Application of Integral Operator for Regularized Least Square Regression

Sun Hongwei *
School of Science, Jinan University, Jinan 250022, P.R. China
School of Mathematical Science, Beijing Normal University, Beijing 100875, P.R. China
E-mail: ss_sunhw@ujn.edu.cn

Qiang Wu
Department of Statistic Science, Institute for Genome Sciences & Policy
Duke University, Durham, NC 27708, USA
E-mail: qiang@stat.duke.edu

July 10, 2007

Abstract

In this paper, we study the consistency of the regularized least square regression in a general reproducing kernel Hilbert spaces. We characterized the compactness of the inclusion map from a reproducing kernel Hilbert space to the space of continuous functions and showed that the capacity based analysis by uniform covering numbers may fail in a very general setting. We prove the consistency and compute the learning rate by means of integral operator. To this end, we studied the properties of the integral operator. The analysis reveals that the essence of this approach is the isomorphism of the square root operator.

*Corresponding author
Keywords: regularized least square regression, integral operator, reproducing kernel Hilbert space, consistency.

1 Introduction

Let $X$ be a metric space, $\rho$ be a probability measure on $Z = X \times Y$ with $Y = \mathbb{R}$. The regression function $f_\rho$ is defined by

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

where $\rho(y|x)$ be the conditional probability measure. In regression problem, the probability measure $\rho$ is unknown but a set of samples $z = \{(x_i, y_i)\}_{i=1}^m$ are available instead. The target of the regression problem is to learn the regression function $f_\rho$, namely, to construct a good approximation from the set of samples $z$. A well known algorithm to solve this problem is the regularized least square method [5,8,14] defined by

$$f_{z,\gamma} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \gamma \| f \|_{\mathcal{H}_K}^2 \right\},$$

(1.1)

where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) associated with the kernel $K(x,y)$ (see [2] or Section 3 below for the definition and properties of RKHS) and $\gamma > 0$ is a regularization parameter. When the input space $X$ is a compact set and the kernel is a Mercer kernel, analysis of this algorithm has been extensively studied; see [4,9–11,15] and the references therein.

In modern sciences, instruments allows the measurement of processes which has motivated the functional data analysis. This leads to the regression problem in function spaces (usually Hilbert ones). When the function spaces are of infinite dimensions the input space $X$ may be not compact even if it is bounded. This has motivated the study of RKHS on non-compact metric spaces [13]. On the other hand, even when $X$ is compact, the well known capacity based consistency analysis (see below) may still fail because the capacity of the RKHS on general functional spaces has not been a well explored area. These motivate us to further study the regularized least square algorithm in a more general setting and in terms of integral operator technique in this paper.

Recall the goodness of the approximation $f_{z,\gamma}$ is measured by $\|f_{z,\gamma} - f_\rho\|_{\rho_X}^2$. As usual we write

$$\|f_{z,\gamma} - f_\rho\|_{\rho_X}^2 = \|f_{z,\gamma} - f_{\gamma}\|_{\rho_X}^2 + \|f_{\gamma} - f_\rho\|_{\rho_X}^2,$$
where
\[ f_\gamma := \arg \min_{f \in \mathcal{H}_K} \{ \| f - f_\rho \|_{\rho_X}^2 + \gamma \| f \|_{\mathcal{K}}^2 \}. \]  
(1.2)

So the error is split into two parts: the first term on the right as the estimation error (or sample error) and the second term as the approximation error. For the estimation error, the best known methods are based on the uniform law of large numbers. They are all capacity based approaches with capacity described by \( V_\gamma \)-dimension \([1]\), covering numbers \([4, 15]\), Rademacher complexities \([3]\), etc. Capacity analysis is rather general and can lead to fast learning rates. However, the drawback is the capacity is difficult to estimate. For example, the covering numbers of RKHS is only well estimated for certain very smooth kernels such as polynomial and Gaussians and the underlying input space \( X \) is finite dimensional compact set \([12, 16, 17]\). Little studies are available when \( X \) is of infinite dimension to our best knowledge. Moreover, concerning the covering numbers, we know those measured by empirical metrics are more suitable to describe the uniform convergence than the those measured by \( L^\infty \)-norm (also called uniform covering number). But it seems only the latter is estimable for the RKHS. Note that the existence of uniform covering number is equivalent to the compactness of the inclusion map: \( J : \mathcal{H}_K \to C(X) \). In Section 2, we will give a characterization for this compactness showing that it is equivalent to the continuity of the kernel as a bi-variable function. Based on this, we construct an example to show that

**Claim 1:** There exist RKHS \( \mathcal{H}_K \subset C(X) \) so that the the ball of radius \( R \) in \( \mathcal{H}_K \), \( B_R \), is not compact in \( C(X) \) even when \( X \) is compact. Hence the uniform covering number is infinite and can not be used for error analysis of learning algorithms.

Another method for the consistency analysis, introduced in a series papers of Smale-Zhou \([9–11]\), is to use the property of the integral operator

\[ L_K(f)(x) = \int_X K(x, t)f(t)d\rho_X(t). \]

Note that in their papers, the input space is restrict to be compact and the kernel must be a Mercer kernel. The main purpose of this paper is to extend this method to a more general setting and obtain satisfactory learning rate under rather weak conditions (see Section 4 below). In our setting, the input space is only a abstract metric space. No compactness or dimensional restriction is needed so that it applies to functional regression setting. The RKHS is also general, even the kernel can be non-continuous (though this is only of theoretical rather than practical interest). To this end, we will study the properties of \( L_K \) under our weak
conditions in Section 3. Then we apply these to the error analysis of regularized least square regression and the main result is the following learning rate:

**Theorem 1.1.** Assume \( \sup_{x \in X} K(x, x) < \infty \). If \( L_K^{-r} f_\rho \in L^2_{\rho_X}(X) \), \( 0 < r \leq 1 \), then by taking \( \gamma = O(m^{-\frac{1}{2(r+1)}}) \) there holds

\[
\|f_{z, \gamma} - f_\rho\|_{\rho_X} = O(m^{-\frac{r}{2(r+1)}})
\]

in probability.

