A Gentle Introduction to Reproducing Kernel Hilbert Spaces

Eloi Tanguy, June 2025

We delve into theoretical considerations around Reproducing Kernel Hilbert Spaces (RKHS), which is a field of Mathematics that is relatively far removed from Optimal Transport on which the rest of this thesis is focused. To ease the reader into the topic, we will briefly introduce the field of RKHS theory. This introductory chapter is based on blackboard talks given by the author at the MAP5 laboratory in Paris, and is intended to be accessible to a wide audience.

1 Reproducing Kernel Hilbert Spaces

There are many different equivalent definitions of a Reproducing Kernel Hilbert Space (RKHS): in particular, it is possible to begin with a kernel and to construct the associated RKHS, or to begin with a Hilbert space with certain properties and to "discover" its kernel. We will focus on the latter viewpoint, and consider a Hilbert space $(H, \langle \cdot, \cdot \rangle_H)$ of functions $\mathcal{X} \longrightarrow \mathbb{R}$, where $\mathcal{X}$ is a set without a particular structure. First, we remind in Definition 1 the definition of a Hilbert space.

Definition 1. A **Hilbert space** is a vector space H over $\mathbb{R}$ equipped with an inner product $\langle \cdot, \cdot \rangle_H$ that is **complete**, which is to say that every Cauchy sequence in H converges (for the topology induce by the norm $\|h\|_H := \sqrt{\langle h, h \rangle_H}$) to an element of H.
A **Cauchy sequence** is a sequence $(h_n)_{n \in \mathbb{N}}$ in H such that for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$, we have $\|h_n - h_m\|_H \leq \varepsilon$.

For the sake of simplicity, we consider spaces of real-valued functions and Hilbert spaces over $\mathbb{R}$, but the definitions can be extended to $\mathbb{C}^d$-valued functions and Hilbert spaces over $\mathbb{C}$. Given a Hilbert space of functions $\mathcal{X} \longrightarrow \mathbb{R}$, the **evaluation map** at a point $x \in \mathcal{X}$, defined by

$$\delta_x := \left\{ \begin{array}{ccc} H & \longrightarrow & \mathbb{R} \\ h & \longmapsto & h(x) \end{array} \right. \in \mathcal{L}(H, \mathbb{R}),$$

where $\mathcal{L}(H, \mathbb{R})$ is the space of linear maps from H to $\mathbb{R}$, is of particular interest in RKHS theory:

Definition 2. A Hilbert space H of functions $\mathcal{X} \longrightarrow \mathbb{R}$ is said to be a **Reproducing Kernel Hilbert Space** (RKHS) if the evaluation map δ_x is continuous for every $x \in \mathcal{X}$. This means that for every $x \in \mathcal{X}$, there exists a constant $C_x \geq 0$ such that for all $h \in H$, we have:

$$|\delta_x(h)| = |h(x)| \leq C_x \|h\|_H.$$

We can write this condition $\delta_x \in H'$, where H' is the space of continuous linear maps from H to $\mathbb{R}$ (the topological dual of H).

In a RKHS, the norm is "strong" in the sense that convergence in H implies pointwise convergence of functions. In other words, if a sequence of functions $(h_n)_{n \in \mathbb{N}} \in H^{\mathbb{N}}$ is such that $\|h_n - h\|_H \xrightarrow[n \longrightarrow +\infty]{} 0$ for some $h \in H$, then for every $x \in \mathcal{X}$, we have:

$$|h_n(x) - h(x)| = |\delta_x(h_n - h)| \leq C_x \|h_n - h\|_H \xrightarrow[n \longrightarrow +\infty]{} 0.$$

A simple example of a RKHS space is the space of "band-limited" functions, which we present in Example 1.

