Shannon Sampling II. Connections to Learning Theory †

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

July 22, 2004

Abstract

We continue our study [12] of Shannon sampling and function reconstruction. In this paper, the error analysis is improved. The problem of function reconstruction is extended to a more general setting with frames beyond point evaluation. Then we show how our approach can be applied to learning theory: a functional analysis framework is presented; sharp, dimension independent probability estimates are given not only for error in the L^2 spaces, but also for the error in the reproducing kernel Hilbert space where the a learning algorithm is performed. Covering number arguments are replaced by estimates of integral operators.

Keywords and Phrases: 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-limited functions or functions with certain

† The first author is partially supported by NSF grant 0325113. The second author is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 103704] and by City University of Hong Kong [Project No. 7001442].

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 $f_{\mathbf{z},\lambda}$ in $\mathcal{H}_K$. Dimension independent exponential probability estimates are given for the error $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$. 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 $J : \mathcal{H} \to C(X)$ is bounded with $\|J\| < \infty$.

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

$$f(x) =< f, E_x >_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}. \tag{2.1}$$

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

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

We shall always assume that $S_{\overline{x}}$ is bounded.

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

$$< f, S_{\overline{x}}^T c >_{\mathcal{H}} =< S_{\overline{x}} f, c >_{\ell^2(\overline{x})} = \sum_{x \in \overline{x}} c_x f(x) =< f, \sum_{x \in \overline{x}} c_x E_x >_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}.$$

It follows that

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

§3. Algorithm

To allow noise, we make the following assumption.

Special Assumption. *The sampled values $y = (y_x)_{x\in\overline{x}}$ have the form:*

For some $f^* \in \mathcal{H}$, and each $x \in \overline{x}$, $y_x = f^*(x) + \eta_x$, where η_x is drawn from ρ_x. (3.1)

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

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

The Markov inequality for a nonnegative random variable ξ asserts that

$$\text{Prob}\{\xi \le \frac{E(\xi)}{\delta}\} \ge 1-\delta, \qquad \forall\ 0<\delta<1. \tag{3.2}$$

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

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

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

Let $\gamma \ge 0$. With the sample $\mathbf{z} := (x, y_x)_{x\in\overline{x}}$, consider the algorithm

Function reconstruction $$\widetilde{f} := \arg\min_{f\in\mathcal{H}} \left\{ \sum_{x\in\overline{x}} \big(f(x) - y_x\big)^2 + \gamma \|f\|^2_{\mathcal{H}} \right\}. \tag{3.3}$$

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

$$\widetilde{f} = Ly, \qquad L := \big(S_{\overline{x}}^T S_{\overline{x}} + \gamma I\big)^{-1} S_{\overline{x}}^T.$$

Proof. Denote

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

Since $\sum_{x\in\overline{x}} \big(f(x)\big)^2 = \|S_{\overline{x}} f\|^2_{\ell^2(\overline{x})} =< S_{\overline{x}}^T S_{\overline{x}} f, f >_{\mathcal{H}}$, we know that for $f \in \mathcal{H}$,

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

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

$$\big(S_{\overline{x}}^T S_{\overline{x}} + \gamma I\big) f_{\mathbf{z},\gamma} = S_{\overline{x}}^T y.$$

This proves Theorem 1. □

The invertibility of the operator $S_{\overline{x}}^T S_{\overline{x}} + \gamma I$ is valid for rich data.

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

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

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

The problem of function reconstruction here is to estimate the error $\|\widetilde{f} - f^*\|_{\mathcal{H}}$. 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

$$\|\widetilde{f} - f^*\|_{\mathcal{H}} \le \frac{\|J\|\sqrt{\sigma^2/\delta}}{\lambda^2_{\overline{x}}}. \tag{3.5}$$

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

$$\|\widetilde{f} - f^*\|_{\mathcal{H}} \le \frac{\|J\|}{\lambda^2_{\overline{x}}}\left(\sqrt{8\sigma^2 \log\frac{1}{\delta}} + \frac{4}{3}M \log\frac{1}{\delta}\right). \tag{3.6}$$

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

$$f_{\overline{x},\gamma} := L\big(S_{\overline{x}} f^*\big). \tag{4.1}$$

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

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

$$\|\widetilde{f} - f_{\overline{x},\gamma}\|^2_{\mathcal{H}} \le \frac{\|\big(S^T_{\overline{x}} S_{\overline{x}} + \gamma I\big)^{-1}\|^2 \|J\|^2 \sigma^2}{\delta}.$$