This result will be proved in Section 4. Though the conditions become weaker, the learning rates is the same as in [11]. We also remark that our analysis verifies the following claim:

**Claim 2:** Though the compactness and spectral decomposition of \( L_K \) plays important role in many previous works (e.g. [4, 11]), they are not essential for the error analysis. The most essential property so that \( L_K \) is admissible for consistency analysis lies on the fact that and \( L^2_K \) is an isometric isomorphism from \( L^2_{\rho_X}(X) \) onto \( H_K \).

## 2 Compactness of inclusion map

Let \( X \) be a metric space and the kernel \( K(x, y) : X \times X \rightarrow R \) be symmetric and positive semidefinite, i.e., for any finite set of distinct points \( \{x_1, \cdots, x_m\} \subset X \), the matrix \( (K(x_i, x_j))_{i,j=1}^m \) is positive semidefinite. The *reproducing kernel Hilbert space (RKHS)* \( H_K \) associated with \( K(x, y) \) is defined (see [2]) to be the closure of the linear space spanned by \( \{K_x := K(x, \cdot) : x \in X\} \) with the inner product \( <\cdot, \cdot>_K \) defined by

\[
<\sum_{i=1}^n \alpha_i K_{x_i}, \sum_{j=1}^m \beta_j K_{y_j}>_K := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j K(x_i, y_j).
\]

The reproducing property takes the form,

\[
f(x) = <f, K_x>_K \quad \forall x \in X, \quad f \in H_K.
\]  

When \( H_K \) contains only continuous function, i.e., \( H_K \subset C(X) \), an inclusion map \( J : H_K \rightarrow C(X) \) can be defined. In the literature the uniform covering
numbers has been extensively used to estimate the sample error for learning algorithms. As is well known, the finiteness of the covering number is equivalent to the compactness of the inclusion map \( J \). In this section, we will study when the inclusion map is compact.

Define \( \eta : X \rightarrow \mathcal{H}_K \) as \( \eta(x) = K_x \), called the feature map of \( \mathcal{H}_K \). The following results illustrate the relations among the continuity of \( K(x,y) \), the continuity of \( \eta \) and the compactness of the inclusion map \( J \).

Lemma 2.1. The kernel function \( K(x,y) \) is continuous if and only if the feature map \( \eta \) is continuous.

Proof. Necessity. For all \( x, x' \in X \),

\[
\| \eta(x) - \eta(x') \|_K = < K_x - K_{x'}, K_x - K_{x'} >_K = K(x,x) - 2K(x,x') + K(x',x').
\]

Since \( K(x,y) \) is continuous, it follows that \( \| \eta(x) - \eta(x') \|_K \to 0 \) as \( x' \to x \) in \( X \). This proves \( \eta(x) \) is continuous.

Sufficiency. Fix \((x_0, y_0)\) in \( X \times X \). For any \((x, y)\) ∈ \( X \times X \),

\[
|K(x,y) - K(x_0,y_0)| = |< \eta(x), \eta(y) >_K - < \eta(x_0), \eta(y_0) >_K | \\
\leq | < \eta(x) - \eta(x_0), \eta(y) >_K | + | < \eta(x_0), \eta(y) - \eta(y_0) >_K |.
\]

By the continuity of \( \eta \) and Schwartz inequality, it is easy to check that \( |K(x,y) - K(x_0,y_0)| \to 0 \) as \( (x, y) \to (x_0, y_0) \). So \( K(x,y) \) is continuous. \( \square \)

Proposition 2.2. Assume \( X \) to be compact. The following two statements are equivalent:

(i) The kernel \( K(x,y) \) is continuous;

(ii) \( \mathcal{H}_K \subset C(X) \) and the inclusion map \( J : \mathcal{H}_K \to C(X) \) is compact.

Proof. (i)⇒(ii). By Lemma 2.1 and the compactness of \( X \), there exists \( M > 0 \) such that \( \| K_x \|_K \leq M \) for each \( x \in X \).

For \( g \in \mathcal{H}_K \) and \( x \in X \), we have

\[
|g(x)| = |< g, K_x >_K | \leq \| g \|_K \| K_x \|_K \leq M \| g \|_K.
\]
This implies that \( \|g\|_\infty \leq M \|g\|_K \) and \( J \) is bounded. Also, for any \( x' \in X \)

\[
|g(x) - g(x')| = |\langle K_x - K'_x, g \rangle_K| \leq \|g\|_K \|\eta(x) - \eta(x')\|_K.
\]

This in connection with the fact that \( \eta \) is uniformly continuous on \( X \) implies that \( \text{the image of } J \text{ on a bounded subset of } \mathcal{H}_K \) is equi-continuous. Consequently, \( J \) is compact by Ascoli’s Theorem (see e.g. [7]).

(ii) \( \Rightarrow \) (i). Let \( B_1 := \{f \in \mathcal{H}_K : \|f\|_K \leq 1\} \). Then for all \( x, x' \in X \),

\[
\|\eta(x) - \eta(x')\|_K = \sup_{f \in B_1} |\langle \eta(x) - \eta(x'), f \rangle_K| = \sup_{f \in B_1} |g(x) - g(x')|.
\]

By the the equi-continuity of \( J(B_1) \), it follows that \( \eta \) is continuous. So \( \mathcal{K}(x, y) \) is continuous by Lemma 2.1. \( \square \)

Proposition 2.2 provide a sufficient condition for the compactness of the inclusion map: \( X \) being compact and \( K \) continuous. This is most common setting for the learning on Euclidian space. Though, exceptions may happen. Firstly, the compactness of the underlying input space may be difficult to verify or even fails. Secondly, we construct an example to show the existence of of RKHS so that the inclusion map \( J : \mathcal{H}_K \to C(X) \) is not compact, even if \( X \) is compact. Though these kinds of space is not useful in practice, it is of theoretical interest and cast the difficulty on the convergence analysis of algorithms with rather general RKHS using uniform covering numbers.