Example 1. We consider the space $H := \left\{ f \in \mathcal{C}^0(\mathbb{R}) \cap L^2(\mathbb{R}) : \operatorname{supp} \hat{f} \subset [-a, a] \right\}$ of continuous functions f on $\mathbb{R}$ verifying $\int_{\mathbb{R}} f^2 < +\infty$ and whose Fourier transform $\hat{f}$ is supported in the interval $[-a, a]$. We equip H with the L^2 inner product: $\langle f, g \rangle_H := \int_{\mathbb{R}} fg$. The space $(H, \langle \cdot, \cdot, \rangle)_H$ is an RKHS: we apply Definition 2 and fix $x \in \mathbb{R}$ and $f \in H$, with the convention that

$$\hat{f} := \omega \longmapsto \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-i\omega x} \mathrm{d}x,$$

we compute using the Fourier inversion formula, the Cauchy-Schwarz inequality, and the Parseval identity:

$$\begin{aligned} |\delta_x(f)| &= \left| \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{f}(\omega) e^{i\omega x} \mathrm{d}\omega \right| \leq \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} |\hat{f}(\omega)| \mathrm{d}\omega \\ &\leq \frac{1}{\sqrt{2\pi}} \sqrt{\int_{-a}^{a} |\hat{f}(\omega)|^2 \mathrm{d}\omega} \sqrt{\int_{-a}^{a} 1^2 \mathrm{d}\omega} = \sqrt{\frac{a}{\pi}} \sqrt{\int_{\mathbb{R}} |\hat{f}(\omega)|^2 \mathrm{d}\omega} = \sqrt{\frac{a}{\pi}} \|f\|_H. \end{aligned}$$

To define the kernel of an RKHS, we will use a well-known result in Hilbert space theory:

Theorem 1. Riesz Representation Theorem [RAG05, Theorem 13.31] Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space, for every continuous linear functional $\ell \in H'$, there exists a unique element $R\ell \in H$ such that for all $h \in H$, we have $\ell(h) = \langle h, R\ell \rangle_H$. The Riesz operator $R : H' \longrightarrow H$ is an isometric isomorphism[a].

[a] i.e. $\|R\ell\|_H = \|\ell\|_{H'} := \sup_{\|h\|_H \leq 1} |\ell(h)|$ and R is linear and bijective.

Using Theorem 1, we can define the kernel of an RKHS using the evaluation maps, which by assumption are continuous:

Definition 3. The **kernel** of an RKHS H is the map $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ defined by:

$$k := (x, y) \in \mathcal{X}^2 \longmapsto \langle R\delta_x, R\delta_y \rangle_H,$$

where $R\delta_x, R\delta_y \in H$ are the Riesz representations of δ_x, δ_y as in Theorem 1.

For convenience, the element $R\delta_x \in H$ is often denoted by $K_x := R\delta_x$. The map $x \longmapsto K_x$ is called the *canonical feature map*.Using the properties of the inner product $\langle \cdot, \cdot \rangle_H$ and of the Riesz representation, the following properties of the kernel k can be deduced:

Proposition 1. The kernel $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ of an RKHS H satisfies the following properties for any $x, y \in \mathcal{X}$:

1) Symmetry: $k(x, y) = k(y, x)$.
2) Positivity:
$$\forall n \in \mathbb{N}^*, \ \forall x_1, \cdots, x_n \in \mathcal{X}, \ \forall a \in \mathbb{R}^n, \ \sum_{i,j} a_i k(x_i, x_j) a_j \geq 0$$
3) Feature identity: $k(\cdot, x) = K_x$.
4) Reproducing property: $\forall h \in H, \ h(x) = \langle h, K_x \rangle_H$.
5) Self-reproducing property: $\langle k(\cdot, x), k(\cdot, y) \rangle_H$.

Proof. For 1), we have $k(x,y) = \langle K_x, K_y \rangle_H = \langle K_y, K_x \rangle_H = k(y,x)$. To show 2), we compute:

$$\sum_{i,j} a_i k(x_i, x_j) a_j = \sum_{i=1}^n a_i \left\langle K_{x_i}, \sum_{j=1}^n a_j K_{x_j} \right\rangle_H = \left\langle \sum_{i=1}^n a_i K_{x_i}, \sum_{j=1}^n a_j K_{x_j} \right\rangle_H = \left\| \sum_{i=1}^n a_i K_{x_i} \right\|_H^2 \geq 0.$$

For 3), since $K_y = R\delta_y$ it holds that $K_x(y) = \delta_y(K_x) = \langle K_x, K_y \rangle_H = k(y,x)$, concluding that $K_x = k(\cdot, x)$. Regarding 4), we use $K_x = R\delta_x$ and obtain $h(x) = \delta_x(h) = \langle h, K_x \rangle_H$. For 5), we apply 4) to $h := K_y$. □

A natural question is whether a suitable function $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ can be a reproducing kernel of an RKHS H. The answer is that the properties 1) and 2) of Proposition 1 are sufficient. Before stating the result, we introduce a notation for such functions:

Definition 4. A function $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ is said to be **symmetric-positive** if it verifies 1) and 2) of Proposition 1. In that case, we write $k \in S^2(\mathcal{X})$.

