Reproducing Kernel Hilbert Spaces - Part III

Scribes: Manoj Subedi and Jayson Cortez

1 Properties of Reproducing Kernel Hilbert Spaces

In this section, we study additional properties of functions in a RKHS that are inherited from the reproducing kernel function.

1.1 Continuity

In this section we examine the continuity properties of a kernel k.

We define a (pseudo)-metric in terms of k and use this to characterize continuity.

Definition 1. Let $\mathbb{X}$ be a topological vector space. A kernel k on $\mathbb{X}$ is *separately continuous* if $k(\cdot,x) : \mathbb{X} \to \mathbb{R}$ is continuous for all $x \in \mathbb{X}$.

Lemma 1. *Let $\mathbb{X}$ be a topological space and k a kernel on $\mathbb{X}$ with reproducing kernel Hilbert space $\mathbb{H}$. Then k is bounded and separately continuous iff every $f \in \mathbb{H}$ is bounded and continuous. In this case, the inclusion map $id : \mathbb{H} \to C_b(\mathbb{X})$ is continuous and*

$$\|id : \mathbb{H} \to C_b(\mathbb{X})\| = \|k\|_\infty.$$

Definition 2. Let k be a kernel on $\mathbb{X}$ with a feature map $\Phi : \mathbb{X} \to \mathbb{H}$. The *kernel metric* is given by;

$$d_k(x,x') = \|\Phi(x) - \Phi(x')\|_{\mathbb{H}} \qquad x,\ x' \in \mathbb{X}.$$

We remark that d_k is a *pseudo-metric* in general ($d_k(x,x') = 0$ for not imply that $x = x'$ in general), and is a metric if Φ is injective. Furthermore, we have

$$d_k(x,x') = \sqrt{k(x,x) - 2k(x,x) + k(x',x')} \tag{1}$$

which shows that the definition of d_k is independent of feature map Φ.

The following lemma shows how the kernel metric can be used to characterize the continuity of the kernel k.

Lemma 2. *Let $(\mathbb{X},\tau)$ be a topological vector space and k a kernel on $\mathbb{X}$ with feature space $\mathbb{H}$ and feature map Φ. The following are equivalent:*

i. k is continuous.
ii. k is separately continuous and $x \mapsto k(x,x)$ is continuous.
iii. Φ is continuous.
iv. The map $id : (\mathbb{X},\tau) \to (\mathbb{X},d_k)$ is continuous.

Proof. *(i) $\Longrightarrow$ (ii).* Trivial.
(ii) $\Longrightarrow$ (iv). By equation (1) and the assumption, we see that $d_k(\cdot,x) : (\mathbb{X},\tau) \to \mathbb{R}$ is continuous for every $x \in \mathbb{X}$. Consequently, $\{x' \in \mathbb{X} : d_k(x',x) < \varepsilon\}$ is open with respect to τ and therefore $id : (\mathbb{X},\tau) \to (\mathbb{X},d_k)$ is continuous.
(iv) $\Longrightarrow$ (iii). This follows from the fact that $\Phi : (\mathbb{X},d_k) \to \mathbb{H}$ is continuous.
(iii) $\Longrightarrow$ (i). Fix $x_1,\ x_1' \in \mathbb{X}$ and $x_2,\ x_2' \in \mathbb{X}$. Then we have

$$\begin{aligned}\left|k(x_1,x_1') - k(x_2,x_2')\right| &\le \left|\langle \Phi(x_1'), \Phi(x_1) - \Phi(x_2)\rangle\right| + \left|\langle \Phi(x_1') - \Phi(x_2'), \Phi(x_2)\rangle\right| \\ &\le \|\Phi(x_1')\| \cdot \|\Phi(x_1) - \Phi(x_2)\| + \|\Phi(x_2)\| \cdot \|\Phi(x_1') - \Phi(x_2')\|.\end{aligned}$$

From this we conclude that k is continuous.

1.2 Compactness

We have seen above that a RKHS over $\mathbb{X}$ is continuously contained in l^∞ if it has a bounded kernel. The following proposition provides an additional condition so that this inclusion is compact.

Proposition 1. *Let k be a kernel on a space $\mathbb{X}$ with RKHS $\mathbb{H}$ and canonical feature map $\Phi : \mathbb{X} \to \mathbb{H}$. If $\Phi(\mathbb{X})$ is compact in $\mathbb{H}$ then the inclusion map given by*

$$id : \mathbb{H} \to l^\infty(\mathbb{X})$$

is also compact.

