Chapter 4, Part 3: RKHSs as Hypothesis Classes

Advanced Topics in Statistical Machine Learning

Eugenio Clerico

Hilary 2026

eugenio.clerico@stats.ox.ac.uk

Recap RKHS

A RKHS $\mathcal{H}$ is a **Hilbert space**, whose elements h are functions $\mathcal{X} \to \mathbb{R}$, such that, for every $x \in \mathcal{X}$, there exists a unique $k_x \in \mathcal{H}$ such that

$$\langle k_x, h \rangle = h(x) \,.$$

The reproducing kernel of $\mathcal{H}$ is

$$K(x, x') = \langle k_x, k_{x'} \rangle \,.$$

Last time:
Start from $(\mathcal{W}, \phi)$, with $\mathcal{W}$ Hilbert space and $\phi : \mathcal{X} \to \mathcal{W}$.
Construct a RKHS $\mathcal{H}$ where $h(x) = \langle w, \phi(x) \rangle$.
The reproducing kernel of $\mathcal{H}$ satisfies $K(x, x') = \langle \phi(x), \phi(x') \rangle$.

This time: We will start from a kernel...

Properties of Reproducing Kernels

If K is a reproducing kernels, then it is:

- **Symmetric:** $K(x, x') = K(x', x)$ (because the inner product is symmetric...)
- A **positive definite** function.

Definition 1

A **positive definite function** is a $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\forall N \geq 1$, $\forall x_1, \ldots, x_N$, the **Gram matrix** $\hat{K}$ (namely the $N \times N$ matrix with elements $\hat{K}_{ij} = K(x_i, x_j)$) is **positive semi-definite**. This means that for any $v \in \mathbb{R}^N$ we have

$$v^\top \hat{K} v \geq 0 \,.$$

Positive definiteness of reproducing kernels

Fix $v = (x_1, \ldots, x_N)^\top$.

$$
\begin{aligned}
v^\top \hat{K} v &= \sum_{i=1}^{N} \sum_{j=1}^{N} v_i \hat{K}_{ij} v_j = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) v_i v_j \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \langle v_i k_{x_i}, v_j k_{x_j} \rangle = \left\langle \sum_{i=1}^{N} v_i k_{x_i}, \sum_{j=1}^{N} v_j k_{x_j} \right\rangle \\
&= \left\| \sum_{i=1}^{N} v_i k_{x_i} \right\|^2 \geq 0 \,.
\end{aligned}
$$

Kernel Functions

Definition 2

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is **symmetric** and **positive definite** is called a **kernel function**.

So far, we have seen that if $\mathcal{H}$ is a RKHS and K its reproducing kernel, then K is a kernel function.
Is the other direction true? Is any kernel function the reproducing kernel of some RKHS?

Moore-Aronszajn Theorem

Theorem 3

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function. Then, there exists a unique (up to isometries) RKHS $\mathcal{H}$ of functions on $\mathcal{X}$, such that K is its reproducing kernel.

So every kernel function can be written as an inner product on some RKHS!

Some Consequences of Moore-Aronszajn Theorem

All the following procedures are valid!

- Start from $(\mathcal{W}, \phi)$, build a RKHS $\mathcal{H}$, find its reproducing kernel K
- Start from $(\mathcal{W}, \phi)$, define a kernel function $K(x, x') = \langle \phi(x), \phi(x') \rangle$, derive the induced RKHS $\mathcal{H}$.
- Start from a kernel function K, find the RKHS $\mathcal{H}$, let $\mathcal{W} = \mathcal{H}$ and $\phi(x) = K_x$.
- Start from a kernel function K, define $\phi(x) = K(x, \cdot)$ and embed these functions in a Hilbert space $\mathcal{W}$ with an inner product such that $\langle \phi(x), \phi(x') \rangle = K(x, x')$, build from $(\mathcal{W}, \phi)$ a RKHS $\mathcal{H}$.
- etc...

