case, the function f_ρ lies in an interpolation space between the range of L_K and $\mathcal{H}_K$ which tends to be arbitrarily close to $\mathcal{H}_K$. The power $(2r-1)/(4r+2)$ for the convergence rate becomes arbitrarily small.*

§3. Probability Estimates by Vector-Valued Bennett Inequalities

We apply the following Bennett inequality for vector-valued random variables to improve some previous probability estimates of $\|f_{\mathbf{z},\lambda} - f_\rho\|$. It is derived from [7, Theorem 3.4] and the elementary inequality $t\log(1+t) \ge 2t - 2\log(1+t)$ for any $t>0$. We thank Yuan Yao for bringing our attention to this reference.

Lemma 1. *Let H be a Hilbert space and $\{\xi_i\}_{i=1}^m$ be m ($m<\infty$) independent random variables with values in H. Suppose that for each i, $\|\xi_i\| \le \widetilde{M} < \infty$ almost surely. Denote $\sigma^2 = \sum_{i=1}^m E(\|\xi_i\|^2)$. Then*

$$Prob\left\{\left\|\frac{1}{m}\sum_{i=1}^m \bigl[\xi_i - E(\xi_i)\bigr]\right\| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon}{2\widetilde{M}} \log\left(1 + \frac{m\widetilde{M}\varepsilon}{\sigma^2}\right)\right\}, \qquad \forall \varepsilon > 0. \tag{3.1}$$

In our situation, $\{\xi_i\}$ are independent drawers of a random variable.

Lemma 2. *Let H be a Hilbert space and ξ be a random variable on (Z, ρ) with values in H. Assume $\|\xi\| \leq \widetilde{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = E(\|\xi\|^2)$. Let $\{z_i\}_{i=1}^m$ be independent random drawers of ρ. For any $0 < \delta < 1$, with confidence $1 - \delta$,*

$$\left\| \frac{1}{m} \sum_{i=1}^m \bigl[\xi_i - E(\xi_i)\bigr] \right\| \leq \frac{2\widetilde{M} \log\bigl(2/\delta\bigr)}{m} + \sqrt{\frac{2\sigma^2(\xi) \log\bigl(2/\delta\bigr)}{m}}. \tag{3.2}$$

Proof. We apply Lemma 1 to the independent random variables $\{\xi(z_i)\}_{i=1}^m$, and know that for any $\varepsilon > 0$

$$\mathrm{Prob}\left\{ \left\| \frac{1}{m} \sum_{i=1}^m \bigl[\xi(z_i) - E(\xi)\bigr] \right\| \geq \varepsilon \right\} \leq 2 \exp\left\{ -\frac{m\varepsilon}{2\widetilde{M}} \log\left(1 + \frac{\widetilde{M}\varepsilon}{\sigma^2(\xi)}\right) \right\}.$$

Observe that

$$\log(1+t) \geq t/(1+t), \qquad \forall t > 0. \tag{3.3}$$

It follows by taking $t = \frac{\widetilde{M}\varepsilon}{\sigma^2(\xi)}$ that

$$\mathrm{Prob}\left\{ \left\| \frac{1}{m} \sum_{i=1}^m \bigl[\xi(z_i) - E(\xi)\bigr] \right\| \geq \varepsilon \right\} \leq 2 \exp\left\{ -\frac{m\varepsilon}{2\widetilde{M}} \left(\frac{\widetilde{M}\varepsilon}{\widetilde{M}\varepsilon + \sigma^2(\xi)} \right) \right\}.$$

The probability on the right side equals $2\exp\bigl\{-\frac{m\varepsilon^2}{2\widetilde{M}\varepsilon + 2\sigma^2(\xi)}\bigr\}$. Choosing $\varepsilon > 0$ for this probability equal to δ is the same as solving the quadratic equation

$$m\varepsilon^2 = \log\bigl(2/\delta\bigr)\bigl(2\widetilde{M}\varepsilon + 2\sigma^2(\xi)\bigr).$$

We find that with confidence $1 - \delta$ there holds

$$\left\| \frac{1}{m} \sum_{i=1}^m \bigl[\xi(z_i) - E(\xi)\bigr] \right\| \leq \frac{2\widetilde{M} \log\bigl(2/\delta\bigr)}{m} + \sqrt{\frac{2\sigma^2(\xi) \log\bigl(2/\delta\bigr)}{m}}.$$

This is the desired bound. □

Now we can prove our main result.

Proof of Theorem 1. By (1.4), write

$$f_{\mathbf{z},\lambda} - f_\lambda = \bigl(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\bigr)^{-1} \left\{ \frac{1}{m} S_{\mathbf{x}}^T y - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} f_\lambda - \lambda f_\lambda \right\}.$$

Observe that

$$\frac{1}{m}S_{\mathbf{x}}^T y - \frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} f_\lambda = \frac{1}{m}\sum_{i=1}^{m}\big(y_i - f_\lambda(x_i)\big)K_{x_i}$$

and by the definition (2.2) of f_λ,

$$\lambda f_\lambda = L_K\big(f_\rho - f_\lambda\big).$$

It follows that for all $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, and $\lambda > 0$,

$$f_{\mathbf{z},\lambda} - f_\lambda = \big(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1}\left\{\frac{1}{m}\sum_{i=1}^{m}\big(y_i - f_\lambda(x_i)\big)K_{x_i} - L_K\big(f_\rho - f_\lambda\big)\right\}. \tag{3.4}$$

