ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Complexity 21 (2005) 337–349

Journal of COMPLEXITY

www.elsevier.com/locate/jco

Mercer theorem for RKHS on noncompact sets

Hongwei Sun

School of Science, Jinan University, Jinan 250022, Peoples's Republic of China

Received 30 September 2003; accepted 22 September 2004

Abstract

Reproducing kernel Hilbert spaces are an important family of function spaces and play useful roles in various branches of analysis and applications including the kernel machine learning. When the domain of definition is compact, they can be characterized as the image of the square root of an integral operator, by means of the Mercer theorem. The purpose of this paper is to extend the Mercer theorem to noncompact domains, and to establish a functional analysis characterization of the reproducing kernel Hilbert spaces on general domains.

Keywords: Mercer kernel; Reproducing kernel Hilbert spaces; Nondegenerate Borel measure; Positive semidefiniteness

1. Introduction

Let (X, d) be a metric space and $K : X \times X \to \mathbf{R}$ be continuous and symmetric. We say that K is a Mercer kernel if it is positive semidefinite, i.e., for any finite set of points $\{x_1, ..., x_m\} \subset X$ and $\{c_1, ..., c_m\} \subset \mathbf{R}$, there holds $\sum_{i,j=1}^{m} c_i c_j K(x_i, x_j) \geqslant 0$.

The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with the Mercer kernel K is defined [1] to be the closure of span$\{K_x := K(x, \cdot) : x \in X\}$ with the inner product given by

$$\langle f, g\rangle_K = \sum_{i=1}^{n}\sum_{j=1}^{m} c_i d_j K(x_i, y_j)$$

E-mail address: Shw_yb@sina.com.

0885-064X/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.jco.2004.09.002

for

$$f = \sum_{i=1}^{n} c_i K_{x_i}, \quad g = \sum_{j=1}^{m} d_j K_{y_j}.$$

The reproducing kernel property takes the form:

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

This property in connection with the continuity of K tells us that $\mathcal{H}_K$ consists of continuous functions on X, that is, $\mathcal{H}_K \subset C(X)$, the space of continuous functions on X.

The reproducing kernel property (1.1) and the Hilbert space structure make the RKHS very applicable in many fields. For example, in kernel matching learning, one often takes a RKHS $\mathcal{H}_K$ to be a hypothesis space [5,2,9] and investigates the learning of a function in $\mathcal{H}_K$ from a set of given samples $\mathbf{z} = (x_i, y_j)_{i=1}^m \subset X \times \mathbf{R}$ by minimizing the empirical error:

$$\inf_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \| f \|_K^2 \right\}. \tag{1.2}$$

Here $\frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$ is the empirical error and $\lambda \| f \|_K^2$ is a penalty term with $\lambda > 0$ being a penalty parameter. For the approximation of the above minimizer to the desired learned function called target function, see [8,11,12,13].

As $\mathcal{H}_K$ is a Hilbert space, the orthogonal projection of an arbitrary function $f \in \mathcal{H}_K$ onto the finite-dimensional space, $\operatorname{span}\{K_{x_i}\}_{i=1}^m$, denoted as $P(f)$, satisfies $\langle f - P(f), K_{x_i} \rangle_K = 0$ for each $1 \leqslant i \leqslant m$. Then the reproducing kernel property (1.1) implies:

$$P(f)(x_i) = \langle P(f), K_{x_i} \rangle_K = \langle f, K_{x_i} \rangle_K = f(x_i).$$

Therefore if f minimizes (1.2), then $P(f)$ also does, hence f must be equal to $P(f)$, i.e., $f = \sum_{i=1}^m c_i K_{x_i} \in \operatorname{span}\{K_{x_i}\}_{i=1}^m$ and the minimization problem (1.2) can be solved by solving a linear system

$$\left[\left(K(x_i, x_j) \right)_{i,j=1}^m + m\lambda I \right] (c_j)_{j=1}^m = (y_i)_{i=1}^m.$$

See [9,10].

When the domain X is compact, the Hilbert space structure of the RKHS $\mathcal{H}_K$ is well understood from a functional analysis point of view, by means of the Mercer theorem.

