Chapter 4, Part 2: Reproducing Kernel Hilbert Spaces (RKHSs)

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

A More Formal Look at Kernels

- Kernels are inner products between feature maps that form a similarity measure between inputs
- We would like to be able to construct valid kernels directly as similarity measures in way that ensures they imply a valid feature map
- What are the conditions on a similarity measure $k \colon \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ to ensure that it is a valid such kernel?
- To answers this, we need to get a lot more technical and introduce the concept of a **reproducing kernel Hilbert spaces (RKHS)**

Simple SVM (motivation)

We want to improve SVMs expressiveness using features. To understand which *ingredients* we need, let's start by a simple SVM. We fix $b = 0$ for simplicity and we assume data separability. Given a feature $\phi : \mathcal{X} \to \mathbb{R}^d$ we have

$$\text{minimise } \tfrac{1}{2}\|w\|^2 \text{ under } y_i w \cdot \phi(x_i) \geq 1, \, \forall i,$$

where $w \in \mathbb{R}^d$.

What if we want to play with infinite dimensional features (now $d = \infty$...)? How can we formalise this?

Inner Product

Instead of $w \in \mathbb{R}^d$, we now only assume $w \in \mathcal{W}$. Here $\mathcal{W}$ is a (possibly infinitely dimensional) **vector space**. To formulate our problem we need a norm and an inner product...

Definition 1 (Inner Product)

Let $\mathcal{W}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is said to be an **inner product** on $\mathcal{W}$ if it is

1. **symmetric**: $\langle w, w' \rangle = \langle w', w \rangle$
2. **(bi)linear**: $\langle \alpha w_1 + \beta w_2, w' \rangle_{\mathcal{W}} = \alpha \langle w_1, w' \rangle_{\mathcal{W}} + \beta \langle w_2, w' \rangle_{\mathcal{W}}$
3. **positive definite**: $\langle w, w \rangle_{\mathcal{W}} > 0$ if $w \neq 0$

A inner product induces a **norm** $\| \cdot \|_{\mathcal{W}}$ on $\mathcal{W}$ via

$$\|w\|_{\mathcal{W}} = \sqrt{\langle w, w \rangle_{\mathcal{W}}}$$

A vector space $\mathcal{W}$ with a inner product is called **pre-Hilbert** space.

Hilbert Space

The optimisation problem $\min_{w \in \mathcal{W}} \frac{1}{2}\|w\|^2$ with $w \in C$ (the constrained set) is looking for a **projection** of the origin on C. For a general pre-Hilbert space $\mathcal{W}$ such projection might not exist!
If C is convex and closed (as in our case), a sufficient condition for existence is **completeness**.

A space is **complete** if every Cauchy sequence converges (e.g., $\mathbb{R}$ is complete, $\mathbb{Q}$ is not!). This means that for every sequence w_n

$$\|w_n - w_m\|_{\mathcal{W}} \to 0 \quad \Longrightarrow \quad \exists w^\star \in \mathcal{W} \,:\, \|w_n - w^\star\|_{\mathcal{W}} \to 0\,.$$

A pre-Hilbert space $\mathcal{W}$ that is complete under the induced norm is called a **Hilbert space**.
You can think of a Hilbert space as a generalisation of Euclidean space that allows for infinite dimensionality while retaining important properties (e.g., existence of projections).

Optimisation problem

Given a Hilbert space $\mathcal{W}$, we can make sense of

$$\text{minimise } \tfrac{1}{2}\|w\|^2 \text{ under } y_i\langle w, \phi(x_i)\rangle_{\mathcal{W}} \geq 1,\ \forall i,$$

where $\phi : \mathcal{X} \to \mathcal{W}$ is the **feature** function.

For the SVMs, we saw that the optimal w was in the space generated by the x_i. Can we expect something similar now?

We define a subspace $\mathcal{W}_0$ of $\mathcal{W}$ as the closure of the linear span[1] of ϕ, namely

$$\mathcal{W}_0 = \overline{\mathrm{Span}\left\{\phi(x), x \in \mathcal{X}\right\}}.$$

The closure of the span is taken to ensure that $\mathcal{W}_0$ is complete, hence a Hilbert space.

[1] Given a set $V \subseteq \mathcal{W}$, $\mathrm{Span}V$ is the set of finite linear combinations of elements of V (namely $w = \sum_{i=1}^{N} c_i v_i$ for some $\{v_1, \ldots, v_N\}$ in V). The closure $\overline{\mathrm{Span}V}$ is obtained by adding all the limit points of sequences of elements of V.