This gives a bound for the error in the $\mathcal{H}_K$-norm

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{1}{\lambda}\Delta, \quad \Delta := \left\|\frac{1}{m}\sum_{i=1}^{m}\big(y_i - f_\lambda(x_i)\big)K_{x_i} - L_K\big(f_\rho - f_\lambda\big)\right\|_K. \tag{3.5}$$

To estimate Δ, we apply Lemma 2 to the random variable $\xi = \big(y - f_\lambda(x)\big)K_x$ on (Z, ρ) with values in the Hilbert space $\mathcal{H}_K$. It satisfies

$$E(\xi) = \int_X K_x \int_Y \big(y - f_\lambda(x)\big)d\rho(y|x)d\rho_X(x) = L_K\big(f_\rho - f_\lambda\big)$$

and $\|\xi\|_K = |y - f_\lambda(x)|\sqrt{K(x,x)}$. Thus $\sigma^2(\xi) \leq \kappa^2 \int_Z (f_\lambda(x) - y)^2 d\rho$ and almost surely

$$\|\xi\|_K \leq \kappa(M + \|f_\lambda\|_\infty) =: \widetilde{M}.$$

It follows from (3.2) that with confidence $1 - \delta$ there holds

$$\Delta \leq \frac{2\kappa(M + \|f_\lambda\|_\infty)\log\big(2/\delta\big)}{m} + \kappa\sqrt{\frac{2\int_Z (f_\lambda(x) - y)^2 d\rho \log\big(2/\delta\big)}{m}}. \tag{3.6}$$

Note that the definition of the regression function yields

$$\int_Z (f(x) - y)^2 d\rho - \int_Z (f_\rho(x) - y)^2 d\rho = \|f - f_\rho\|_\rho^2, \qquad \forall f : X \to Y. \tag{3.7}$$

Recall the definition (2.1) of f_λ. Taking $f = 0$ yields $\|f_\lambda - f_\rho\|_\rho^2 + \lambda\|f_\lambda\|_K^2 \leq \|f_\rho\|_\rho^2$. Hence

$$\|f_\lambda - f_\rho\|_\rho \leq \|f_\rho\|_\rho \qquad \text{and} \qquad \|f_\lambda\|_K \leq \|f_\rho\|_\rho/\sqrt{\lambda}. \tag{3.8}$$

By (3.8), we have $\|f_\lambda - f_\rho\|_\rho \le M$ and $\|f_\lambda\|_K \le \frac{M}{\sqrt{\lambda}}$. It follows from (3.7) with $f=0$ and $f = f_\lambda$ that $\int_Z (f_\rho(x) - y)^2 d\rho \le \int_Z (0-y)^2 d\rho \le M^2$, thereby $\int_Z (f_\lambda(x)-y)^2 d\rho \le 2M^2$; and from (1.2) that $\|f_\lambda\|_\infty \le \kappa \|f_\lambda\|_K \le \kappa M/\sqrt{\lambda}$. Therefore, with confidence $1-\delta$ we have

$$\Delta \le \frac{2\kappa M(1+\kappa/\sqrt{\lambda})\log(2/\delta)}{m} + 2\kappa M\sqrt{\frac{\log(2/\delta)}{m}}. \tag{3.9}$$

If $\frac{\kappa}{\sqrt{m\lambda}} \le \frac{1}{3\log(2/\delta)}$, the above estimate can be bounded further as

$$\Delta \le \frac{2\kappa M\log(2/\delta)}{m} + \frac{2\kappa M\log(2/\delta)}{\sqrt{m}}\frac{\kappa}{\sqrt{m\lambda}} + \frac{2\kappa M\log(2/\delta)}{\sqrt{m}}\frac{1}{\sqrt{\log(2/\delta)}} \le \frac{6\kappa M\log(2/\delta)}{\sqrt{m}}.$$

This yields the desired bound when $\frac{\kappa}{\sqrt{m\lambda}} \le \frac{1}{3\log(2/\delta)}$.

When $\frac{\kappa}{\sqrt{m\lambda}} > \frac{1}{3\log(2/\delta)}$, we have $\frac{6\kappa M\log(2/\delta)}{\sqrt{m\lambda}} \ge \frac{2M}{\sqrt{\lambda}}$. In this case, we use (3.8) and the trivial bound $\|f_{\mathbf{z},\lambda}\|_K \le M/\sqrt{\lambda}$ seen from (1.3) by taking $f=0$. Then there holds $\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \le 2M/\sqrt{\lambda}$ with probability 1. So the desired inequality also holds in the second case. This proves Theorem 1. □

To get the total error estimates stated in Theorem 2, we need bounds for the approximation error $\|f_\lambda - f_\rho\|$. Recall [11, Theorem 4 and equation (7.10)].

