Topics in Deep Learning Theory (Spring 2025)

Lecture 3: Reproducing Kernel Hilbert Spaces

Instructor: Lei Wu Date: April 19, 2025

Abstract

Reproducing Kernel Hilbert Spaces (RKHS) play a pivotal role not only in machine learning theory but also in statistics, computational mathematics, and functional analysis. These structured Hilbert spaces use a kernel-induced inner product, offering a unifying viewpoint for kernel methods. A central feature of RKHS is the Representer Theorem, which ensures that seeking solutions within these infinite-dimensional spaces can be reduced to a finite-dimensional problem, making the computations tractable.

In this lecture, we delve into the theoretical properties of RKHS from **multiple perspectives**. Particularly, we stress that, to build a better intuition, it is helpful to view an RKHS as the **approximation space** of ridge-regularized linear models, expressed as $\sum_{j=1}^m \theta_j \varphi_j(x)$, in the limit as $m \to \infty$. This perspective bridges finite-dimensional linear models (like classical regression) with their infinite-dimensional counterparts (kernel ridge regression). Please also read [Bach, 2024, Section 7.1-7.2] and [Wainwright, 2019, Section 12] for more discussion.

1 Functional Analysis Background

We will make use of a few concepts from functional analysis and here we review what we need.

Definition 1.1 (Function Space)**.** Let $\mathcal{X}$ be the input domain. A function space $\mathcal{F}$ is a space whose elements are functions, e.g. $f : \mathcal{X} \to \mathbb{R}$. We will focus on linear spaces of functions in the sense that if $f, g \in \mathcal{F}$, then $r_1 f + r_2 g \in \mathcal{F}$ for any $r_1, r_2 \in \mathbb{R}$.

Definition 1.2. An inner product is a function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that satisfies the following properties for every $f, g \in \mathcal{F}$:

1. Symmetric: $\langle f, g \rangle = \langle g, f \rangle$.
2. Linear: $\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$ for any $r_1, r_2 \in \mathbb{R}$.
3. Positive-definite: $\langle f, f \rangle \geq 0$ for all $f \in \mathcal{F}$ and $\langle f, f \rangle = 0$ iff $f = 0$.

Definition 1.3. A norm is a nonnegative function $\| \cdot \| : \mathcal{F} \to \mathbb{R}$ such that for all $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

- Positivity: $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
- Positive homogeneity: $\|\alpha f\| = |\alpha| \|f\|$.
- Triangular inequality: $\|f + g\| \leq \|f\| + \|g\|$;

Lemma 1.4. *Let $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ be an inner product space. Let $\|f\| = \sqrt{\langle f, f \rangle}$. Then, $\| \cdot \|$ is a norm.*

Proof. It is trivial to verify the positivity and positive homogeneity. What we need is to verify the triangular inequality. Noting

$$\begin{aligned}\|f+g\|^2 &= \langle f+g, f+g\rangle = \|f\|^2 + \|g\|^2 + 2\langle f, g\rangle \\ (\|f\| + \|g\|)^2 &= \|f\|^2 + \|g\|^2 + 2\|f\|\|g\|,\end{aligned}$$

we only need to verify the Cauchy-Schwartz inequality

$$\langle f, g\rangle \leq \|f\|\|g\|.$$

To this end, consider

$$\begin{aligned}\|f+\lambda g\|^2 &= \|f\|^2 + 2\lambda\langle f, g\rangle + \lambda^2\|g\|^2 \\ &= \left(\|f\| + \lambda\frac{\langle f, g\rangle}{\|f\|}\right)^2 + \lambda^2\left(\|g\|^2 - \frac{\langle f, g\rangle}{\|f\|}\right)\end{aligned}$$

As the above quantity is non-negative for any $\lambda \in \mathbb{R}$. We must have

$$\|g\|^2 - \frac{\langle f, g\rangle}{\|f\|} \geq 0,$$

which establishes the Cauchy-Schwartz inequality, thereby the triangular inequality.

□

Note that while the dot product in $\mathbb{R}^d$ is an excellent example, an inner product is more general than this, and requires only those properties given above.