If $|\eta_x| \le M$ for some $M \ge 0$ and each $x \in \overline{x}$, then for every $\varepsilon > 0$, we have

$$Prob_y\left\{\|\widetilde{f} - f_{\overline{x},\gamma}\|^2_{\mathcal{H}} \le \|L\|^2\sigma^2(1+\varepsilon)\right\} \ge 1 - \exp\left\{-\frac{\varepsilon\sigma^2}{2M^2}\log\big(1+\varepsilon\big)\right\}.$$

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

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

But

$$S^T_{\overline{x}}\big(y - S_{\overline{x}} f^*\big) = \sum_{x \in \overline{x}} \big(y_x - f^*(x)\big) E_x.$$

Hence

$$\|S^T_{\overline{x}}\big(y - S_{\overline{x}} f^*\big)\|^2_{\mathcal{H}} = \sum_{x \in \overline{x}} \sum_{x' \in \overline{x}} \big(y_x - f^*(x)\big)\big(y_{x'} - f^*(x')\big) < E_x, E_{x'} >_{\mathcal{H}} .$$

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

$$E\big(\|S^T_{\overline{x}}\big(y - S_{\overline{x}} f^*\big)\|^2_{\mathcal{H}}\big) = \sum_{x \in \overline{x}} \sigma_x^2 < E_x, E_x >_{\mathcal{H}} .$$

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

$$E\big(\|\widetilde{f} - f_{\overline{x},\gamma}\|^2_{\mathcal{H}}\big) \leq \|\big(S^T_{\overline{x}} S_{\overline{x}} + \gamma I\big)^{-1}\|^2 \|J\|^2 \sigma^2.$$

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 $\{\eta_x^2\}_{x \in \overline{x}}$. Special Assumption tells us that $E(\eta_x) = 0$, which implies $E(\eta_x^2) = \sigma_x^2$. Then we see that for every $\varepsilon > 0$,

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

Here we have used the condition $|\eta_x| \leq M$, which implies $\big|\eta_x^2 - \sigma_x^2\big| \leq M^2$. Also, $\sum_{x \in \overline{x}} \sigma^2(\eta_x^2) \leq \sum_{x \in \overline{x}} \big(E(\eta_x^4)\big) \leq M^2 \sigma^2 < \infty$.

The desired bound then follows from $\|\widetilde{f} - f_{\overline{x},\gamma}\|^2_{\mathcal{H}} \leq \|L\|^2 \|\{\eta_x\}\|^2_{\ell^2(\overline{x})}$. □

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

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

$$\|\big(S^T_{\overline{x}} S_{\overline{x}} + \gamma I\big)^{-1}\| \leq \frac{1}{\lambda^2_{\overline{x}} + \gamma}.$$

For the operator L, we have

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

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

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

Taking inner products on both sides with u, we have

$$< S_{\overline{x}} u, S_{\overline{x}} u >_{\ell^2(\overline{x})} + \gamma \|u\|_{\mathcal{H}}^2 =< v, u >_{\mathcal{H}} \leq \|v\|_{\mathcal{H}} \|u\|_{\mathcal{H}}.$$

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

$$< S_{\overline{x}} u, S_{\overline{x}} u >_{\ell^2(\overline{x})} = \|S_{\overline{x}} u\|_{\ell^2(\overline{x})}^2 \geq \lambda_{\overline{x}}^2 \|u\|_{\mathcal{H}}^2.$$

It follows that

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

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

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

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

§5. Integration Error

Recall that $f_{\overline{x},\gamma} = L\big(S_{\overline{x}} f^*\big) = \big(S_{\overline{x}}^T S_{\overline{x}} + \gamma I\big)^{-1} S_{\overline{x}}^T S_{\overline{x}} f^*$. Then

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

This in connection with Proposition 1 proves the following proposition.

Proposition 2. *If $S_{\overline{x}}^T S_{\overline{x}} + \gamma I$ is invertible, then*

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

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

$$\|\widetilde{f} - f^*\|_{\mathcal{H}} \le \frac{\|J\| \sqrt{\sigma^2/\delta}}{\lambda_{\overline{x}}^2}.$$

For the poor data case $\lambda_{\overline{x}} = 0$, we need to estimate $\gamma\big(S_{\overline{x}}^T S_{\overline{x}} + \gamma I\big)^{-1} f^*$ according to (5.1).

