



Shannon sampling II: Connections to learning theory[☆]

Steve Smale^{a,*}, Ding-Xuan Zhou^b

^a Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago, IL 60637, USA

^b Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China

Received 22 July 2004; revised 29 January 2005; accepted 10 March 2005

Available online 3 August 2005

Communicated by Charles K. Chui

Abstract

We continue our study [S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305] of Shannon sampling and function reconstruction. In this paper, the error analysis is improved. Then we show how our approach can be applied to learning theory: a functional analysis framework is presented; dimension independent probability estimates are given not only for the error in the L^2 spaces, but also for the error in the reproducing kernel Hilbert space where the learning algorithm is performed. Covering number arguments are replaced by estimates of integral operators.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Shannon sampling; Function reconstruction; Learning theory; Reproducing kernel Hilbert space; Frames

1. Introduction

This paper considers regularization schemes associated with the least square loss and Hilbert spaces $\mathcal{H}$ of continuous functions. Our target is to provide a unified approach for two topics: interpolation theory, or more generally, function reconstruction in Shannon sampling theory with $\mathcal{H}$ being a space of band-

[☆] The first author is partially supported by NSF grant 0325113. The second author is supported partially by a grant from the Research Grants Council of Hong Kong [Project No. CityU 1087/02P].

* Corresponding author.

E-mail addresses: smale@math.berkeley.edu (S. Smale), mazhou@cityu.edu.hk (D.-X. Zhou).

limited functions or functions with certain decay; and regression problem in learning theory with $\mathcal{H}$ being a reproducing kernel Hilbert space $\mathcal{H}_K$.

First, we improve the probability estimates in [12] with a simplified development. Then we apply the technique for function reconstruction to learning theory. In particular, we show that a regression function f_ρ can be approximated by a regularization scheme $f_{z,\lambda}$ in $\mathcal{H}_K$. Dimension independent exponential probability estimates are given for the error $\|f_{z,\lambda} - f_\rho\|_K$. Our error bounds provide clues to the asymptotic choice of the regularization parameter γ or λ .

2. Sampling operator

Let $\mathcal{H}$ be a Hilbert space of continuous functions on a complete metric space X and the inclusion $J: \mathcal{H} \rightarrow C(X)$ is bounded with $\|J\| < \infty$.

Then for each $x \in X$, the point evaluation functional $f \rightarrow f(x)$ is bounded on $\mathcal{H}$ with norm at most $\|J\|$. Hence there exists an element $E_x \in \mathcal{H}$ with $\|E_x\|_{\mathcal{H}} \leq \|J\|$ such that

$$f(x) = \langle f, E_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}. \quad (2.1)$$

Let $\bar{x}$ be a discrete subset of X . Define the sampling operator $S_{\bar{x}}: \mathcal{H} \rightarrow \ell^2(\bar{x})$ by

$$S_{\bar{x}}(f) = (f(x))_{x \in \bar{x}}.$$

We shall always assume that $S_{\bar{x}}$ is bounded. This holds naturally when $\bar{x}$ is finite.

Denote $S_{\bar{x}}^T$ as the adjoint of $S_{\bar{x}}$. For each $c \in \ell^2(\bar{x})$, there holds

$$\langle f, S_{\bar{x}}^T c \rangle_{\mathcal{H}} = \langle S_{\bar{x}} f, c \rangle_{\ell^2(\bar{x})} = \sum_{x \in \bar{x}} c_x f(x) = \left\langle f, \sum_{x \in \bar{x}} c_x E_x \right\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

It follows that

$$S_{\bar{x}}^T c = \sum_{x \in \bar{x}} c_x E_x, \quad \forall c \in \ell^2(\bar{x}).$$

3. Algorithm

To allow noise, we make the following assumption.

Assumption. The sampled values $y = (y_x)_{x \in \bar{x}}$ have the form for some $f^* \in \mathcal{H}$:

$$\text{For each } x \in \bar{x}, y_x = f^*(x) + \eta_x, \text{ where } \eta_x \text{ is independently drawn from } \rho_x. \quad (3.1)$$

Here for each $x \in X$, ρ_x is a probability measure with zero mean, and its variance σ_x^2 satisfies $\sigma^2 := \sum_{x \in \bar{x}} \sigma_x^2 < \infty$.

Note that $\sum_{x \in \bar{x}} (f^*(x))^2 = \|S_{\bar{x}} f^*\|_{\ell^2(\bar{x})}^2 \leq \|S_{\bar{x}}\|^2 \|f^*\|_{\mathcal{H}}^2 < \infty$.

The Markov inequality for a nonnegative random variable ξ asserts that

$$\text{Prob} \left\{ \xi \leq \frac{E(\xi)}{\delta} \right\} \geq 1 - \delta, \quad \text{for all } 0 < \delta < 1. \quad (3.2)$$

It tells us that for every $\varepsilon > 0$,

$$\text{Prob}\{\|\{\eta_x\}\|_{\ell^2(\bar{x})}^2 > \varepsilon\} \leq E(\|\{\eta_x\}\|_{\ell^2(\bar{x})}^2)/\varepsilon = \frac{\sigma^2}{\varepsilon}.$$

By taking $\varepsilon \rightarrow \infty$, we see that $\{\eta_x\} \in \ell^2(\bar{x})$ and hence $y \in \ell^2(\bar{x})$ in probability.

Let $\gamma \geq 0$. With the sample $\mathbf{z} := (x, y_x)_{x \in \bar{x}}$, consider the minimization problem

$$\text{Function reconstruction} \quad \tilde{f} := \arg \min_{f \in \mathcal{H}} \left\{ \sum_{x \in \bar{x}} (f(x) - y_x)^2 + \gamma \|f\|_{\mathcal{H}}^2 \right\}. \quad (3.3)$$

Theorem 1. If $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ is invertible, then $\tilde{f}$ exists, is unique and

$$\tilde{f} = Ly, \quad L := (S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1} S_{\bar{x}}^T.$$

Proof. Denote

$$\mathcal{E}_{\mathbf{z}}(f) := \sum_{x \in \bar{x}} (f(x) - y_x)^2.$$

Since $\sum_{x \in \bar{x}} (f(x))^2 = \|S_{\bar{x}} f\|_{\ell^2(\bar{x})}^2 = \langle S_{\bar{x}}^T S_{\bar{x}} f, f \rangle_{\mathcal{H}}$, we know that for $f \in \mathcal{H}$,

$$\mathcal{E}_{\mathbf{z}}(f) + \gamma \|f\|_{\mathcal{H}}^2 = \langle (S_{\bar{x}}^T S_{\bar{x}} + \gamma I) f, f \rangle_{\mathcal{H}} - 2 \langle S_{\bar{x}}^T y, f \rangle_{\mathcal{H}} + \|y\|_{\ell^2(\bar{x})}^2.$$

Taking the functional derivative [10] for $f \in \mathcal{H}$, we see that any minimizer $\tilde{f}$ of (3.3) satisfies

$$(S_{\bar{x}}^T S_{\bar{x}} + \gamma I) \tilde{f} = S_{\bar{x}}^T y.$$

This proves Theorem 1. $\square$

Thus (3.3) with the invertibility of the operator $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ becomes algorithmic. The invertibility condition is valid for rich data.

Definition 1. We say that $\bar{x}$ provides *rich data* (with respect to $\mathcal{H}$) if

$$\lambda_{\bar{x}} := \inf_{f \in \mathcal{H}} \|S_{\bar{x}} f\|_{\ell^2(\bar{x})} / \|f\|_{\mathcal{H}} \quad (3.4)$$

is positive. It provides *poor data* if $\lambda_{\bar{x}} = 0$.