Definition 1.5. A Hilbert space is a **complete**, (possibly) infinite-dimensional linear space endowed with an inner product. Let $\mathcal{H}$ be a Hilbert space. Denote by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ the associated inner product and norm.

The most popular finite-dimensional Hilbert space is the Euclidean space $\mathbb{R}^d$ equiped with the $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$. Another popular Hilbert space is $L^2(\mu)$ induced by the inner product

$$\langle f, g\rangle_{L^2(\mu)} = \int f(x)g(x)\,\mathrm{d}\mu(x) = \mathbb{E}_{x\sim\mu}[f(x)g(x)],$$

where μ is a probability distribution over $\mathcal{X}$.

While this tells us what a Hilbert space is, it is not intuitively clear why we need this mechanism, or what we gain by using it. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces of functions. In particular, the fact that a Hilbert space is complete will guarantee the convergence of certain algorithms. More importantly, the presence of an inner product allows us to make use of orthogonality and projections, which will later become important.

Definition 1.6 (Linear functional)**.** Let $\mathcal{F}$ be a linear function space. $A : \mathcal{F} \mapsto \mathbb{R}$ is said to be a linear functional if for any $\alpha, \beta \in \mathbb{R}$ and $f, g \in \mathcal{F}$, we have

$$A(\alpha f + \beta g) = \alpha A(f) + \beta A(g)$$

Definition 1.7. Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a normed function space. A linear functional $A : \mathcal{F} \mapsto \mathbb{R}$ is said to be continuous if there exist a constant $C > 0$ such that for any $f, g \in \mathcal{F}$, we have

$$|A(f-g)| \leq C\|f-g\|_{\mathcal{F}}.$$

The norm of A is defined by

$$\|A\| = \sup_{\|f\|_{\mathcal{F}} \leq 1} |A(f)|.$$

Obviously,

$$|A(f-g)| \leq \|A\|\|f-g\|_{\mathcal{F}}.$$

Theorem 1.8 (Riesz representation theorem)**.** *Suppose $\mathcal{H}$ to be a Hilbert space. For any continuous linear functional A, there exist a unique vector $f_A \in \mathcal{H}$, called the Riesz representation of A, such that*

$$A(g) = \langle f_A, g\rangle_{\mathcal{H}} \qquad \forall g \in \mathcal{H}.$$

2 Definitions of RKHS and KRR

2.1 Reproducing kernel Hilbert spaces

We begin by introducing two key types of functions that form the foundation of RKHS theory.

Definition 2.1 (Positive semidefinie (PSD) function)**.** A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is said to be PSD if (1) k is symmetric, i.e., $k(x, x') = k(x', x)$ for any $x, x' \in \mathcal{X}$, and (2) for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $K = (k(x_i, x_j))_{i,j} \in \mathbb{R}^{n\times n}$ is PSD.

Definition 2.2 (Kernel)**.** A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is said to be a kernel if k is symmetric and there exists a Hilbert space $\mathcal{H}$ and a feature map $\varphi : \mathcal{X} \mapsto \mathcal{H}$ such that

$$k(x, x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}}.$$

It is obvious that every kernel is PSD. To see this, consider any $\alpha \in \mathbb{R}^n$ and $x_1, \ldots, x_n \in \mathcal{X}$. We have:

$$\sum_{i,j=1}^{n} \alpha_i\alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i\alpha_j\langle\varphi(x_i), \varphi(x_j)\rangle_{\mathcal{H}} = \left\|\sum_{i=1}^{n} \alpha_i\varphi(x_i)\right\|_{\mathcal{H}}^2 \geq 0. \tag{1}$$

Definition 2.3 (RKHS)**.** Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$. It is said to be a reproducing kernel Hilbert space (RKHS) if there is a *reproducing kernel* $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.
- *Reproducing property:* $\forall x \in X$, $f \in \mathcal{H}$, $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

The next theorem plays a fundamental role in RKHS theory.