Example 1. For \( n \in \mathbb{Z}_+ \), let

\[
\phi_n(x) := \begin{cases} 
4n^2(n + 1)^2(\frac{1}{n} - x)(x - \frac{1}{n+1}), & \frac{1}{n+1} \leq x \leq \frac{1}{n}, \\
0, & \text{otherwise}.
\end{cases}
\]

Obviously \( \phi_n(x) \in C(X) \) with \( X = [0, 1] \). Consider the set \( \mathcal{H} := \{\sum_{n=1}^\infty a_n \phi_n : \sum_{n=1}^\infty a_n^2 < \infty\} \) and define the inner product by

\[
\langle \sum_{n=1}^\infty a_n \phi_n , \sum_{n=1}^\infty b_n \phi_n \rangle_{\mathcal{H}} = \sum_{n=1}^\infty a_n b_n.
\]

Then \( \mathcal{H} \) is a Hilbert space with \( \{\phi_n(x)\} \) being an orthonormal basis. If \( \sum_{n=1}^\infty a_n \phi_n \in \mathcal{H} \), then

\[
\|\sum_{n=1}^\infty a_n \phi_n\|_\infty = \max_n |a_n| \leq \sqrt{\sum_{n=1}^\infty a_n^2} = \|\sum_{n=1}^\infty a_n \phi_n\|_{\mathcal{H}}.
\]

This implies that the point evaluation functional on \( \mathcal{H} \) is bounded. So \( \mathcal{H} \) is an RKHS. The above formula also implies that \( \sum_{k=1}^n a_k \phi_k \) uniformly converges to \( \sum_{k=1}^\infty a_k \phi_k \in C(X) \). So \( \mathcal{H} \subset C(X) \).
Let us find out the expression of $K(x, y)$. For any fixed $x \in X$, suppose that

$$K_x(y) = \sum_{k=1}^{\infty} b_k(x) \phi_k(y).$$

Then by the reproducing kernel property,

$$\phi_n(x) = < K_x, \phi_n >_H = < \sum_{k=1}^{\infty} b_k(x) \phi_k(y), \phi_n(y) >_H = b_n(x).$$

Therefore,

$$K(x, y) = \sum_{k=1}^{\infty} \phi_k(x) \phi_k(y). \quad (2.2)$$

This kernel is not continuous. To see this, let $x_n = y_n = \frac{2n + 1}{2n(n + 1)}$, $n \in \mathbb{Z}_+$. Then $K(x_n, y_n) = 1$ but $K(0, 0) = 0$. Consequently, $K(x, y)$ is not continuous at $(0, 0)$. By Proposition 2.2, the inclusion map $J$ is not compact. In fact, this is apparent by notice that $B_1 = \{ f : f \in \mathcal{H}, \| f \|_H \leq 1 \} \supset \{ \phi_n : n \in \mathbb{Z}_+ \}$ is not equi-continuity in $C(X)$.

In Sections 3 and 4 we will study the integral operator and consistency of learning algorithm where we are only interested in bounded functions. The following conclusion provide a characterization of when an RKHS contains only bounded functions.

**Proposition 2.3.** The following assertions are equivalent:

(a) Each $f \in \mathcal{H}_K$ is a bounded function.

(b) The feature map $\eta(x)$ is uniformly bounded, i.e.,

$$\sup_{x \in X} \| \eta(x) \|_K = \sup_{x \in X} \| K_x \|_K = \sup_{x \in X} \sqrt{K(x, x)} < \infty.$$ 

(c) Any orthonormal basis $\{ \varphi_\lambda \}_{\lambda \in \mathcal{I}}$ satisfies $\sup \{ \sum_{\lambda \in \mathcal{I}} \varphi_\lambda^2(x), x \in X \} < \infty$.

**Proof.** Suppose (a) holds, thus for any $f \in \mathcal{H}_K$

$$\sup_{x \in X} | < f, K_x >_K | = \sup_{x \in X} | f(x) | < \infty.$$
By the well known uniform bounded theorem (see e.g. [7]), we have \( \sup_{x \in X} \| K_x \|_K < \infty \), this proves (b).

On the other hand, it (b) holds, then for any \( g \in \mathcal{H}_K \),

\[
\sup_{x \in X} |g(x)| = \sup_{x \in X} | < g, K_x >_K | \leq \sup_{x \in X} \| K_x \|_K \| g \|_K < \infty.
\]

Hence (a) is true. This proves the equivalence between (a) and (b).

The equivalence between (b) and (c) is an immediate consequence of the following equation:

\[
\sum_{\lambda \in I} \varphi^2_\lambda(x) = \sum_{\lambda \in I} < K_x, \varphi_\lambda >^2_K = \| K_x \|^2_K = K(x, x).
\]

The proof is completed. \( \square \)

3 Properties of the integral Operator \( L_K \)

In the sequel, we use the notation \( \mathcal{H} \) for \( \mathcal{H}_K \) for simplicity. We need the following basic assumptions:

**Assumption 1.** The feather map \( \eta \) of the RKHS \( \mathcal{H} \) is uniformly bounded, i.e.,

\[
\mathcal{K} := \sup_{x \in X} \sqrt{K(x, x)} < \infty.
\]

**Assumption 2.** For any \( f \in \mathcal{H} \), \( f(x) \) is \( \rho_X \)-measurable.

By Proposition 2.3, Assumption 1 let us consider only bounded functions. Assumption 2 let us avoid consider the tricky measurement issues. In particular, it always holds if \( \mathcal{H} \subset C(X) \).