The problem of function reconstruction here is to estimate the error $\|\tilde{f} - f^*\|_{\mathcal{H}}$. In this paper we shall show in Corollary 2 below that in the rich data case, with $\gamma = 0$, for every $0 < \delta < 1$, with probability $1 - \delta$, there holds

$$\|\tilde{f} - f^*\|_{\mathcal{H}} \leq \frac{\|J\| \sqrt{\sigma^2 / \delta}}{\lambda_{\bar{x}}^2}. \quad (3.5)$$

This estimate does not require the boundedness of the noise ρ_x . Moreover, under the stronger condition (see [12]) that $|\eta_x| \leq M$ for each $x \in \bar{x}$, we shall use the McDiarmid inequality and prove in Theorem 5 below that for every $0 < \delta < 1$, with probability $1 - \delta$,

$$\|\tilde{f} - f^*\|_{\mathcal{H}} \leq \frac{\|J\|}{\lambda_{\bar{x}}^2} \left(\sqrt{8\sigma^2 \log \frac{1}{\delta}} + \frac{4}{3} M \log \frac{1}{\delta} \right). \quad (3.6)$$

The two estimates, (3.5) and (3.6), improve the bounds in [12]. It turns out that Theorem 4 in [12] is a consequence of the remark which follows it about the Markov inequality. Conversations with David McAllester were important to clarify this point.

4. Sample error

Define

$$f_{\bar{x},\gamma} := L(S_{\bar{x}} f^*). \quad (4.1)$$

The sample error takes the form $\|\tilde{f} - f_{\bar{x},\gamma}\|_{\mathcal{H}}^2$.

Theorem 2. *If $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ is invertible and Assumption holds, for every $0 < \delta < 1$, with probability $1 - \delta$, there holds*

$$\|\tilde{f} - f_{\bar{x},\gamma}\|_{\mathcal{H}}^2 \leq \frac{\|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\|^2 \|J\|^2 \sigma^2}{\delta}.$$

If $|\eta_x| \leq M$ for some $M \geq 0$ and each $x \in \bar{x}$, then for every $\varepsilon > 0$, we have

$$\text{Prob}_y \left\{ \|\tilde{f} - f_{\bar{x},\gamma}\|_{\mathcal{H}}^2 \leq \|L\|^2 \sigma^2 (1 + \varepsilon) \right\} \geq 1 - \exp \left\{ -\frac{\varepsilon \sigma^2}{2M^2} \log(1 + \varepsilon) \right\}.$$

Proof. Write $\|\tilde{f} - f_{\bar{x},\gamma}\|_{\mathcal{H}}^2$ as

$$\|L(y - S_{\bar{x}} f^*)\|_{\mathcal{H}}^2 \leq \|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\|^2 \|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}}^2.$$

But

$$S_{\bar{x}}^T (y - S_{\bar{x}} f^*) = \sum_{x \in \bar{x}} (y_x - f^*(x)) E_x.$$

Hence

$$\|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}}^2 = \sum_{x \in \bar{x}} \sum_{x' \in \bar{x}} (y_x - f^*(x)) (y_{x'} - f^*(x')) \langle E_x, E_{x'} \rangle_{\mathcal{H}}.$$

By the independence of the samples and $E(y_x - f^*(x)) = 0$, $E\{(y_x - f^*(x))^2\} = \sigma_x^2$, its expected value is

$$E(\|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}}^2) = \sum_{x \in \bar{x}} \sigma_x^2 \langle E_x, E_x \rangle_{\mathcal{H}}.$$

Now $\langle E_x, E_x \rangle_{\mathcal{H}} = \|E_x\|_{\mathcal{H}}^2 \leq \|J\|^2$. So the expected value of the sample error can be bounded as

$$E(\|\tilde{f} - f_{\bar{x},\gamma}\|_{\mathcal{H}}^2) \leq \|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\|^2 \|J\|^2 \sigma^2.$$

The first desired probability estimate follows from the Markov inequality (3.2).

For the second estimate, we apply Theorem 3 from [12] (with $w \equiv 1$) to the random variables $\{\eta_x^2\}_{x \in \bar{x}}$. Assumption tells us that $E(\eta_x) = 0$, which implies $E(\eta_x^2) = \sigma_x^2$. So we see that for every $\varepsilon > 0$,

$$\text{Prob}_y \left\{ \sum_{x \in \bar{x}} \{\eta_x^2 - \sigma_x^2\} > \varepsilon \right\} \leq \exp \left\{ -\frac{\varepsilon}{2M^2} \log \left(1 + \frac{M^2 \varepsilon}{\sum_{x \in \bar{x}} \sigma^2(\eta_x^2)} \right) \right\}.$$

Here we have used the condition $|\eta_x| \leq M$, which implies $|\eta_x^2 - \sigma_x^2| \leq M^2$. Also, $\sum_{x \in \bar{x}} \sigma^2(\eta_x^2) \leq \sum_{x \in \bar{x}} E(\eta_x^4) \leq M^2 \sigma^2 < \infty$. The desired bound then follows from the inequality $\|\tilde{f} - f_{\bar{x}, \gamma}\|_{\mathcal{H}}^2 \leq \|L\|^2 \|\{\eta_x\}\|_{\ell^2(\bar{x})}^2$ after replacing ε by $\varepsilon \sigma^2$. $\square$

Remark. When $\bar{x}$ contains m elements, we can take $\sigma^2 \leq m M^2 < \infty$.

Proposition 1. *The sampling operator $S_{\bar{x}}$ satisfies*

$$\|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\| \leq \frac{1}{\lambda_{\bar{x}}^2 + \gamma}.$$

For the operator L , we have

$$\|L\| \leq \frac{\|S_{\bar{x}}\|}{\lambda_{\bar{x}}^2 + \gamma}.$$

Proof. Let $v \in \mathcal{H}$ and $u = (S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}v$. Then

$$(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)u = v.$$

Taking inner products on both sides with u , we have

$$\langle S_{\bar{x}} u, S_{\bar{x}} u \rangle_{\ell^2(\bar{x})} + \gamma \|u\|_{\mathcal{H}}^2 = \langle v, u \rangle_{\mathcal{H}} \leq \|v\|_{\mathcal{H}} \|u\|_{\mathcal{H}}.$$

The definition of the richness $\lambda_{\bar{x}}$ tells us that

$$\langle S_{\bar{x}} u, S_{\bar{x}} u \rangle_{\ell^2(\bar{x})} = \|S_{\bar{x}} u\|_{\ell^2(\bar{x})}^2 \geq \lambda_{\bar{x}}^2 \|u\|_{\mathcal{H}}^2.$$

It follows that

$$(\lambda_{\bar{x}}^2 + \gamma) \|u\|_{\mathcal{H}}^2 \leq \|v\|_{\mathcal{H}} \|u\|_{\mathcal{H}}.$$

Hence $\|u\|_{\mathcal{H}} \leq (\lambda_{\bar{x}}^2 + \gamma)^{-1} \|v\|_{\mathcal{H}}$. This is true for every $v \in \mathcal{H}$. So the bound for the first operator follows. The second inequality is trivial. $\square$

Corollary 1. *If $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ is invertible and Assumption holds, then for every $0 < \delta < 1$, with probability $1 - \delta$, there holds*

$$\|\tilde{f} - f_{\bar{x}, \gamma}\|_{\mathcal{H}}^2 \leq \frac{\|J\|^2 \sigma^2}{(\lambda_{\bar{x}}^2 + \gamma)^2 \delta}.$$

5. Integration error

Recall that $f_{\bar{x}, \gamma} = L(S_{\bar{x}} f^*) = (S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1} S_{\bar{x}}^T S_{\bar{x}} f^*$. It can be written as

$$f_{\bar{x}, \gamma} = (S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1} (S_{\bar{x}}^T S_{\bar{x}} + \gamma I - \gamma I) f^* = f^* - \gamma (S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1} f^*. \quad (5.1)$$

This in connection with Proposition 1 proves the following proposition.

Proposition 2. Suppose $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ is invertible. We have

$$\|f_{\bar{x},\gamma} - f^*\|_{\mathcal{H}} \leq \frac{\gamma \|f^*\|_{\mathcal{H}}}{\lambda_{\bar{x}}^2 + \gamma}.$$

In particular, in the rich data case, we may take $\gamma = 0$ and obtain $f_{\bar{x},\gamma} = f^*$. Hence the following estimate is a consequence of Corollary 1.

