1 Reproducing Kernel Hilbert Space

Definition 1. A Hilbert space is a complete and separable inner product space:

1. Want convergence in normed space, \( \| f_n - f \|_H < \epsilon \) for all \( n > N_\epsilon \) and any given \( \epsilon > 0 \) (cf. asymptotic convergence, but obviously this does not imply consistency),

2. (Complete.) Every Cauchy sequence converges (hence bounding the Cauchy sequence meaningfully shows the learning progress),

3. (Separable.) Exists a countable dense subset which intersects every non-empty open set (hence working with the countable subset is sufficient),

4. (Which norm?) The inner product induces a norm.

Definition 2. Suppose that \( \mathcal{H} \) consists of functions \( f : \mathcal{X} \mapsto \mathbb{R} \). For a fixed \( x \in \mathcal{X} \), the evaluation functional is written \( \delta_x : f \mapsto f(x) \).

Lemma 1 (Riesz representation). For every linear functional \( f \) in the dual space of \( \mathcal{H} \), there exists a unique \( y \in \mathcal{X} \) such that
\[
\forall x \in \mathcal{X} : \quad f(x) = \langle x, y \rangle.
\]

Note that the evaluation functional is an element of the dual space of \( \mathcal{H} \) and hence can be expressed as an inner product of \( \mathcal{H} \).
Definition 3. A reproducing kernel Hilbert space (RKHS) is a Hilbert space with bounded evaluation functionals $\delta_x$ for all $x \in \mathcal{X}$.

Theorem 1. Suppose that $f_n$ are in an RKHS and $\lim_{n \to \infty} \|f_n - f\|_H = 0$ (convergence in norm). Then, $\lim_{n \to \infty} |f_n(x) - f(x)| = 0, \forall x \in \mathcal{X}$ (convergence point-wise).

Proof. It is equivalent to show
$$\lim_{n \to \infty} |f_n(x) - f(x)| = \lim_{n \to \infty} |\delta_x f_n - \delta_x f| = 0$$
for all $x \in \mathcal{X}$. We have
$$|\delta_x f_n - \delta_x f| \leq \|\delta_x\| \|f_n - f\|_H.$$ The operator norm is defined as $\|\delta_x\| := \sup_{f \in \mathcal{H}} \|\delta_x f\| / \|f\|_H$. This holds true since the evaluation function is bounded in an RKHS by definition, and $\lim_{n \to \infty} |f_n(x) - f(x)| = 0$. 

Note that norm convergence does not generally imply point-wise convergence for an arbitrary Hilbert space.

The above theorem carries an important consequence — the true function $f$ can be learned by finding an $f_n$ that is close to $f$ in norm. For this task, one needs to answer two questions: 1) Which norm to use? and 2) How to perform the approximation? The introduction of RKHS provides simple answers to these questions. We leverage RKHS to establish

- a Hilbert space $\mathcal{H}$ of all linear combinations $\sum_i a_i \phi(x_i)$ which assumed to contain $f$, where $\phi : \mathcal{X} \mapsto \mathcal{H}$ maps the data point to a function in the Hilbert space. The inner product of $f = \sum_i a_i \phi(x_i)$ and $g = \sum_j b_j \phi(x_j)$ is written $\langle f, g \rangle_\mathcal{H} = \sum_i \sum_j a_i b_j \langle \phi(x_i), \phi(x_j) \rangle = \sum_i \sum_j a_i b_j k(x_i, x_j)$. An RKHS specifies the kernel function $k$. This answers the first question.
- solving for $f = \sum_i a_i \phi(x_i)$ amounts to finding $a_i$. The reproducing property of RKHS yields linear systems $\langle f, \phi(x_i) \rangle_\mathcal{H} = \sum_j a_j k(x_j, x_i) = f(x_i)$ at each data point $x_i$.

These points are elaborated through a discussion of the RKHS below.
Definition 4. From Riesz representation lemma, \( f(x) \) can be written in the form

\[
f(x) = \delta_x(f) = \langle f, y_x \rangle.
\]

Importantly, observe that \( y_x \in \mathcal{H} \) does not depend on \( f \), and is a function of \( x \in \mathcal{X} \). Here, let \( k(x, \cdot) := y_x \), then \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is called the reproducing kernel which satisfies

\[
\forall f \in \mathcal{H}, \ x \in \mathcal{X} : \ f(x) = \langle f, k(x, \cdot) \rangle.
\]

This is referred to as the reproducing property.

Of particular note is that \( k_x := k(x, \cdot) \in \mathcal{H} \), the reproducing property yields

\[
k(x, y) := k_x(y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = \langle k(y, \cdot), k(x, \cdot) \rangle_{\mathcal{H}} = k(y, x). \tag{1}
\]

Theorem 2. The reproducing kernel of an RKHS exists and is unique.

Proof. From Riesz representation lemma, \( k(x, \cdot) \) exists and is unique for all \( x \in \mathcal{X} \). In addition, the RKHS is a Hilbert space with a specific inner product by definition. Thus, the unique reproducing kernel is given by (1).

Definition 5. Let \( \mathcal{X} \) be a non-empty set. A function \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is called a kernel if there exists a Hilbert space \( \mathcal{H} \) and a map \( \phi: \mathcal{X} \rightarrow \mathcal{H} \) such that for all \( x_i, x_j \in \mathcal{X} \):

\[
k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}.
\]

We call \( \phi(\cdot) \) a feature map and \( \mathcal{H} \) a feature space of \( k \).