Theorem 2 (Moore-Aronszajn Theorem [Aro50])**.** Let $k \in S^2(\mathcal{X})$ be a symmetric-positive function. Then there exists a unique (up to isometry) RKHS $(H, \langle \cdot, \cdot \rangle_H)$ with kernel k.

The idea of the proof of Theorem 2 is to begin with the space

$$H_0 := \left\{ \sum_{i=1}^n a_i k(\cdot, x_i),\ n \in \mathbb{N},\ a \in \mathbb{R}^n,\ (x_1, \cdots, x_n) \in \mathcal{X}^n \right\},$$

equipped with the inner product:

$$\left\langle \sum_{i=1}^n a_i k(\cdot, x_i), \sum_{j=1}^m b_j k(\cdot, y_j) \right\rangle_{H_0} = \sum_{i,j} a_i k(x_i, y_j) b_j.$$

The space $(H_0, \langle \cdot, \cdot \rangle_{H_0})$ is only a pre-Hilbert space, and the technicality of the proof resides in studying its Hilbertian completion (limits of Cauchy sequences in H_0). A self-sufficient proof of Theorem 2 can be found in Jean-Philippe Vert's course notes. Some commonly used kernels are:

- The Gaussian (or RBF) kernel: $k(x,y) := \exp(-\|x-y\|_2^2/s^2)$
- The polynomial kernel: $k(x,y) := (c + \langle x, y \rangle_2)^d$
- The Laplace kernel: $k(x,y) := \exp(-\|x-y\|_2/s)$
- Radial kernels for some finite positive measure μ on $\mathbb{R}_+$: $k(x,y) := \int_0^{+\infty} e^{-t\|x-y\|_2^2} \mathrm{d}\mu(t)$
- Taylor kernels for coefficients $(a_n) \in \mathbb{R}_+^{\mathbb{N}}$ with sufficient decay: $k(x,y) := \sum_{n=0}^{+\infty} a_n \langle x, y \rangle^n$

2 Kernel Interpolation

In this section, we explain the concept of kernel interpolation, which allows for the approximation of a function $f : \mathcal{X} \longrightarrow \mathbb{R}$ whose values are known at a finite set of points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$. In other words, given values (y_i) and points (x_i), we want to find the simplest possible function of an RKHS H verifying $h(x_i) = y_i$ at each i. Again, we focus on the case of a real-valued functions for simplicity. Mathematically, the goal is to find a solution of the following exact interpolation problem:

$$\underset{\substack{h \in H \\ \forall i \in [\![1,n]\!],\ h(x_i) = y_i}}{\operatorname{argmin}} \quad \|h\|_H^2. \tag{2.1}$$

Minimising the norm $\|h\|_H$ can be understood as a way of finding the simplest possible function in H verifying the interpolation conditions. To motivate this intuition, we refer to [CS08, Theorem

4.48], which states that the norm associated to the RKHS induce by the Gaussian kernel on a bounded set of $\mathbb{R}^d$ dominates all Sobolev norms. In this setting, a small norm implies that the function is very regular, which corresponds to our intuitive term "simple function".

Before tackling the problem of Eq. (2.1), we first study the constraints in the case where a solution exists: we provide a characterisation of the condition that two functions $h_1, h_2 \in H$ be equal on a finite set of points $\{x_1, \dots, x_n\}$.