To see this, let μ be a nondegenerate Borel measure on (X, d). Then the integral operator L_K on $L^2(X, \mu)$ defined by

$$L_K f(x) = \int_X K(x, y) f(y) \, d\mu(y) \tag{1.3}$$

is compact, positive and symmetric. It has at most countably many positive eigenvalues $\{\lambda_i\}_{i=1}^\infty$ and corresponding orthonormal eigenfunctions $\{\phi_i\}_{i=1}^\infty$. The Mercer theorem [7]

asserts that:

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y),$$

where the series converges absolutely and uniformly on $X \times X$. Here one needs to assume that μ is nondegenerate in the sense that $\mu(S) > 0$ for any nonempty open set $S \subset X$, i.e., the complement of any set of measure zero is dense in X. For a simple proof of the Mercer theorem, when $X = [0, 1]$ and $d\mu = dx$, see [6]. The same proof works for general nondegenerate measures μ, as pointed out by Cucker and Smale [2,3].

An interesting consequence of the Mercer theorem is that $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ forms an orthonormal basis of $\mathcal{H}_K$. This was proved in [2,4].

Recall that the definition of the RKHS does not require the compactness of the domain X. It is a natural question to understand the Hilbert space structure from a functional analysis point of view for the general domain. To this end, we ask whether the Mercer theorem holds, and need to find a nice orthonormal basis if the Mercer theorem is valid. The purpose of this paper is to answer these questions when X is not necessarily compact.

Finally, we mention that attempts have been made in the community of learning theory to understand the learning of functions on unbounded domains. It is expected that our results would provide some theoretical backgrounds for this direction.

2. Mercer theorem on noncompact domains

In this section, we establish a Mercer theorem on a general domain, and discuss the Hilbert space structure of the RKHS $\mathcal{H}_K$ under some assumptions. In the next section we show how to check these assumptions.

Let (X, d) be a metric space, and μ be a nondegenerate Borel measure on X, that means for every open set $U \subset X$, $\mu(U) > 0$. Assume a (sequence) compactness structure for X: $X = \bigcup_{n=1}^{+\infty} X_n$, where $X_1 \subset X_2 \subset \cdots \subset X_n \subset \cdots$, and each X_n is compact with finite measure: $\mu(X_n) < +\infty$. Moreover, any compact subset of X is contained in X_i for some i.

Let $K : X \times X \to \mathbf{R}$ be a Mercer kernel. Define the integral operator L_K on $L^2(X, \mu)$ as

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

Concerning the kernel K and the measure μ we assume the following:

Assumption 1. $K_x \in L^2(X, \mu)$ for every $x \in X$.

Assumption 2. L_K is a bounded and positive operator on $L^2(X, \mu)$, and for every $g \in L^2(X, \mu)$, $L_K(g) \in C(X)$.

Assumption 3. L_K has at most countably many positive eigenvalue $\{\lambda_i\}_{i=1}^{\infty}$, and corresponding orthonormal eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$.

The above assumptions in connection with the reproducing property of the RKHS yield the following.

Lemma 1. *If $f \in C(X)$ is supported on X_n for some $n \in \mathbf{N}$, then $L_K(f) \in \mathcal{H}_K$ and for $h \in \mathcal{H}_K$, holds*

$$\langle L_K(f), h\rangle_K = \int_X f(x)h(x)\,d\mu(x). \tag{2.1}$$

Proof. Since f is supported on X_n, we have

$$L_k(f)(x) = \int_X K(x, y)f(y)\,d\mu(y) = \int_{X_n} K(x, y)f(y)\,d\mu(y).$$

Take a sequence $\{\delta_k > 0\}_{k\in\mathbf{N}}$ such that $\lim_{k\to\infty}\delta_k = 0$. For each k, the compactness of X_n enables us to partition X_n into subsets $\{X_{k,i}\}_{i=1}^{m_k}$ such that $X_{k,i} \cap X_{k,j} = \emptyset$ for $i \neq j$, $\bigcup_{i=1}^{m_k} X_{k,i} = X_n$, and the diameter of each $X_{k,i}$ is at most δ_k. This can be obtained by taking a finite subcovering of the open balls with radius δ_k centered at points in X_n.

Choose a set of points $\{y_{k,i}\}_{i=1}^{m_k}$ such that $y_{k,i} \in X_{k,i}$. Then for each function $g \in C(X_n)$, there holds

$$\lim_{k\to\infty}\sum_{i=1}^{m_k} g(y_{k,i})\mu(X_{k,i}) = \int_{X_n} g(y)\,d\mu(y).$$

In fact, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that $|g(x) - g(y)| < \varepsilon$ whenever $d(x, y) \leqslant \delta$. When $\delta_k \leqslant \delta$, we have

$$\begin{aligned}
&\left|\sum_{i=1}^{m_k} g(y_{k,i})\mu(X_{k,i}) - \int_{X_n} g(y)\,d\mu(y)\right| \\
&\quad= \left|\sum_{i=1}^{m_k}\int_{X_{k,i}} g(y_{k,i}) - g(y)\,d\mu(y)\right| \\
&\quad\leqslant \sum_{i=1}^{m_k}\varepsilon\mu(X_{k,i}) = \varepsilon\mu(X_n).
\end{aligned}$$