Definition 6. A positive definite kernel \( k(\cdot, \cdot) \) is a symmetric function satisfying

\[
\sum_i \sum_j c_i c_j k(x_i, x_j) \geq 0, \quad \forall c_i, c_j \in \mathbb{R} \text{ and } \forall x_i, x_j \in \mathcal{X}.
\]

Equivalent, let \( K \) be a Gram matrix with elements \( K_{ij} = k(x_i, x_j) \). The definition of positive definiteness is equivalent to the positive semidefiniteness of \( K \):

\[
c^T K c \geq 0 \quad \text{with} \quad c = [c_1 \ c_2 \ \cdots]^T.
\]
Lemma 2. Inner products and kernels are positive definite functions.

Proof. For any inner product \( \langle x_i, x_j \rangle_H := \langle \phi(x_i), \phi(x_j) \rangle \), one can verify that
\[
\sum_i \sum_j a_i a_j \langle x_i, x_j \rangle_H = \left\langle \sum_i a_i \phi(x_i), \sum_i a_i \phi(x_i) \right\rangle \\
\geq 0,
\]
holds true for all \( a_i, a_j \in \mathbb{R} \).

This also implies that every reproducing kernel \( k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle \) is positive definite by replacing \( x_i = k(x_i, \cdot) \) and \( x_j = k(x_j, \cdot) \) in the above lemma.

It can be seen that \( k(\cdot, \cdot) \) may not be bilinear; however, it shares some properties of an inner product.

Lemma 3. For a positive definite \( k(\cdot, \cdot) \) and \( x_i, x_j \in X \), it holds that
\[
k(x_i, x_j)^2 \leq k(x_i, x_i) k(x_j, x_j).
\]

Proof. Let \( \lambda_i \) and \( v_i \) be the eigenvalues and eigenvalues of \( K \), one can equivalently show
\[
(e_i^T Ke_j)^2 \leq e_i^T Ke_i e_j^T Ke_j \\
\left( \sum_t \lambda_t e_i^T v_t e_j^T v_t \right)^2 \leq \left( \sum_t \lambda_t e_i^T v_t e_i^T v_t \right) \left( \sum_t \lambda_t e_j^T v_t e_j^T v_t \right) \\
\left[ \sum_t (\sqrt{\lambda_t} e_i^T v_t) (\sqrt{\lambda_t} e_j^T v_t) \right]^2 \leq \left[ \sum_t (\sqrt{\lambda_t} e_i^T v_t) \right] \left[ \sum_t (\sqrt{\lambda_t} e_j^T v_t) \right].
\]
The last line is the Cauchy-Schwarz inequality, and \( K \) has nonnegative eigenvalues.

We have seen that each RKHS has a unique kernel (Theorem 2), the converse is also true.

Theorem 3. Every kernel has a unique RKHS.
Sketch proof. We can show that the RKHS of linear combinations \( \sum_i a_i k(x_i, \cdot) \) is the one. Suppose that
\[
f = \sum_i a_i k(x_i, \cdot) \quad \text{and} \quad g = \sum_i b_i k(x_i, \cdot).
\]
Clearly, \( k \) is a reproducing kernel for this RKHS as \( \langle f, k(x, \cdot) \rangle_H = \sum_i a_i k(x_i, x) = f(x) \). It also follows that
\[
\langle f, g \rangle_H = \sum_i \sum_j a_i b_j k(x_i, x_j) = \sum_j b_j f(x_j) = \sum_i a_i g(x_i).
\]
The remainder is to show that the completion of the RKHS is equivalent to itself. \( \square \)

**Corollary 1.** Consider a subspace of an RKHS \( \mathcal{H} \) induced by \( \mathcal{X} = \{x_1, x_2, \cdots, x_n\} \), which consists of all linear combinations of \( g_1, g_2, \cdots, g_m \in \mathcal{H} \). Let \( k \) be the reproducing kernel of \( \mathcal{H} \), then the subspace is also an RKHS with the same reproducing kernel.

**Proof.** For any \( f = \sum_{i=1}^m a_i g_i \) in the subspace, we have \( \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^m a_i \langle g_i, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \) for all \( x \in \mathcal{X} \). Thus, \( k \) is the reproducing kernel of the subspace. \( \square \)

**Theorem 4** (Representer theorem). Given an empirical risk function \( V(f, \mathcal{D}) \) on data \( \mathcal{D} \) and a real-valued monotonic increasing function \( g(\cdot) \). Let \( \mathcal{H} \) be an RKHS, the solution for the minimization problem
\[
\min_{f \in \mathcal{H}} V(f, \mathcal{D}) + g(\|f\|)
\]
ads a representation of the form
\[
f_\star = \sum_i \alpha_i k(\cdot, x_i), \quad \forall x_i \in \mathcal{D}.
\]

**Proof sketch.** An arbitrary function \( f \in \mathcal{H} \) can be decomposed into the component in the RKHS induced by the data and the orthogonal component \( v \):
\[
f = \sum_i \alpha_i k(\cdot, x_i) + v.
\]
Since $\langle v, k(\cdot, x_i) \rangle = 0$, the orthogonal component $v$ does not impact the estimation of $\alpha_i$. Furthermore,

$$g(\|f\|) = g\left(\left\| \sum_i \alpha_i k(\cdot, x_i) + v \right\|\right)$$

$$= g\left(\sqrt{\left\| \sum_i \alpha_i k(\cdot, x_i) \right\|^2 + \|v\|^2}\right)$$

$$\geq g\left(\left\| \sum_i \alpha_i k(\cdot, x_i) \right\|\right).$$

Thus, the induced component alone yields the optimal solution, i.e., $v = 0$. $\Box$