Corollary 2. If $\lambda_{\bar{x}} > 0$ and $\gamma = 0$, for every $0 < \delta < 1$, with probability $1 - \delta$, there holds

$$\|\tilde{f} - f^*\|_{\mathcal{H}} \leq \frac{\|J\| \sqrt{\sigma^2/\delta}}{\lambda_{\bar{x}}^2}.$$

For the poor data case $\lambda_{\bar{x}} = 0$, we need to estimate $\gamma(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1} f^*$ according to (5.1). Recall that for a positive self-adjoint linear operator $\mathcal{L}$ on a Hilbert space $\mathcal{H}$, there holds

$$\|\gamma(\mathcal{L} + \gamma I)^{-1} f\|_{\mathcal{H}} = \|\gamma(\mathcal{L} + \gamma I)^{-1}(f - \mathcal{L}g + \mathcal{L}g)\|_{\mathcal{H}} \leq \|f - \mathcal{L}g\|_{\mathcal{H}} + \gamma \|g\|_{\mathcal{H}}$$

for every $g \in \mathcal{H}$. Taking the infimum over $g \in \mathcal{H}$, we have

$$\|\gamma(\mathcal{L} + \gamma I)^{-1} f\|_{\mathcal{H}} \leq \mathcal{K}(f, \gamma) := \inf_{g \in \mathcal{H}} \{\|f - \mathcal{L}g\|_{\mathcal{H}} + \gamma \|g\|_{\mathcal{H}}\}, \quad \forall f \in \mathcal{H}, \gamma > 0. \quad (5.2)$$

This is the K -functional between $\mathcal{H}$ and the range of $\mathcal{L}$. Thus, when f is in the closure of the range of $\mathcal{L}$ in $\mathcal{H}$, we have $\lim_{\gamma \rightarrow 0} \|\gamma(\mathcal{L} + \gamma I)^{-1} f\|_{\mathcal{H}} = 0$. If f is in the range of $\mathcal{L}^r$ for some $0 < r \leq 1$, then $\|\gamma(\mathcal{L} + \gamma I)^{-1} f\|_{\mathcal{H}} \leq 2\|\mathcal{L}^{-r} f\|_{\mathcal{H}} \gamma^r$. See, e.g., [11].

Using (5.2) for the operator $\mathcal{L} = S_{\bar{x}}^T S_{\bar{x}}$, we can use a K -functional between $\mathcal{H}$ and the range of $S_{\bar{x}}^T S_{\bar{x}}$ to get the convergence rate.

Proposition 3. Define f_{γ}^* as

$$f_{\gamma}^* := \arg \inf_{g \in \mathcal{H}} \{\|f^* - S_{\bar{x}}^T S_{\bar{x}} g\|_{\mathcal{H}} + \gamma \|g\|_{\mathcal{H}}\}, \quad \gamma > 0.$$

There holds

$$\|f_{\bar{x},\gamma} - f^*\|_{\mathcal{H}} \leq \|f^* - S_{\bar{x}}^T S_{\bar{x}} f_{\gamma}^*\|_{\mathcal{H}} + \gamma \|f_{\gamma}^*\|_{\mathcal{H}}.$$

In particular, if f^* lies in the closure of the range of $S_{\bar{x}}^T S_{\bar{x}}$, then $\lim_{\gamma \rightarrow 0} \|f_{\bar{x},\gamma} - f^*\|_{\mathcal{H}} = 0$. If f^* is in the range of $(S_{\bar{x}}^T S_{\bar{x}})^r$ for some $0 < r \leq 1$, then $\|f_{\bar{x},\gamma} - f^*\|_{\mathcal{H}} \leq 2\|(S_{\bar{x}}^T S_{\bar{x}})^{-r} f^*\|_{\mathcal{H}} \gamma^r$.

Compared with Corollary 2, Proposition 3 in connection with Corollary 1 gives an error estimate for the poor data case when f^* is in the range of $(S_{\bar{x}}^T S_{\bar{x}})^r$. For every $0 < \delta < 1$, with probability $1 - \delta$, there holds

$$\|\tilde{f} - f^*\|_{\mathcal{H}} \leq \frac{\|J\| \sqrt{\sigma^2}}{\gamma \sqrt{\delta}} + 2\|(S_{\bar{x}}^T S_{\bar{x}})^{-r} f^*\|_{\mathcal{H}} \gamma^r.$$

6. More general setting of function reconstruction

From (2.1) we see that the boundedness of $S_{\bar{x}}$ is equivalent to the Bessel sequence property of the family $\{E_x\}_{x \in \bar{x}}$ of elements in $\mathcal{H}$, i.e., there is a positive constant B such that

$$\sum_{x \in \bar{x}} |\langle f, E_x \rangle_{\mathcal{H}}|^2 \leq B \|f\|_{\mathcal{H}}^2, \quad \forall f \in \mathcal{H}. \quad (6.1)$$

Moreover, $\bar{x}$ provides rich data if and only if this family forms a *frame* of $\mathcal{H}$, i.e., there are two positive constants $A \leq B$ called frame bounds such that

$$A \|f\|_{\mathcal{H}}^2 \leq \sum_{x \in \bar{x}} |\langle f, E_x \rangle_{\mathcal{H}}|^2 \leq B \|f\|_{\mathcal{H}}^2, \quad \forall f \in \mathcal{H}.$$

In this case, the operator $S_{\bar{x}}^T S_{\bar{x}}$ is called the *frame operator*. Its inverse is usually difficult to compute, but it satisfies the reconstruction property:

$$f = \sum_{x \in \bar{x}} \langle f, (S_{\bar{x}}^T S_{\bar{x}})^{-1} E_x \rangle_{\mathcal{H}} E_x, \quad \forall f \in \mathcal{H}.$$

For these basic facts about frames, see [17].

The function reconstruction algorithm studied in the previous sections can be generalized to a setting with a *Bessel sequence* $\{E_x\}_{x \in \bar{x}}$ in $\mathcal{H}$ satisfying (6.1). Here the point evaluation (2.1) is replaced by the functional $\langle f, E_x \rangle_{\mathcal{H}}$ and the minimization becomes

$$\tilde{f} := \arg \min_{f \in \mathcal{H}} \left\{ \sum_{x \in \bar{x}} (\langle f, E_x \rangle_{\mathcal{H}} - y_x)^2 + \gamma \|f\|_{\mathcal{H}}^2 \right\}. \quad (6.2)$$

The sample values in Assumption now take the form $y_x = \langle f^*, E_x \rangle_{\mathcal{H}} + \eta_x$. If we replace the sampling operator $S_{\bar{x}}$ by the operator from $\mathcal{H}$ to $\ell^2(\bar{x})$ mapping f to $(\langle f, E_x \rangle_{\mathcal{H}})_{x \in \bar{x}}$, then (6.2) can be analyzed in the same way as above and all the error bounds hold true. Concrete examples for this generalized setting can be found in the literature of image processing, inverse problems [6] and sampling theory [1]: the Fredholm integral equation of the first kind, the moment problem, and the function reconstruction from weighted-averages.

One can even consider more general function reconstruction schemes: replacing the least-square loss in (6.2) by some other loss function and $\|\cdot\|_{\mathcal{H}}$ by some other norm. For example, if we choose Vapnik's ε -insensitive loss: $|t|_{\varepsilon} := \max\{|t| - \varepsilon, 0\}$, and a function space $\tilde{\mathcal{H}}$ included in $\mathcal{H}$ (such as a Sobolev space in L^2), then a function reconstruction scheme becomes

$$\tilde{f} := \arg \min_{f \in \tilde{\mathcal{H}}} \left\{ \sum_{x \in \bar{x}} |\langle f, E_x \rangle_{\mathcal{H}} - y_x|_{\varepsilon} + \gamma \|f\|_{\tilde{\mathcal{H}}}^2 \right\}. \quad (6.3)$$

When $\{E_x\}_{x \in \bar{x}}$ is a Bessel sequence in $\mathcal{H}$ but not a frame (corresponding to the poor data case), the scheme (6.3) can be solved by a quadratic convex optimization problem but not by a linear operator in general. We do not expect to analyze this scheme in a linear functional analysis framework. The error would involve not only the variance, the Bessel sequence, and the regularization parameter, but also the choice of the parameter ε . It would be interesting to derive error bounds for the function reconstruction scheme (6.3).

The rich data requirement is reasonable for function reconstruction such as sampling theory [13]. On the other hand, in learning theory, the situation of poor data or poor frame bounds ($A \rightarrow 0$ as the number of points in $\bar{x}$ increases) often happens. For such situations, we take $\bar{x}$ to be random samples of some probability distribution.