It follows that

$$L_K(f)(x) = \lim_{k\to+\infty}\sum_{j=1}^{m_k} f(y_{k,j})\mu(X_{k,j})K(x, y_{k,j}), \quad \forall x \in X. \tag{2.2}$$

In the same way, we have

$$\lim_{s,t\to+\infty}\sum_{i=1}^{m_s}\sum_{j=1}^{m_t} f(y_{s,i})f(y_{t,j})\mu(X_{s,i})\mu(X_{t,j})K(y_{s,i},\ y_{t,j})$$

$$= \int_{X_n \times X_n} f(x)K(x, y)f(y)\,d\mu(x)\,d\mu(y)$$
$$= \int_{X \times X} f(x)K(x, y)f(y)\,d\mu(x)\,d\mu(y).$$

Let $\Phi_k(x) = \sum_{j=1}^{m_k} f(y_{k,j})\mu(X_{k,j})K(x, y_{k,j})$. Then $\Phi_k \in \mathcal{H}_K$. We have

$$\|\Phi_s - \Phi_t\|_K^2 = \langle \Phi_s, \Phi_s\rangle_K - 2\langle \Phi_s, \Phi_t\rangle_K + \langle \Phi_t, \Phi_t\rangle_K. \tag{2.3}$$

Here

$$\langle \Phi_s, \Phi_t\rangle_K = \sum_{i=1}^{m_s}\sum_{j=1}^{m_t} f(y_{s,i})f(y_{t,j})\mu(X_{s,i})\mu(X_{t,j})K(y_{s,i}, y_{t,j})$$

which tends to $\int_{X\times X} f(x)K(x, y)f(y)\,d\mu(x)\,d\mu(y)$ as $s, t \to +\infty$. Also,

$$\langle \Phi_s, \Phi_s\rangle_K \to \int_{X\times X} f(x)K(x, y)f(y)\,d\mu(x)\,d\mu(y)$$

as $s \to +\infty$. So $\{\Phi_k\}$ is a Cauchy sequence in $\mathcal{H}_K$ and has a limit $\Phi \in \mathcal{H}_K$. By (2.2), for each $x \in X$, $\lim_{k\to\infty} \Phi_k(x) = L_K(f)(x)$. Therefore $L_K(f) = \Phi \in \mathcal{H}_K$.

The function $h \in \mathcal{H}_K$ is continuous on X_n for each $n \in \mathbf{N}$. Since $\lim_{k\to\infty} \Phi_k = L_K(f)$ in $\mathcal{H}_K$, we have

$$\langle L_K(f), h\rangle_K = \lim_{k\to+\infty}\langle \Phi_k, h\rangle_K = \lim_{k\to+\infty}\sum_{i=1}^{m_k} f(y_{k,i})\mu(X_{k,i})h(y_{k,i})$$

which equals $\int_{X_n} f(x)h(x)\,d\mu(x) = \int_X f(x)h(x)\,d\mu(x)$. This proves Lemma 1. □

Define

$$C_B(X) = \{f \in C(X) : f \text{ is supported on } X_n \text{ for some } n\}.$$

It is easy to see that $C_B(X) \subset L^2(X, \mu)$ and $C_B(X)$ is dense in $L^2(X, \mu)$.

Lemma 2. *Under Assumptions* 1 *and* 2, *for any* $g \in L^2(X, \mu)$ *we have* $L_K(g) \in \mathcal{H}_K$ *and*

$$\|L_K(g)\|_K^2 = \langle L_K(g), g\rangle_{L^2(X,\mu)}. \tag{2.4}$$

Also, for any $h \in \mathcal{H}_K \cap L^2(X, \mu)$, *there holds*

$$\langle L_K(g), h\rangle_K = \langle g, h\rangle_{L^2(X,\mu)}. \tag{2.5}$$

Proof. Since $g \in L^2(X, \mu)$, there is a sequence $\{g_n\} \subset C_B(X)$ such that $g_n \to g$ in $L^2(X, \mu)$. By Lemma 1, $L_K(g_n) \in \mathcal{H}_K$. Moreover,

$$\|L_K(g_n - g_m)\|_K^2 = \left\langle \int_X (g_n(s) - g_m(s))K(x, s)\,d\mu(s),\right.$$

$$\begin{aligned}
&\left\langle \int_X (g_n(s) - g_m(s))K(x,s)\,d\mu(s) \right\rangle_K \\
&= \int_{X\times X} (g_n(s) - g_m(s))K(t,s)(g_n(t) - g_m(t))\,d\mu(s)\,d\mu(t) \\
&= \langle L_K(g_n - g_m), g_n - g_m\rangle_{L^2} = \|L_K^{\frac{1}{2}}(g_n - g_m)\|_{L^2}^2 \\
&\leqslant \|L_K^{\frac{1}{2}}\|^2 \|g_n - g_m\|_{L^2}^2 \to 0 \quad (\text{as } n, m \to \infty). \qquad (2.6)
\end{aligned}$$