Under these two assumptions, we have that for every \( f \in \mathcal{H} \),

\[
\| f \|_{\rho_X}^2 = \int_X |f(t)|^2 d\rho_X = \int_X | < f, K_t >_K |^2 d\rho_X
\]

\[
\leq \int_X \| f \|^2_K \| K_t \|^2_K d\rho_X \leq \mathcal{K}^2 \| f \|^2_K.
\]

(3.1)

So \( \mathcal{H} \subset L^2_{\rho_X}(X) \). In particular, for each \( x \in X \), there is \( K_x \in L^2_{\rho_X}(X) \) and \( \| K_x \|^2_{\rho_X} \leq \mathcal{K}^2 \| K_x \|^2_K \leq \mathcal{K}^4 \). Thus for each \( f \in L^2_{\rho_X}(X) \) and \( x \in X \) we can define

\[
L_K(f)(x) = \int_X K(x, t)f(t)d\rho_X(t).
\]

It turns out that \( L_K \) is bounded linear operator from \( L^2_{\rho_X}(X) \) to itself. Moreover, it has the following properties.
Proposition 3.1. The following conclusions hold.

(1) For any \( f \in L^2_{\rho_X}(X) \), \( L_K(f) \in \mathcal{H} \) and \( \|L_K(f)\|_K \leq \mathcal{K}\|f\|_{\rho_X} \).

(2) For all \( f \in L^2_{\rho_X}(X) \) and \( g \in \mathcal{H} \),
\[
< L_K(f), g >_K = < f, g >_{\rho_X}.
\]

(3) \( L_K \) is a positive compact operator from \( L^2_{\rho_X}(X) \) to \( L^2_{\rho_X}(X) \).

(4) The restriction \( L_K|_{\mathcal{H}} \) of \( L_K \) on \( \mathcal{H} \) is a positive bounded linear operator from \( \mathcal{H} \) into itself. Moreover, for \( f \in \mathcal{H} \), \( L_K(f) = 0 \) in \( \mathcal{H} \) if and only if \( f = 0 \) almost everywhere relative to the probability measure \( \rho_X \).

Let \( \mathcal{H}_0 = \{ f \in \mathcal{H} : f = 0 \ a.e. \ \rho_X \} \) and \( \mathcal{H}_1 = \mathcal{H}_0^\perp \) be the orthogonal complement of \( \mathcal{H}_0 \) in \( \mathcal{H} \). From Proposition 3.1, we know that \( \mathcal{H}_0 = \ker(L_K|_{\mathcal{H}}) \) and hence is closed. Now by \( \overline{\mathcal{H}_1} \) we denote the closure of \( \mathcal{H}_1 \) in \( L^2_{\rho_X}(X) \). Obviously, \( \overline{\mathcal{H}_1} = \mathcal{H} \).

Proposition 3.2. (1) The restriction \( L_K|_{\overline{\mathcal{H}_1}} \) of \( L_K \) on \( \overline{\mathcal{H}_1} \) is an injective, positive and compact operator from \( (\overline{\mathcal{H}_1}, \| \cdot \|_{\rho_X}) \) into \( (\overline{\mathcal{H}_1}, \| \cdot \|_{\rho_X}) \).

(2) Let \( \{\lambda_s\}_{s=1}^\infty \) be the set of nonzero eigenvalues of \( L_K|_{\overline{\mathcal{H}_1}} \) satisfying \( \lambda_1 \geq \lambda_2 \geq \ldots \) and \( \{\phi_s\}_{s \geq 1} \) the corresponding unit eigenvectors. Then \( \{\sqrt{\lambda_s}\phi_s\}_{s \geq 1} \) is an orthonormal basis of \( \mathcal{H}_1 \).

Proposition 3.3. \( L_K^{\frac{1}{2}} f \in \mathcal{H}_1 \) for any \( f \in L^2_{\rho_X}(X) \), and \( L_K^{\frac{1}{2}} \) is an isometric isomorphism from \( (\overline{\mathcal{H}_1}, \| \cdot \|_{\rho_X}) \) onto \( (\mathcal{H}_1, \| \cdot \|_K) \).

The proofs of Propositions 3.1, 3.2 and 3.3 can be found in Appendix.

4 Regularized least-square regression

In this section, we study the consistency of the least-square regularized regression in a general RKHS. Except for Assumptions 1 and 2 in last section, we need an
additional assumption.

**Assumption 3.** $f_\rho \in L^2_{\rho_X}(X)$ and $\sigma^2 = \int_Z (y - f_\rho(x))^2 dp$, the variance of $\rho$, is finite.

This assumption is natural from a practical point of view. In fact, $y$ is usually uniform bounded in practice which implies the finiteness of the variance.

In order to investigate how $f_{\gamma, \gamma}$ approximates $f_\rho$ as $\gamma \to 0$ and $m \to \infty$, we need to estimate the sample error and approximation error as defined in Section 1.

Let us first estimate the approximation error $\|f_\gamma - f_\rho\|_{\rho_X}$. We will also study the approximation in terms of $\mathcal{H}_K$ norm, i.e., upper bound for $\|f_\gamma - f_\rho\|_K$ when $f_\rho \in \mathcal{H}$.