Lemma 3. *Define f_λ by (2.2). If $L_K^{-r} f_\rho \in L^2_{\rho_X}$, then*

$$\|f_\lambda - f_\rho\|_\rho^2 + \lambda\|f_\lambda\|_K^2 \le \lambda^{2r}\|L_K^{-r}f_\rho\|_\rho^2, \qquad \text{if} \quad 0 < r \le \frac{1}{2} \tag{3.10}$$

and

$$\|f_\lambda - f_\rho\|_K \le \lambda^{r-\frac{1}{2}}\|L_K^{-r}f_\rho\|_\rho, \qquad \text{if} \quad \frac{1}{2} < r \le 1. \tag{3.11}$$

Moreover, for $0 < r \le 1$, there holds

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|L_K^{-r} f_\rho\|_\rho. \tag{3.12}$$

The bound (3.10) estimates the regularization error [10]. It is only used for the proof of Corollary 3 below.

Proof of Theorem 2. Combining Theorem 1 with (3.11), we find that with confidence $1-\delta$, the total error satisfies

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \le \|f_{\mathbf{z},\lambda} - f_\lambda\|_K + \|f_\lambda - f_\rho\|_K \le 2\log(2/\delta)\left\{\frac{3\kappa M}{\sqrt{m\lambda}} + \lambda^{r-\frac{1}{2}}\|L_K^{-r}f_\rho\|_\rho\right\}.$$

Minimize the right hand side over $\lambda > 0$ to obtain

$$\lambda = \big(3\kappa M / \|L_K^{-r} f_\rho\|_\rho\big)^{\frac{2}{1+2r}} \left(\frac{1}{m}\right)^{\frac{1}{1+2r}}.$$

With this choice of λ, the bound becomes (2.3). This proves Theorem 2. □

§4. Distributions with Small Variances

In Theorem 1, we only assume that $|y| \le M$ almost surely. That is, for almost every $x \in X$, the conditional distribution $\rho(\cdot|x)$ is supported on $[-M, M]$. Notice that the mean of $\rho(\cdot|x)$ is $f_\rho(x)$ and the variance is $\int_Y (f_\rho(x) - y)^2 d\rho(y|x)$. It is natural to define the variance of ρ as the average variance of the conditional distributions.

Definition 1. *The* **variance** *of ρ is defined to be*

$$\sigma_\rho^2 = \int_Z (f_\rho(x) - y)^2 d\rho = \int_X \int_Y (f_\rho(x) - y)^2 d\rho(y|x) d\rho_X(x).$$

If some conditions are assumed on the variance (not only boundedness), Theorem 1 can be improved, as follows.

Theorem 3. *Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \le M$ almost surely. Then for any $0 < \delta < 1$, with confidence $1 - \delta$ we have*

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \le 2\kappa \log\big(2/\delta\big) \left\{ \frac{\sqrt{\sigma_\rho^2} + \|f_\lambda - f_\rho\|_\rho}{\sqrt{m}\lambda} + \frac{M + \kappa\|f_\lambda\|_K}{m\lambda} \right\}.$$

Proof. Applying (3.7) and (1.2) to (3.6), we get

$$\Delta \le \frac{2\kappa(M + \kappa\|f_\lambda\|_K)\log\big(2/\delta\big)}{m} + \kappa\sqrt{\frac{2\log\big(2/\delta\big)\big(\sigma_\rho^2 + \|f_\lambda - f_\rho\|_\rho\big)^2}{m}}.$$

Since $\sqrt{2\log(2/\delta)} < 2\log\big(2/\delta\big)$, our conclusion follows. □

Notice the similarity between the first term $2\kappa \log\big(2/\delta\big)\sqrt{\sigma_\rho^2}/(\sqrt{m}\lambda)$ of the bound in Theorem 3 and the error estimate $6\kappa M \log\big(2/\delta\big)/(\sqrt{m}\lambda)$ of Theorem 1 when the variance σ_ρ^2 is not small.

When the variance vanishes (i.e., when the distribution is noise-free), Theorem 3 provides better error analysis than Theorem 1: $\|f_\lambda - f_\rho\|_\rho \to 0$ if f_ρ can be approximated by $\mathcal{H}_K$ in $L^2_{\rho_X}$, and the second term of the bound in Theorem 3 is of higher order. One example of noise-free situation is the PAC learning (Probably Approximately Correct).

Corollary 3. *Let* $\mathbf{z}$ *be randomly drawn according to* ρ *satisfying* $y = f_\rho(x)$ *(i.e.,* $\sigma_\rho^2 = 0$*) and* $|y| \leq M$ *almost everywhere. Assume that* f_ρ *is in the range of* L_K^r *for some* $\frac{1}{2} < r < 1$*. Take* $\lambda = \left(\frac{2\kappa M}{m\|L_K^{-r} f_\rho\|_\rho}\right)^{\frac{2}{1+2r}}$*. For any* $0 < \delta < 1$*, with confidence* $1 - \delta$ *we have*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq 4 \log\big(2/\delta\big)\big(2\kappa M\big)^{\frac{2r-1}{2r+1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{2}{2r+1}} \left(\frac{1}{m}\right)^{\frac{2r-1}{2r+1}},$$

provided that m *is large enough in the following sense*

$$m \geq \big(2\kappa M\big)^{\frac{4r}{2r-1}} \big(M + \kappa\|L_K^{-1/2} f_\rho\|_\rho\big)^{\frac{2+4r}{2r-1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{2}{1-2r}}. \tag{4.1}$$