Proof. Since $\Phi(\mathbb{X})$ is compact, then k is bounded and the space $(\mathbb{X},d_k)$ is compact with respect to the kernel metric d_k. Let $C(\mathbb{X},d_k)$ be the space of functions $f : \mathbb{X} \to \mathbb{R}$ that are continuous with respect to d_k. For $x,\ x' \in \mathbb{X}$ and $f \in \mathbb{H}$ we have

$$\left|f(x) - f(x')\right| = \left|\langle f, \Phi(x) - \Phi(x')\rangle\right| \le \|f\|_{\mathbb{H}} \cdot d_k(x,x'),$$

showing that f is continuous on $(\mathbb{X},d_k)$. It follows that the unit ball $B_{\mathbb{H}} \subset \mathbb{H}$ is equicontinuous and bounded. By Arzela-Ascoli Theorem, $\overline{B_{\mathbb{H}}}$ is compact in $\mathbb{C}(\mathbb{X},d_k)$ and, hence, in $l^\infty(\mathbb{X})$ since $\mathbb{C}(\mathbb{X},d_k) \subset l^\infty(\mathbb{X})$. This shows that $id : \mathbb{H} \to l^\infty(\mathbb{X})$ is compact.

Next, we provide a sufficient condition for the separability of a RKHS $\mathbb{H}$.

Proposition 2. *Let $\mathbb{X}$ be a separable topological space and k a continuous kernel on $\mathbb{X}$. Then the RKHS $\mathbb{H}$ of k is separable.*

Proof. By Proposition 1, the canonical feature map $\Phi : \mathbb{X} \to \mathbb{H}$ is continuous, thereby implying that $\Phi(\mathbb{X})$ is separable. It follows that vector space

$$\mathbb{H}_{pre} := \{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in k \quad x_1, ..., x_n \in \mathbb{X}\} \tag{2}$$

is also separable. We observed in the proof of a previous theorem that $\mathbb{H}_{pre}$ is dense in $\mathbb{H}$. The separability of $\mathbb{H}$ follows by completion.

2 Mercer's Theorem

This theorem shows the existence of a series representation for continuous kernels that are defined on a compact domain.

Theorem 1. *(Mercer's) Let $\mathbb{X}$ be compact metric space and $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be continuous. Let μ be a finite Borel measure with $supp(\mu) = \mathbb{X}$. Then there exists a countable orthonormal sequence $(e_i)_{i \in I} \subset \mathbb{H}$ and a family $(\lambda_i)_{i \in I} \subset \mathbb{R}$ converging to* 0 *such that*

$$k(x, x') = \sum_{i \in I} \lambda_i e_i(x) e_i(x') \quad x, x' \in \mathbb{X} \tag{3}$$

with absolute and uniform convergence. Here we assume that

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots$$

Remarks:

- Equation 3 implies that $\Phi : \mathbb{X} \mapsto \ell^2$ given by $\Phi(x) = (\sqrt{\lambda_i} e_i(x))_{i \in I}, x \in \mathbb{X}$, is a feature map. In fact $k(x, x') = \langle \Phi(x'), \Phi(x) \rangle$.
- With the assumptions of Mercer's theorem, if $(a_i)_{i \in I} \subset \ell^2(I)$ and $x \in X$, $J \subset I$, then

$$\sum_{i \in J} \left| a_i \sqrt{\lambda_i} e_i(x) \right| \leq \left(\sum_{i \in J} a_i^2 \right)^{1/2} \left(\sum_{i \in J} \lambda_i e_i^2(x) \right)^{1/2} = \|(a_i)\|_{\ell^2(I)} \cdot \sqrt{k(x,x)}.$$

Theorem 2. *(Mercer's Representation theorem for RKHS) With the assumptions from previous theorem, let*

$$H := \left\{ \sum_{i \in I} a_i \sqrt{\lambda_i}\, e_i : (a_i) \in \ell^2(I) \right\}.$$

For

$$f = \sum a_i \sqrt{\lambda_i} e_i \in H, \quad g = \sum b_i \sqrt{\lambda_i} e_i \in H$$

set

$$\langle f, g \rangle_H = \sum_{i \in I} a_i b_i.$$

Then H equipped with $\langle \cdot, \cdot \rangle_H$ is the RKHS of the kernel of k. Furthermore, $T_k^{1/2} : L^2(\mu) \to H$ given by

$$T_k^{1/2} f = \sum_{i \in I} \lambda_i^{1/2} \langle f, e_i \rangle e_i$$

is an isometric isomorphism.