**Proposition 4.1.** Suppose that $L^{-r}_K f_\rho \in L^2_{\rho_X}(X)$. Then if $0 \leq r \leq 1$,

$$\|f_\gamma - f_\rho\|_{\rho_X} \leq \gamma^r \|L^{-r}_K f_\rho\|_{\rho_X}$$

and, if $\frac{1}{2} \leq r \leq \frac{3}{2}$

$$\|f_\gamma - f_\rho\|_K \leq \gamma^{r-\frac{1}{2}} \|L^{-r}_K f_\rho\|_{\rho_X}.$$  

**Proof.** By setting the functional derivative of (1.2) to 0, we obtain $f_\gamma = (L_K + \gamma I)^{-1} L_K f_\rho$. Then

$$f_\gamma - f_\rho = (L_K + \gamma I)^{-1} L_K f_\rho - f_\rho = -\gamma(L_K + \gamma I)^{-1} f_\rho.$$

We have

$$\|f_\gamma - f_\rho\|_{\rho_X} = \gamma \|(L_K + \gamma I)^{-1} L_K^r \|L^{-r}_K f_\rho\|_{\rho_X} \leq \gamma \|(L_K + \gamma I)^{-1} L_K^r \|L^{-r}_K f_\rho\|_{\rho_X}.$$  

In order to estimate $\|(L_K + \gamma I)^{-1} L_K^r \|$, we investigate the function $\varphi(t) = \frac{r}{\gamma + t}$, $0 \leq r \leq 1$. If $0 < r < 1$, it is easy to check its maximum is obtained at $t_0 = \frac{\gamma}{1-r}$. So

$$\max_{t \geq 0} \varphi(t) = \gamma^{r-1} r (1-r)^{1-r} \leq \gamma^{r-1}.$$  

Trivial argument shows this is also true for $r = 0$ and $r = 1$. By the Gelfand Theorem [6], $\|(L_K + \gamma I)^{-1} L_K^r \| \leq \gamma^{r-1}$. This in connection with (4.3) proves (4.1).

To prove (4.2), suppose that $\frac{1}{2} \leq r \leq \frac{3}{2}$. Then $f_\rho \in L^2_K(L^2_{\rho_X}(X)) \subset L^\frac{1}{2}(L^2_{\rho_X}(X)) = \mathcal{H}_1$ and

$$\|f_\gamma - f_\rho\|_K^2 = \gamma^2 < (L_K + \gamma I)^{-1} f_\rho, (L_K + \gamma I)^{-1} f_\rho>_K$$

$$= \gamma^2 < L^\frac{1}{2}_K(L_K + \gamma I)^{-1} L^\frac{1}{2}_K f_\rho, L^\frac{1}{2}_K(L_K + \gamma I)^{-1} L^\frac{1}{2}_K f_\rho>_K.$$  

(4.4)
Let \( g = (L_K + \gamma I)^{-1}L_K^{\frac{1}{2}}f_\rho \) and \( g_p \) be the projection of \( g \) on \( \overline{H}_1 \). Then there is a sequence \( \{g_n\} \) in \( \mathcal{H}_1 \) which converges to \( g_p \) in the \( L_{\rho X}^2(X) \) norm. By proposition 3.3, that \( L_K^{\frac{1}{2}} \) is an isometric isomorphism of \( \overline{H}_1 \) onto \( \mathcal{H}_1 \). So \( L_K^{\frac{1}{2}}g_n \) converges to \( L_K^{\frac{1}{2}}g_p \) in \( \mathcal{H} \). Hence

\[
\|f_\gamma - f_\rho\|_K^2 = \gamma^2 < L_K^{\frac{1}{2}}g_p, L_K^{\frac{1}{2}}g_p >_K
\]
\[
= \lim_{n \to \infty} \gamma^2 < L_K^{\frac{1}{2}}g_n, L_K^{\frac{1}{2}}g_n >_K
\]
\[
= \lim_{n \to \infty} \gamma^2 < g_n, g_n >_{\rho X}
\]
\[
= \gamma^2 < g_p, g_p >_{\rho X}
\]
\[
\leq \gamma^2 < (L_K + \gamma I)^{-1}L_K^{\frac{1}{2}}f_\rho, (L_K + \gamma I)^{-1}L_K^{\frac{1}{2}}f_\rho >_{\rho X}
\]
\[
= \gamma^2 < (L_K + \gamma I)^{-1}L_K^{\frac{1}{2}}L_K^{r}f_\rho, (L_K + \gamma I)^{-1}L_K^{r}L_K^{\frac{1}{2}}f_\rho >_{\rho X}
\]
\[
\leq \gamma^2 \|(L_K + \gamma I)^{-1}L_K^{\frac{1}{2}}\|_2 \cdot \|L_K^{r}f_\rho\|^2_{\rho X}
\]
\[
\leq \gamma^{2r-1} \|L_K^{r}f_\rho\|^2_{\rho X}. \quad (4.5)
\]

In the last step, we have used the Gelfanfd Theorem and the fact that \( \max_{t \geq 0} \psi(t) \leq \gamma^{r-\frac{3}{2}} \) for the function \( \psi(t) = \frac{e^{-t}}{\gamma + t} \) with \( \frac{1}{2} \leq r \leq \frac{3}{2} \). \( \square \)

Notice that the assumptions are similar as in [10]. However, it has slight different meaning. As for a general RKHS, the assumption \( r \geq 1/2 \) implies \( f_\rho \in \mathcal{H}_1 \) but not only \( \mathcal{H} \). In fact, assume only \( f_\rho \in \mathcal{H} \), by Proposition 3.2, we can write \( f_\rho = \sum_{s \geq 1} \alpha_s \sqrt{\lambda_s} \phi_s + f_0 \) where \( f_0 \) is the projection of \( f_\rho \) on \( \mathcal{H}_0 \) and \( \sum_{s \geq 1} \alpha_s^2 < \infty \). Then

\[
\|f_\rho - f_\gamma\|_K^2 = \|\gamma (\gamma I + L_K)^{-1}f_\rho\|_K^2 = \| \sum_{s \geq 1} \alpha_s \sqrt{\lambda_s} \frac{\gamma}{\gamma + \lambda_s} \phi_s + f_0 \|^2_2 = \sum_{s \geq 1} \alpha_s^2 \frac{\gamma^2}{(\gamma + \lambda_s)^2} + \|f_0\|^2_2.
\]

Thus \( \|f_\rho - f_\gamma\|_K^2 \) tends to \( \|f_0\|^2_2 \) (maybe not 0) as \( \gamma \to 0 \). This means that \( f_\gamma \) approaches \( f_\rho \) in \( \mathcal{H} \)-norm only when \( f_\rho \in \mathcal{H}_1 \).