Theorem 2.4 (Moore-Aronszajn theorem)**.** *Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a psd function. Let*

$$\mathcal{H}^0 = \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha \in \mathbb{R}^n \text{ and } x_i \in \mathcal{X} \text{ for } i \in [n] \right\} \tag{2}$$

and endow it with the inner product: for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), g = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j)$,

$$\langle f, g \rangle_{\mathcal{H}^0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j). \tag{3}$$

Then, the pointwise closure $\mathcal{H}_k = \overline{\mathcal{H}^0}$ is a RKHS with k as its reproducing kernel.

The formal proof is deferred to 7.1. However, the intuition behind the construction of $\mathcal{H}_k^0$ is straightforward to grasp. By definition, $k(x, \cdot) \in \mathcal{H}_k$ for any $x \in \mathcal{X}$, and since $\mathcal{H}_k^0$ consists of finite linear combinations of such functions, linearity ensures $\mathcal{H}_k^0 \subseteq \mathcal{H}_k$. The inner product is naturally defined to align with the reproducing property in $\mathcal{H}_k$:

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}_k^0} &= \left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{m} \beta_j k(x'_j, \cdot) \right\rangle_{\mathcal{H}^0} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \langle k(x_i, \cdot), k(x'_j, \cdot) \rangle_{\mathcal{H}^0} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j),
\end{aligned}$$

where the second equality follows from the linearity of the inner product, and the final equality anticipates the reproducing property $\langle k(x_i, \cdot), k(x'_j, \cdot) \rangle_{\mathcal{H}^0} = k(x_i, x'_j)$.

One important implication of Moore-Aronszajn theorem is as follows:

Theorem 2.5. *The notions of PSD functions, kernels, and reproducing kernels are equivalent.*

Proof. First, (1) implies that kernels are PSD functions. Second, Moore-Aronszajn theorem guarantees that PSD functions are reproducing kernels. Third, for any reproducing kernel k, we can construct a feature map by setting $\varphi(x) = k(x, \cdot)$ and $\mathcal{H} = \mathcal{H}_k$. Thus, $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}_k}$, suggesting that every reproducing kernel is also a kernel. Thus, the three concepts are equivalent. □

Given this equivalence, we will not explicitly distinguish between these terms and will simply refer to them as kernels in the remainder of this lecture.

Lemma 2.6. *Let $\mathcal{H}$ be a RKHS. Then, its reproducing kernel k is unique.*

Proof. For any two reproducing kernels k_1, k_2, we have

$$\langle f, k_1(\cdot, x) - k_2(\cdot, x) \rangle_{\mathcal{H}} = f(x) - f(x) = 0, \forall x \in X, \forall f \in \mathcal{H}.$$

Taking $f = k_1(\cdot, x) - k_2(\cdot, x)$ lead to $\|k_1(\cdot, x) - k_2(\cdot, x)\|_{\mathcal{H}}^2 = 0, \forall x \in \mathcal{X}$. Hence, $k_1 = k_2$. □

Lemma 2.7. *For any reproducing kernel k, there is a unique RKHS with k as its reproducing kernel. Thus, this is the one constructed by Moore-Aronszajn theorem.*

The proof is lengthy and deferred to Section 7.2.

2.2 kernel ridge regression

The kernel ridge regression (KRR) using a RKHS as its hypothesis space and fitting data using the following strategy:

$$\hat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}_k}^2. \tag{4}$$

This is an infinite-dimensional optimizaiton. The representer theorem shows that it can be reduced to a finite dimensional problem:

Theorem 2.8 (Representer Theorem)**.** *For any $\lambda > 0$, there exists $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ such that $\hat{f}_\lambda = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$.*

Proof. Define $V = \operatorname{span}\{k(x_i, \cdot) : i = 1, \ldots, n\}$. For any $f \in \mathcal{H}$, decompose $f = f_\parallel + f_\perp$, where $f_\parallel \in V$ and $f_\perp \in V^\perp$, the orthogonal complement of V in $\mathcal{H}$. By the reproducing property, for any $i \in [n]$,

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_\parallel, k(x_i, \cdot)\rangle_{\mathcal{H}} + \langle f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_\parallel, k(x_i, \cdot)\rangle_{\mathcal{H}} = f_\parallel(x_i),$$

since $f_\perp \in V^\perp$ implies $\langle f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = 0$.