Proof. Since $r > \frac{1}{2}$, the range of L_K^r is a subset of the range of $L_K^{1/2}$. By (3.10) with r replaced by $1/2$, we find that

$$\lambda \|f_\lambda\|_K^2 \leq \lambda \|L_K^{-1/2} f_\rho\|_\rho^2.$$

This implies that

$$\|f_\lambda\|_K \leq \|L_K^{-1/2} f_\rho\|_\rho.$$

Using the assumption $y = f_\rho(x)$ almost surely and (3.12), we know from Theorem 3 that for any $0 < \delta < 1$, with confidence $1 - \delta$

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{2\kappa \log\big(2/\delta\big)}{\sqrt{m}\lambda} \left\{\lambda^r \|L_K^{-r} f_\rho\|_\rho + \frac{M}{\sqrt{m}} + \frac{\kappa \|L_K^{-1/2} f_\rho\|_\rho}{\sqrt{m}}\right\}.$$

Balancing the two terms $\lambda^r \|L_K^{-r} f_\rho\|_\rho$ and $\big(M + \kappa\|L_K^{-1/2} f_\rho\|_\rho\big)/\sqrt{m}$, we see that for

$$\lambda \leq \left\{\big(M + \kappa\|L_K^{-1/2} f_\rho\|_\rho\big)/\|L_K^{-r} f_\rho\|_\rho\right\}^{1/r} \big(1/m\big)^{1/(2r)} \tag{4.2}$$

there holds with confidence $1 - \delta$

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \leq \frac{4\kappa M \log\big(2/\delta\big)}{m\lambda}.$$

This in connection with (3.11) tells us that with confidence $1 - \delta$

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq 2 \log\big(2/\delta\big) \left\{\lambda^{r - \frac{1}{2}} \|L_K^{-r} f_\rho\|_\rho + \frac{2\kappa M}{m\lambda}\right\}.$$

Again, balancing the above two terms, we know that for $\lambda = \left(\frac{2\kappa M}{m\|L_K^{-r} f_\rho\|_\rho}\right)^{\frac{2}{1+2r}}$, the error $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ is bounded by $8 \log\big(2/\delta\big)\kappa M/(m\lambda)$ with confidence $1 - \delta$. With this choice of λ, when m satisfies the restriction (4.1), we know that (4.2) holds. This verifies the desired bound for $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$. □

In the case that f_ρ lies in the range of L_K, we have for noise-free distributions the convergence rate of $O(m^{-1/3})$ for $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$.

§5. Application to Classification Algorithms

One application of our error analysis in $\mathcal{H}_K$ is for binary classification algorithms[2]. If we label the two classes by $\{1, -1\}$, we can consider ρ as a distribution supported on $X \times \{1, -1\}$. A **binary classifier** f is a function from X to $\{1, -1\}$, and it assigns a label $f(x) \in \{1, -1\}$ for each point $x \in X$. Since $\rho(\cdot|x)$ is supported only on two points $\{1, -1\}$, we have $f_\rho(x) = \int_{\rm I\!R} y d\rho(y|x) = P(y = 1|x) - P(y = -1|x)$. It follows that

$$P(y = \operatorname{sgn}(f_\rho(x))|x) \geq P(y \neq \operatorname{sgn}(f_\rho(x))|x).$$

Note that for $y \in \{1, -1\}$, $y \neq \operatorname{sgn}(f_\rho(x))$ is the same as $|y - \operatorname{sgn}(f_\rho(x))| = 2$. Thus, for each $x \in X$, the class $y = \operatorname{sgn}(f_\rho(x))$ has larger probability. This shows that the best classifier, called the **Bayes rule**, is given by

$$\operatorname{sgn}(f_\rho(x)) = \begin{cases} 1, & \text{if } P(y = 1|x) \geq P(y = -1|x), \\ -1, & \text{if } P(y = 1|x) < P(y = -1|x). \end{cases} \tag{5.1}$$

The distance between a classifier f and the Bayes rule is measured in L^2 by

$$\|f - \operatorname{sgn}(f_\rho)\|_\rho = \left(\int_X \big(f(x) - \operatorname{sgn}(f_\rho)(x)\big)^2 d\rho_X \right)^{1/2}.$$

If $f : X \to {\rm I\!R}$ is a real-valued function, it generates a classifier $\operatorname{sgn}(f) : X \to \{1, -1\}$ by taking $\operatorname{sgn}(f)(x) = \operatorname{sgn}(f(x))$ which equals 1 if $f(x) \geq 0$ and -1 otherwise. Denote the **misclassification set** of the classifier $\operatorname{sgn}(f)$ as

$$X_f = X \setminus \widehat{X}_f, \text{ where } \widehat{X}_f = \{x \in X : \operatorname{sgn}(f)(x) = \operatorname{sgn}(f_\rho)(x)\}.$$

It is easy to see that

$$\|\operatorname{sgn}(f) - \operatorname{sgn}(f_\rho)\|_\rho^2 = 4\rho_X(X_f).$$

In the following, we show that $\operatorname{sgn}(f)$ approximates the Bayes rule $\operatorname{sgn}(f_\rho)$ well if f is a good approximation of f_ρ in L^∞. To this end, we introduce a function motivated by the Tsybakov condition [12] with noise exponent $q (0 < q \leq \infty)$: for some constant $c_q > 0$,

$$\rho_X\big(\{x \in X : 0 < |f_\rho(x)| \leq c_q t\}\big) \leq t^q, \qquad \forall t > 0. \tag{5.2}$$

[2] Conversations in Genova with Caponnetto, De Vito, Rosasco, and Verri were helpful in developing this section.

