Reproducing Kernel Hilbert Spaces

Lecturer: Peter L. Bartlett *Scribe: Choongsoon Bae*

1 Reproducing Kernel Hilbert Spaces

Definition. A *Hilbert Space* is an inner product space which is separable and complete respect to the norm defined by the inner product.

Definition. $\langle \cdot, \cdot \rangle$ is an inner product *iff*

1. It is *bilinear*
$$\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle \ \ \forall u, v, w \in \mathcal{X} \ \ \forall \alpha, \beta \in \mathbb{R}$$
2. It is *symmetric*
$$\langle u, v \rangle = \langle v, u \rangle \ \ \forall u, v \in \mathcal{X}$$
3. It is *positive definite*
$$\langle u, u \rangle \geq 0, \ \ \forall u \in \mathcal{X}$$
$$\langle u, u \rangle = 0 \Leftrightarrow u = 0$$

Definition. A metric space is complete if all Cauchy sequences have limits within the space.

Some examples of Hilbert Spaces are:

1. The vector space $\mathbb{R}^n$ with the usual dot product $\langle u, v \rangle = u^T v$.
2. The space $L_2([a, b])$ of *square integrable* functions over an interval $[a, b]$ with $\langle f, g \rangle = \int_a^b f(x)g(x)dx$.
3. ℓ_2 with $\langle f, g \rangle = \sum_{i=1}^{\infty} f_i g_i$

We will be dealing with Hilbert Spaces for which there is a reproducing kernel.

Definition. A Hilbert Space $\mathscr{H}$ is a *Reproducing Kernel Hilbert Space* (RKHS) if there is a kernel function $k(\cdot, \cdot)$ such that:
$$\forall f \in \mathscr{H}, \quad f(x) = \langle f(\cdot), k(x, \cdot) \rangle$$
$$\mathscr{H} = \overline{\mathrm{span}\{k(x, \cdot) : x \in \mathcal{X}\}}$$

Note that the delta function has the reproducing property because:
$$f(x) = \int f(z)\delta(z - x)dx$$

But since the delta function is not square integrable, it does not follow that L_2 is an RKHS. Let us now define a kernel.

Definition. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *kernel* if the following two conditions are satisfied:

1. k is symmetric, *i.e.* $k(u, v) = k(v, u)$
2. k is positive semi-definite *i.e.* for all $x_1, \ldots, x_n \in \mathcal{X}$, the "Gram matrix" $\mathbf{K}$ (defined by $\mathbf{K}_{ij} = k(x_i, x_j)$) is positive semi-definite. This means that $\forall \mathbf{u} \in \mathbb{R}^n,\ \mathbf{u}^T\mathbf{K}\mathbf{u} \geq 0$.

We shall see how we can use a (positive semi-definite) kernel to define an RKHS. Before we do that, some properties of a kernel are worth noting:

- $k(x, x) \geq 0$ (Consider the case when $n = 1$)
- $k(u, v)^2 \leq k(u, u).k(v, v)$

To see why the second property holds, consider the case when $n = 2$:

$$\forall w_1, w_2 \quad (w_1 w_2) \begin{bmatrix} k(u,u) & k(u,v) \\ k(v,u) & k(v,v) \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \geq 0$$

$$\Rightarrow \forall w_1, w_2 \quad w_1^2 k(u,u) + 2 w_1 w_2 k(u,v) + w_2^2 k(v,v) \geq 0$$

This implies that this quadratic expression involving w_1, w_2 has a non-positive discriminant. Therefore,

$$(2k(u,v))^2 - 4k(u,u)k(v,v) \leq 0$$

Any k can be thought of as an inner product of $\Phi(u)$ and $\Phi(v)$ for some transformation[1] $\Phi : \mathcal{X} \to \mathscr{H}$. To see this, start with a kernel k, and define the "reproducing kernel feature map" $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ as:

$$\Phi(x) = k(\cdot, x)$$

Consider the vector space:

$$\begin{aligned} & \operatorname{span}\{\Phi(x) : x \in \mathcal{X}\} \\ = \quad & \{f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \ : \ n \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}\} \end{aligned}$$

For $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^n \beta_i k(\cdot, x'_j)$, define $\langle f, g \rangle = \sum_{i,j} \alpha_i \beta_j k(x_i, x'_j)$. k has the reproducing property:

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{n} \alpha_i k(x, x_i) = f(x)$$

We can also express this by saying that k is the representer of evaluation. As defined above $\langle f, g \rangle$ is an inner product. A few of its properties are listed below.

1. It is symmetric (because k is symmetric).
2. It is bilinear, because
$$\langle f, g \rangle = \sum_i \alpha_i g(x_i) = \sum_j \beta_j f(x'_j)$$

[1] The space to which Φ maps to is often referred to as the *feature space*.