Thus, the data-fitting term $\sum_{i=1}^n (y_i - f(x_i))^2 = \sum_{i=1}^n (y_i - f_\parallel(x_i))^2$ depends only on $f_\parallel$. However, the regularization term satisfies

$$\|f\|_{\mathcal{H}}^2 = \|f_\parallel\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_\parallel\|_{\mathcal{H}}^2.$$

Thus, $\hat{f}_\lambda$ must lie in V, i.e., there exists a $\alpha \in \mathbb{R}^n$ such that $\hat{f}_\lambda = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. □

Noting $\| \sum_{i=1}^n \alpha k(x_i, \cdot)\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K \alpha$, then $\hat{f}_\lambda = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ with $\hat{\alpha}$ is given by minizing the following problem:

$$J(\alpha) = \frac{1}{n}\|K\alpha - y\|^2 + \lambda \alpha^\top K \alpha.$$

Kernel methods. General kernel methods employ the hypothesis $f_\alpha(x) := \sum_{i=1}^n \alpha_i k(x_i, \cdot)$, which extends beyond the scope of regression. This approach is versatile and can be applied to tasks such as classification and unsupervised learning.

3 The perspective of evaluation functional

Definition 3.1. Let $\mathcal{F}$ be a function space. For any $x \in \mathcal{X}$, the evaluation functional $L_x : \mathcal{F} \mapsto \mathbb{R}$ is defined by

$$L_x(f) = f(x).$$

Lemma 3.2. *For a RKHS $\mathcal{H}_k$, the evaluation functional $L_x : \mathcal{H}_k \mapsto \mathbb{R}$ is continuous.*

Proof. For any $x \in X$ and $f, g \in \mathcal{H}$,

$$\begin{aligned} |L_x(f) - L_x(g)| &= |f(x) - g(x)| = |\langle f - g, k(x, \cdot)\rangle_{\mathcal{H}_k}| \\ &\leq \|k(x, \cdot)\|_{\mathcal{H}_k} \|f - g\|_{\mathcal{H}_k}, \end{aligned}$$

where the last step follows from the Cauchy-Schwartz inequality. This means that $\|L_x\| \leq \|k(x, \cdot)\|_{\mathcal{H}_k} < \infty$, as $k(x, \cdot) \in \mathcal{H}_k$. □

An important implication is that the convergence in norm implies the pointwise convergence. If $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}_k} = 0$, then

$$|f_n(x) - f(x)| \le \|L_x\| \|f_n - f\|_{\mathcal{H}_k} \to 0 \qquad \text{as } n \to \infty.$$

This is a major difference between a RKHS and a general Hilbert space. For instance, for $L^2(\mu)$, the norm convergence does not imply the pointwise convergence.

This continuity of the evaluation functiona can be used as an equivalent definition of RKHS.

Theorem 3.3. *Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$. Then, $\mathcal{H}$ is a RKHS if and only if the evaluation functional is continuous.*

Proof. If L_x is continuous, by Riesz representation theorem, there exist $K_x \in \mathcal{H}$ such that

$$L_x(f) = \langle K_x, f \rangle_{\mathcal{H}}.$$

Define the kernel:

$$k(x, x') = \langle K_x, K_{x'} \rangle_{\mathcal{H}} = K_{x'}(x) = K_x(x'),$$

for which

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \langle f, K_x \rangle = f(x), \quad \forall f \in \mathcal{H}.$$

This means $k(\cdot, \cdot)$ is a reproducing kernel of $\mathcal{H}$. □

4 The perspective of approximation space and feature map

In this section, we show that KRR can be viewed as linear ridge regression in the feature space and consequently, RKHS can be viewed as the *approximation space* of linear ridge regression in the feature space. Let $\varphi_1, \varphi_2, \dots, \varphi_m, \dots$ be infinitely many features and $\varphi(x) := (\varphi_1(x), \varphi_2(x), \dots, \varphi_m(x), \dots)$ be the full feature map satisfying

Assumption 4.1. $\sum_{j=1}^{\infty} \varphi_j(x)^2 < \infty$ *for any* $x \in \mathcal{X}$.