This means that $\{L_K(g_n)\}$ is a Cauchy sequence in $\mathcal{H}_K$ and has a limit $f \in \mathcal{H}_K$. This in connection with the reproducing kernel property (1.1) implies that for each $m \in \mathbf{N}$,

$$\sup_{x\in X_m} |L_K(g_n)(x) - f(x)| \leqslant \|L_K(g_n) - f\|_K \sup_{x\in X_m} K(x,x) \to 0 \quad (\text{as } n \to \infty).$$

Hence $\{L_K(g_n)\}$ converges to f uniformly on X_m. By Assumptions 2, $L_K(g_n)$, $L_K(g)$ are all continuous on X and $\lim_{n\to\infty} L_K(g_n) = L_K(g)$ in $L^2(X,\mu)$. Since μ is nondegenerate, $L_K(g_n) \to L_K(g)$ almost everywhere on X_m for each $m \in \mathbf{N}$. Thus, $L_k(g) = f$ almost everywhere on X_m. But $L_k(g)$ and f are both continuous on X_m, we have $L_k(g) = f$ on each X_m and hence on X. Therefore $L_K(g) \in \mathcal{H}_K$. By (2.1)

$$\begin{aligned}
\langle L_K(g), h\rangle_K &= \lim_{n\to+\infty} \langle L_K(g_n), h\rangle_K \\
&= \lim_{n\to+\infty} \int_X h(y)g_n(y)\,d\mu(y) \\
&= \langle h, g\rangle_{L^2(X,\mu)}
\end{aligned}$$

and

$$\|L_K(g)\|_K^2 = \langle L_K(g), L_K(g)\rangle_K = \langle L_K(g), g\rangle_{L^2(X,\mu)}.$$

Thus, both (2.4) and (2.5) hold.

We first claim that $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is an orthonormal system.

Theorem 1. *Under Assumptions* 1–3, $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ *is an orthonormal system in* $\mathcal{H}_K$.

Proof. Since $\phi_i = \frac{1}{\lambda_i} L_K(\phi_i)$, by Lemma 2, $\phi_i \in \mathcal{H}_K \cap L^2(X,\mu)$. Then (2.5) yields

$$\langle \sqrt{\lambda_i}\phi_i, \sqrt{\lambda_j}\phi_j\rangle_K = \left\langle L_K(\phi_i), \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}}\phi_j \right\rangle_K = \left\langle \phi_i, \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}}\phi_j \right\rangle_{L^2(X,\mu)} = \delta_{ij}.$$

This proves our statement. □

Using Theorem 1, we can now state the Mercer theorem on noncompact domains.

Theorem 2. *Suppose Assumptions* 1–3 *hold. Then*

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y), \qquad (2.7)$$

where the series converges absolutely and uniformly on $Y_1 \times Y_2$ with Y_1 and Y_2 being any compact subsets of X.

Proof. For an arbitrarily fixed point $x \in X$, $K_x \in \mathcal{H}_K \cap L^2(X, \mu)$. By Theorem 1, the orthogonal projection of K_x onto $\overline{\text{span}}\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ equals

$$\sum_{i=1}^{\infty} \langle K_x, \sqrt{\lambda_i}\phi_i\rangle_K \sqrt{\lambda_i}\phi_i(y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y). \tag{2.8}$$

Moreover,

$$\left\langle \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i - K_x, \sqrt{\lambda_j}\phi_j \right\rangle_K = 0, \quad \forall j \in \mathbf{N}. \tag{2.9}$$

Notice that as functions of the variable y, series (2.8) converges in $\mathcal{H}_K$ and in $L^2(X, \mu)$. Set K_1 as

$$K_1(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y) - K(x, y).$$

Then $(K_1)_x \in \mathcal{H}_K \cap L^2(X, \mu)$ as a function of the variable y. By (2.9),

$$\begin{aligned} 0 = \langle (K_1)_x, \sqrt{\lambda_j}\phi_j\rangle_K &= \left\langle K_1(x, \cdot), \frac{1}{\sqrt{\lambda_j}} \int_X K(\cdot, t)\phi_j(t)\, d\mu(t) \right\rangle_K \\ &= \frac{1}{\sqrt{\lambda_j}} \int_X K_1(x, t)\phi_j(t)\, d\mu(t). \end{aligned} \tag{2.10}$$