Proof. It is straightforward to verify that H is a Hilbert Space, i.e., H is a complete inner product space under the norm $\langle \cdot, \cdot \rangle_H$. For $x \in \mathbb{X}$, by Mercer's Theorem we have that

$$k(\cdot, x) = \sum_{i \in I} \sqrt{\lambda_i} e_i(x) \sqrt{\lambda_i} e_i(\cdot)$$

showing that $k(\cdot, x) \in H$. For $f = \sum_{i \in I} a_i \sqrt{\lambda_i} e_i \in H$, we have

$$\langle f, k(\cdot, x) \rangle_H = \sum_{i \in I} a_i \sqrt{\lambda_i} e_i(x) = f(x), \qquad x \in \mathbb{X}$$

showing that k is the reproducing kernel of k.

Let us now focus on $T_k^{1/2}$. Fix $f \in L^2(\mu)$. Since $(e_i)_{i \in I}$ is an orthonormal basis of $L^2(\mu)$, we can write $f = \sum_{i \in I} \langle f, e_i \rangle_{L^2(\mu)} e_i$. Hence, by Parseval's formula,

$$\|f\|_{L^2(\mu)}^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2,$$

showing that $(\langle f, e_i \rangle)_{i \in I} \subset \ell^2(I)$. It follows that

$$T_k^{1/2} f = \sum_{i \in I} \langle f, e_i \rangle \sqrt{\lambda}_i e_i \in H.$$

Moreover,

$$\|T_k^{1/2} f\|^2 = \sum_{i \in I} |\langle f, e_i \rangle|^2 = \|f\|_{L^2}^2(\mu),$$

implying that $T_k^{1/2}$ is an isometry on H and hence injective. To show that $T_k^{1/2}$ is surjective, fix $f \in H$. By the definition of H there is a sequence $(a_i)_{i \in I} \subset \ell^2(I)$ such that $f(x) = \sum_{i \in I} a_i \sqrt{\lambda}_i e_i(x)$. We obviously have that $g := \sum_{i \in I} a_i e_i \in L^2(\mu)$ and, thus, $a_i = \langle g, e_i \rangle_{L^2}$. Thus,

$$T_k^{1/2} g(x) = \sum_{i \in I} \langle g, e_i \rangle_{L^2} \sqrt{\lambda}_i e_i(x) = \sum_{i \in I} a_i \sqrt{\lambda}_i e_i(x) = f(x),$$

proving that $T_k^{1/2}$ is surjective.

3 Universal kernels

We have seen that SVMs are based on minimization problems over RKHS. We will eventually see that the ”size” of the RKHS is a critical issue on the generalization ability of an SVM since we want a solution space large enough to give accurate solutions, yet not too large to avoid over-fitting.

Definition 3. A continuous kernel k on a compact metric space $\mathbb{X}$ is *universal* if the RKHS $\mathbb{H}$ of k is dense in $C(\mathbb{X})$, i.e., for every $g \in C(\mathbb{X})$ and all $\varepsilon > 0$, there exists an $f \in \mathbb{H}$ such that

$$\|f - g\|_\infty \leq \varepsilon.$$

Definition 4. Let k be a kernel on a metric space $\mathbb{X}$ with RKHS $\mathbb{H}$. We say that k *separates* the disjoint sets $A,\ B \subset \mathbb{X}$, if there exists an $f \in \mathbb{H}$ such that $f(x) > 0$ for all $x \in A$, and $f(x) < 0$ for all $x \in B$. We say that k separates all finite (or compact) sets if k separates all finite (or compact) disjoints sets $A,\ B \subset \mathbb{X}$.

Proposition 3. *Let $\mathbb{X}$ be a compact metric space and k a universal kernel on $\mathbb{X}$. Then k separates all compact sets in $\mathbb{X}$.*

Proof. Let $A,\ B \subset \mathbb{X}$ be disjoint and compact. For any $x \in \mathbb{X}$, set

$$g(x) = \frac{d(x,B)}{d(x,A) + d(x,B)} - \frac{d(x,A)}{d(x,A) + d(x,B)} \tag{4}$$

where $d(x,C) = \inf_{x' \in C} d(x,x')$, $x \in \mathbb{X}$ and $C \subset \mathbb{X}$. If $x \in A$ then $g(x) = 1$, and if $x \in B$ then $g(x) = -1$. Note that g is continuous by the continuity of the metric d. Let $\mathbb{H}$ be the RKHS of k. Since k is universal, we can find $f \in \mathbb{H}$ such that $\|f - g\|_\infty < 1/2$ implying that $f(x) > 1/2$ if $x \in A$, and $f(x) < 1/2$ if $x \in B$.