3. It is positive semi-definite.

$$\langle f, f \rangle = \alpha^T \mathbf{K} \alpha \geq 0 \text{ (because the Gram matrix } \mathbf{K} \text{ is positive semi-definite)}$$

4. $\langle f, g \rangle^2 \leq \langle f, f \rangle \langle g, g \rangle$

   PROOF. $\forall a \in \mathbb{R}$ $\langle af + g, af + g \rangle = a^2 \langle f, f \rangle + 2a \langle f, g \rangle + \langle g, g \rangle \geq 0$
   Thus, $\langle f, g \rangle^2 - \langle f, f \rangle \langle g, g \rangle \leq 0$ □

5. We have,

$$\begin{aligned} |f(x)|^2 &= \langle k(\cdot, x), f \rangle^2 \\ &\leq \langle k(\cdot, x), k(\cdot, x) \rangle \langle f, f \rangle \\ &= k(x, x) \langle f, f \rangle \end{aligned}$$

   Therefore, if $\langle f, f \rangle = 0$ then f is identically zero.

We've define an inner product space. Complete it to make it a *Hilbert Space* with $\| f \|^2 = \langle f, f \rangle$
We now define an RKHS as follows.

Definition. For a compact subset $\mathcal{X}$ of $\mathbb{R}^d$ and Hilbert space $\mathscr{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, we say that $\mathscr{H}$ is a reproducing kernel Hilbert space if $\exists k : \mathcal{X}^2 \to \mathbb{R}$, such that:

1. k has the reproducing property.
$$\langle k(\cdot, x), f \rangle = f(x)$$

2. k spans $\mathscr{H}$.
$$\overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}} = \mathscr{H}$$

Starting from this definition, it is necessary that k is symmetric and positive semi-definite. Another characterization is via the following theorem of Mercer.

Theorem 1.1. For a compact subset $\mathcal{X} \subseteq \mathbb{R}^d$, and a continuous symmetric function $k : \mathcal{X}^2 \to \mathbb{R}$, suppose that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ defined on L_2 as:

$$T_k f(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) dx$$

is positive definite, that is,

$$\forall f \in L_2(\mathcal{X}), \quad \int_{\mathcal{X}^2} f(u) f(v) k(u, v) du\, dv \geq 0$$

then T_k has eigenfunctions $\psi_i \in L_2(\mathcal{X})$ (say $|\psi_i|_{L_2} = 1$) with associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and we can write for $u, v \in \mathcal{X}$,

$$k(u, v) = \sum_{i=1}^{\infty} \lambda_i \psi_i(u) \psi_i(v)$$

Furthermore, the series converges uniformly *i.e.* $\lim_{n \to \infty} \sup_{u,v} \left| k(u, v) - \sum_{i=1}^{n} \lambda_i \psi_i(u) \psi_i(v) \right| = 0.$

This theorem shows that we are not too far from the finite dimensional case. Following is the analogue in the finite case, $\mathcal{X} = \{x_1, \ldots, x_n\}$, with counting measure. $\mathbf{K}_{ij} = k(x_i, x_j)$ and identify $f : \mathcal{X} \to \mathbb{R}$ with $\tilde{f}_i = f(x_i)$.[2]

$$T_k f = \sum k(\cdot, x_i)\tilde{f}_i$$

$$\forall \tilde{f} \quad \tilde{f}^T \mathbf{K} \tilde{f} \geq 0$$

$\mathbf{K}$ is symmetric positive semi-definite $\Rightarrow \mathbf{K} = \sum_{i=1}^n \lambda_i v_i v_i^T$ and so we have,

$$k(x_i, x_j) = \mathbf{K}_{ij} = \left(V \Lambda V^T\right)_{ij} = \sum_{t=1}^n \lambda_t v_{ti} v_{tj} = \sum_{t=1}^n \lambda_t \psi_t(x_i) \psi_t(x_j)$$

The feature map here is $\psi(x) = (\psi_1(x), \ldots)^T$.

Let us now make note of a few important things.

- We have seen two representations of $k(u, v)$ as an inner product $k(u, v) = \langle \Phi(u), \Phi(v) \rangle$.
  1. $\Phi_1(u) = k(\cdot, u)$ with inner product such that $\langle k(\cdot, u), k(\cdot, v) \rangle = k(u, v)$.
  2. $\Phi_2(u) = \left(\sqrt{\lambda_1}\psi_1(u), \sqrt{\lambda_2}\psi_2(u), \ldots\right)$ with inner product in ℓ_2.
- Computing a positive semi-definite kernel $k(u, v)$ is equivalent to computing $\langle \Phi(u), \Phi(v) \rangle$.
- We can approximate $\Phi(u)$ by a finite dimensional map because:

$$\lim_{n \to \infty} \sup_{u,v} \left| k(u, v) - \sum_{i=1}^n \lambda_i \psi_i(u) \psi_i(v) \right| = 0$$

- If we write $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ then $\|f\|_{\mathscr{H}}^2 = \alpha^T K \alpha$. That is, minimizing $\|w\|^2 + \frac{c}{n} \sum (1 - y_i w^T x_i)_+$ can be viewed as minimizing:

$$\underbrace{\|f\|_{\mathscr{H}}^2}_{\text{"complexity penalty"}} + \underbrace{\frac{c}{n} \sum (1 - y_i f(x_i))_+}_{\text{cost of } f \text{ on training data}}$$

  We want to fit the data well without getting a function with large norm.

[2] A function over a finite set can be viewed as a vector of length equal to the cardinality of the set.