Proposition 2. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$ For any $h_1, h_2 \in H$, we have:

$$\forall i \in [\![1, n]\!],\ h_1(x_i) = h_2(x_i) \iff h_1 - h_2 \in W^\perp,\ W := \operatorname{Span}\left(k(\cdot, x_i)\right)_{i=1}^n$$

Proof. By the reproducing property of k (Proposition 1 item 4)), we have for each $i \in [\![1, n]\!]$:

$$(h_1 - h_2)(x_i) = 0 \iff \langle h_1 - h_2, k(\cdot, x_i)\rangle_H = 0 \iff h_1 - h_2 \in k(\cdot, x_i)^\perp,$$

concluding the proof by intersecting over $i \in [\![1, n]\!]$. □

Thanks to the characterisation of Proposition 2, we can reformulate the interpolation problem of Eq. (2.1) as a problem over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$. This way, the infinite-dimensional problem of Eq. (2.1) can be reduced to a finite-dimensional problem over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$.

Theorem 3. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$ and $(y_1, \dots, y_n) \in \mathbb{R}^n$. Consider the matrix $K \in \mathbb{R}^{n\times n}$ of entries $K_{i,j} := k(x_i, x_j)$, and assume that K is invertible. Then there is a unique solution to the exact kernel interpolation problem of Eq. (2.1), which is:

$$h^* = \sum_{i=1}^n a_i k(\cdot, x_i),\ \text{where}\ \ a = K^{-1}y, \tag{2.2}$$

with $y := (y_1, \cdots, y_n) \in \mathbb{R}^n$.

Proof. For existence, take $a := K^{-1}y$ and let $h_0 := \sum_{i=1}^n a_i k(\cdot, x_i)$. Since $Ka = y$, the function h_0 satisfies the interpolation conditions $h_0(x_i) = y_i$ for all $i \in [\![1, n]\!]$. Now we see from Proposition 2 that $h \in H$ verifies the constraints if and only if $h - h_0 \in W^\perp$, which can be re-written as $h = h_0 + h_\perp$ for some $h_\perp \in W^\perp$. We then have by orthogonality $\|h\|_H^2 = \|h_0\|_H^2 + \|h_\perp\|_H^2$, showing that for any $h \in H$ verifying the constraints, we have $\|h\|_H^2 \geq \|h_0\|_H^2$, which shows that h_0 is indeed a solution of the interpolation problem Eq. (2.1).

For uniqueness, if $h \in H$ verifies the constraints, we have seen that we can write $h = h_0 + h_\perp$ for some $h_\perp \in W^\perp$. If $h \notin W$ then $h_\perp \neq 0$ and thus $\|h\|_H^2 > \|h_0\|_H^2$, which shows that a solution h must be in W. Writing such a solution as $h = \sum_{i=1}^n b_i k(\cdot, x_i)$ for some $b \in \mathbb{R}^n$, using the constraints we observe that $Kb = y$, which implies that $a = b$ by invertibility, showing that the expression in Eq. (2.2) is the unique solution. □

The condition of invertibility of the matrix K in Theorem 3 is crucial, as it ensures that the interpolation problem has a unique solution. For some specific kernels, this condition can be guaranteed for distinct points $(x_i)_{i=1}^n$ (see [Mic86, Theorem 2.3] or [Set22, Theorem 3.5] for example). In general, it is natural to resort to a regularisation technique, which consists in replacing the exact interpolation problem Eq. (2.1) by the following approximate interpolation problem:

$$\operatorname*{argmin}_{h\in H}\ \lambda\|h\|_H^2 + \sum_{i=1}^n (h(x_i) - y_i)^2, \tag{2.3}$$

which we interpret as a kernel regression problem. Using the same tools as in Theorem 3, we will reduce this problem to optimisation over the coefficients $a \in \mathbb{R}^n$ of a function $h \in W$:

Theorem 4. For $(x_1, \dots, x_n) \in \mathcal{X}^n$, $(y_1, \dots, y_n) \in \mathbb{R}^n$ and $\lambda > 0$. A function $h \in H$ is a solution of the kernel regression problem of Eq. (2.3) if and only if it can be written $h = \sum_{i=1}^n a_i^* k(\cdot, x_i)$ for some $a^* \in \mathbb{R}^n$ which is a solution to:

$$\operatorname*{argmin}_{a \in \mathbb{R}^n} \lambda a^\top K a + \|Ka - y\|_2^2, \tag{2.4}$$

where $K \in \mathbb{R}^{n \times n}$ is the matrix of entries $K_{i,j} := k(x_i, x_j)$ and $y := (y_1, \dots, y_n) \in \mathbb{R}^n$. A vector $a \in \mathbb{R}^n$ is a solution of Eq. (2.4) if and only if it verifies:

$$K\left((K + \lambda I)a - y\right) = 0. \tag{2.5}$$

Proof. We begin with the second statement about solutions of Eq. (2.4). By the symmetry and positivity properties of k (Proposition 1 items 1) and 2)), the matrix K is symmetric positive semi-definite, and thus the energy $J := a \longmapsto \lambda a^\top K a + \|Ka - y\|_2^2$ is convex. Furthermore, it is differentiable and we compute $\nabla J(a) = 2\lambda K a + 2K(Ka - y)$, and Eq. (2.5) is equivalent to the condition $\nabla J(a) = 0$. Since K is symmetric positive semi-definite, the matrix $K + \lambda I$ is symmetric positive definite thanks to the assumption $\lambda > 0$, and thus the the vector $a_0 := (K + \lambda I)^{-1} y$ is a solution of Eq. (2.4).

Now we consider the element $h_0 := \sum_{i=1}^n a_i^{(0)} k(\cdot, x_i) \in W$. For any $h \in H$, take its orthogonal projection $h_W := P_W(h)$ onto W (using the closedness and convexity of W and the Hilbert projection Theorem ([Rud87, Theorem 4.11])). We have $\|h_W\|_H \leq \|h\|_H$ and since $h - h_W \in W^\perp$, we deduce from Proposition 2 that $h_W(x_i) = h(x_i)$ for all $i \in [\![1, n]\!]$. This shows that h_W has lower cost than h in the kernel regression problem:

$$\lambda \|h_W\|_H^2 + \sum_{i=1}^n (h_W(x_i) - y_i)^2 \leq \lambda \|h\|_H^2 + \sum_{i=1}^n (h(x_i) - y_i)^2,$$

with a strict inequality if $h \notin W$, and thus if a solution of the kernel regression problem exists, it must be in W. Given the definition of W, we conclude that $h \in H$ is a solution of Eq. (2.3) if and only if it is in W and is a solution of Eq. (2.3), which is equivalent to being of the form $h = \sum_{i=1}^n a_i k(\cdot, x_i)$ for some $a \in \mathbb{R}^n$ solution of Eq. (2.4), concluding the proof. □

3 Kernel Mean Embedding and Maximum Mean Discrepancy

A cornerstone application of RKHS theory in Machine Learning is the Kernel Mean Embedding (KME), which embeds a probability measure μ into an RKHS H. The norm of of the difference between two embeddings is then a measure of discrepancy between the two measures, which is called the Maximum Mean Discrepancy (MMD) [Gre+06; Smo+07; Mua+17]. In this section, we provide a simple definition of the KME without the use of Bochner integrals (see [Hyt+16, Section 1.2.a], or [CS08, Appendix A.5.4] for references on this notion). We begin with a technical lemma that will allow us to define the KME. Given a measurable space $\mathcal{X}$, we say that a kernel $k : \mathcal{X}^2 \longrightarrow \mathbb{R}$ is measurable if for every $x \in \mathcal{X}$, the function $k(\cdot, x) : \mathcal{X} \longrightarrow \mathbb{R}$ is measurable. We will write $\mathcal{P}_k(\mathcal{X})$ the set of probability measures μ on $\mathcal{X}$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\mu(\mathrm{d}x) < +\infty$.

Lemma 1. Let $(H, \langle \cdot, \cdot \rangle_H)$ be a RKHS with a measurable kernel k. Let $\mu \in \mathcal{P}_k(\mathcal{X})$, then the linear map defined by:

$$\ell_\mu := \left\{ \begin{array}{ccc} H & \longrightarrow & \mathbb{R} \\ h & \longmapsto & \int_{\mathcal{X}} h \mathrm{d}\mu \end{array} \right. \tag{3.1}$$

is continuous.