Restriction to $\mathcal{W}_0$

We claim that the solution of the optimisation problem

$$\text{minimise } \tfrac{1}{2}\|w\|^2 \text{ under } y_i\langle w, \phi(x_i)\rangle_{\mathcal{W}} \geq 1,\ \forall i,$$

is in $\mathcal{W}_0$.

For every $w \in \mathcal{W}$ we have the orthogonal decomposition

$$w = w_0 + w_\perp\,,$$

where $w_0 \in \mathcal{W}_0$, and $\langle w_\perp, \phi(x)\rangle_{\mathcal{W}} = 0$, $\forall x \in \mathcal{X}$. If w satisfies the constraints, then w_0 satisfies the constraints as well! And

$$\|w_0\|^2 \leq \|w_0\|^2 + \|w_\perp\|^2 = \|w\|^2\,!$$

So w_0 is always *better* than w, and the optimal w is in $\mathcal{W}_0$!
Hence, we can **restrict** w to be in $\mathcal{W}_0$.

Functional space

Note that ϕ only appears in the problem inside inner products. We can define

$$h_w : \mathcal{X} \to \mathbb{R}\,, \quad x \mapsto \langle w, \phi(x) \rangle_{\mathcal{W}}\,.$$

For our classification problem, we are actually interested in a classifier in the form

$$x \mapsto \operatorname{sign} h_w(x)\,.$$

So, rather than looking for the optimal $w \in \mathcal{W}_0$, we could think of looking for the optimal function h_w. We define

$$\mathcal{H} = \{f_w \,:\, w \in \mathcal{W}_0\}\,.$$

$\mathcal{H}$ is a vector space. Let's now see that it can inherit from $\mathcal{W}_0$ a Hilbert space structure.

$\mathcal{H} \cong \mathcal{W}_0$

The mapping $\Xi : \mathcal{W}_0 \to \mathcal{H}$, $w \mapsto h_w$ is a linear **surjection**.

Let's now check that Ξ is also **injective**. Say that $h_w = h_{w'}$. Then $\langle h_w - h_{w'}, \phi(x) \rangle = h_w(x) - h_{w'}(x) = 0$, $\forall x$. So $w - w' \in \mathcal{W}_0^\perp$. But $w - w' \in \mathcal{W}_0$, and $\mathcal{W}_0 \cap \mathcal{W}_0^\perp = \{0\}$. So, $w = w'$.

Since Ξ is a **linear bijection**, we can define a inner product on $\mathcal{H}$

$$\langle h, h' \rangle_{\mathcal{H}} = \langle w_h, w_{h'} \rangle_{\mathcal{W}} ,$$

where $w_f = \Xi^{-1}(f)$.

With this inner product, $\mathcal{H}$ is a **Hilbert space** isomorphic to $\mathcal{W}_0$ (i.e., we can effectively identify them via the mapping Ξ, which preserves the inner product).

Functional formulation

We can restate our original simple SVM problem on the functional Hilbert space $\mathcal{H}$:

$$\text{minimise } \tfrac{1}{2}\|h\|_{\mathcal{H}}^2 \text{ under } y_i h(x_i) \geq 0,\ \forall i$$

This is completely equivalent to our original problem, but now we are tacking it explicitly in terms of functions $h \in \mathcal{H}$!

Let's start from functions...

So far we first phrased everything in $\mathcal{W}$ for a given feature function, then realised that we could treat the problem at the level of functions directly.

Can we start from a Hilbert space $\mathcal{H}$ whose elements are functions $\mathcal{X} \to \mathbb{R}$, without having to worry about the feature function ϕ or the space $\mathcal{W}$?

Can we obtain any $\mathcal{H}$ in this way?

Well, not really. The $\mathcal{H}$ that we construct in this way has some *nice* continuity property that a general Hilbert space of functions might not have...