We turn to estimate the sample error \( f_{x,\gamma} - f_\gamma \). We use the following expression that has been obtained in [10, Theorem 1]

\[
f_{x,\gamma} - f_\gamma = \left( \frac{1}{m} S_x^T S_x + \gamma I \right)^{-1} \{ \frac{1}{m} \sum_{i=1}^m (y_i - f_\gamma(x_i)) K_{x_i} - L_K (f_\rho - f_\gamma) \}.
\]

Then

\[
\|f_{x,\gamma} - f_\gamma\|_K \leq \frac{1}{\gamma} \left( \frac{1}{m} \sum_{i=1}^m (y_i - f_\gamma(x_i)) K_{x_i} - L_K (f_\rho - f_\gamma) \right)\|_K
\]
\[
= \frac{1}{\gamma} \left( \frac{1}{m} \sum_{i=1}^m \xi(z_i) - L_K (f_\rho - f_\gamma) \right)\|_K,
\]

(4.7)
here \( \xi(z) = (y - f_\gamma(x))K_x \) is a random variable from \( Z \) to \( \mathcal{H} \). In our case, we have not assumed that \( Z \) is randomly drawn according to \( \rho \) satisfying \( y \) being uniform bounded almost surely, so we apply The Chebyshev’s inequality for random variable of vector values.

**Lemma 4.2.** Let \( \theta \) be a random variable with values in Hilbert space \( \mathcal{H} \), has finite mean \( \mu \) and variance \( \sigma^2_\theta := \int_Z \| \theta(z) - \mu \|^2 d\rho \), where \( (Z, \rho) \) is the probability space. Then, for any \( \varepsilon > 0 \),

\[
\text{Prob}\{ \| \frac{1}{m} \sum_{i=1}^{m} \theta(z_i) - \mu \| \geq \varepsilon \} \leq \frac{\sigma^2_\theta}{m\varepsilon^2} \quad (4.8)
\]

We have not found a reference for this inequality though it probably has appeared in the literature. For completeness we give a proof in the appendix.

It is easy to check that \( E(\xi) = L_K(f_\rho - f_\gamma) \) and, if \( L_K^{-r}f_\rho \in L^2_{\rho_X}(X) \),

\[
\sigma^2_\xi = \int_Z \| (y - f_\gamma(x))K_x - L_K(f_\rho - f_\gamma) \|^2 d\rho \\
= \int_Z (y - f_\gamma(x))^2K(x,x)d\rho - <f_\rho - f_\gamma, L_K(f_\rho - f_\gamma)>_{\rho_X} \\
\leq \int_Z (y - f_\rho(x))^2K(x,x)d\rho + \int_X (f_\rho(x) - f_\gamma(x))^2K(x,x)d\rho_X \\
\leq \mathcal{K}^2(\| f_\rho - f_\gamma \|^2_{\rho_X} + \sigma^2) \\
\leq \mathcal{K}^2(\gamma^{2r}\| L_K^{-r}f_\rho \|^2_{\rho_X} + \sigma^2) \quad (4.9)
\]

The following result is an immediate conclusion of application of Lemma 4.2 to the random variable \( \xi \).

**Proposition 4.3.** If \( L_K^{-r}f_\rho \in L^2_{\rho_X}(X) \) for some \( r > 0 \) and \( \gamma > 0 \), then for every \( \varepsilon > 0 \)

\[
\text{Prob}\{ \| f_{x,\gamma} - f_\gamma \|_{\mathcal{K}} \geq \varepsilon \} \leq \frac{\mathcal{K}^2(\gamma^{2r}\| L_K^{-r}f_\rho \|^2_{\rho_X} + \sigma^2)}{m\gamma^2\varepsilon^2} \quad (4.10)
\]

Combining Proposition 4.1 and Proposition 4.3, it is not difficult to prove the following conclusion.
Proposition 4.4. If $L_K^{-r}f_\rho \in L_{\rho X}^2(X)$ and $\gamma > 0$, then, for every $0 < \delta < 1$, with probability $1 - \delta$, there holds

$$\|f_{z,\gamma} - f_\rho\|_{\rho X} \leq \gamma^r \|L_K^{-r}f_\rho\|_{\rho X} + \frac{KM^*}{\gamma \sqrt{m \delta}}$$  \hspace{1cm} (4.11)

when $0 < r \leq 1$, and

$$\|f_{z,\gamma} - f_\rho\|_K \leq \gamma^{-\frac{1}{2}} \|L_K^{-r}f_\rho\|_{\rho X} + \frac{M^*}{\gamma \sqrt{m \delta}}$$  \hspace{1cm} (4.12)

when $\frac{1}{2} < r \leq \frac{3}{2}$, where $M^*$ is a constant defined by

$$M^* = K \sqrt{\gamma^2 \|L_K^{-r}f_\rho\|_{\rho X}^2 + \sigma^2}$$

The following corollary gives the approximation rate of $f_{z,\gamma}$ converging to $f_\rho$. Theorem 1.1 is just a restate of its first conclusion.

