Background on RKHS

Luca Raffo
EPFL, Institute of Mathematics - luca.raffo@epfl.ch
Capital Fund Management - luca.raffo@cfm.com

June 2025

Contents

1 RKHS **1**
1.1 Basics and representer theorem 1
1.2 MMD 6

1 RKHS

An RKHS (*reproducing kernel hilbert space*) is an Hilbert space of functions $\mathcal{H}$ which has some interesting properties. We will see how we can leverage on these spaces to define the MMD distance.

We suggest consulting [2] or [1] for a more profound understanding.

1.1 Basics and representer theorem

In this subsection we review the main definitions of RKHS theory and state the important *representer theorem.*

Definition 1.1. A Hilbert space $\mathcal{H}$ is a complete, possibly infinite-dimensional linear space endowed with a inner product.

In the following we will only work with functions defined on $\mathbb{R}^d$. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces.

A norm in $\mathcal{H}$ can be naturally defined from the given inner product, as $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$. Our norm will always be assumed to be the one arising from the latter. Furthermore, we always assume that $\mathcal{H}$ is separable (contains a countable dense subset) so that $\mathcal{H}$ has a countable orthonormal basis[1].

The main tool when dealing with these spaces is the *Rieszs representaiton theorem*.

Theorem 1.2 (Riesz Representation Theorem I)**.** Let $\mathcal{H}$ be a Hilbert space over $\mathbb{R}$, and let $\Lambda : \mathcal{H} \to \mathbb{R}$ be a continuous linear functional. Then, there exists a unique element $g \in \mathcal{H}$ such that for all $f \in \mathcal{H}$:

$$\Lambda(f) = \langle f, g\rangle_{\mathcal{H}}.$$

Moreover, $\|\Lambda\| = \|g\|_{\mathcal{H}}$.

This result allows us to represent any continuous linear functional on $\mathcal{H}$ as an inner product with a unique element in $\mathcal{H}$.

Theorem 1.3 (Riesz Representation Theorem II)**.** Let $\mathcal{H}$ be a Hilbert space, and let $h \in \mathcal{H}$. Then:

$$\sup_{\|f\|_{\mathcal{H}}\leq 1} \langle f, h\rangle_{\mathcal{H}} = \|h\|_{\mathcal{H}}.$$

That is, the operator norm of the linear functional $f \mapsto \langle f, h\rangle_{\mathcal{H}}$ is equal to the norm of h in $\mathcal{H}$.

Example 1.4. Let us consider the class of square integrable functions on the interval $[a, b]$, denoted by $L^2[a, b]$. We define the inner product as

$$\langle f, g\rangle = \int_a^b f(x)g(x)\,dx$$

with associated norm

$$\|f\| = \left(\int_a^b f^2(x)\,dx\right)^{1/2}$$

[1] This is just a technical condition used in the proof of the *representer theorem*.

It can be checked that this space is complete, so it is a Hilbert space. However, there is one problem with the functions in this space. Consider trying to evaluate the function $f(x)$ at the point $x = k$. There exists a function g in the space defined as follows:

$$g(x) = \begin{cases} c & \text{if } x = k \\ f(x) & \text{otherwise.} \end{cases}$$

Because it differs from f only at one point, g is clearly still square-integrable, and moreover, $\|f - g\| = 0$. In fact, we can set the constant c (or, more generally, the value of $g(x)$ at any finite number of points) to an arbitrary real value. What this means is that a condition on the integrability of the function is not strong enough to guarantee that we can use it predictively, since prediction requires evaluating the function at a particular data value. This characteristic is what will differentiate reproducing kernel Hilbert spaces from ordinary Hilbert spaces, as we discuss in the following.

Definition 1.5. An *evaluation functional* over a Hilbert space of functions $\mathcal{H}$ is a linear functional

$$\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$$

that evaluates each function in the space at the point t, or

$$\mathcal{F}_t[f] = f(t) \quad \text{for all } f \in \mathcal{H}.$$

Definition 1.6. A Hilbert space $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) if the evaluation functionals are bounded, i.e., if for all t there exists some $M > 0$ such that

$$|\mathcal{F}_t[f]| = |f(t)| \leq M\|f\|_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H},$$

which means that the norm topology is stronger than the topology induced by pointwise convergence.

While this condition might seem obscure or specific, it is actually quite general and is the weakest possible condition that ensures us both the existence of an inner product and the ability to evaluate each function in the space at every point in the domain.

In practice, it is difficult to work with this definition directly. We would like to establish an equivalent notion that is more useful in practice. To do this, we will need the *reproducing kernel* from which the reproducing kernel Hilbert space takes its name.

Firstly, from the definition of the reproducing kernel Hilbert space, we get that evaluations are linear and bounded functionals, hence we can apply the Riesz representation theorem.