Consider the "discrete" model: $f_m(x; \theta) = \sum_{j=1}^{m} \theta_j \varphi_j(x)$ and the corresponding ridge regression

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left(f_m(x_i; \theta) - y_i\right)^2 + \lambda \|\theta\|^2. \tag{5}$$

Then, a natural question what is the *approximation space* of this ridge-regularized linear model, i.e., the function spaces where this method can approximate and estimate well.

Proposition 4.2 (Inverse approximation theorem)**.** *Suppose Assumption (4.1) holds. If there exists a sequence $(\theta^{(m)} \in \mathbb{R}^m)_{m=1}^{\infty}$ with $\|\theta^{(m)}\|_2 \le B$, such that $\lim_{m\to\infty} f_m(\cdot; \theta^{(m)})(x) = f^*(x)$ for all $x \in \mathcal{X}$. Then, there must exist a $\theta^* \in \ell^2$ such that $f^*(x) = \sum_{j=1}^{m} \varphi_j(x)\theta_j^*$ and $\|\theta^*\| \le B$.*

This proposition is highly intuitive; however, its rigorous proof requires some functional analysis arguments, which we defer to Section 7.3. Essentially, it establishes that the approximation space of a ridge-regularized linear model is given by

$$G_\varphi = \left\{ x \mapsto \sum_{j=1}^{\infty} \theta_j \varphi_j(x) : \|\theta\|_2 < \infty \right\}$$

We next show that G_φ is exactly the RKHS $\mathcal{H}_{k_\varphi}$ associated with the kernel:

$$k_\varphi(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\ell^2}.$$

We proceed by considering a general feature map $\varphi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is an arbitrary Hilbert space, extending beyond the special case discussed above with $\mathcal{H} = \ell^2$. Specifically, we define the kernel function as

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}.$$

Definition 4.3. Let $\mathcal{F}_\varphi = \{f(x; \beta) = \langle \beta, \varphi(x) \rangle_{\mathcal{H}} : \beta \in \mathcal{H}\}$. For $f \in \mathcal{F}_\varphi$, define

$$\|f\|_{\mathcal{F}_\varphi} = \inf_{f = \langle \beta, \varphi(\cdot) \rangle_{\mathcal{H}}} \|\beta\|_{\mathcal{H}}.$$

Note that for a given f, its "representation" β is not necessarily unique. Taking the infimum ensures that the norm is independent of the specific choice of β. Nevertheless, one can ignore this and simply treat $\|f(\cdot; \beta)\|_{\mathcal{F}_\varphi} = \|\beta\|_{\mathcal{H}}$.

Lemma 4.4. *$\|\cdot\|_{\mathcal{F}_\varphi}$ is indeed a well-defined norm.*

Proof. Assume $f_1 = \langle \beta_1, \varphi(\cdot) \rangle_{\mathcal{H}}$, $f_2 = \langle \beta_2, \varphi(\cdot) \rangle_{\mathcal{H}}$. Then,

$$\lambda_1 f_1 + \lambda_2 f_2 = \langle \lambda_1 \beta_1 + \lambda_2 \beta_2, \varphi(\cdot) \rangle_{\mathcal{H}}.$$

By the definition,

$$\|\lambda_1 f_1 + \lambda_2 f_2\|_{\mathcal{F}_\varphi} \leq \|\lambda_1 \beta_1 + \lambda_2 \beta_2\|_{\mathcal{H}} \leq |\lambda_1| \|\beta_1\|_{\mathcal{H}} + |\lambda_2| \|\beta_2\|_{\mathcal{H}}.$$

Taking infimum over β_1 and β_2 yields

$$\|\lambda_1 f_1 + \lambda_2 f_2\|_{\mathcal{F}_\varphi} \leq |\lambda_1| \|f_1\|_{\mathcal{F}_\varphi} + |\lambda_2| \|f_2\|_{\mathcal{F}_\varphi}.$$