Proof. First, we show that ℓ_μ is well-defined. Thanks to the Moore-Aronszajn Theorem, we can write any $h \in H$ as a limit (in H and in particular pointwise) of functions $h_n \in H$ of the form $h_n = \sum_{i=1}^{m_n} a_i^{(n)} k(\cdot, x_i^{(n)})$, and thus h is measurable (see [CS08, Lemma 4.24] for additional details).

We now show that for any $h \in H$, we have $h \in L^1(\mu)$, using the reproducing property (Proposition 1 item 4)), the Cauchy-Schwarz inequality and the assumption $\mu \in \mathcal{P}_k(\mathcal{X})$:

$$\begin{aligned}\int_{\mathcal{X}} |h(x)| \mathrm{d}\mu(x) &= \int_{\mathcal{X}} |\langle h, k(\cdot, x)\rangle_H| \mathrm{d}\mu(x) \\ &\leq \int_{\mathcal{X}} \|h\|_H \|k(\cdot, x)\|_H \mathrm{d}\mu(x) \\ &= \|h\|_H \int_{\mathcal{X}} \sqrt{k(x,x)} \mathrm{d}\mu(x) < +\infty. \end{aligned} \tag{3.2}$$

The left hand-side term exists in $\mathbb{R}_+ \cup \{+\infty\}$ by measurability, and the computations above show that it is finite. Thus, ℓ_μ is well-defined, and by linearity of integration, it is clearly linear. We now show continuity using Eq. (3.2), which yields that for every $h \in H$, we have $|\ell_\mu(h)| \leq \|h\|_H \int_{\mathcal{X}} \sqrt{k(x,x)} \mathrm{d}\mu(x)$, which proves (Lipschitz) continuity. □

Thanks to Lemma 1, we can define the Kernel Mean Embedding of a probability measure μ as the Riesz representation of ℓ_μ.

Definition 5. Let $(H, \langle \cdot, \cdot \rangle_H)$ be an RKHS with a measurable kernel k, and let $\mu \in \mathcal{P}_k(\mathcal{X})$. The Kernel Mean Embedding of μ is the element of H defined by: $\mathrm{M}(\mu) := R\ell_\mu \in H$, where R is the Riesz representation operator and $\ell_\mu \in H'$ is defined in Eq. (3.1).

The KME allows comparison of two probability measures using the norm of the difference of their embeddings in H:

Definition 6. Let $(H, \langle \cdot, \cdot \rangle_H)$ be an RKHS with a measurable kernel k and $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$. The Maximum Mean Discrepancy (MMD) between μ and ν is defined as:

$$\mathrm{MMD}(\mu, \nu) := \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H. \tag{3.3}$$

We can rewrite the MMD as an Integral Probability Metric [Mül97]:

Proposition 3. For probability measures $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$, we have:

$$\mathrm{MMD}(\mu, \nu) = \sup_{h \in H, \|h\|_H \leq 1} \int_{\mathcal{X}} h(x) \mathrm{d}\mu(x) - \int_{\mathcal{X}} h(x) \mathrm{d}\nu(x). \tag{3.4}$$

Proof. By definition of the KME, we have:

$$\sup_{h \in H, \|h\|_H \leq 1} \int_{\mathcal{X}} h(x) \mathrm{d}\mu(x) - \int_{\mathcal{X}} h(x) \mathrm{d}\nu(x) = \sup_{h \in H, \|h\|_H \leq 1} \langle h, \mathrm{M}(\mu) - \mathrm{M}(\nu) \rangle_H,$$

and the supremum of the right hand-side is attained for $h^* = \frac{\mathrm{M}(\mu) - \mathrm{M}(\nu)}{\|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H}$, called the "witness function" of the MMD. We then compute:

$$\langle h^*, \mathrm{M}(\mu) - \mathrm{M}(\nu) \rangle_H = \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H = \mathrm{MMD}(\mu, \nu),$$

concluding the proof. □

It is clear that the MMD verifies the non-negativity, symmetry and triangle inequality axioms for a distance, and that if $\mu = \nu$ then $\mathrm{MMD}(\mu, \nu) = 0$. The converse is not true in general, and is seen in literature as a property of the kernel k used to define the MMD (see [Sri+10; SFL11]):

Definition 7. A measurable kernel k is said to be characteristic if for any $\mu, \nu \in \mathcal{P}_k(\mathcal{X})$, we have $\mathrm{MMD}(\mu, \nu) = 0$ if and only if $\mu = \nu$ (for the MMD associated to k).