This in connection with Assumptions 2 and 3 implies that

$$L_K((K_1)_x) = 0. \tag{2.11}$$

In particular, we have

$$0 = \int_X K(x, y)K_1(x, y)\, d\mu(y) = \int_X \{K_1(x, y)\}^2\, d\mu(y). \tag{2.12}$$

It tells us that the set $X_x := \{y \in X : K_1(x, y) = 0\}$ is the complement of a set of measure zero. Since μ is nondegenerate, X_x is dense in X. As functions of the single variable y, both $K(x, y)$ and $\sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$ are in $\mathcal{H}_K$, hence are continuous on X. It follows that $(K_1)_x$ is also continuous on X. But it vanishes on the dense subset X_x. Therefore, $(K_1)_x \equiv 0$, and

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y), \quad \forall x, y \in X. \tag{2.13}$$

In particular,

$$K(x, x) = \sum_{i=1}^{\infty} \lambda_i (\phi_i(x))^2. \tag{2.14}$$

As $K(x,x)$ and $\phi_i(x)$ are continuous on X, series (2.14) converges uniformly on any compact subset X_1. By the Schwartz inequality

$$\left|\sum_{i=m}^{n}\lambda_i\phi_i(x)\phi_i(y)\right|^2 \leqslant \left\{\sum_{i=m}^{n}|\lambda_i\phi_i(x)\phi_i(y)|\right\}^2 \leqslant \left[\sum_{i=m}^{n}\lambda_i|\phi_i(x)|^2\right]\left[\sum_{i=m}^{n}\lambda_i|\phi_i(y)|^2\right]. \tag{2.15}$$

Then we see that the series $\sum_{i=1}^{\infty}\lambda_i\phi_i(x)\phi_i(y)$ converges absolutely and uniformly on $Y_1\times Y_2$ with Y_1 and y_2 being any compact subsets of X. This proves Theorem 2. □

A nice corollary of the Mercer theorem is that the orthonormal system $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is complete.

Theorem 3. *Under Assumptions* 1–3, $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ *form an orthonormal basis of* $\mathcal{H}_K$.

Proof. By the proof of Theorem 2,

$$K(x,y)=\sum_{i=1}^{+\infty}\lambda_i\phi_i(x)\phi_i(y) \tag{2.16}$$

and for each fixed $x\in X$, the series converges to $K(x,y)$ in $\mathcal{H}_K$.

Suppose $h\in\mathcal{H}_K$, and $\langle h,\phi_i\rangle_K=0$ for each i, then for each $x\in X$,

$$h(x)=\langle K(x,\cdot),h\rangle_K=\sum_{i=1}^{+\infty}\lambda_i\phi_i(x)\langle\phi_i,h\rangle_K=0 \tag{2.17}$$

which means $h=0$, so the orthonormal system $\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is complete and forms an orthonormal basis of $\mathcal{H}_K$. The proof of Theorem 3 is complete. □

By Theorem 3, the Hilbert space structure of RKHS $\mathcal{H}_K$ is well understood, and we can easily get the following corollary.

Corollary 1. *Under Assumptions* 1–3, $\mathcal{H}_K$ *is the range of* $L_K^{1/2}$, *where* $L_K^{1/2}:\overline{\mathcal{D}}_K\to H_k$ *is an isometric isomorphism, with* $\overline{\mathcal{D}}_K$ *being the closure of* $\mathcal{D}_K:=\operatorname{span}\{K_x:x\in X\}$ in $L^2(X,\mu)$.

Proof. By the proof of Theorem 2, $\overline{\mathcal{D}}_K\subseteq\overline{\operatorname{span}}\{\phi_1,\phi_2,\ldots\}$. If f is orthogonal to $\overline{\mathcal{D}}_K$, then $\langle f,K_x\rangle_{L^2}=0$ for every $x\in X$. This implies $L_K(f)=0$. It follows that $\langle f,\phi_i\rangle_{L^2}=\langle L_K(f),\frac{1}{\lambda_i}\phi_i\rangle_{L^2}=0$ for each $i\in\mathbf{N}$. So $\overline{\mathcal{D}}_K=\overline{\operatorname{span}}\{\phi_1,\phi_2,\ldots\}$. For $f=\sum_{i=1}^{+\infty}\alpha_i\phi_i\in\overline{\mathcal{D}}_K$, $L_K^{1/2}(f)=\sum_{i=1}^{+\infty}\alpha_i\sqrt{\lambda_i}\phi_i$, thus $\|L_K^{1/2}(f)\|_K=\|f\|_{L^2}$ by Theorem 3. Hence Corollary 1 holds. □

3. The integral operator and $\mathcal{H}_K$

In this section we show how to fulfill the conditions concerning the operator L_K assumed in Section 2.