Corollary 4.5. If $L_K^{-r}f_\rho \in L_{\rho X}^2(X)$ with $0 < \gamma \leq \left\{\frac{\sqrt{3\sigma}}{\|L_K^{-r}f_\rho\|_{\rho X}}\right\}^\frac{1}{r}$, then for every $0 < \delta < 1$, with probability $1 - \delta$, there hold

(i) if $0 < r \leq 1$,

$$\|f_{z,\gamma} - f_\rho\|_{\rho X} \leq \left\{\|L_K^{-r}f_\rho\|_{\rho X} + \frac{2K^2\sigma}{\sqrt{\delta}}\right\} m^{-\frac{r}{2(r+1)}}, \quad \gamma = m^{-\frac{1}{2(r+1)}}; \hspace{1cm} (4.13)$$

(ii) if $\frac{1}{2} < r \leq \frac{3}{2}$,

$$\|f_{z,\gamma} - f_\rho\|_K \leq \left\{\frac{2K\sigma}{\sqrt{\delta}} + \|L_K^{-r}f_\rho\|_{\rho X}\right\} m^{-\frac{r-1}{4(r+1)}}, \quad \gamma = m^{-\frac{1}{2(r+1)}}. \hspace{1cm} (4.14)$$

Proof. Since $0 < \gamma \leq \left\{\frac{\sqrt{3\sigma}}{\|L_K^{-r}f_\rho\|_{\rho X}}\right\}^\frac{1}{r}$, then $M^* = K \sqrt{\gamma^{2r} \|L_K^{-r}f_\rho\|_{\rho X}^2 + \sigma^2} \leq 2K\sigma$, by Proposition (4.12), for every $0 < \delta < 1$, with probability $1 - \delta$, there holds,

$$\|f_{z,\gamma} - f_\rho\|_K \leq \gamma^{-\frac{1}{2}} \|L_K^{-r}f_\rho\|_{\rho X} + \frac{2K\sigma}{\gamma \sqrt{m \delta}}$$

Let $\gamma = m^{-\frac{1}{2(r+1)}}$, thus (4.14) holds. In the same way, we can prove (4.13), so Proposition 4.5 holds. \qed

Acknowledgement. This work is supported by the Research Grants Council of the Hong Kong [Project No. CityU 103405]. The author would like to thank Professor D.X.Zhou for his useful suggestions and comments.
Appendix: Proofs

Proof of Proposition 3.1. (1) Fixed $f \in L^2_{\rho_X}(X)$, the equation $T_f(g) = \langle g, f >_{\rho_X}$, $g \in \mathcal{H}$, defines a linear functional $T_f$ on $\mathcal{H}$. Moreover, by the equation (3.1),

$$|T_f(g)| = |\langle f, g >_{\rho_X}| \leq K\|f\|_{\rho_X}\|g\|_K.$$ 

So $T_f$ is a bounded linear functional on $\mathcal{H}$. Consequently, there exists $P_f \in \mathcal{H}$ such that

$$T_f(g) = \langle P_f, g >_K, \quad g \in \mathcal{H}.$$ 

From

$$P_f(x) = \langle P_f, K_x >_K = T_f(K_x) = \int_X K(x,y)f(y)d\rho_X(y),$$

we see that $L_K(f) \in \mathcal{H}$ and

$$\|L_K(f)\|_K = \|P_f\|_K = \|T_f\| \leq K \cdot \|f\|_{\rho_X}.$$ 

(2) \forall f \in L^2_{\rho_X}(X), g \in \mathcal{H},

$$<f, g >_{\rho_X} = T_f(g) = \langle P_f, g >_K = \langle L_K(f), g >_K.$$ 

(3) We have $K(x,y) \in L^2_{\rho_X}(X \times X)$ by noting that

$$\int_X \int_X |K(x,y)|^2d\rho_X(y)d\rho_X(x) = \int_X \|K_x\|^2_{\rho_X}d\rho_X(x) \leq \int_X K^2\|K_x\|^2_Kd\rho_X(x) \leq K^4.$$ 

So $L_K$ is compact (see e.g. [7, page 216]). To show the positivity, let $f \in L^2_{\rho_X}(X)$. Then by the parts (1) and (2), $< L_K(f), f >_{\rho_X} = < L_K(f), L_K(f) >_K \geq 0$.

(4) The first statement immediately follows from the part (1) and the equation (3.1). The second follows from the parts (1) and (2).

The proof is complete.

Proof of Proposition 3.2. For $f \in \overline{\mathcal{H}_1}$, we have $L_K(f) \in \mathcal{H} \subseteq \mathcal{H} = \overline{\mathcal{H}_1}$. So $L_K|_{\overline{\mathcal{H}_1}}$ is a linear map from $\overline{\mathcal{H}_1}$ into itself. Now the positivity and compactness of $L_K|_{\overline{\mathcal{H}_1}}$ follows from Proposition 3.1(3). To see the injectivity, let $g \in \overline{\mathcal{H}_1}$ and suppose $L_Kg = 0$ in $L^2_{\rho_X}(X)$. Then $L_Kg \in \mathcal{H}_0$. Take a sequence $g_n \in \mathcal{H}_1$ which converges to $g$ in $L^2_{\rho_X}$. Then

$$< g, g >_{\rho_X} = \lim_n < g, g_n >_{\rho_X} = \lim_n < L_kg, g_n >_K = 0.$$ 

So $g = 0$ a.e. $\rho_X$. This proves (1).
Since $L_K\phi_s = \lambda_s \phi_s$, it follows that $\phi_s \in \mathcal{H}$. To see further that $\phi_s \in \mathcal{H}_1$, we let $g \in \mathcal{H}_0$. Then
\[
< \lambda_s \phi_s, g >_K = < L_k \phi_s, g >_K = < \phi_s, g >_{\rho_X} = 0.
\]
So $\phi_s \in \mathcal{H}^1_0 = \mathcal{H}_1$. The remainder of the proof is routine. □