Definition 2. *The* **Tsybakov function** *associated with the probability distribution* ρ *on* $X \times \{1, -1\}$ *is defined to be the function* $T = T_\rho : [0, 1] \to [0, 1]$ *given by*

$$T(L) = meas_\rho f_\rho^{-1}([-L, L]) = \rho_X\big(\{x \in X : f_\rho(x) \in [-L, L]\}\big), \qquad L \in [0, 1]. \tag{5.3}$$

The Tsybakov function T_ρ measures different qualities of the condition of the binary classification problem defined by ρ on $X \times \{1, -1\}$. The following list of properties follows immediately from the definition.

Proposition 1. *Let* ρ *be a probability distribution on* $X \times \{1, -1\}$*, and* T *given by (5.3).*

(1) $T(1) = 1$.

(2) $\lim_{L \to 0+} T(L) = T(0) = \rho_X\big(f_\rho^{-1}(0)\big)$.

(3) For $0 < q < \infty$*, (5.2) holds only and only if* $T(L) - T(0) = O(L^q)$.

(4) (5.2) with $q = \infty$ *holds only and only if* $T(L) \equiv T(0)$ *on* $[0, c_\infty)$.

The set $f_\rho^{-1}(0)$ is called the **decision boundary**, which is a submanifold in general if f_ρ is smooth.

We say that ρ has (hard) **margin** $\tau > 0$ if $T(L) \equiv 0$ on $[0, \tau)$.

Proposition 2. *For any measurable function* $f : X \to \mathbb{R}$*, we have*

$$\|sgn(f) - sgn(f_\rho)\|_\rho^2 \le 4T\big(\|f - f_\rho\|_\infty\big) \tag{5.4}$$

and

$$\|sgn(f) - sgn(f_\rho)\|_\rho^2 \le 4T\big(\|f - f_\rho\|_\rho / \sqrt{\delta}\big) + 4\delta, \qquad \forall\ 0 < \delta < 1. \tag{5.5}$$

Proof. The left side of (5.4) equals $4\rho_X\big(X_f\big)$. But for each $x \in X_f$, we have

$$|f_\rho(x)| \le |f(x) - f_\rho(x)| \le \|f - f_\rho\|_\infty. \tag{5.6}$$

It means that the set X_f is a subset of (or equal to) $\{x \in X : |f_\rho(x)| \le \|f - f_\rho\|_\infty\}$. The ρ_X-measure of the latter equals $T\big(\|f - f_\rho\|_\infty\big)$ according to the definition of the Tsybakov function. Hence our first statement holds true.

To prove the second statement, we apply the Markov inequality $\text{Prob}\{\xi > \varepsilon\} \le E(\xi)/\varepsilon$ for the nonnegative random variable $\xi = (f(x) - f_\rho(x))^2$ on (X, ρ_X). For any $0 < \delta < 1$

there is some subset $U \subset X$ with $\rho_X(U) \geq 1-\delta$ such that $\|f - f_\rho\|_{L^\infty_{\rho_X}(U)} \leq \|f - f_\rho\|_\rho/\sqrt{\delta}$. Then

$$|f_\rho(x)| \leq |f(x) - f_\rho(x)| \leq \|f - f_\rho\|_{L^\infty_{\rho_X}(U)} \leq \|f - f_\rho\|_\rho/\sqrt{\delta}, \qquad \forall x \in X_f \cap U.$$

Thus $\rho_X\big(X_f \cap U\big) \leq \rho_X\big(\{x \in X : |f_\rho(x)| \leq \|f - f_\rho\|_\rho/\sqrt{\delta}\}\big) = T\big(\|f - f_\rho\|_\rho/\sqrt{\delta}\big)$. But $\rho_X\big(X_f \setminus U\big) \leq \delta$. So $\|\mathrm{sgn}(f) - \mathrm{sgn}(f_\rho)\|_\rho^2 = 4\rho_X\big(X_f\big)$ can be bounded as in (5.5). □

Remark. *When ρ has hard margin $\tau > 0$, $T(L) = 0$ for $L < \tau$. So it is sufficient to consider the case $\|f - f_\rho\|_\infty \geq \tau$ in Proposition 2.*

Applying Corollary 2 to Proposition 2 yields the following result.

Theorem 4. *Let $\mathbf{z}$ be randomly drawn from a probability distribution ρ on $X \times \{1, -1\}$. If $\|L_K^{-1} f_\rho\|_\rho < \infty$ and $\lambda = \big(3\kappa/\|L_K^{-1} f_\rho\|_\rho\big)^{2/3} m^{-1/3}$, then with confidence $1-\delta$,*

$$\|sgn(f_{\mathbf{z},\lambda}) - sgn(f_\rho)\|_\rho^2 \leq 4T\bigg(6\log\big(2/\delta\big)\kappa^{4/3}\|L_K^{-1} f_\rho\|_\rho^{2/3}\big(1/m\big)^{1/6}\bigg).$$

Definition 3. *Let $0 < q < \infty$ and ρ be a probability distribution on $X \times \{1, -1\}$. We define the q-coefficient as follows (if it is finite)*

$$a_q = a_{q,\rho} = \sup_{0<L<1} \frac{T(L)}{L^q}. \tag{5.7}$$

The Tsybakov condition (5.2) is the same as $a_q < \infty$ if $T(0) = 0$.