Recall that for a positive self-adjoint linear operator $\mathcal{L}$ on a Hilbert space $\mathcal{H}$, there holds

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

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

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

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

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

Proposition 3. *Define f_γ^* as*

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

then there holds

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

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

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 $\big(S_{\overline{x}}^T S_{\overline{x}}\big)^r$. For every

$0 < \delta < 1$, with probability $1 - \delta$, there holds

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

§6. More General Setting of Function Reconstruction

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

$$\sum_{x\in\overline{x}} \big| < f, E_x >_{\mathcal{H}} \big|^2 \leq B\|f\|_{\mathcal{H}}^2, \qquad \forall f \in \mathcal{H}. \tag{6.1}$$

Moreover, $\overline{x}$ 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

$$A\|f\|_{\mathcal{H}}^2 \leq \sum_{x\in\overline{x}} \big| < f, E_x >_{\mathcal{H}} \big|^2 \leq B\|f\|_{\mathcal{H}}^2, \qquad \forall f \in \mathcal{H}.$$

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

$$f = \sum_{x\in\overline{x}} < f, \left(S_{\overline{x}}^T S_{\overline{x}}\right)^{-1} E_x >_{\mathcal{H}} E_x, \qquad \forall f \in \mathcal{H}.$$

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** $\{E_x\}_{x\in\overline{x}}$ in $\mathcal{H}$ satisfying (6.1). Here the point evaluation (2.1) is replaced by the functional $< f, E_x >_{\mathcal{H}}$ and the algorithm becomes

$$\widetilde{f} := \arg\min_{f\in\mathcal{H}} \left\{ \sum_{x\in\overline{x}} \big(< f, E_x >_{\mathcal{H}} - y_x \big)^2 + \gamma\|f\|_{\mathcal{H}}^2 \right\}. \tag{6.2}$$

The sample values are given by $y_x = < f^*, E_x >_{\mathcal{H}} + \eta_x$. If we replace the sampling operator $S_{\overline{x}}$ by the operator from $\mathcal{H}$ to $\ell^2(\overline{x})$ mapping f to $\big(< f, E_x >_{\mathcal{H}} \big)_{x\in\overline{x}}$, then the algorithm can be analyzed in the same 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: $|t|_\epsilon := \max\{|t| - \epsilon, 0\}$, and a function space $\widetilde{\mathcal{H}}$ included in $\mathcal{H}$ (such as a Sobolev space in L^2), then a function reconstruction scheme becomes

$$\widetilde{f} := \arg\min_{f\in\widetilde{\mathcal{H}}}\left\{\sum_{x\in\overline{x}}\big|< f, E_x >_{\mathcal{H}} - y_x\big|_\epsilon + \gamma\|f\|^2_{\widetilde{\mathcal{H}}}\right\}.$$

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 ($A \to 0$ as the number of points in $\overline{x}$ increases) often happens. For such situations, we take $\overline{x}$ 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 $f : X \to Y$ is given by $\mathcal{E}(f) := \int_Z \big(f(x) - y\big)^2 d\rho$. The function minimizing the error is called the **regression function** and is given by

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

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 $\mathbf{z} := \big\{(x_i, y_i)\big\}_{i=1}^m$ drawn independently according to ρ. This purpose is achieved in Corollaries 3, 4, and 5 below. Here we consider kernel based learning algorithms.

Let $K : X \times X \to \mathbb{R}$ 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 [2] of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in$

$X\}$ with the inner product $< \cdot, \cdot >_K$ satisfying $< K_x, K_y >_K = K(x, y)$. The reproducing property takes the form

$$< K_x, g >_K = g(x), \qquad \forall x \in X, g \in \mathcal{H}_K. \tag{7.1}$$

The learning algorithm we study here is a regularized one:

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

We shall investigate how $f_{\mathbf{z},\lambda}$ approximates f_ρ and how the choice of the regularization parameter λ 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 weel 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 algorithm can be written in the same way as (3.3). To see this, we denote $\overline{x} = \{x_i\}_{i=1}^m, \rho_x = \rho(\cdot|x) - f_\rho(x)$ for $x \in X$. Then $E_x = K_x$ for $x \in \overline{x}$. Special 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 learning scheme (7.2) becomes

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

Therefore, Theorem 1 still holds and we have

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

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

Denote $\kappa := \sqrt{\sup_{x \in X} K(x, x)}$ and $f_\rho|_{\overline{x}} := \big(f_\rho(x)\big)_{x \in \overline{x}}$. Define