We now focus on the case of discrete measures and take $\mu := \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}_k(\mathcal{X})$. By linearity of the Riesz operator we have:

$$\mathrm{M}(\mu) = R\ell_\mu = \sum_{i=1}^n a_i R\ell_{\delta_{x_i}} = \sum_{i=1}^n a_i R\left(h \longmapsto h(x_i)\right) = \sum_{i=1}^n a_i k(\cdot, x_i),$$

where we used Proposition 1 item 3). Given now two empirical probability measures $\mu := \sum_{i=1}^n a_i \delta_{x_i}$, $\nu := \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}_k(\mathcal{X})$, we have by the self-reproducing property (Proposition 1 item 5)):

$$\begin{aligned}
\mathrm{MMD}^2(\mu,\nu) &= \|\mathrm{M}(\mu) - \mathrm{M}(\nu)\|_H^2 \\
&= \left\langle \sum_{i=1}^n a_i k(\cdot, x_i) - \sum_{j=1}^n b_j k(\cdot, y_j), \sum_{i'=1}^n a_{i'} k(\cdot, x_{i'}) - \sum_{j'=1}^n b_{j'} k(\cdot, y_{j'}) \right\rangle_H \\
&= \sum_{i=1}^n \sum_{i'=1}^n a_i k(x_i, x_{i'}) a_{i'} - 2\sum_{i=1}^n \sum_{j=1}^m a_i k(x_i, y_j) b_j + \sum_{j=1}^m \sum_{j'=1}^m b_j k(y_j, y_{j'}) b_{j'}.
\end{aligned}$$

Writing the vectors $a := (a_1, \cdots, a_n) \in \mathbb{R}^n$ $b := (b_1, \cdots, b_m) \in \mathbb{R}^m$ and the matrices $K_{xx} := [k(x_i, x_{i'})]_{i,i'} \in \mathbb{R}^{n\times n}$, $K_{xy} := [k(x_i, y_j)]_{i,j} \in \mathbb{R}^{n\times m}$ and $K_{yy} := \left[k(y_j, y_{j'})\right]_{j,j'} \in \mathbb{R}^{m\times m}$, we can rewrite the MMD as:

$$\mathrm{MMD}^2(\mu,\nu) = a^\top K_{xx} a - 2a^\top K_{xy} b + b^\top K_{yy} b.$$

References

[Aro50] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404 (cit. on p. 3).

[CS08] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008 (cit. on pp. 3, 5).

[Gre+06] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. "A kernel method for the two-sample-problem". In: *Advances in neural information processing systems* 19 (2006) (cit. on p. 5).

[Hyt+16] Tuomas Hytönen, Jan Van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach spaces*. Vol. 12. Springer, 2016 (cit. on p. 5).

[Mic86] Charles A Micchelli. "Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions". In: *Constr. Approx* 2 (1986), pp. 11–22 (cit. on p. 4).

[Mua+17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2 (2017), pp. 1–141 (cit. on p. 5).

[Mül97] Alfred Müller. "Integral probability metrics and their generating classes of functions". In: *Advances in applied probability* 29.2 (1997), pp. 429–443 (cit. on p. 6).

[RAG05] Steven Roman, S Axler, and FW Gehring. *Advanced linear algebra*. Vol. 3. Springer, 2005 (cit. on p. 2).

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill, Inc., 1987 (cit. on p. 5).

[Set22] Michio Seto. *A Fock space approach to the theory of strictly positive kernels*. 2022. arXiv: 2208.02980 [math.FA] (cit. on p. 4).

[Smo+07] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. "A Hilbert space embedding for distributions". In: *International conference on algorithmic learning theory*. Springer. 2007, pp. 13–31 (cit. on p. 5).

[SFL11] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." In: *Journal of Machine Learning Research* 12.7 (2011) (cit. on p. 6).

[Sri+10] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561 (cit. on p. 6).