NB: The last approach is essentially the proof of Moore-Aronszajn theorem!

Proof's sketch of Moore-Aronszajn Theorem (1/3)

- For every x, define the function $\phi_x : \mathcal{X} \to \mathbb{R}$ as $\phi_x(x') = K(x, x')$.
- Let $\mathcal{F} = \mathrm{Span}\{\phi_x : x \in \mathcal{X}\}$, namely the set of functions $f = \sum_{i=1}^{N} a_i \phi_{x_i}$ (for any $N \geq 1$, any reals $a_1 \dots a_N$, any $x_1 \dots x_N$ in $\mathcal{X}$). $\mathcal{F}$ is a vector space.
- For $f = \sum_{i=1}^{N} a_i \phi_{x_i}$ and $g = \sum_{j=1}^{M} b_j \phi_{y_j}$, define $\langle f, g \rangle_{\mathcal{F}} = \sum_{i,j} a_i b_j K(x_i, y_j)$.[1]

[1] To see that this is well defined, we need to ensure that if h can be written both as $f = \sum_{i=1}^{N} a_i \phi_{x_i}$ and $h = \sum_{i=1}^{N'} a'_i \phi_{x'_i}$, then $\sum_{i,j} a_i b_j K(x_i, y_j) = \sum_{i,j} a'_i b_j K(x'_i, y_j)$. This is true since for every fixed y we have $\sum_i a_i K(x_i, y) = \sum_i a_i \phi_{x_i}(y) = f(y) = \sum_i a'_i \phi_{x'_i}(y) = \sum_i a'_i K(x'_i, y)$.

Proof's sketch of Moore-Aronszajn Theorem (2/3)

- $\langle \cdot, \cdot \rangle_F$ is linear in both entries (exercise).
- $\langle \cdot, \cdot \rangle_F$ is symmetric (trivial).
- $\langle \cdot, \cdot \rangle_F$ is positive definite:
  - Fix $f = \sum_{i=1}^{n} a_i \phi_{x_i} \in F$.
  - For $y \in \mathcal{X}$ and $t \in \mathbb{R}$, consider $(x_1 \dots x_n, y)$ and $(a_1 \dots a_n, t)$.
  - K PD: $\sum_{ij} a_i a_j K(x_i, x_j) + 2t \sum_i a_i K(x_i, y) + t^2 K(y, y) \geq 0$.
  - $\sum_{ij} a_i a_j K(x_i, x_j) = \langle f, f \rangle_F$; $\sum_i a_i K(x_i, y) = f(y)$.
  - $\langle f, f \rangle_F + 2tf(y) + t^2 K(y, y) \geq 0$, $\forall y, t$.
  - From $t = 0$, $\langle f, f \rangle_F \geq 0$, forall f.
  - If $\langle f, f \rangle_F = 0$, we need $f(y) = 0$ for inequality to hold $\forall t$!
  - We conclude that $\langle f, f \rangle_F > 0$ if $f \neq 0$.
- So, $\langle \cdot, \cdot \rangle_F$ is an inner product!

Proof's sketch of Moore-Aronszajn Theorem (3/3)

- So far we have constructed a pre-Hilbert space $\mathcal{F}$.
- We can pick a Hilbert space $\mathcal{W}$ that contains $\mathcal{F}$ as a subspace and preserves its inner product.[2] Namely each $f \in \mathcal{F}$ is also an element in $\mathcal{W}$, and for any f, g in $\mathcal{F}$, $\langle f, g \rangle_{\mathcal{F}} = \langle f, g \rangle_{\mathcal{W}}$.
- So now we have a Hilbert space $\mathcal{W}$ and we can define the feature map $\phi : \mathcal{X} \to \mathcal{W}$, $x \mapsto \phi_x$.
- We can proceed as in last lecture and construct a RKHS $\mathcal{H}$ from $(\mathcal{W}, \phi)$.
- We showed last lecture that the reproducing kernel of $\mathcal{H}$ satisfies $K_{\mathcal{H}}(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{W}}$. So, $K_{\mathcal{H}} = K$!
- We won't prove uniqueness (although it's not too hard!).