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

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

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

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

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

It is a good approximation of the integral operator $L_K : L^2_{\rho_X} \to \mathcal{H}_K$ 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. As operators on $\mathcal{H}_K$, $L_{K,\overline{x}} S_{\overline{x}}$ approximates L_K well. In fact, it was shown in [5] that $E\big(\|L_{K,\overline{x}} S_{\overline{x}} - L_K\|_{\mathcal{H}_K \to \mathcal{H}_K}\big) \le \frac{\kappa^2}{\sqrt{m}}$. To get sharper error bounds, we need to get estimates for the operators with domain $L^2_{\rho_X}$.

Lemma 1. *Let $\overline{x} \in X^m$ be randomly drawn according to ρ_X. Then for any $f \in L^2_{\rho_X}$,*

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

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

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

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

The function $f_{\overline{x},\lambda}$ may be considered as an approximation of f_λ where f_λ is defined by

$$f_\lambda := \big(L_K + \lambda I\big)^{-1} L_K f_\rho. \tag{7.3}$$

In fact, f_λ is a minimizer of the optimization problem:

$$f_\lambda := \arg\min_{f \in \mathcal{H}_K} \big\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \big\} = \arg\min_{f \in \mathcal{H}_K} \big\{ \mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^2 \big\}. \tag{7.4}$$

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

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

and

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

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

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

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

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

$$\tilde{f}_\lambda:=\big(L_{K,\overline{x}}S_{\overline{x}}+\lambda I\big)^{-1}L_Kf_\rho. \tag{7.5}$$

Since

$$f_{\overline{x},\lambda}-\tilde{f}_\lambda=\big(L_{K,\overline{x}}S_{\overline{x}}+\lambda I\big)^{-1}\big(S_{\overline{x}}^Tf_\rho|_{\overline{x}}-L_Kf_\rho\big), \tag{7.6}$$

Lemma 1 tells us that

$$E\big(\|f_{\overline{x},\lambda}-\tilde{f}_\lambda\|_K\big)\le\frac{1}{\lambda}E\big(\|S_{\overline{x}}^Tf_\rho|_{\overline{x}}-L_Kf_\rho\|_K\big)\le\frac{\kappa\|f_\rho\|_\rho}{\sqrt{m\lambda}}.$$

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

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

Hence

$$\|\tilde{f}_\lambda-f_\lambda\|_K\le\frac{1}{\lambda}\|L_Kf_\lambda-L_{K,\overline{x}}S_{\overline{x}}f_\lambda\|_K. \tag{7.7}$$

Applying Lemma 1 again, we see that

$$E\big(\|\tilde{f}_\lambda-f_\lambda\|_K\big)\le\frac{1}{\lambda}E\big(\|L_Kf_\lambda-L_{K,\overline{x}}S_{\overline{x}}f_\lambda\|_K\big)\le\frac{\kappa\|f_\lambda\|_\rho}{\sqrt{m\lambda}}.$$

Note that f_λ is a minimizer of (7.4). Taking $f=0$ yields $\|f_\lambda-f_\rho\|_\rho^2+\lambda\|f_\lambda\|_K^2\le\|f_\rho\|_\rho^2$. Hence

$$\|f_\lambda\|_\rho\le 2\|f_\rho\|_\rho\qquad\text{and}\qquad\|f_\lambda\|_K\le\|f_\rho\|_\rho/\sqrt{\lambda}. \tag{7.8}$$

Therefore, our second estimate follows. □

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

Theorem 4. *Define f_λ by (7.3). If $L_K^{-r} f_\rho \in L^2_{\rho_X}$ for some $0 < r \le 1$, then*

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|L_K^{-r} f_\rho\|_\rho. \tag{7.9}$$

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

$$\|f_\lambda - f_\rho\|_K \le \lambda^{r-\frac{1}{2}} \|L_K^{-r} f_\rho\|_\rho. \tag{7.10}$$

We follow the same line as we did in [11]. Estimates similar to (7.9) 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 $0 < r < s$,

$$\inf_{b\in\mathcal{H}} \left\{ \|b-a\|^2 + \gamma \|A^{-s} b\|^2 \right\} \le \gamma^{r/s} \|A^{-r} a\|^2. \tag{7.11}$$

(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 (7.9) was given in [5]. Here we provide a complete proof because the idea is used for verifying (7.10).