It is well known that if L_K is compact and positive, then L_K has at most countably many positive eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, and corresponding orthonormal eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$. Hence Assumptions 2 and 3 are satisfied. So we first investigate when L_K is compact and positive. For the purpose of Theorems 2 and 3, we also want to know when $L_K(L^2(X,\mu)) \subset C(X)$.

Let (X, d) be a metric space, μ be a Borel measure on X, and $K : X \times X \to \mathbf{R}$ be a Mercer kernel satisfying

$$\|K\| := \int_X \int_X (K(x,y))^2 \, d\mu(x) \, d\mu(y) < +\infty. \tag{3.1}$$

For the propositions given in this section, we need Assumption 1 and (3.1) only (but not Assumptions 2 or 3).

Proposition 1. *If Assumption 1 and* (3.1) *hold, then L_K is bounded, compact and positive.*

Proof. The boundedness of L_K with $\|L_K\| \leqslant \sqrt{\|K\|}$ follows from (3.1) and the Schwartz inequality:

$$\begin{aligned}\|L_K g\|^2_{L^2(X,\mu)} &\leqslant \int_X \left\{ \int_X |K(x,y)|^2 \, d\mu(y) \int_X |g(y)|^2 \, d\mu(y) \right\} d\mu(x) \\ &= \|g\|^2_{L^2(X,\mu)} \|K\|.\end{aligned}$$

The positivity of L_K is a consequence of the positive semidefiniteness of the kernel K.

Let us now prove that L_K is compact. We shall approximate L_K by a sequence of finite rank operators.

Let $\{\phi_i\}_{i=1}^{\infty}$ be an orthonormal basis of $L^2(X,\mu)$. Fixed a point $x \in X$. Then we have $\sum_{i=1}^{+\infty} \langle K_x, \phi_i \rangle^2_{L^2(X,\mu)} \leqslant \|K_x\|^2_{L^2(X,\mu)} < \infty$ and the series expansion in $L^2(X,\mu)$:

$$K(x,y) = K_x(y) = \sum_{i=1}^{+\infty} \langle K_x, \phi_i \rangle_{L^2(X,\mu)} \phi_i(y). \tag{3.2}$$

Set $K_n(x,y) = \sum_{i=1}^{n} \langle K_x, \phi_i \rangle_{L^2(X,\mu)} \phi_i(y)$. Since $\langle K_x, \phi_i \rangle_{L^2(X,\mu)} = L_K(\phi_i)$, L_{K_n} is a finite rank operator. For each $x \in X$,

$$\begin{aligned}|(L_K - L_{K_n})(g)(x)|^2 &= \left| \int_X (K(x,y) - K_n(x,y)) g(y) \, d\mu(y) \right|^2 \\ &\leqslant \int_X |K(x,y) - K_n(x,y)|^2 \, d\mu(y) \int_X |g(y)|^2 \, d\mu(y).\end{aligned}$$

Then

$$\|(L_K - L_{K_n})(g)\|^2 \leqslant \int_X |g(y)|^2 \, d\mu(y) \int_X \sum_{i=n+1}^{+\infty} |\langle K_x, \phi_i \rangle_{L^2(X,\mu)}|^2 \, d\mu(x).$$

It follows that

$$\|L_K - L_{K_n}\|^2 \leqslant \int_X \sum_{i=n+1}^{+\infty} |\langle K_x, \phi_i\rangle_{L^2(X,\mu)}|^2 \, d\mu(x). \tag{3.3}$$

Consider the sequence of functions in the integrand. For any $n \in \mathbf{N}$,

$$\sum_{i=n+1}^{+\infty} |\langle K_x, \phi_i\rangle_{L^2(X,\mu)}|^2 \leqslant \sum_{i=1}^{+\infty} |\langle K_x, \phi_i\rangle_{L^2(X,\mu)}|^2 \leqslant \|K_x\|^2_{L^2(X,\mu)}.$$

That means the sequence of functions of the variable x is dominated by an integrable function:

$$\int_X \|K_x\|^2_{L^2(X,\mu)} \, d\mu(x) = \int_X \int_X (K(x, y))^2 \, d\mu(x) \, d\mu(y) = \|K\| < \infty.$$