7. Learning theory

From now on we assume that X is compact. Let ρ be a probability measure on $Z := X \times Y$ with $Y = \mathbb{R}$. The error for a function $f: X \rightarrow Y$ is given by $\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho$. The function minimizing the error is called the *regression function* and is given by

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

Here $\rho(y|x)$ is the conditional distribution at x induced by ρ . The marginal distribution on X is denoted as ρ_X . We assume that $f_\rho \in L^2_{\rho_X}$. Denote $\|f\|_\rho = \|f\|_{L^2_{\rho_X}}$ and $\sigma^2(\rho)$ as the variance of ρ .

The purpose of the regression problem in learning theory [3,7,9,14,15] is to find good approximations of the regression function from a set of random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn independently according to ρ . This purpose is achieved in Corollaries 3, 4, and 5 below. Here we consider kernel based learning algorithms.

Let $K: X \times X \rightarrow \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, \dots, x_\ell\} \subset X$, 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 [2] of the linear span of the set of functions $\{K_x = K(x, \cdot): x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, f \in \mathcal{H}_K. \quad (7.1)$$

The optimization problem we study here is a regularized one with some $\lambda > 0$

$$\text{Learning scheme} \quad f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}. \quad (7.2)$$

We shall investigate how $f_{\mathbf{z},\lambda}$ approximates f_ρ and how the choice of the regularization parameter $\lambda = \lambda(m)$ leads to (optimal) convergence rates. The convergence in $L^2_{\rho_X}$ has been considered in [4,5,18]. The purpose of this section is to present a simple functional analysis approach, and to provide the convergence rates in the space $\mathcal{H}_K$ as well as sharper, dimension independent probability estimates in $L^2_{\rho_X}$.

The reproducing kernel property (7.1) tells us that the minimizer of (7.2) lies in $\mathcal{H}_{K,\mathbf{z}} := \text{span}\{K_{x_i}\}_{i=1}^m$ by projection onto this subspace. Thus, the optimization problem can be written in the same way as (3.3). To see this, we denote $\bar{x} = \{x_i\}_{i=1}^m$, $\rho_x = \rho(\cdot|x) - f_\rho(x)$ for $x \in X$. Then $E_x = K_x$ for $x \in \bar{x}$. Assumption

holds, and (3.1) is true except that $f^* \in \mathcal{H}$ is replaced by $f^* = f_\rho$. Denote $y = (y_i)_{i=1}^m$. The scheme (7.2) becomes a learning algorithm

$$f_{z,\lambda} := \arg \min_{f \in \mathcal{H}_{K,z}} \left\{ \sum_{x \in \bar{x}} (f(x) - y_x)^2 + \gamma \|f\|_K^2 \right\}, \quad \gamma = m\lambda.$$

Therefore, Theorem 1 still holds and we have

$$f_{z,\lambda} = (S_{\bar{x}}^T S_{\bar{x}} + m\lambda I)^{-1} S_{\bar{x}}^T y.$$

This implies the expression (see, e.g., [3]) that $f_{z,\lambda} = \sum_{i=1}^m c_i K_{x_i}$ with $c = (c_i)_{i=1}^m$ satisfying $((K(x_i, x_j))_{i,j=1}^m + m\lambda I)c = y$.

Denote $\kappa = \sqrt{\sup_{x \in X} K(x, x)}$ and $f_\rho|_{\bar{x}} = (f_\rho(x))_{x \in \bar{x}}$. Following our analysis for the scheme (3.3), define

$$f_{\bar{x},\lambda} = (S_{\bar{x}}^T S_{\bar{x}} + m\lambda I)^{-1} S_{\bar{x}}^T f_\rho|_{\bar{x}}.$$

Observe that $S_{\bar{x}}^T : \mathbb{R}^m \rightarrow \mathcal{H}_{K,z}$ is given by $S_{\bar{x}}^T c = \sum_{i=1}^m c_i K_{x_i}$. Then $S_{\bar{x}}^T S_{\bar{x}}$ satisfies

$$S_{\bar{x}}^T S_{\bar{x}} f = \sum_{x \in \bar{x}} f(x) K_x = m L_{K,\bar{x}} S_{\bar{x}}(f), \quad f \in \mathcal{H}_{K,z},$$

where $L_{K,\bar{x}} : \ell^2(\bar{x}) \rightarrow \mathcal{H}_K$ is defined as

$$L_{K,\bar{x}} c = \frac{1}{m} \sum_{i=1}^m c_i K_{x_i}.$$

It is a good approximation of the integral operator $L_K : L_{\rho_X}^2 \rightarrow \mathcal{H}_K$ defined by

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

The operator L_K can also be defined as a self-adjoint operator on $\mathcal{H}_K$ or on $L_{\rho_X}^2$. We shall use the same notion L_K for these operators defined on different domains. As operators on $\mathcal{H}_K$, $L_{K,\bar{x}} S_{\bar{x}}$ approximates L_K well. In fact, it was shown in [5] that

$$E(\|L_{K,\bar{x}} S_{\bar{x}} - L_K\|_{\mathcal{H}_K \rightarrow \mathcal{H}_K}) \leq \frac{\kappa^2}{\sqrt{m}}. \quad (7.3)$$

To get sharper error bounds in Theorem 3, we need estimates for $\|L_{K,\bar{x}}(f_\rho|_{\bar{x}}) - L_K f_\rho\|_K$. Since the function $f_\rho \notin \mathcal{H}_K$ in general, we need the following improvement of (7.3) with domain $L_{\rho_X}^2$.

Lemma 1. Let $\bar{x} \in X^m$ be randomly drawn according to ρ_X . For any $f \in L_{\rho_X}^2$,

$$E(\|L_{K,\bar{x}}(f|_{\bar{x}}) - L_K f\|_K) = E\left(\left\|\frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i} - L_K f\right\|_K\right) \leq \frac{\kappa \|f\|_{\rho}}{\sqrt{m}}.$$

Proof. Define ξ to be the $\mathcal{H}_K$ -valued random variable $\xi = f(x)K_x$ over (X, ρ_X) . Then $\frac{1}{m} \sum_{i=1}^m f(x_i) \times K_{x_i} - L_K f = \frac{1}{m} \sum_{i=1}^m \xi(x_i) - E(\xi)$. We know that

$$\left\{ E \left(\left\| \frac{1}{m} \sum_{i=1}^m \xi(x_i) - E(\xi) \right\|_K \right)^2 \right\} \leq E \left(\left\| \frac{1}{m} \sum_{i=1}^m \xi(x_i) - E(\xi) \right\|_K^2 \right) = \frac{1}{m} (E(\|\xi\|_K^2) - \|E(\xi)\|_K^2)$$

which is bounded by $\kappa^2 \|f\|_\rho^2 / m$. $\square$

The function $f_{\bar{x}, \lambda}$ may be considered as an approximation of f_λ where

$$f_\lambda := (L_K + \lambda I)^{-1} L_K f_\rho. \quad (7.4)$$

In fact, f_λ is a minimizer of the optimization problem

$$f_\lambda = \arg \min_{f \in \mathcal{H}_K} \{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \} = \arg \min_{f \in \mathcal{H}_K} \{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^2 \}. \quad (7.5)$$

Theorem 3. Let $\mathbf{z}$ be randomly drawn according to ρ . Then

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z}, \lambda} - f_{\bar{x}, \lambda}\|_K) \leq \frac{\kappa \sqrt{\sigma^2(\rho)}}{\sqrt{m\lambda}}$$

and

$$E_{\bar{x} \in X^m} (\|f_{\bar{x}, \lambda} - f_\lambda\|_K) \leq \frac{3\kappa \|f_\rho\|_\rho}{\sqrt{m\lambda}}.$$

Proof. The same proof as that of Theorem 2 and Proposition 1 shows that

$$E_y (\|f_{\mathbf{z}, \lambda} - f_{\bar{x}, \lambda}\|_K^2) \leq \frac{\kappa^2 \sum_{i=1}^m \sigma_{x_i}^2}{(\lambda_{\bar{x}}^2 + m\lambda)^2}.$$

But $E_{\bar{x}} (\sum_{i=1}^m \sigma_{x_i}^2) = m\sigma^2(\rho)$. So the first statement follows.