Proof of Theorem 4. If $\{\lambda_i, \psi_i\}_{i\ge1}$ are the normalized eigenpairs of the integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$, then $\|\sqrt{\lambda_i}\psi_i\|_K = 1$ when $\lambda_i > 0$.

Write $f_\rho = L_K^r g$ for some $g = \sum_{i\ge1} d_i \psi_i$ with $\|\{d_i\}\|_{\ell^2} = \|g\|_\rho < \infty$. Then $f_\rho = \sum_{i\ge1} \lambda_i^r d_i \psi_i$ and by (7.3),

$$f_\lambda - f_\rho = \big(L_K + \lambda I\big)^{-1} L_K f_\rho - f_\rho = -\sum_{i\ge1} \frac{\lambda}{\lambda_i + \lambda} \lambda_i^r d_i \psi_i.$$

It follows that

$$\|f_\lambda - f_\rho\|_\rho = \left\{ \sum_{i\ge1} \left(\frac{\lambda}{\lambda_i+\lambda} \lambda_i^r d_i \right)^2 \right\}^{1/2} = \lambda^r \left\{ \sum_{i\ge1} \left(\frac{\lambda}{\lambda_i+\lambda} \right)^{2(1-r)} \left(\frac{\lambda_i}{\lambda+\lambda_i} \right)^{2r} d_i^2 \right\}^{1/2}.$$

This is bounded by $\lambda^r \|\{d_i\}\|_{\ell^2} = \lambda^r \|g\|_\rho = \lambda^r \|L_K^{-r} f_\rho\|_\rho$. Hence (7.9) holds.

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

$$\|f_\lambda - f_\rho\|_K^2 = \sum_{\lambda_i>0} \left(\frac{\lambda}{\lambda_i+\lambda} \lambda_i^{r-\frac{1}{2}} d_i \right)^2 = \lambda^{2r-1} \sum_{i\ge1} \left(\frac{\lambda}{\lambda_i+\lambda} \right)^{3-2r} \left(\frac{\lambda_i}{\lambda+\lambda_i} \right)^{2r-1} d_i^2.$$

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 (7.10) has been verified. □

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*. Assume* $L_K^{-r} f_\rho \in L^2_{\rho_X}$ *for some* $\frac{1}{2} < r \leq 1$*, then*

$$E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_K\big) \leq C_{\rho,K}\bigg\{\frac{1}{\sqrt{m\lambda}} + \lambda^{r-\frac{1}{2}}\bigg\}, \tag{7.12}$$

where $C_{\rho,K} := \kappa\sqrt{\sigma^2(\rho)} + 3\kappa\|f_\rho\|_\rho + \|L_K^{-r} f_\rho\|_\rho$ *is independent of the dimension. Hence*

$$\lambda = m^{-\frac{1}{1+2r}} \Longrightarrow E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_K\big) \leq 2C_{\rho,K}\big(\frac{1}{m}\big)^{\frac{2r-1}{4r+2}}. \tag{7.13}$$

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* $L_K^{-\frac{1}{2}} f_\rho \in L^2_{\rho_X}$*. To get convergence rates we assume a stronger condition* $L_K^{-r} f_\rho \in L^2_{\rho_X}$ *for some* $\frac{1}{2} < r \leq 1$*. The optimal rate derived from Corollary 3 is achieved by* $r = 1$*. In this case,* $E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_K\big) = \big(\frac{1}{m}\big)^{\frac{1}{6}}$*. The norm* $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ *can not be bounded by the error* $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$*, hence our bound for the* $\mathcal{H}_K$*-norm error is new in learning theory.*

Corollary 4. *Let* $\mathbf{z}$ *be randomly drawn according to* ρ*. If* $L_K^{-r} f_\rho \in L^2_{\rho_X}$ *for some* $0 < r \leq 1$*, then*

$$E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho\big) \leq C'_{\rho,K}\bigg\{\frac{1}{\sqrt{m\lambda}} + \lambda^{r}\bigg\}, \tag{7.14}$$

where $C'_{\rho,K} := \kappa^2\sqrt{\sigma^2(\rho)} + 3\kappa^2\|f_\rho\|_\rho + \|L_K^{-r} f_\rho\|_\rho$*. Thus,*

$$\lambda = m^{-\frac{1}{2+2r}} \Longrightarrow E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho\big) \leq 2C'_{\rho,K}\big(\frac{1}{m}\big)^{\frac{r}{2r+2}}. \tag{7.15}$$