In addition, let $\|f\|_{\mathcal{F}_\varphi} = 0$. By definition, for any $\varepsilon > 0$, there exist β_ε such that $f = \langle \beta_\varepsilon, \varphi(\cdot) \rangle_{\mathcal{H}}$ and $\|\beta_\varepsilon\|_{\mathcal{H}} \leq \varepsilon$. Hence, for any $x \in \mathcal{X}$,

$$|f(x)| = |\langle \beta_\varepsilon, \varphi(x) \rangle_{\mathcal{H}}| \leq \|\beta_\varepsilon\|_{\mathcal{H}} \|\varphi(x)\|_{\mathcal{H}} \leq \varepsilon \|\varphi(x)\|_{\mathcal{H}}.$$

Taking $\varepsilon \to 0$, we obtain $f(x) = 0$ for any $x \in \mathcal{X}$. □

Definition 4.5. Let $\mathcal{F}_\varphi$ be the function space defined in Definition 4.3. For any $f, g \in \mathcal{F}_\varphi$, define $\langle f, g \rangle_{\mathcal{F}_\varphi} = \frac{\|f+g\|^2_{\mathcal{F}_\varphi} - \|f-g\|^2_{\mathcal{F}_\varphi}}{4}$.

It is easy to verify that $\mathcal{F}_\varphi$ is indeed well-defined inner product.

Lemma 4.6. *For any $f, g \in \mathcal{F}_\varphi$, there exists $\beta_f, \beta_g \in \mathcal{H}$ such that $f = \langle \beta_f, \varphi(\cdot) \rangle_{\mathcal{H}}, g = \langle \beta_g, \varphi(\cdot) \rangle_{\mathcal{H}}$ and*

$$\langle f, g \rangle_{\mathcal{F}} = \langle \beta_f, \beta_g \rangle_{\mathcal{H}}.$$

Proof. Taking β_f, β_g such that

$$\|f\|^2_{\mathcal{F}_\varphi} = \|\beta_f\|^2_{\mathcal{H}}$$
$$\|g\|^2_{\mathcal{F}_\varphi} = \|\beta_g\|^2_{\mathcal{H}}.$$

Hence,

$$\langle f, g\rangle_{\mathcal{F}_\varphi} = \frac{\|\beta_f + \beta_g\|^2_{\mathcal{H}} - \|\beta_f - \beta_g\|^2_{\mathcal{H}}}{4} = \langle \beta_f, \beta_g\rangle_{\mathcal{H}}.$$

□

The above lemma shows that the inner product of the functions are equivalent to the inner product of the corresponding representation.

Lemma 4.7. *For any $x \in \mathcal{X}$, $\|k(\cdot, x)\|_{\mathcal{F}_\varphi} = \|\varphi(x)\|^2_{\mathcal{H}} < \infty$.*

Proof. Note that $k(x, x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}}$. If $\beta_x \in \mathcal{H}$ is an representation of $k(x, \cdot)$, i.e.,

$$k(x, \cdot) = \langle \beta_x, \varphi(\cdot)\rangle_{\mathcal{H}},$$

then, we have $\langle \beta_x - \varphi(x), \varphi(x')\rangle_{\mathcal{H}} = 0$ for any $x' \in \mathcal{X}$. This means that $\beta_x - \varphi(x) \perp \text{span}\{\varphi(x')\}$. Hence,

$$\|\beta_x\|^2_{\mathcal{H}} = \|\beta_x - \varphi(x) + \varphi(x)\|^2_{\mathcal{H}} = \|\beta_x - \varphi(x)\|^2_{\mathcal{H}} + \|\varphi(x)\|^2_{\mathcal{H}} \geq \|\varphi(x)\|^2_{\mathcal{H}}.$$

Therefore, we have $\|k(x, \cdot)\|^2_{\mathcal{F}_\varphi} = \|\varphi(x)\|^2_{\mathcal{H}}$. □

Theorem 4.8 (Feature perspective of RKHS)**.** $(\mathcal{F}_\varphi, \langle \cdot, \cdot\rangle_{\mathcal{F}_\varphi}) = (\mathcal{H}_k, \langle \cdot, \cdot\rangle_{\mathcal{H}_k})$.