To see the second statement we write $f_{\bar{x}, \lambda} - f_\lambda$ as $f_{\bar{x}, \lambda} - \tilde{f}_\lambda + \tilde{f}_\lambda - f_\lambda$, where

$$\tilde{f}_\lambda := (L_{K, \bar{x}} S_{\bar{x}} + \lambda I)^{-1} L_K f_\rho. \quad (7.6)$$

Since

$$f_{\bar{x}, \lambda} - \tilde{f}_\lambda = (L_{K, \bar{x}} S_{\bar{x}} + \lambda I)^{-1} \left(\frac{1}{m} S_{\bar{x}}^T f_\rho|_{\bar{x}} - L_K f_\rho \right), \quad (7.7)$$

applying Lemma 1 to $f = f_\rho$ tells us that

$$E (\|f_{\bar{x}, \lambda} - \tilde{f}_\lambda\|_K) \leq \frac{1}{\lambda} E \left(\left\| \frac{1}{m} S_{\bar{x}}^T f_\rho|_{\bar{x}} - L_K f_\rho \right\|_K \right) \leq \frac{\kappa \|f_\rho\|_\rho}{\sqrt{m\lambda}}.$$

To estimate $\tilde{f}_\lambda - f_\lambda$, we write $L_K f_\rho$ as $(L_K + \lambda I) f_\lambda$. Then

$$\tilde{f}_\lambda - f_\lambda = (L_{K, \bar{x}} S_{\bar{x}} + \lambda I)^{-1} (L_K + \lambda I) f_\lambda - f_\lambda = (L_{K, \bar{x}} S_{\bar{x}} + \lambda I)^{-1} (L_K f_\lambda - L_{K, \bar{x}} S_{\bar{x}} f_\lambda).$$

Hence

$$\|\tilde{f}_\lambda - f_\lambda\|_K \leq \frac{1}{\lambda} \|L_K f_\lambda - L_{K, \bar{x}} S_{\bar{x}} f_\lambda\|_K. \quad (7.8)$$

Applying Lemma 1 again, we see that

$$E(\|\tilde{f}_\lambda - f_\lambda\|_K) \leq \frac{1}{\lambda} E(\|L_K f_\lambda - L_{K,\bar{x}} S_{\bar{x}} f_\lambda\|_K) \leq \frac{\kappa \|f_\lambda\|_\rho}{\sqrt{m\lambda}}.$$

Note that f_λ is a minimizer of (7.5). Taking $f = 0$ yields $\|f_\lambda - f_\rho\|_\rho^2 + \lambda \|f_\lambda\|_K^2 \leq \|f_\rho\|_\rho^2$. Hence

$$\|f_\lambda\|_\rho \leq 2\|f_\rho\|_\rho \quad \text{and} \quad \|f_\lambda\|_K \leq \|f_\rho\|_\rho / \sqrt{\lambda}. \quad (7.9)$$

Therefore, our second estimate follows. $\square$

The last step is to estimate the approximation error $\|f_\lambda - f_\rho\|$.

Theorem 4. Define f_λ by (7.4). If $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $0 < r \leq 1$, then

$$\|f_\lambda - f_\rho\|_\rho \leq \lambda^r \|L_K^{-r} f_\rho\|_\rho. \quad (7.10)$$

When $\frac{1}{2} < r \leq 1$, we have

$$\|f_\lambda - f_\rho\|_K \leq \lambda^{r-\frac{1}{2}} \|L_K^{-r} f_\rho\|_\rho. \quad (7.11)$$

We follow the same line as we did in [11]. Estimates similar to (7.10) can be found [3, Theorem 3 (1)]: for a self-adjoint strictly positive compact operator A on a Hilbert space $\mathcal{H}$, there holds for $0 < r < s$,

$$\inf_{b \in \mathcal{H}} \{ \|b - a\|^2 + \gamma \|A^{-s} b\|^2 \} \leq \gamma^{r/s} \|A^{-r} a\|^2. \quad (7.12)$$

(A mistake was made in [3] when scaling from $s = 1$ to general $s > 0$: r should be < 1 in the general situation.) A proof of (7.10) was given in [5]. Here we provide a complete proof because the idea is used for verifying (7.11).

Proof of Theorem 4. If $\{\lambda_i, \psi_i\}_{i \geq 1}$ are the normalized eigenpairs of the integral operator $L_K : L_{\rho_X}^2 \rightarrow L_{\rho_X}^2$, then $\|\sqrt{\lambda_i} \psi_i\|_K = 1$ when $\lambda_i > 0$.

Write $f_\rho = L_K^r g$ for some $g = \sum_{i \geq 1} d_i \psi_i$ with $\|\{d_i\}\|_{\ell^2} = \|g\|_\rho < \infty$. Then $f_\rho = \sum_{i \geq 1} \lambda_i^r d_i \psi_i$ and by (7.4),

$$f_\lambda - f_\rho = (L_K + \lambda I)^{-1} L_K f_\rho - f_\rho = - \sum_{i \geq 1} \frac{\lambda}{\lambda_i + \lambda} \lambda_i^r d_i \psi_i.$$

It follows that

$$\|f_\lambda - f_\rho\|_\rho = \left\{ \sum_{i \geq 1} \left(\frac{\lambda}{\lambda_i + \lambda} \lambda_i^r d_i \right)^2 \right\}^{1/2} = \lambda^r \left\{ \sum_{i \geq 1} \left(\frac{\lambda}{\lambda_i + \lambda} \right)^{2(1-r)} \left(\frac{\lambda_i}{\lambda + \lambda_i} \right)^{2r} d_i^2 \right\}^{1/2}.$$

This is bounded by $\lambda^r \|\{d_i\}\|_{\ell^2} = \lambda^r \|g\|_\rho = \lambda^r \|L_K^{-r} f_\rho\|_\rho$. Hence (7.10) holds.

When $r > \frac{1}{2}$, we have

$$\|f_\lambda - f_\rho\|_K^2 = \sum_{\lambda_i > 0} \left(\frac{\lambda}{\lambda_i + \lambda} \lambda_i^{r-\frac{1}{2}} d_i \right)^2 = \lambda^{2r-1} \sum_{i \geq 1} \left(\frac{\lambda}{\lambda_i + \lambda} \right)^{3-2r} \left(\frac{\lambda_i}{\lambda + \lambda_i} \right)^{2r-1} d_i^2.$$

This is again bounded by $\lambda^{2r-1} \|\{d_i\}\|_{\ell^2}^2 = \lambda^{2r-1} \|L_K^{-r} f_\rho\|_\rho^2$. The second statement (7.11) has been verified. $\square$

Combining Theorems 3 and 4, we find the expected value of the error $\|f_{\mathbf{z},\lambda} - f_\rho\|$. By choosing the optimal parameter in this bound, we get the following convergence rates.

Corollary 3. *Let $\mathbf{z}$ be randomly drawn according to ρ . Denote $\Sigma(\rho) = \kappa\sqrt{\sigma^2(\rho)} + 3\kappa\|f_\rho\|_\rho$. Suppose $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $\frac{1}{2} < r \leq 1$. We have*

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leq \frac{\Sigma(\rho)}{\sqrt{m\lambda}} + \lambda^{r-\frac{1}{2}} \|L_K^{-r} f_\rho\|_\rho. \quad (7.13)$$

It follows that when $\lambda = (\Sigma(\rho)/\|L_K^{-r} f_\rho\|_\rho)^{\frac{2}{2r+1}} (1/m)^{\frac{1}{1+2r}}$,

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z},\lambda} - f_\rho\|_K) \leq 2(\Sigma(\rho))^{\frac{2r-1}{2r+1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{2}{2r+1}} \left(\frac{1}{m}\right)^{\frac{2r-1}{4r+2}}. \quad (7.14)$$

Remark. Corollary 3 provides estimates for the $\mathcal{H}_K$ -norm error of $f_{\mathbf{z},\lambda} - f_\rho$. So we require $f_\rho \in \mathcal{H}_K$ which is equivalent to $L_K^{-\frac{1}{2}} f_\rho \in L_{\rho_X}^2$. To get convergence rates we assume a stronger condition $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $\frac{1}{2} < r \leq 1$. The optimal rate derived from Corollary 3 is achieved by $r = 1$. In this case, $E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z},\lambda} - f_\rho\|_K) = O((1/m)^{\frac{1}{6}})$. Note that the norm $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ cannot be bounded by the excess error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$.