Remark. *The convergence rate (7.15) for the* $L^2_{\rho_X}$*-norm is obtained by optimizing the regularization parameter* λ *in (7.14). The sharp rate derived from Corollary 4 is* $\big(\frac{1}{m}\big)^{\frac{1}{4}}$*, which is achieved by* $r = 1$*.*

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

$$E_{\mathbf{z}\in Z^m}\big(\mathcal{E}(f_{\mathbf{z},\lambda})\big) \leq \bigg(1 + \frac{2\kappa^2}{m\lambda}\bigg)^2 \inf_{f\in\mathcal{H}_K}\bigg\{\mathcal{E}(f) + \frac{\lambda}{2}\|f\|_K^2\bigg\}. \tag{7.16}$$

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

$$\mathcal{D}(\lambda) := \inf_{f\in\mathcal{H}_K}\big\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_K^2\big\} = \inf_{f\in\mathcal{H}_K}\big\{\|f - f_\rho\|_\rho^2 + \lambda\|f\|_K^2\big\}, \tag{7.17}$$

then the bound (7.16) can be restated as

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

One can then derive the convergence rate $\big(\frac{1}{m}\big)^{\frac{1}{4}}$ in expectation when $f_\rho\in\mathcal{H}_K$ and $\mathcal{E}(f_\rho)>0$. In fact, (7.11) with $\mathcal{H}=L^2_{\rho_X}$, $A=L_K$ holds for $r=s=1/2$, which yields the best rate for the regularization error $\mathcal{D}(\lambda)\le\|f_\rho\|_K^2\lambda$. One can thus get $E_{\mathbf{z}\in Z^m}\big(\big\|f_{\mathbf{z},\lambda}-f_\rho\big\|_\rho^2\big)=\big(\frac{1}{m}\big)^{\frac{1}{2}}$ by taking $\lambda=1/\sqrt{m}$. Applying (3.2), one can have the probability estimate $\big\|f_{\mathbf{z},\lambda}-f_\rho\big\|_\rho\le\frac{C}{\delta}\big(\frac{1}{m}\big)^{\frac{1}{4}}$ 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 $1-\delta$,

$$\big|\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\lambda)\big|\le\frac{M\kappa^2}{\sqrt{m}\lambda}\Big(1+\frac{\kappa}{\sqrt{\lambda}}\Big)\Big(1+\sqrt{2\log\big(2/\delta\big)}\Big).\tag{7.18}$$

Convergence rates were also derived in [5, Corollary 1] by combining (7.18) with (7.9): when f_ρ lies in the range of L_K, for any $0<\delta<1$, with confidence $1-\delta$, there holds

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

Thus the confidence is improved from $1/\delta$ to $\log(2/\delta)$, while the rate is weakened to $\big(\frac{1}{m}\big)^{\frac{1}{5}}$. In the next section we shall verify the same confidence estimate while the sharp rate is kept. Our approach is short and neat, without involving the leave-one-out technique. Moreover, we can derive convergence rates in the space $\mathcal{H}_K$.

§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 $\mathbf{t}=(t_1,\cdots,t_m)\in\Omega^m$ and $t_i'\in\Omega$, we denote $\mathbf{t}^i:=(t_1,\cdots,t_{i-1},t_i',t_{i+1},\cdots,t_m)$.

Lemma 2. *Let $\{t_i,t_i'\}_{i=1}^m$ be i.i.d. drawers of the probability distribution ρ on Ω, and $F:\Omega^m\to\mathbb{R}$ be a measurable function.*

(1) If for each i there is c_i such that $\sup_{\mathbf{t}\in\Omega^m, t_i\in\Omega}\big|F(\mathbf{t}) - F(\mathbf{t}^i)\big| \le c_i$, then

$$Prob_{\mathbf{t}\in\Omega^m}\left\{F(\mathbf{t}) - E_{\mathbf{t}}\big(F(\mathbf{t})\big) \ge \varepsilon\right\} \le \exp\left\{-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right\}, \qquad \forall \varepsilon > 0. \tag{8.1}$$

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

$$Prob_{\mathbf{t}\in\Omega^m}\left\{F(\mathbf{t}) - E_{\mathbf{t}}\big(F(\mathbf{t})\big) \ge \varepsilon\right\} \le \exp\left\{-\frac{\varepsilon^2}{2\big(B\varepsilon/3 + \sum_{i=1}^m \sigma_i^2(F)\big)}\right\}, \quad \forall \varepsilon > 0, \tag{8.2}$$

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

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

First, we show how the probability estimate for function reconstruction stated in Theorem 2 can be improved.