Continuity of the evaluation functional

Consider a functional space $\mathcal{H}$, constructed as before, starting from $\mathcal{W}$ and ϕ. For each fixed $x \in \mathcal{X}$, define the evaluation operator

$$E_x : \mathcal{H} \to \mathbb{R}, \quad h \mapsto h(x).$$

- This is a linear operator:
  $E_x(\alpha_1 h_1 + \alpha_2 h_2) = \alpha_1 h_1(x) + \alpha_2 h_2(x) =$
  $\alpha_1 E_x(h_1) + \alpha_2 E_x(h_2)$.
- It is also **continuous** (Lipschitz constant $\|\phi(x)\|_{\mathcal{W}}$):
  $|E_x(h_1) - E_x(h_2)| = |\langle w_{h_1} - w_{h_2}, \phi(x)\rangle_{\mathcal{W}}| \leq$
  $\|\phi(x)\|_{\mathcal{W}} \|h_1 - h_2\|_{\mathcal{H}}$ by Cauchy Schwarz, as
  $\|h_1 - h_2\|_{\mathcal{H}} = \|w_{h_1} - w_{h_2}\|_{\mathcal{W}}$. This implies
  $h_n \to h \implies |E_x(h_n) - E_x(h)| \leq \|\phi(x)\|_{\mathcal{W}} \|h_n - h\|_{\mathcal{H}} \to 0$.

Note that continuity of E_x simply means that convergence in $\mathcal{H}$ implies pointwise convergence:

$$h_n \to h \implies h_n(x) \to h(x) \quad \forall x.$$

Reproducing Kernel Hilbert Spaces

Definition 2 (RKHS)

A **reproducing kernel Hilbert space** (RKHS) is a Hilbert space $\mathcal{H}$, whose elements are functions $\mathcal{X} \to \mathbb{R}$, where convergence in norm implies pointwise convergence.
More explicitly, this is to say that, for every $x \in \mathcal{X}$, the evaluation functional $E_x : \mathcal{H} \to \mathbb{R}$, $h \mapsto h(x)$ is continuous.

When we construct our $\mathcal{H}$ starting from a Hilbert space $\mathcal{W}$ and a feature map ϕ, we always obtain a RKHS!
We next will see that every time we are given a RKHS $\mathcal{H}$, we can find a space $\mathcal{W}$ and a feature map ϕ from which we can *derive* it.

Riesz Representation Theorem

Theorem 3 (Riesz)

Let $\mathcal{H}$ be a Hilbert space, and $L : \mathcal{H} \to \mathbb{R}$ a continuous linear function. Then, there exists a unique element $h_L \in \mathcal{H}$, such that

$$L(h) = \langle h_L, h \rangle_{\mathcal{H}}\,, \quad \forall h \in \mathcal{H}\,.$$

Consider a RKHS $\mathcal{H}$. For every x, the evaluation functional $h \mapsto h(x)$ is a continuous linear function $\mathcal{H} \to \mathbb{R}$. By Riesz theorem, there is an element $k_x \in \mathcal{H}$ such that $h(x) = \langle h, k_x \rangle_{\mathcal{H}}$.

Set $\mathcal{W} = \mathcal{H}$, and define the feature function $\phi : x \mapsto k_x$. Starting from $\mathcal{W}$ and ϕ, the construction that we have done earlier rederive exactly $\mathcal{H}$! So every RKHS can be obtained in this way!

Reproducing Kernels

Let $\mathcal{H}$ be a RKHS. Fix x and let $k_x \in \mathcal{H}$ be the Riesz representer of E_x. Then k_x is a function $\mathcal{X} \to \mathbb{R}$ (as it is an element of $\mathcal{H}$). Pick $x' \in \mathcal{X}$. We have $k_x(x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$. We define the **reproducing kernel** associated with $\mathcal{H}$ as the function

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad x, x' \mapsto \langle k_x, k_{x'} \rangle_{\mathcal{H}} .$$

Reproducing kernels will play an essential name in the study of RKHS (as the name RKHS suggests...). More on this next lecture!

Note that if we start from $\mathcal{W}$ and ϕ, and derive $\mathcal{H}$ (with kernel K). We have that $\langle \phi(x), \phi(x') \rangle_{\mathcal{W}} = K(x, x')$. Indeed, for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$ we have $h(x) = \langle w_h, \phi(x) \rangle_{\mathcal{W}} = \langle h, h_{\phi(x)} \rangle_{\mathcal{H}}$. So, $h_{\phi(x)} = k_x$ by uniqueness of the Riesz representer.

Further Reading

- An alternative (relatively gentle) lecture from Ulrike von Luxburg: `https://youtu.be/EoM_DF3VA08`
- An alternative (less gentle but more in depth) lecture from Arthur Gretton: `https://youtu.be/alrKls6BORc` (note this goes into some things we are yet to cover and some things we will not cover at all)
- For a more detailed and mathsy, but still introductory, perspective on RKHS: Paulsen and Raghupathi, **An Introduction to the Theory of Reproducing Kernel Hilbert Spaces**, Cambridge University Press, 2016.