Proposition 1.7. If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists a function $k_t \in \mathcal{H}$ (called the *representer of* t) with the *representing property*

$$\mathcal{F}_t[f] = \langle k_t, f \rangle_{\mathcal{H}} = f(t) \quad \text{for all } f \in \mathcal{H}.$$

This allows us to represent our linear evaluation functional by taking the inner product with an element of $\mathcal{H}$. Since k_t is a function in $\mathcal{H}$, by the representing property, for each $x \in X$ we can write

$$k_t(x) = \langle k_t, k_x \rangle_{\mathcal{H}}.$$

We take this to be the definition of reproducing kernel in $\mathcal{H}$.

Definition 1.8. The *reproducing kernel of* $\mathcal{H}$ is a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined by

$$k(t, x) := k_t(x).$$

Definition 1.9. A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *reproducing kernel* if it is symmetric, i.e. $k(x, y) = k(y, x)$, and positive definite:

$$\sum_{i,j=1}^{n} c_i c_j k(t_i, t_j) \geq 0$$

for any $n \in \mathbb{N}$ and choice of $t_1, \ldots, t_n \in X$ and $c_1, \ldots, c_n \in \mathbb{R}$.

Having this general notion of a reproducing kernel is important because it allows us to define an RKHS in terms of its reproducing kernel, rather than attempting to derive the kernel from the definition of the function space directly. The following theorem formally establishes the relationship between the RKHS and a reproducing kernel.

Proposition 1.10. A RKHS defines a corresponding reproducing kernel. Conversely, a reproducing kernel defines a unique RKHS.

Let us move to the main result for this subsection. Let us suppose having N data points $(x_i, y_i)_{i=1,\dots,N}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We want to find the function in $\mathcal{H}$ that best interpoles the data while not being too complex, which formally is translated into:

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

for a loss function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and a constant $\lambda \in \mathbb{R}$. Optimizing through a vector space with an infinite number of dimension is a priori undoable on a computer, but we have the following theorem, known as *representer theorem* which allows us to move directly to a finite dimension optimization problem.

Theorem 1.11. The minimizer over the RKHS $\mathcal{H}$, of the regularized empirical loss functional

$$f^* := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$

can be represented by the expression

$$f^*(x) = \sum_{i=1}^{N} c_i K(x_i, x),$$

for some N-tuple $(c_1, \dots, c_N) \in \mathbb{R}^N$. Hence, minimizing over the (possibly infinite-dimensional) Hilbert space boils down to minimizing over $\mathbb{R}^{N \times d}$.

For the proof we suggest reading [1].

As a fimal remark, we show some examples of commonly used RKHS.

1. *Linear kernel* $k(x, x') = x \cdot x'$
2. *Gaussian kernel* $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right), \quad \sigma > 0$
3. *Polynomial kernel* $k(x, x') = (x \cdot x' + 1)^d, \quad d \in \mathbb{N}$

1.2 MMD

We now introduce the *Maximum Mean Discrepancy* (MMD), a distance between probability measures defined via RKHS embeddings.

Definition 1.12. Let $\mathcal{H}$ be a RKHS over $\mathbb{R}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The *Maximum Mean Discrepancy* (MMD) between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$\mathrm{MMD}_{\mathcal{H}}(\mu, \nu) := \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{y \sim \nu}[f(y)] \right).$$

This quantity measures how well functions in $\mathcal{H}$ can distinguish between μ and ν. Crucially, it admits a closed-form expression due to the RKHS structure.

Proposition 1.13. Let $\mu_k := \mathbb{E}_{x \sim \mu}[k(x, \cdot)] \in \mathcal{H}$ denote the *mean embedding* of μ in the RKHS $\mathcal{H}$. Then,

$$\mathrm{MMD}_{\mathcal{H}}(\mu, \nu) = \|\mu_k - \nu_k\|_{\mathcal{H}}.$$

Proof. For any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq 1$, we have:

$$\mathbb{E}_{x \sim \mu}[f(x)] = \langle f, \mu_k \rangle_{\mathcal{H}}, \quad \mathbb{E}_{y \sim \nu}[f(y)] = \langle f, \nu_k \rangle_{\mathcal{H}}.$$

Hence,

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\langle f, \mu_k - \nu_k \rangle_{\mathcal{H}} \right) = \|\mu_k - \nu_k\|_{\mathcal{H}},$$

by the Riesz representation theorem. □

The MMD can thus be computed explicitly using the kernel function:

$$\mathrm{MMD}^2_{\mathcal{H}}(\mu, \nu) = \mathbb{E}_{x, x' \sim \mu}[k(x, x')] + \mathbb{E}_{y, y' \sim \nu}[k(y, y')] - 2\mathbb{E}_{x \sim \mu, y \sim \nu}[k(x, y)].$$

Regarding empirical estimation, if $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m} \sum_{j=1}^m \delta_{y_j}$, the empirical (biased) estimator of MMD^2 is:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) + \frac{1}{m^2} \sum_{i,j=1}^m k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m k(x_i, y_j).$$

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[2] Jonathan H. Manton and Pierre-Olivier Amblard. A primer on reproducing kernel hilbert spaces, 2015.