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

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

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

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

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

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

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

$$E_y(F) \le \sqrt{E_y\big(F^2\big)} = \sqrt{\sum_{x\in\overline{x}} \sigma_x^2 < E_x, E_x >_{\mathcal{H}}} \le \|J\|\sqrt{\sigma^2}. \tag{8.3}$$

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

$$\big|F(y) - F(y^{x_0})\big| = \big|\|S_{\overline{x}}^T\big(y - S_{\overline{x}}f^*\big)\|_{\mathcal{H}} - \|S_{\overline{x}}^T\big(y^{x_0} - S_{\overline{x}}f^*\big)\|_{\mathcal{H}}\big| \le \|S_{\overline{x}}^T\big(y - y^{x_0}\big)\|_{\mathcal{H}}.$$

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

$$\begin{aligned} E_{y_{x_0}}\left(\big|F(y) - E_{y_0}\big(F(y)\big)\big|^2\right) &\le \int\left(\int\big|y_{x_0} - y'_{x_0}\big|\|J\|d\rho_{x_0}(y'_{x_0})\right)^2 d\rho_{x_0}(y_{x_0}) \\ &\le \int\int\left(y_{x_0} - y'_{x_0}\right)^2\|J\|^2 d\rho_{x_0}(y'_{x_0})d\rho_{x_0}(y_{x_0}) \le 4\|J\|^2\sigma_{x_0}^2. \end{aligned}$$

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

$$\mathrm{Prob}_{y\in Y^{\overline{x}}}\left\{F(y) - E_y\big(F(y)\big) \ge \varepsilon\right\} \le \exp\left\{-\frac{\varepsilon^2}{2\big(2M\|J\|\varepsilon/3 + 4\|J\|^2\sigma^2\big)}\right\}.$$

Solving the quadratic equation

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

gives the probability estimate

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

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

Then we 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 to a probability estimate form.

Lemma 3. *Let $\overline{x} \in X^m$ be randomly drawn according to ρ_X. Then for any $f \in L^\infty_{\rho_X}$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

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

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

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

For $j \in \{1, \cdots, m\}$, we have

$$\big|F(\overline{x}) - F(\overline{x}^j)\big| \le \|\frac{1}{m}\big(f(x_j) - f(x_j')\big)K_{x_j}\|_K \le \frac{\kappa}{m}\big|f(x_j) - f(x_j')\big|.$$

It follows that $\big|F(\overline{x}) - E_{x_j}\big(F(\overline{x})\big)\big| \le \frac{2\kappa\|f\|_\infty}{m} =: B$. Moreover,

$$\begin{aligned} E_{x_j}\bigg(F(\overline{x}) - E_{x_j}\big(F(\overline{x})\big)\bigg)^2 &\le \int_X \bigg(\int_X \frac{\kappa}{m}\big|f(x_j) - f(x_j')\big|d\rho_X(x_j')\bigg)^2 d\rho_X(x_j) \\ &\le \frac{\kappa^2}{m^2}\int_X\int_X 2\big|f(x_j)\big|^2 + 2\big|f(x_j')\big|^2 d\rho_X(x_j')d\rho_X(x_j) \le \frac{4\kappa^2\|f\|_\rho^2}{m^2}. \end{aligned}$$

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

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

$$\mathrm{Prob}_{\overline{x}\in X^m}\bigg\{F(\overline{x}) - E_{\overline{x}}(F(\overline{x})) \ge \varepsilon\bigg\} \le \exp\bigg\{-\frac{\varepsilon^2}{2\big(\frac{2\kappa\|f\|_\infty\varepsilon}{3m} + \frac{4\kappa^2\|f\|_\rho^2}{m}\big)}\bigg\}.$$

Solving a quadratic equation again, we see that with confidence $1-\delta$, we have

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

Lemma 1 says that $E_{\overline{x}}(F(\overline{x})) \le \frac{\kappa\|f\|_\rho}{\sqrt{m}}$. Then our conclusion follows. □

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

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_K \le \frac{\kappa M\log\big(4/\delta\big)}{\sqrt{m}\lambda}\bigg(30 + \frac{4\kappa}{3\sqrt{m\lambda}}\bigg).$$

Proof. Since $|y| \le M$ almost everywhere, we know that $\|f_\rho\|_\rho \le \|f_\rho\|_\infty \le M$.