Applying our error analysis in $\mathcal{H}_K$, we get from Theorem 2 with $M = 1$ and Proposition 2 the following error bound for the classifier $\mathrm{sgn}(f_{\mathbf{z},\lambda})$.

Corollary 4. *Let $\mathbf{z}$ be randomly drawn according to a probability distribution ρ on $X \times \{1, -1\}$ having $a_q < \infty$ for some $0 < q < \infty$. Assume that f_ρ is in the range of L_K^r for some $\frac{1}{2} < r \leq 1$. Take the regularization parameter as $\lambda = \big(3\kappa/\|L_K^{-r} f_\rho\|_\rho\big)^{\frac{2}{1+2r}} m^{-\frac{1}{1+2r}}$. For any $0 < \delta < 1$, with confidence $1-\delta$,*

$$\|sgn(f_{\mathbf{z},\lambda}) - sgn(f_\rho)\|_\rho \leq \widetilde{C}\sqrt{a_q}\big(\log\big(2/\delta\big)\big)^{q/2}\left(\frac{1}{m}\right)^{\frac{q(2r-1)}{8r+4}}.$$

where $\widetilde{C} = 2^{\frac{q}{2}+1} 3^{q(2r-1)/(4r+2)} \kappa^{2qr/(2r+1)} \|L_K^{-r} f_\rho\|_\rho^{\frac{q}{1+2r}}$.

For a fixed q, the above error bound is proportional to $\sqrt{a_q}$. So we see that a_q describes well the behavior of the distribution ρ for the classification purpose.

Another way to measure the error of a classifier sgn(f) is the misclassification error defined by

$$\mathcal{R}(\mathrm{sgn}(f)) = \mathrm{Prob}\{\mathrm{sgn}(f)(x) \neq y\} = \frac{1}{4}\int_Z (y - \mathrm{sgn}(f)(x))^2 d\rho.$$

One can easily see that the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(\mathrm{sgn}(f_\rho))$ equals

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(\mathrm{sgn}(f_\rho)) = \int_{X_f} |f_\rho(x)| d\rho_X.$$

Hence it can be bounded as

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(\mathrm{sgn}(f_\rho)) \leq \int_{X_f} |f(x) - f_\rho(x)| d\rho_X \leq \|f - f_\rho\|_\rho.$$

This estimate may give very small excess misclassification error, even if the distribution is badly posed ($T(0) \approx 1$, or $T(L)$ is large even for reasonably small L).

§6. Error Analysis in $L^2_{\rho_X}$

One might estimate the error of $f_{\mathbf{z},\lambda} - f_\rho$ in $L^2_{\rho_X}$ by bounds in $\mathcal{H}_K$ (given in Theorem 1) and the relation (1.2). In this way, one obtains $\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho \leq \frac{6\kappa^2 M \log\left(2/\delta\right)}{\sqrt{m}\lambda}$ with confidence $1-\delta$. However, better error bounds of type $O(1/\sqrt{m\lambda})$ are in a preliminary draft of a paper of Andrea Caponnetto and Ernesto de Vito entitled "Fast rates for regularized least-squares algorithm". We are indebted to Lorenzo Rosasco for pointing this out to us and indicating how our (3.4) leads to the same rate.

The detailed results and analysis follow.

Theorem 5. *Let* $\mathbf{z}$ *be randomly drawn according to* ρ *satisfying* $|y| \leq M$ *almost surely. Then for any* $0 < \delta < 1$, *with confidence* $1-\delta$ *there holds*

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho \leq \frac{12\kappa M \log\left(4/\delta\right)}{\sqrt{m\lambda}}$$

provided that

$$\lambda \geq \frac{8\kappa^2 \log\left(4/\delta\right)}{\sqrt{m}}. \tag{6.1}$$

Before proving Theorem 5, we explain some ideas.

The main observation for the improvement of error bounds in L^2 is to apply the relation $\|g\|_\rho = \|L_K^{1/2} g\|_K$ to the proof of Theorem 1. With that (3.4) yields

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho = \left\| L_K^{1/2} \big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} \Big\{ \frac{1}{m} \sum_{i=1}^m \big(y_i - f_\lambda(x_i)\big) K_{x_i} - L_K\big(f_\rho - f_\lambda\big) \Big\} \right\|_K$$

and

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_\rho \le \left\| L_K^{1/2} \big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} \right\| \Delta, \tag{6.2}$$

where the norm is the operator norm of the operator $L_K^{1/2}\big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1}$ from $\mathcal{H}_K$ to $\mathcal{H}_K$. In addition to the estimate of Δ given by (3.9) in the proof of Theorem 1, we need to bound this operator norm. Since $\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}$ is a good approximation of L_K, one expects to bound this norm with confidence, similar to

$$\|L_K^{1/2}\big(L_K + \lambda I\big)^{-1}\| = \|L_K^{1/2}\big(L_K + \lambda I\big)^{-1/2}\big(L_K + \lambda I\big)^{-1/2}\| \le 1/\sqrt{\lambda}. \tag{6.3}$$