[2] This is always possible given a pre-Hilbert space, e.g. taking its completion.

RKHSs as Hypothesis Classes

Can we use an RKHS as a hypothesis class for (regularised) empirical risk minimization (ERM)?

A typical and general setup would be that we are looking for the function h^* in the RKHS $\mathcal{H}$ which solves

$$h^\star = \arg\min_{f \in \mathcal{H}} \hat{R}(h) + r\left(\|h\|_{\mathcal{H}}^2\right),$$

for empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i))$, a loss function $L \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and any increasing function r.

Representer Theorem

Theorem 4 (Representer Theorem)

If there exists a minimiser $h^\star$ for the optimisation problem

$$\inf_{h \in \mathcal{H}} \left(\hat{R}(h) + r(\|h\|_{\mathcal{H}}^2) \right) ,$$

where $\hat{R}$ is the empirical risk and $r : \mathbb{R}_+ \to \mathbb{R}$ is a strictly increasing function, then $h^\star$ takes the form

$$h^\star = \sum_{i=1}^{n} a_i K(x_i, \cdot) .$$

Here, K is the reproducing kernel of $\mathcal{H}$, x_i are the input data points on which $\hat{R}$ evaluates the loss, and a_i are some real coefficients.

Representer Theorem Proof

- Let $k_x = K(x, \cdot)$ and $\mathcal{H}_0 = \mathrm{Span}\{k_{x_1} \dots k_{x_n}\}$.
- Define $\mathcal{H}_0^\perp = \{h \in \mathcal{H} : \langle h, h' \rangle_\mathcal{H} = 0\,,\ \forall h' \in \mathcal{H}_0\}$.
- For every $h \in \mathcal{H}$ we have a unique decomposition $h = h_0 + h_\perp$, with $h_0 \in \mathcal{H}_0$ and $h_\perp \in \mathcal{H}_0^\perp$.
- Since $\langle h_\perp, k_{x_i} \rangle_\mathcal{H} = 0$, we have $h(x_i) = \langle h, k_{x_i} \rangle_\mathcal{H} = \langle h_0, k_{x_i} \rangle_\mathcal{H} = h_0(x_i)$.
- $\hat{R}$ depends on h only through the $h(x_i)$, so $\hat{R}(h) = \hat{R}(h_0)$.
- r is strictly increasing and $\|h\|_\mathcal{H}^2 = \|h_0\|_\mathcal{H}^2 + \|h_\perp\|_\mathcal{H}^2$, so $r(h) \geq r(h_0)$ (with equality only if $h = h_0$).
- We conclude that $h^\star \in \mathcal{H}_0$!

A note on the existence of the solution

When we introduced Hilbert spaces we said we wanted **completeness** for solution of optimisation problems to exist. A similar justification can be given to **continuity** of $x \mapsto h(x)$ in RKHSs. Indeed, this continuity ensures that if the loss is nice ($\hat{y} \mapsto L(y, \hat{y})$ is convex and continuous), then the empirical risk is also nice ($h \mapsto \hat{R}(h)$ is also convex and continuous). In particular, for r strictly increasing, this ensures that $h^\star$ exists and is unique!

The definition of RKHS can be seen as a natural way to makes sense of the minimisation problem $\min_{h \in H}(\hat{R}(h) + r(\|h\|_H^2))$, ensuring that the solution exists for a well behaved loss.