Also, for each fixed $x \in X$,

$$\lim_{n\to\infty} \sum_{i=n+1}^{+\infty} |\langle K_x, \phi_i\rangle_{L^2(X,\mu)}|^2 = 0.$$

Therefore, by the dominated convergence theorem, we have

$$\lim_{n\to\infty} \int_X \sum_{i=n+1}^{+\infty} |\langle K_x, \phi_i\rangle_{L^2(X,\mu)}|^2 \, d\mu(x) = 0.$$

Thus $\|L_K - L_{K_n}\| \to 0$, and L_K is compact. This proves Proposition 1. □

The converse of the positivity of L_K is also true.

Proposition 2. *Suppose K satisfies* (3.1). *Then L_K is positive if and only if K is positive semidefinite*.

The proof of Proposition 2 is trivial, but it is necessary that μ is nondegenerate.

Proposition 3. *If Assumption* 1 *holds and $k(x) := \int_X |K(x, y)|^2 \, d\mu(y)$ is bounded on each X_i, then for every $g \in L^2(X, \mu)$, $L_K(g) \in C(X)$.*

Proof. Let $g \in L^2(X, \mu)$. By the dominated convergence theorem,

$$\lim_{m\to\infty} \int_{X\setminus X_m} |g(y)|^2 \, d\mu(y) = 0.$$

Let $x_0 \in X$. We show that $L_K(g)$ is continuous at x_0. To this end, let $U(x_0)$ be a bounded neighborhood of x_0 and $\{x_n\} \subset U(x_0)$ be a sequence tending to x_0. Then $U(x_0) \subseteq X_{i_0}$ for some i_0. Denote $M := \sup_{x\in X_{i_0}} k(x)^{\frac{1}{2}} < \infty$. Then

$$|L_K(g)(x_n) - L_K(g)(x_0)|$$

$$
\begin{aligned}
&\leqslant \int_{X_m} |K(x_n, y) - K(x_0, y)||g(y)|d\mu(y) \\
&\quad + \int_{X\setminus X_m} |K(x_n, y) - K(x_0, y)||g(y)|d\mu(y) \\
&\leqslant \left[\int_{X_m} |K(x_n, y) - K(x_0, y)|^2 \, d\mu(y)\right]^{\frac{1}{2}} \left[\int_{X_m} |g(y)|^2 \, d\mu(y)\right]^{\frac{1}{2}} \\
&\quad + \left[\int_{X\setminus X_m} |K(x_n, y) - K(x_0, y)|^2 d\mu(y)\right]^{\frac{1}{2}} \left[\int_{X\setminus X_m} |g(y)|^2 \, d\mu(y)\right]^{\frac{1}{2}} \\
&\leq \left[\int_{X_m} |K(x_n, y) - K(x_0, y)|^2 \, d\mu(y)\right]^{\frac{1}{2}} \|g\|_{L^2(X,\mu)} \\
&\quad + 2M \left[\int_{X\setminus X_m} |g(y)|^2 \, d\mu(y)\right]^{\frac{1}{2}}.
\end{aligned}
$$

As K is uniformly continuous on the compact set $X_{i_0} \times X_m$, we know that

$$\lim_{n\to\infty} \int_{X_m} |K(x_n, y) - K(x_0, y)|^2 \, d\mu(y) = 0.$$

Therefore,

$$\lim_{n\to\infty} L_K(g)(x_n) - L_K(g)(x_0) = 0.$$

This proves the continuity of $L_K(g)$. □

Proposition 4. *If Assumption* 1 *and* (3.1) *hold, then* $\mathcal{H}_K \subset L^2(X, \mu)$.

Proof. Since $\mathcal{D}_K \subset L^2(X, \mu) \cap \mathcal{H}_K$ and $\mathcal{D}_K$ is dense in $\mathcal{H}_K$, we only need to compare the norm of $L^2(X, \mu)$ and the norm of $\mathcal{H}_K$.

For fixed $f = \sum_{k=1}^m \alpha_k K_{y_k} \in \mathcal{H}_K$, there hold

$$\|f\|_K^2 = \sum_{i,j=1}^m \alpha_i \alpha_j K(y_i, y_j) \tag{3.4}$$

and

$$
\begin{aligned}
\|f\|_{L^2}^2 &= \int_X \left(\sum_{k=1}^m \alpha_k K(x, y_k)\right)^2 d\mu(x) \\
&= \sum_{i,j=1}^m \alpha_i \alpha_j \int_X K(x, y_i) K(x, y_j) \, d\mu(x).
\end{aligned} \tag{3.5}
$$

Let $b = \frac{1}{2}\|L_K^{\frac{1}{2}}\|^{-2}$, and $K_1(x, y) = K(x, y) - b\int_X K(t, x)K(t, y) \, d\mu(t)$.