**Proof of Proposition 3.3.** We prove the conclusion in three steps.

**Step 1.** By Proposition 3.1(4), the range of $L_K|_{\mathcal{H}_1}$ is dense in $\mathcal{H}_1$. let $f \in \mathcal{H}_1$, there exists $f_n \in \mathcal{H}_1$, $n = 1, 2, \cdots$; s.t. $L_K(f_n) \to f$ in $\mathcal{H}_1$. By (3.1), this means $L_K f_n \to f$ and $L_K(L_K^{-1} f_n) \to L_K^1 f$ in $L^2_{\rho_X}(X)$.

\[
\left\| L_K \left( L_K^{1} f_n - L_K^{1} f_m \right) \right\|^2_K = < L_K^{1/2} (f_n - f_m), L_K^{1/2} (f_n - f_m) >_{\rho_X} = \left\| L_K (f_n - f_m) \right\|^2_{\rho_X}. \tag{A.1}
\]

Then $L_K(L_K^{1/2} f_n)$, $n = 1, 2, \ldots$ is a Cauchy sequence in $\mathcal{H}_1$. So $L_K(L_K^{1/2} f_n) \to g$ in $\mathcal{H}_1$ and in $L^2_{\rho_X}(X)$. Therefore $L_K^{1/2}(f) = g$ in $L^2_{\rho_X}(X)$ and $L_K^{1/2}(f) \in \mathcal{H}_1$.

**Step 2.** Now $(L_K|_{\mathcal{H}_1})^{1/2}$ and $L_K^{1/2}|_{\mathcal{H}_1}$ both are operators from $\mathcal{H}_1$ to $\mathcal{H}_1$, in this part, we prove $(L_K|_{\mathcal{H}_1})^{1/2} = L_K^{1/2}|_{\mathcal{H}_1}$. By Proposition 3.1 and the result of Step 1, $L_K^{1/2}|_{\mathcal{H}_1}$ is a positive bounded operator in $L^2_{\rho_X}$.

Consider the operator $(L_K|_{\mathcal{H}_1})^{1/2}$. For any $f \in \mathcal{H}_1$
\[
\left\| (L_K|_{\mathcal{H}_1})^{1/2} f \right\|^2_{\rho_X} \leq \mathcal{K}^2 \left\| (L_K|_{\mathcal{H}_1})^{1/2} f \right\|^2_K = \mathcal{K}^2 < L_K|_{\mathcal{H}_1} f, f >_K = \mathcal{K}^2 \left\| f \right\|^2_{\rho_X}.
\]

and
\[
< (L_K|_{\mathcal{H}_1})^{1/2} f, f >_{\rho_X} = < (L_K|_{\mathcal{H}_1})^{1/2} f, L_K f >_K = < (L_K|_{\mathcal{H}_1})^{1/2} f, (L_K|_{\mathcal{H}_1})^{1/2} ((L_K|_{\mathcal{H}_1})^{1/2} f) >_K \geq 0. \tag{A.2}
\]

So $(L_K|_{\mathcal{H}_1})^{1/2}$ is also a positive bounded operator in $L^2_{\rho_X}$.

Moreover, the two operators both are the square root of $L_K$. By the theory of Functional Calculus (see e.g. [6]), $(L_K|_{\mathcal{H}_1})^{1/2} = L_K^{1/2}|_{\mathcal{H}_1}$.

**Step 3.** Let $f \in L^2_{\rho_X}(X)$, without loss of generality, we can suppose $f \in \overline{\mathcal{H}_1} = \overline{\mathcal{H}}$, there is $f_n \in \mathcal{H}_1$, $n = 1, 2, \cdots$; s.t. $f_n \to f$ in $L^2_{\rho_X}(X)$.

\[
\left\| L_K^{1/2} (f_n - f_m) \right\|^2_K = \left\| (L_K|_{\mathcal{H}_1})^{1/2} (f_n - f_m) \right\|^2_K
= < (L_K|_{\mathcal{H}_1})^{1/2} (f_n - f_m), (L_K|_{\mathcal{H}_1})^{1/2} (f_n - f_m) >_K
= < L_K (f_n - f_m), (f_n - f_m) >_K = \left\| f_n - f_m \right\|^2_{\rho_X}.
\]
Then $L^1_K(f_n)$ is a Cauchy sequence in $\mathcal{H}$. Since $L^1_K(f_n) \to L^1_K(f)$ in $L^2_\rho(X)$, there hold $L^1_K(f) \in \mathcal{H}_1$ and
\[
\|L^1_K(f)\|_K = \lim_{n \to \infty} \|L^1_K(f_n)\|_K = \lim_{n \to \infty} \|f_n\|_\rho_X = \|f\|_\rho_X.
\] (A.3)

By the fact $R(L^1_K|_{\mathcal{H}_1}) = R((L_K|_{\mathcal{H}_1})^{1/2})$ is dense in $\mathcal{H}_1$, $L^1_K$ is an isometric isomorphism from $\overline{\mathcal{H}_1}$ onto $\mathcal{H}_1$. □

**Proof of Lemma 4.2.** By the Chebyshev’s inequality, we have,
\[
\text{Prob}\{\|\frac{1}{m} \sum_{i=1}^{m} \theta(z_i) - \mu\|_{\mathcal{H}} \geq \varepsilon\} \leq \frac{1}{\varepsilon^2} E\{\|\frac{1}{m} \sum_{i=1}^{m} \theta(z_i) - \mu\|_{\mathcal{H}}^2\}.
\]
But
\[
E\{\|\frac{1}{m} \sum_{i=1}^{m} \theta(z_i) - \mu\|_{\mathcal{H}}^2\} = \frac{1}{m^2} \sum_{i,j=1}^{m} E < \theta(z_i) - \mu, \theta(z_j) - \mu >_{\mathcal{H}}
\]
\[
= \frac{1}{m^2} \sum_{i,j=1}^{m} \delta_{ij}\sigma^2_\theta = \frac{1}{m}\sigma^2_\theta.
\]
Therefore, the inequality (4.8) is proved. □

**References**