Recall the function $\tilde{f}_\lambda$ defined by (7.5). It satisfies (7.6). Hence

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

Applying Lemma 3 to the function f_ρ, we have with confidence $1-\delta$,

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

In the same way, by Lemma 3 with the function f_λ and (7.7), we find

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

By (7.8), we have $\|f_\lambda\|_\rho \le 2M$ and

$$\|f_\lambda\|_\infty \le \kappa\|f_\lambda\|_K \le \frac{\kappa M}{\sqrt{\lambda}}.$$

Therefore, with confidence $1-\delta$, there holds

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

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

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

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

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

to the random variable $\xi(x) = \int_Y \big(y-f_\rho(x)\big)^2 d\rho(y|x)$. It satisfies $0\le\xi\le 4M^2$, $E(\xi) = \sigma^2(\rho)$, and $\sigma^2(\xi)\le 4M^2\sigma^2(\rho)$. Then we see that

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

Hence

$$\sqrt{\sigma^2} \le \sqrt{m\sigma^2(\rho)} + M\sqrt{3\log\big(1/\delta\big)} + \big(8mM^2\sigma^2(\rho)\log\big(1/\delta\big)\big)^{1/4}$$

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

$$\|f_{\mathbf{z},\lambda} - f_{\overline{x},\lambda}\|_K \le \frac{4\kappa\sqrt{2\sigma^2(\rho)\log\big(1/\delta\big)}}{\sqrt{m}\lambda} + \frac{34\kappa M\log\big(1/\delta\big)}{3m\lambda}.$$

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

$$\frac{\kappa M}{\sqrt{m}\lambda}\left\{\frac{13\log\big(1/\delta\big)}{\sqrt{m}} + 3 + 3\sqrt{8\log\big(1/\delta\big)} + \frac{4\sqrt{2\sigma^2(\rho)\log\big(1/\delta\big)}}{M} + \frac{4\kappa\log\big(1/\delta\big)}{3\sqrt{m\lambda}}\right\}$$
$$\le \frac{\kappa M}{\sqrt{m}\lambda}\sqrt{\log\big(1/\delta\big)}\left\{13\sqrt{\frac{\log\big(1/\delta\big)}{m}} + \frac{3}{\log 2} + 6\sqrt{2} + \frac{4\sqrt{2\sigma^2(\rho)}}{M} + \frac{4\kappa}{3}\sqrt{\frac{\log\big(1/\delta\big)}{m\lambda}}\right\}.$$

But $\sigma^2(\rho) \le M^2$. Then our conclusion follows. □

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* $|y| \le M$ *almost everywhere. If* f_ρ *is in the range of* L_K*, then for any* $0 < \delta < 1$*, with confidence* $1 - \delta$ *we have*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \le \widetilde{C}\left(\frac{\big(\log\big(4/\delta\big)^2}{m}\right)^{\frac{1}{6}} \qquad \textit{by taking} \qquad \lambda = \left(\frac{\big(\log\big(4/\delta\big)^2}{m}\right)^{\frac{1}{3}} \tag{8.5}$$

and

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_\rho \le \widetilde{C}\left(\frac{\big(\log\big(4/\delta\big)^2}{m}\right)^{\frac{1}{4}} \qquad \textit{by taking} \qquad \lambda = \left(\frac{\big(\log\big(4/\delta\big)^2}{m}\right)^{\frac{1}{4}}, \tag{8.6}$$

where $\widetilde{C}$ *is a constant independent of the dimension:*

$$\widetilde{C} := 30\kappa M + 2\kappa^2 M + \|L_K^{-1} f_\rho\|_\rho.$$

References

[1] A. Aldroubi and K. Gröchenig, Non-uniform sampling and reconstruction in shift-invariant spaces, SIAM Review **43** (2001), 585–620.

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

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

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

[5] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Foundat. Comput. Math., to appear.

[6] H. W. Engl, M. Hanke, and A. Neubauer, Regularization of Inverse Problems, Mathematics and Its Applications, **375**, Kluwer, Dordrecht, 1996.

[7] T. Evgeniou, M. Pontil, and 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, 1998.

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

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

[12] S. Smale and 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, John Wiley & Sons, 1998.

[15] G. Wahba, Spline Models for Observational Data, SIAM, 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 Comp. **15** (2003), 1397–1437.