Corollary 4. *Let $\mathbf{z}$ be randomly drawn according to ρ . Denote $\Sigma(\rho) = \kappa\sqrt{\sigma^2(\rho)} + 3\kappa\|f_\rho\|_\rho$. Assume $L_K^{-r} f_\rho \in L_{\rho_X}^2$ for some $0 < r \leq 1$. We have*

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho) \leq \frac{\kappa \Sigma(\rho)}{\sqrt{m\lambda}} + \lambda^r \|L_K^{-r} f_\rho\|_\rho. \quad (7.15)$$

In particular, if we take $\lambda = (\kappa \Sigma(\rho)/\|L_K^{-r} f_\rho\|_\rho)^{\frac{1}{r+1}} (1/m)^{\frac{1}{2+2r}}$, there holds

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho) \leq 2(\kappa \Sigma(\rho))^{\frac{r}{r+1}} \|L_K^{-r} f_\rho\|_\rho^{\frac{1}{r+1}} \left(\frac{1}{m}\right)^{\frac{r}{2r+2}}. \quad (7.16)$$

Remark. The convergence rate (7.16) for the $L_{\rho_X}^2$ -norm is obtained by optimizing the regularization parameter λ in (7.15). The sharp rate derived from Corollary 4 is $O((1/m)^{\frac{1}{4}})$, which is achieved by $r = 1$.

Our bound for the $\mathcal{H}_K$ -norm error stated in Corollary 3 is new in learning theory.

Let us now compare our error bounds in L^2 with the existing results. In [18], a leave-one-out technique was used to derive the expected value of learning schemes. For the scheme (7.2), the result can be expressed as

$$E_{\mathbf{z} \in Z^m} (\mathcal{E}(f_{\mathbf{z},\lambda})) \leq \left(1 + \frac{2\kappa^2}{m\lambda}\right)^2 \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \quad (7.17)$$

Notice that $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2$. Denote the regularization error (see [12]) as

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^2 \} = \inf_{f \in \mathcal{H}_K} \{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \}. \quad (7.18)$$

The bound (7.17) can be restated as

$$E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho^2) \leq \mathcal{D}(\lambda/2) + (\mathcal{E}(f_\rho) + \mathcal{D}(\lambda/2)) \left(\frac{4\kappa^2}{m\lambda} + \frac{4\kappa^4}{(m\lambda)^2} \right).$$

One can then derive the convergence rate $(1/m)^{\frac{1}{4}}$ in expectation when $f_\rho \in \mathcal{H}_K$ and $\mathcal{E}(f_\rho) > 0$. In fact, (7.12) with $\mathcal{H} = L_{\rho_X}^2$, $A = L_K$ holds for $r = s = 1/2$, which yields the best rate for the regularization error $\mathcal{D}(\lambda) \leq \|f_\rho\|_K^2 \lambda$. By taking $\lambda = 1/\sqrt{m}$, one can thus get $E_{\mathbf{z} \in Z^m} (\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho^2) = O((1/m)^{\frac{1}{2}})$, the same as Corollary 4. Applying (3.2), one can have the probability estimate $\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho \leq (C/\delta)(1/m)^{\frac{1}{4}}$ for the confidence $1 - \delta$.

In [5], a functional analysis approach was employed for the error analysis of the scheme (7.2). The main result asserts that for any $0 < \delta < 1$, with confidence $1 - \delta$,

$$|\mathcal{E}(f_{\mathbf{z}, \lambda}) - \mathcal{E}(f_\lambda)| \leq \frac{M\kappa^2}{\sqrt{m\lambda}} \left(1 + \frac{\kappa}{\sqrt{\lambda}} \right) \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right). \quad (7.19)$$

Convergence rates were also derived in [5, Corollary 1] by combining (7.19) with (7.10): when f_ρ lies in the range of L_K , for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho \leq C \left(\frac{\log(2/\delta)}{m} \right)^{\frac{1}{5}}, \quad \text{if } \lambda = \left(\frac{\log(2/\delta)}{m} \right)^{\frac{1}{5}}.$$

Thus the confidence is improved from $1/\delta$ to $\log(2/\delta)$, while the rate is weakened to $(1/m)^{\frac{1}{5}}$. In the next section we shall show that $\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho \leq C \sqrt{\log(4/\delta)} (1/m)^{\frac{1}{4}}$ with confidence $1 - \delta$, thus improving the confidence estimate for the best rate known so far. Our approach is short and neat, without involving the leave-one-out technique.

8. Probability estimates by McDiarmid inequalities

In this section we apply some McDiarmid inequalities to improve the probability estimates derived from expected values by the Markov inequality.

Let (Ω, ρ) be a probability space. For $\mathbf{t} = (t_1, \dots, t_m) \in \Omega^m$ and $t'_i \in \Omega$, we denote $\mathbf{t}^i := (t_1, \dots, t_{i-1}, t'_i, t_{i+1}, \dots, t_m)$.

Lemma 2. Let $\{t_i, t'_i\}_{i=1}^m$ be i.i.d. drawers of the probability distribution ρ on Ω , and $F: \Omega^m \rightarrow \mathbb{R}$ be a measurable function.

(1) If for each i there is c_i such that $\sup_{\mathbf{t} \in \Omega^m, t'_i \in \Omega} |F(\mathbf{t}) - F(\mathbf{t}^i)| \leq c_i$, then

$$\text{Prob}_{\mathbf{t} \in \Omega^m} \{ F(\mathbf{t}) - E_{\mathbf{t}}(F(\mathbf{t})) \geq \varepsilon \} \leq \exp \left\{ - \frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right\}, \quad \forall \varepsilon > 0. \quad (8.1)$$

(2) If there is $B \geq 0$ such that $\sup_{\mathbf{t} \in \Omega^m, 1 \leq i \leq m} |F(\mathbf{t}) - E_{t_i}(F(\mathbf{t}))| \leq B$, then

$$\text{Prob}_{\mathbf{t} \in \Omega^m} \{F(\mathbf{t}) - E_{\mathbf{t}}(F(\mathbf{t})) \geq \varepsilon\} \leq \exp \left\{ -\frac{\varepsilon^2}{2(B\varepsilon/3 + \sum_{i=1}^m \sigma_i^2(F))} \right\}, \quad \forall \varepsilon > 0, \quad (8.2)$$

where $\sigma_i^2(F) := \sup_{\mathbf{z} \setminus \{t_i\} \in \Omega^{m-1}} E_{t_i} \{(F(\mathbf{t}) - E_{t_i}(F(\mathbf{t})))^2\}$.

The first inequality is the McDiarmid inequality, see [8]. The second inequality is its Bernstein form which can be found in [16].

First, we show how the probability estimate for the function reconstruction stated in Theorem 2 can be improved, replacing $1/\delta$ by $\log(1/\delta)$.