Geometrical Interpretation:
Assume that there exists a universal kernel k on $\mathbb{H}_0$, with the feature map $\Phi_0 : \mathbb{X} \to \mathbb{H}_0$. Let $\mathbb{X}$ be a compact metric space and $\{x_1, ..., x_n\} \subset \mathbb{X}$. By proposition 3, for every choice of labels $\{y_1, ..., y_n\} \subset \{-1, 1\}$ we can find $f \in \mathbb{H}$ satisfying $y_i f(x_i) > 0 \quad \forall i = 1, ...n$. This f can be represented by $f = \langle w, \Phi_0(\cdot)\rangle$ for a suitable $w \in \mathbb{H}_0$. As a result, the mapped training points $\langle \Phi_0(x_1), y_1\rangle, ..., \langle \Phi_0(x_n), y_n\rangle$ can be correctly separated in $\mathbb{H}_0$ by the hyperplane defined by w.

Theorem 3. *(Test for universality) Let $\mathbb{X}$ be a compact metric space and k a continuous kernel on $\mathbb{X}$ with the property that $k(x,x) > 0$ for all $x \in \mathbb{X}$. Suppose that we have an injective feature map $\Phi : \mathbb{X} \to \ell^2$ and denote $\Phi(x) = (\phi_1(x), \phi_2(x), ..., \phi_k(x), ...)$, $x \in \mathbb{X}$. If $\mathscr{A} = span\{\phi_n : n \in \mathbb{N}\}$ is an algebra, then k is universal.*

We first recall the Stone-Weierstrass Theorem:

Let $(\mathbb{X}, \tau)$ be a compact metric space and $\mathscr{A} \subset C(\mathbb{X})$ be an algebra. Then $\mathscr{A}$ is dense in $C(\mathbb{X})$ if both $\mathscr{A}$ does not vanish, i.e., for all $x \in \mathbb{X}$, there exists an $f \in \mathscr{A}$ with $f(x) \neq 0$, and $\mathscr{A}$ separates points, i.e., for all $x, y \in \mathbb{X}$ with $x \neq y$, there exists an $f \in \mathscr{A}$ with $f(x) \neq f(y)$.

Proof. We first observe that the algebra $\mathscr{A} = span\{\phi_n : n \in \mathbb{N}\}$ does not vanish since $\|(\phi_n(x))\|_{\ell 2}^2 = k(x,x) > 0$ for all $x \in \mathbb{X}$. Moreover, k is continuous and, thus, by Lemma 2, every $\phi_n : \mathbb{X} \to \mathbb{R}$ is continuous. This shows that $\mathscr{A} \subset C(X)$. Moreover, the injectivity of Φ implies that $\mathscr{A}$ separates points and, thus, Stone-Weierstass Theorem implies that $\mathscr{A}$ is dense in $C(\mathbb{X})$. Now fix a $g \in C(X)$ and an $\varepsilon > 0$. Then there exists a function $f \in \mathscr{A}$ of the form

$$f = \sum_{j=1}^{m} \alpha_j \phi_{n_j} \quad with \quad \|f - g\|_\infty \leq \varepsilon.$$

For $n \in \mathbb{N}$, we define $w_n := \alpha_j$ if there is an index j with $n_j = n$ and $w_n := 0$ otherwise. This yields $w := (w_n) \in \ell_2$ and $f = \langle w, \Phi(\cdot) \rangle_{\ell_2}$, and thus k is universal.

Examples of universal Kernels

- Polynomial: $k(x,x') = f(\langle x, x' \rangle)$ where $f(t) = \sum_{k=0}^{\infty} a_k t^k, \quad a_k > 0$.
- Exponential: $k(x,x') = exp(\langle x, x' \rangle)$.
- Gaussian: $k_\gamma(x,x') = exp(-\gamma^2 \|x - x'\|_L^2)$

References

1. G. Folland, *Real Analysis: Modern Techniques and Their Applications* A Wiley-Interscience Publication, Canada, 1999.
2. I. Steinwart, A. Christmann, *Support Vector Machines*, Springer Science (2008).
3. B. Scholkopf, A. J. Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press Cambridge, Massachusetts (2002).