To realize the above expectation, we write $\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I$ as

$$L_K + \lambda I - \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big) = \Big\{ I - \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big)\big(L_K + \lambda I\big)^{-1} \Big\} \big(L_K + \lambda I\big).$$

It follows that

$$L_K^{1/2}\big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} = L_K^{1/2}\big(L_K + \lambda I\big)^{-1} \Big\{ I - \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big)\big(L_K + \lambda I\big)^{-1} \Big\}^{-1} \tag{6.4}$$

if the last inverse exists. To verify the invertibility and estimate the norm, we use the identity $S_{\mathbf{x}}^T S_{\mathbf{x}} = \sum_{i=1}^m K_{x_i} \langle \cdot, K_{x_i} \rangle_K$ and find that

$$\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} \big(L_K + \lambda I\big)^{-1} = \frac{1}{m} \sum_{i=1}^m \xi(x_i).$$

Here ξ is the random variable on (X, ρ_X) given by

$$\xi(x) = K_x \langle \cdot, K_x \rangle_K \big(L_K + \lambda I\big)^{-1}, \qquad x \in X. \tag{6.5}$$

The values of ξ are rank-one operators on $\mathcal{H}_K$. To apply probability inequalities for random variables with values in Hilbert spaces for estimating $\|\frac{1}{m} \sum_{i=1}^m \xi(x_i) - E(\xi)\|$, as in [4] we consider ξ to be a random variable with values in $HS(\mathcal{H}_K)$, the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}_K$, with inner product $\langle A, B\rangle_{HS} = \mathrm{Tr}(B^T A)$. Here Tr denotes the trace of a (trace-class) linear operator. The space $HS(\mathcal{H}_K)$ is a subspace of the space of bounded linear operators on $\mathcal{H}_K$, denoted as $(L(\mathcal{H}_K), \|\cdot\|)$, with the norm relations

$$\|A\| \leq \|A\|_{HS}, \qquad \|AB\|_{HS} \leq \|A\|_{HS}\|B\|. \tag{6.6}$$

Lemma 4. *Let* $\mathbf{x}$ *be a sample drawn from* (X, ρ_X). *With confidence* $1-\delta$, *we have*

$$\left\| \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big)\big(L_K + \lambda I\big)^{-1} \right\|_{HS} \leq \frac{4\kappa^2 \log\big(2/\delta\big)}{\sqrt{m}\lambda}.$$

Proof. Consider the random variable ξ defined by (6.5) with values in $HS(\mathcal{H}_K)$. For $x \in X$ and $f \in \mathcal{H}_K$, the reproducing property (1.1) ensures

$$\big(\xi(x)\big)\big(f\big) = K_x \langle \big(L_K + \lambda I\big)^{-1}(f), K_x\rangle_K = K_x\big(L_K + \lambda I\big)^{-1}(f)(x).$$

Hence

$$E\big(\xi\big)(f) = E_x\big(\xi(x)(f)\big) = E_x\Big(K_x\big(L_K + \lambda I\big)^{-1}(f)(x)\Big) = \Big(L_K\big(L_K + \lambda I\big)^{-1}\Big)(f).$$

This means $E(\xi) = L_K\big(L_K + \lambda I\big)^{-1}$ and thereby

$$\big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big)\big(L_K + \lambda I\big)^{-1} = E(\xi) - \frac{1}{m}\sum_{i=1}^{m} \xi(x_i). \tag{6.7}$$

Now we apply Lemma 2 to ξ with $H = HS(\mathcal{H}_K)$. For $x \in X$, (6.6) tells us that

$$\|\xi(x)\|_{HS} \leq \|A_x\|_{HS}/\lambda,$$

where A_x is the self-adjoint rank-one linear operator $A_x = K_x\langle\cdot, K_x\rangle_K$. An intermediate step in the proof of Lemma 2 of [4] shows that $\|A_x\|_{HS} = K(x,x) \leq \kappa^2$. Therefore, $\|\xi\|_{HS} \leq \kappa^2/\lambda$, $\sigma^2(\xi) \leq \kappa^4/\lambda^2$ and our conclusion follows from Lemma 2 and (6.7). □

We are in a position to prove the error bound in L^2, stated in Theorem 5.

Proof of Theorem 5. Applying Lemma 4 with δ replaced by $\delta/2$, we know that there is a subset U_1 of Z^m, with measure at least $1-\delta/2$, such that

$$\left\| \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big) \big(L_K + \lambda I\big)^{-1} \right\|_{HS} \leq \frac{4\kappa^2 \log(4/\delta)}{\sqrt{m}\lambda}, \qquad \forall \mathbf{z} \in U_1.$$

This in connection with (6.6) implies that for λ satisfying (6.1) and $\mathbf{z} \in U_1$,

$$\left\| \big(L_K - \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}\big) \big(L_K + \lambda I\big)^{-1} \right\| \leq \frac{1}{2}.$$