Theorem 5. Suppose $S_{\bar{x}}^T S_{\bar{x}} + \gamma I$ is invertible and Assumption holds. Under the condition that $|y_x - f^*(x)| \leq M$ for each $x \in \bar{x}$, we have for every $0 < \delta < 1$, with probability $1 - \delta$,

$$\begin{aligned} \|\tilde{f} - f_{\bar{x}, \gamma}\|_{\mathcal{H}} &\leq \|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\| \|J\| \left(\sqrt{\sigma^2} + \sqrt{8\sigma^2 \log \frac{1}{\delta}} + \frac{4}{3} M \log \frac{1}{\delta} \right) \\ &\leq \frac{\|J\|}{\lambda_{\bar{x}}^2 + \gamma} \left(\sqrt{\sigma^2} + \sqrt{8\sigma^2 \log \frac{1}{\delta}} + \frac{4}{3} M \log \frac{1}{\delta} \right). \end{aligned}$$

Proof. Write $\|\tilde{f} - f_{\bar{x}, \gamma}\|_{\mathcal{H}}$ as

$$\|L(y - S_{\bar{x}} f^*)\|_{\mathcal{H}} \leq \|(S_{\bar{x}}^T S_{\bar{x}} + \gamma I)^{-1}\| \|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}}.$$

Consider the function $F: \ell^2(\bar{x}) \rightarrow \mathbb{R}$ defined by

$$F(y) = \|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}}.$$

Recall from the proof of Theorem 2 that $F(y) = \|\sum_{x \in \bar{x}} (y_x - f^*(x)) E_x\|_{\mathcal{H}}$ and

$$E_y(F) \leq \sqrt{E_y(F^2)} = \sqrt{\sum_{x \in \bar{x}} \sigma_x^2 \langle E_x, E_x \rangle_{\mathcal{H}}} \leq \|J\| \sqrt{\sigma^2}. \quad (8.3)$$

Then we can apply the McDiarmid inequality. Let $x_0 \in \bar{x}$ and y'_{x_0} be a new sample at x_0 . We have

$$|F(y) - F(y^{x_0})| = \left| \|S_{\bar{x}}^T (y - S_{\bar{x}} f^*)\|_{\mathcal{H}} - \|S_{\bar{x}}^T (y^{x_0} - S_{\bar{x}} f^*)\|_{\mathcal{H}} \right| \leq \|S_{\bar{x}}^T (y - y^{x_0})\|_{\mathcal{H}}.$$

The bound equals $\|(y_{x_0} - y'_{x_0}) E_{x_0}\|_{\mathcal{H}} \leq |y_{x_0} - y'_{x_0}| \|J\|$. Since $|y_x - f^*(x)| \leq M$ for each $x \in \bar{x}$, it can be bounded by $2M\|J\|$, which can be taken as B in Lemma 2 (2). Also,

$$\begin{aligned} E_{y_{x_0}}(|F(y) - E_{y_{x_0}}(F(y))|^2) &\leq \int \left(\int |y_{x_0} - y'_{x_0}| \|J\| d\rho_{x_0}(y'_{x_0}) \right)^2 d\rho_{x_0}(y_{x_0}) \\ &\leq \int \int (y_{x_0} - y'_{x_0})^2 \|J\|^2 d\rho_{x_0}(y'_{x_0}) d\rho_{x_0}(y_{x_0}) \leq 4\|J\|^2 \sigma_{x_0}^2. \end{aligned}$$

This yields $\sum_{x_0 \in \bar{x}} \sigma_{x_0}^2(F) \leq 4\|J\|^2 \sigma^2$. Thus Lemma 2 (2) tells us that for every $\varepsilon > 0$,

$$\text{Prob}_{y \in Y^{\bar{x}}} \{F(y) - E_y(F(y)) \geq \varepsilon\} \leq \exp \left\{ -\frac{\varepsilon^2}{2(2M\|J\|\varepsilon/3 + 4\|J\|^2 \sigma^2)} \right\}.$$

Choose the smallest ε so that the probability bound equals δ . That is, solve the quadratic equation

$$\frac{\varepsilon^2}{2(2M\|J\|\varepsilon/3 + 4\|J\|^2\sigma^2)} = \log \frac{1}{\delta}.$$

We find the probability estimate

$$F(y) \leq E_y(F) + \|J\| \left(\sqrt{8\sigma^2 \log \frac{1}{\delta}} + \frac{4}{3} M \log \frac{1}{\delta} \right)$$

for the confidence $1 - \delta$. This in connection with (8.3) proves Theorem 5. $\square$

Turn to the learning theory estimates. The purpose is to improve the bound in Theorem 3 by applying the McDiarmid inequality. To this end, we refine Lemma 1 from the expected value to a probability estimate form.

Lemma 3. Let $\bar{x} \in X^m$ be randomly drawn according to ρ_X . For any $f \in L_{\rho_X}^\infty$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\left\| \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i} - L_K f \right\|_K \leq \frac{4\kappa \|f\|_\infty}{3m} \log \frac{1}{\delta} + \frac{\kappa \|f\|_\rho}{\sqrt{m}} \left(1 + \sqrt{8 \log \frac{1}{\delta}} \right).$$

Proof. Define a function $F : X^m \rightarrow \mathbb{R}$ as

$$F(\bar{x}) = F(x_1, \dots, x_m) = \left\| \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i} - L_K f \right\|_K.$$

For $j \in \{1, \dots, m\}$, we apply the triangle inequality and obtain

$$|F(\bar{x}) - F(\bar{x}^j)| \leq \left\| \frac{1}{m} (f(x_j) - f(x'_j)) K_{x_j} \right\|_K \leq \frac{\kappa}{m} |f(x_j) - f(x'_j)|.$$

It follows that $|F(\bar{x}) - E_{x_j}(F(\bar{x}))| \leq (2\kappa \|f\|_\infty)/m =: B$. Moreover,

$$\begin{aligned} E_{x_j}(F(\bar{x}) - E_{x_j}(F(\bar{x})))^2 &\leq \int_X \left(\int_X \frac{\kappa}{m} |f(x_j) - f(x'_j)| d\rho_X(x'_j) \right)^2 d\rho_X(x_j) \\ &\leq \frac{\kappa^2}{m^2} \int_X \int_X 2|f(x_j)|^2 + 2|f(x'_j)|^2 d\rho_X(x'_j) d\rho_X(x_j) \leq \frac{4\kappa^2 \|f\|_\rho^2}{m^2}. \end{aligned}$$

So we have $\sum_{j=1}^m \sigma_j^2(F) \leq (4\kappa^2 \|f\|_\rho^2)/m$.

Thus we can apply Lemma 2 (2) to the function F and find that

$$\text{Prob}_{\bar{x} \in X^m} \{F(\bar{x}) - E_{\bar{x}}(F(\bar{x})) \geq \varepsilon\} \leq \exp \left\{ - \frac{\varepsilon^2}{2 \left(\frac{2\kappa \|f\|_\infty \varepsilon}{3m} + \frac{4\kappa^2 \|f\|_\rho^2}{m} \right)} \right\}.$$

Solving a quadratic equation again by setting the probability bound to be δ , we see that with confidence $1 - \delta$,

$$F(\bar{x}) \leq E_{\bar{x}}(F(\bar{x})) + \frac{4\kappa \|f\|_\infty}{3m} \log \frac{1}{\delta} + \frac{\kappa \|f\|_\rho}{\sqrt{m}} \sqrt{8 \log \frac{1}{\delta}}.$$

Lemma 1 says that $E_{\bar{x}}(F(\bar{x})) \leq (\kappa \|f\|_{\rho})/\sqrt{m}$. Our conclusion follows. $\square$

Theorem 6. Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \leq M$ almost surely. Then for any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\|f_{\mathbf{z},\lambda} - f_{\lambda}\|_K \leq \frac{\kappa M \log(4/\delta)}{\sqrt{m\lambda}} \left(36 + \frac{4\kappa}{3\sqrt{m\lambda}} \right).$$

Proof. Since $|y| \leq M$ almost surely, we know that $\|f_{\rho}\|_{\rho} \leq \|f_{\rho}\|_{\infty} \leq M$.