Now we want to prove that L_{K_1} is a positive operator. Notice that

$$L_{K_1}(g)(x) = L_K(g)(x) - bL_K(L_K(g))(x).$$

Hence

$$(L_{K_1}(g), g) = (L_K(g), g) - b(L_K(g), L_K(g)) \tag{3.6}$$

and

$$b(L_K(g), L_K(g)) = b\|L_K^{\frac{1}{2}}(L_K^{\frac{1}{2}}(g))\|^2 \leqslant \frac{1}{2}\|(L_K^{\frac{1}{2}}(g))\|^2 \leqslant \frac{1}{2}(L_K(g), g).$$

So $(L_{K_1}(g), g) \geqslant \frac{1}{2}(L_K(g), g) \geqslant 0$. By Proposition 2, K_1 is positive semidefinite. This implies

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j K(y_i, y_j) \geqslant b \sum_{i,j=1}^{m} \alpha_i \alpha_j \int_X K(x, y_i)K(x, y_j)\, d\mu(x).$$

That is,

$$\|f\|_K \geqslant \sqrt{b}\|f\|_{L^2}. \tag{3.7}$$

Thus we have $\mathcal{H}_K \subset L^2(X, \mu)$. □

4. Example of Gaussian kernels

In this section we give the example with the Gaussian kernels.

Example. *Let* $X = \mathbf{R}^n$, $K(x, y) = e^{-\frac{(x-y)^2}{c^2}}$ *with* $c > 0$. *If* $r \in L^2(\mathbf{R}^n)$ *is positive almost everywhere and* $d\mu = r(x)\, dx$, *then Assumption* 1 *and* (3.1) *hold. Hence Theorems* 1–3 *are valid.*

Proof. Let $K_x(t) = K(x, t) = e^{-\frac{(x-t)^2}{c^2}}$. Then

$$\int_{\mathbf{R}^n} K_x^2(t)\, d\mu(t) = \int_{\mathbf{R}^n} K_x^2(t) r(t)\, dt \leqslant \int_{\mathbf{R}^n} K_x(t) r(t)\, dt \leqslant \|K_x\|_2 \|r\|_2 < \infty.$$

Therefore $K_x \in L^2_\mu(\mathbf{R}^n)$ for each $x \in \mathbf{R}^n$ and Assumption 1 holds.

Set $A = \int_{\mathbf{R}^n} e^{-\frac{x^2}{c^2}}\, dx$. Then $0 < A < +\infty$ and

$$\begin{aligned}\int_{\mathbf{R}^n}\int_{\mathbf{R}^n} K^2(x, y)\, d\mu(y)\, d\mu(x) &\leqslant \int_{\mathbf{R}^n} r(x) \int_{\mathbf{R}^n} e^{-\frac{(x-y)^2}{c^2}} r(y)\, dy\, dx \\ &= \int_{\mathbf{R}^n} r(x) \int_{\mathbf{R}^n} e^{-\frac{t^2}{c^2}} r(x - t)\, dt\, dx\end{aligned}$$

$$= \int_{\mathbf{R}^n} e^{-\frac{t^2}{c^2}} \int_{\mathbf{R}^n} r(x) r(x-t)\, dx\, dt$$
$$\leqslant \int_{\mathbf{R}^n} e^{-\frac{t^2}{c^2}} \|r\|_2^2\, dt \leq \|r\|_2^2 A < \infty.$$

This verifies (3.1). Hence our statements hold true. □

Acknowledgments

The work of this paper was partially supported by City University of Hong Kong under Project No. 7001442. The author would like to thank Dr. D.X. Zhou and the referees for their helpful suggestions.

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[2] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Soc. 39 (2001) 1–49.
[3] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: on the bias-variance problem, Found. Comput. Math. 2 (2002) 413–428.
[4] F. Cucker, D.X. Zhou, Learning Theory, Cambridge University Press, Cambridge.
[5] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[6] H. Hochstadt, Integral Equations, Wiley, New York, 1973.
[7] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations, Philos. Trans. Royal Soc. London 209 (1909) 415–446.
[8] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[9] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[10] G. Wahba, Spline Models for Observational Data, Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
[11] F. Girosi, An equivalence between sparse approximation and support vector machines neural computation, Neural Comput. 10 (1998) 1455–1480.
[12] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.
[13] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.