It follows that the last inverse in (6.4) exists, and combining (6.4) with (6.3) gives

$$\left\| L_K^{1/2} \big(\frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\big)^{-1} \right\| \leq 2 \left\| L_K^{1/2} \big(L_K + \lambda I\big)^{-1} \right\| \leq \frac{2}{\sqrt{\lambda}}, \qquad \forall \mathbf{z} \in U_1. \tag{6.8}$$

Recall (3.9) in the proof of Theorem 1. Replacing δ by $\delta/2$, we see that there is another subset U_2 of Z^m, with measure at least $1-\delta/2$, such that for $\mathbf{z} \in U_2$,

$$\Delta \leq \frac{2\kappa M (1 + \kappa/\sqrt{\lambda}) \log(4/\delta)}{m} + 2\kappa M \sqrt{\frac{\log(4/\delta)}{m}}.$$

Under the restriction (6.1), we have

$$\Delta \leq \frac{6\kappa M \log(4/\delta)}{\sqrt{m}}, \qquad \forall \mathbf{z} \in U_2. \tag{6.9}$$

Finally, we combine (6.2) with (6.8) and (6.9), and find that for $\mathbf{z} \in U_1 \cap U_2$, a subset of measure at least $1-\delta$, the desired error bound holds true. □

To get rates for the total error in L^2, we take the regularization parameter

$$\lambda = \lambda(m) = \begin{cases} \log\big(4/\delta\big)\big(12\kappa M/\|L_K^{-r} f_\rho\|_\rho\big)^{2/(1+2r)} \big(1/m\big)^{1/(1+2r)}, & \text{if } r > 1/2, \\ 8\kappa^2 \log\big(4/\delta\big)/\sqrt{m}, & \text{if } r \leq 1/2. \end{cases}$$

Corollary 5. *Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \leq M$ almost surely. Assume that f_ρ is in the range of L_K^r for some $0 < r \leq 1$. For $m \geq C_r$ and any $0 < \delta < 1$, with confidence $1-\delta$,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \leq \begin{cases} 2\log\big(4/\delta\big)\big(12\kappa M\big)^{2r/(1+2r)} \|L_K^{-r} f_\rho\|_\rho^{1/(1+2r)} \big(\frac{1}{m}\big)^{r/(1+2r)}, & \textit{if } r > 1/2, \\ \log\big(4/\delta\big)\Big(8M + 8^r \kappa^{2r} \|L_K^{-r} f_\rho\|_\rho\Big)\big(\frac{1}{m}\big)^{r/2}, & \textit{if } r \leq 1/2. \end{cases}$$

where λ is chosen as above, $C_r = 1$ for $r \leq 1/2$ and

$$C_r = \big(\|L_K^{-r} f_\rho\|_\rho/(12\kappa M)\big)^{4/(2r-1)}(8\kappa^2)^{(2+4r)/(2r-1)}, \qquad \text{if } r > 1/2.$$

Proof. Take $\lambda = t\log\big(4/\delta\big)$ with $t > 0$ satisfying $t \geq 8\kappa^2/\sqrt{m}$. Then (6.1) is valid. By Theorem 5 and Lemma 3, with confidence $1-\delta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \leq \frac{12\kappa M \log\big(4/\delta\big)}{\sqrt{m\lambda}} + \lambda^r \|L_K^{-r} f_\rho\|_\rho \leq \log\big(4/\delta\big)\left\{\frac{12\kappa M}{\sqrt{mt}} + t^r\|L_K^{-r} f_\rho\|_\rho\right\}.$$

The bound on the right side is optimized by minimizing over t

$$t = \big(12\kappa M/\|L_K^{-r} f_\rho\|_\rho\big)^{2/(1+2r)}\left(\frac{1}{m}\right)^{1/(1+2r)}.$$

Choose this value for t when $r > 1/2$. For $r \leq 1/2$, we choose $t = 8\kappa^2/\sqrt{m}$. The error bounds are verified. □

The above error bounds are kernel independent, except the requirement $\|L_K^{-r} f_\rho\|_\rho < \infty$. They may be improved when some extra information about the kernel such as its regularity is available. See [14].

References

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

[2] F. Cucker and S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. **39** (2001), 1–49.

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

[4] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. **5** (2005), 59–85.

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

[6] P. Niyogi and F. Girosi, Generalization bounds for function approximation from scattered noisy data, Adv. Comput. Math. **10** (1999), 51–80.

[7] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. **22** (1994), 1679–1706.

[8] T. Poggio and S. Smale, The mathematics of learning: dealing with data, Notices Amer. Math. Soc. **50** (2003), 537–544.

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

[10] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. **41** (2004), 279–305.

[11] S. Smale and D. X. Zhou, Shannon sampling II. Connections to learning theory, Appl. Comput. Harmonic Anal., to appear.

[12] A. B. Tsybakov, Optimal aggregation of classifiers in statistical learning, Ann. Stat. **32** (2004), 135–166.

[13] G. Wahba, Spline Models for Observational Data, SIAM, 1990.

[14] Q. Wu, Y. Ying, and D. X. Zhou, Learning rates of least-square regularized regression, submitted to Found. Comput. Math.

[15] Y. Yurinsky, Sums and Gaussian Vectors, Lecture Notes in Mathematics Vol. **1617**, Springer-Verlag, Berlin, 1995.

[16] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comp. **15** (2003), 1397–1437.

[17] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49** (2003), 1743–1752.