Recall the function $\tilde{f}_{\lambda}$ defined by (7.6). It satisfies (7.7). Hence

$$\|f_{\bar{x},\lambda} - \tilde{f}_{\lambda}\|_K \leq \frac{1}{\lambda} \left\| \frac{1}{m} \sum_{i=1}^m f_{\rho}(x_i) K_{x_i} - L_K f_{\rho} \right\|_K.$$

Applying Lemma 3 to the function f_{ρ} , we find that with confidence $1 - \delta$,

$$\|f_{\bar{x},\lambda} - \tilde{f}_{\lambda}\|_K \leq \frac{4\kappa M}{3m\lambda} \log \frac{1}{\delta} + \frac{\kappa M}{\sqrt{m\lambda}} \left(1 + \sqrt{8 \log \frac{1}{\delta}} \right).$$

In the same way, by Lemma 3 with the function f_{λ} and (7.8), we find

$$\text{Prob}_{\bar{x} \in X^m} \left\{ \|\tilde{f}_{\lambda} - f_{\lambda}\|_K \leq \frac{4\kappa \|f_{\lambda}\|_{\infty}}{3m\lambda} \log \frac{1}{\delta} + \frac{\kappa \|f_{\lambda}\|_{\rho}}{\sqrt{m\lambda}} \left(1 + \sqrt{8 \log \frac{1}{\delta}} \right) \right\} \geq 1 - \delta.$$

According to (7.9), $\|f_{\lambda}\|_{\rho} \leq 2M$ and $\|f_{\lambda}\|_{\infty} \leq \kappa \|f_{\lambda}\|_K \leq (\kappa M)/\sqrt{\lambda}$. Therefore, with confidence $1 - \delta$, there holds

$$\|\tilde{f}_{\lambda} - f_{\lambda}\|_K \leq \frac{4\kappa^2 M}{3m\lambda\sqrt{\lambda}} \log \frac{1}{\delta} + \frac{2\kappa M}{\sqrt{m\lambda}} \left(1 + \sqrt{8 \log \frac{1}{\delta}} \right).$$

Finally, we apply Theorem 5. For each $\bar{x} \in X^m$, there holds with confidence $1 - \delta$,

$$\|f_{\mathbf{z},\lambda} - f_{\bar{x},\lambda}\|_K \leq \frac{\kappa}{m\lambda} \left(\sqrt{\sigma^2} + \sqrt{8\sigma^2 \log \frac{1}{\delta}} + \frac{4}{3} M \log \frac{1}{\delta} \right). \quad (8.4)$$

Here $\sigma^2 = \sum_{i=1}^m \sigma_{x_i}^2$. Apply the Bernstein inequality

$$\text{Prob}_{\bar{x} \in X^m} \left\{ \frac{1}{m} \sum_{i=1}^m \xi(x_i) - E(\xi) \geq \varepsilon \right\} \leq \exp \left\{ -\frac{m\varepsilon^2}{2(B\varepsilon/3 + \sigma^2(\xi))} \right\}$$

to the random variable $\xi(x) = \int_Y (y - f_{\rho}(x))^2 d\rho(y|x)$. It satisfies $0 \leq \xi \leq 4M^2$, $E(\xi) = \sigma^2(\rho)$, and $\sigma^2(\xi) \leq 4M^2\sigma^2(\rho)$. Also, $E(\xi) = \sigma^2(\rho)$ and $(1/m) \sum_{i=1}^m \xi(x_i) = (1/m) \sum_{i=1}^m \sigma_{x_i}^2$. Solving the quadratic equation for the probability bound equal to δ , we see that

$$\text{Prob}_{\bar{x} \in X^m} \left\{ \frac{1}{m} \sum_{i=1}^m \sigma_{x_i}^2 \leq \sigma^2(\rho) + \frac{8M^2 \log(1/\delta)}{3m} + \sqrt{\frac{8M^2\sigma^2(\rho) \log(1/\delta)}{m}} \right\} \geq 1 - \delta.$$

Hence with confidence $1 - \delta$,

$$\sqrt{\sigma^2} \leq \sqrt{m\sigma^2(\rho)} + M \sqrt{3 \log \frac{1}{\delta}} + \left(8mM^2\sigma^2(\rho) \log \frac{1}{\delta} \right)^{1/4}$$

which is bounded by $2\sqrt{m\sigma^2(\rho)} + 2M\sqrt{3\log(1/\delta)}$. Together with (8.4), we see that with probability $1 - 2\delta$ in Z^m , we have the bound

$$\|f_{\mathbf{z},\lambda} - f_{\tilde{\mathbf{x}},\lambda}\|_K \leq \frac{2\kappa\sqrt{\sigma^2(\rho)}}{\sqrt{m\lambda}} + \frac{5\kappa M\sqrt{\log(1/\delta)}}{m\lambda} \left(1 + \sqrt{8\log\frac{1}{\delta}}\right).$$

Combining the above three bounds for $\|f_{\tilde{\mathbf{x}},\lambda} - \tilde{f}_\lambda\|_K$, $\|\tilde{f}_\lambda - f_\lambda\|_K$, and $\|f_{\mathbf{z},\lambda} - f_{\tilde{\mathbf{x}},\lambda}\|_K$, we know that for $0 < \delta < 1/4$, with confidence $1 - 4\delta$, $\|f_{\mathbf{z},\lambda} - f_\lambda\|_K$ is bounded by

$$\begin{aligned} & \frac{\kappa M}{\sqrt{m\lambda}} \left\{ \frac{20\log(1/\delta)}{\sqrt{m}} + 3 + 3\sqrt{8\log\frac{1}{\delta}} + \frac{5\sqrt{2\sigma^2(\rho)\log(1/\delta)}}{M} + \frac{4\kappa\log(1/\delta)}{3\sqrt{m\lambda}} \right\} \\ & \leq \frac{\kappa M}{\sqrt{m\lambda}} \sqrt{\log\frac{1}{\delta}} \left\{ 20\sqrt{\frac{\log(1/\delta)}{m}} + \frac{3}{\log 2} + 6\sqrt{2} + \frac{5\sqrt{2\sigma^2(\rho)}}{M} + \frac{4\kappa}{3}\sqrt{\frac{\log(1/\delta)}{m\lambda}} \right\}. \end{aligned}$$

But $\sigma^2(\rho) \leq M^2$. So our conclusion follows. $\square$

We are in a position to state our convergence rates in both $\|\cdot\|_K$ and $\|\cdot\|_\rho$ norms.

Corollary 5. Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \leq M$ almost surely. If f_ρ is in the range of L_K , then for any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \leq 2\|L_K^{-1}f_\rho\|_\rho^{2/3} \left(40\kappa M \log\frac{4}{\delta}\right)^{1/3} \left(\frac{1}{m}\right)^{\frac{1}{6}} \quad (8.5)$$

for $m \geq (\kappa^2\|L_K^{-1}f_\rho\|_\rho)/(360M \log(4/\delta))$, by taking $\lambda = (40\kappa M \log(4/\delta)/\|L_K^{-1}f_\rho\|_\rho)^{2/3}(1/m)^{1/3}$; and

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \leq 2\|L_K^{-1}f_\rho\|_\rho^{1/2} \left(40\kappa M \log\frac{4}{\delta}\right)^{1/2} \left(\frac{1}{m}\right)^{\frac{1}{4}} \quad (8.6)$$

for $m \geq (\kappa^2\|L_K^{-1}f_\rho\|_\rho^{2/3})/(180(M \log(4/\delta))^{2/3})$, by taking $\lambda = (40\kappa M \log(4/\delta)/\|L_K^{-1}f_\rho\|_\rho)^{1/2}(1/m)^{1/4}$.

References

- [1] A. Aldroubi, K. Gröchenig, Non-uniform sampling and reconstruction in shift-invariant spaces, SIAM Rev. 43 (2001) 585–620.
- [2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
- [3] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
- [4] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory, Found. Comput. Math. 2 (2002) 413–428.
- [5] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.
- [6] H.W. Engl, M. Hanke, A. Neubauer, Regularization of Inverse Problems, Mathematics and Its Applications, vol. 375, Kluwer Academic, Dordrecht, 1996.
- [7] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
- [8] C. McDiarmid, Concentration, in: Probabilistic Methods for Algorithmic Discrete Mathematics, Springer-Verlag, Berlin, 1998, pp. 195–248.
- [9] P. Niyogi, The Informational Complexity of Learning, Kluwer Academic, Boston, 1998.

- [10] T. Poggio, S. Smale, The mathematics of learning: Dealing with data, *Notices Amer. Math. Soc.* 50 (2003) 537–544.
- [11] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.* 1 (2003) 17–41.
- [12] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* 41 (2004) 279–305.
- [13] M. Unser, Sampling—50 years after Shannon, *Proc. IEEE* 88 (2000) 569–587.
- [14] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
- [15] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
- [16] Y. Ying, McDiarmid inequalities of Bernstein and Bennett forms, Technical Report, City University of Hong Kong, 2004.
- [17] R.M. Young, *An Introduction to Non-Harmonic Fourier Series*, Academic Press, New York, 1980.
- [18] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comput.* 15 (2003) 1397–1437.