Representer Theorem Implications

The critical part of this result is that $h^\star$ is a linear combination of the feature mappings of our training data

- We can work with complex RKHS hypothesis classes while knowing that our solution will still take a simple form
- There is a very clear direct dependency of the functions we learn from the kernel we choose
- For a fixed kernel, the complexity of $h^\star$ is restricted for a given n, helping to prevent overfitting: we learn more complex functions as and when we see more data
- A downside is that we need to retain all our data to make predictions and this prediction will cost at least $O(n)$: can make kernel methods unsuitable for large datasets

Example: Kernel SVMs

We can express the primal problem for a kernel-SVM (fixing $b = 0$ for simplicity)

$$\min_{h \in \mathcal{H}} \left(\frac{1}{2} \|h\|_{\mathcal{H}}^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i h(x_i)) \right)$$

which is the form required by the representer theorem.

We know from before that this leads to the decision function $\hat{y}(x) = \text{sign}(h(x))$, where, because $b = 0$,

$$h(x) = \langle w, k(x, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x)$$

and we see that this is of the required form with factors $\alpha_i y_i$

Kernel Methods are Powerful

Our results so far demonstrate a number of advantages of kernel methods:

- RKHS spaces are a general and powerful class of function spaces: virtually all "well–behaved" function spaces can be expressed as an RKHS
- We can use kernels to perform ERM with an RKHS as our hypothesis class, allowing for very wide ranges of predictors to be learned in a **nonparametric** manner
- Many kernel methods permit simple, or even analytic, solutions to the ERM because of their basis in linear models

Kernel Methods have Drawbacks

But also some major drawbacks:

- Choosing the right kernel can be extremely important: our predictor will depend directly on this choice.
- Especially when solving numerically the optimisation problem, the convergence rate to the right solution depends heavily on the choice of the kernel.
- Choosing the right kernel (and thus RKHS) can be difficult: some RKHSs are actually very restrictive, while choosing an overly broad RKHS will lead to poor generalisation (due to overfitting)
- They tend to have relatively poor computational scaling in the size of the data (compared with, e.g., deep learning): they are based on pairwise similarities and thus have at best $O(n^2)$ scaling at training time and $O(n)$ at test time and often much

Example of RKHS: RBF kernel

Consider the RBF kernel $K(x,y) = e^{-(x-y)^2/2}$.
This kernel admits a feature representation in ℓ^2 (the space of square summable sequences $\ell^2 = \{w = (w_t) : \sum_{t=0}^{\infty} w_t^2 < \infty\}$).
With inner product $\langle w, w' \rangle_{\ell^2} = \sum_{t=0}^{\infty} w_t w'_t$, ℓ^2 is a Hilbert space.

$$\phi(x) = \left(e^{-x^2/2}, e^{-x^2/2}x, e^{-x^2/2}\frac{x^2}{\sqrt{2!}}, e^{-x^2/2}\frac{x^3}{\sqrt{3!}}, \dots\right)$$

$\phi(x)$ is in ℓ^2 because $\sum_{t=0}^{\infty}\left(e^{-x^2/2}\frac{x^t}{\sqrt{t!}}\right)^2 = e^{-x^2}e^{x^2} = 1$.
We have $\langle \phi(x), \phi(y) \rangle_{\ell^2} = K(x,y)$.
It turns out that $\mathcal{W}_0 = \mathrm{Span}(\phi(x),\, x \in \mathbb{R}) = \ell^2$, so the RKHS $\mathcal{H}$ is

$$\mathcal{H} = \left\{h_w : x \mapsto \sum_{t=0}^{\infty} w_t e^{-x^2/2}\frac{x^t}{\sqrt{t!}}\,,\ w \in \ell^2\right\}.$$

The inner product is $\langle h_w, h_{w'} \rangle_{\mathcal{H}} = \langle w, w' \rangle_{\ell^2} = \sum_{t=0}^{\infty} w_t w'_t$.

Further reading

For more rigorous and detailed proofs (+ several examples!) see Paulsen and Raghupathi, **An Introduction to the Theory of Reproducing Kernel Hilbert Spaces**, Cambridge University Press, 2016 (in Sec 2.2 thm 2.14 is Moore-Aronszajn thm; in Sec 8.6 thm 8.7 is a version of the representer thm).