Proof. By the uniqueness of RKHS, we only need to verify that $(\mathcal{F}_\varphi, \langle \cdot, \cdot\rangle_{\mathcal{F}_\varphi})$ is a RKHS with k as its reproducing kernel. First, Lemma 4.7 establishes that $k(\cdot, x) \in \mathcal{F}_\varphi$ for any $x \in \mathcal{X}$. For any $f \in \mathcal{F}_\varphi$, assume $f(x) = \langle \beta_f, \varphi(x)\rangle_{\mathcal{H}}$ and $\|\beta_f\|^2_{\mathcal{H}} = \|f\|^2_{\mathcal{F}_\varphi}$. Then, we have

$$\langle f, k(\cdot, x)\rangle_{\mathcal{F}_\varphi} = \langle \beta_f, \varphi(x)\rangle_{\mathcal{H}} = f(x),$$

establishing the reproducing property. Thus, we complete the proof. □

Theorem 4.8 is foundational to our subsequent analysis, demonstrating that an explicit characterization of the corresponding RKHS is achievable whenever a feature map can be constructed. We will illustrate this concept in greater detail with examples in Sections 5 and 6.

A direct consequence of Theorem 4.8 is the following conclusion, which further establishes that performing linear ridge regression with $f(\cdot; \beta)$ is equivalent to KRR as defined in (4).

Proposition 4.9. *Let $f(x; \beta) = \langle \varphi(x), \beta\rangle_{\mathcal{H}}$ be a linear model. Let*

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{\beta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i; \beta) - y_i)^2 + \lambda \|\beta\|^2_{\mathcal{H}}.$$

Then, we have $f(\cdot; \hat{\beta}_\lambda) = \hat{f}_\lambda$, where the later is solution of KRR (4).

Proof. Note that $f(\cdot; \beta) \in \mathcal{H}_k$ and $\|\beta\|_{\mathcal{H}} \geq \|f(\cdot; \beta)\|_{\mathcal{H}_k}$, moreover by definition, for any $f \in \mathcal{H}^s$, there exists a $\beta \in \mathcal{H}$ such that $f = f(\cdot; \beta)$ with $\|\beta\|_{\mathcal{H}} = \|f\|_{\mathcal{H}_k}$. Combining these facts, we complete the proof. □

5 The perspective of spectral decomposition

For a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, we define an integral operator $\mathcal{T}_k : L^2(\mu) \mapsto L^2(\mu)$ as follows

$$\mathcal{T}_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') \, \mathrm{d}\mu(x')$$

Theorem 5.1 (Mercer's theorem)**.** *Let k be a continuous kernel on a* ***compact*** *set $\mathcal{X}$. There exist an orthonormal basis $\{e_j\}_{j=1}^{\infty}$ of $L^2(\mu)$ such that $\forall x, x' \in \mathcal{X}$,*

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x'). \tag{6}$$

The convergence is uniform on $\mathcal{X} \times \mathcal{X}$ and absolute for each $(x, x') \in \mathcal{X} \times \mathcal{X}$.

In fact, the existence of a spectral decomposition requires only the condition $\int k(x, x) \, \mathrm{d}\mu(x) < \infty$, which ensures that $\mathcal{T}_k$ has a finite trace and thereby, is compact. However, the compactness only guarantees the covergence in (6) is in $L^2(\mu \times \mu)$. The Mercer's theorem provides a stronger guarantee, establishing a convergence in $C(\mathcal{X} \times \mathcal{X})$. Note that $(\lambda_j)_{j \geq 1}$ and $(e_j)_{j \geq 1}$ are the eigenvalues and eigenfunctions of the integral operator $\mathcal{T}_k$ in the sense that

- $\mathcal{T}_k e_j = \lambda_j e_j$, i.e., $\mathbb{E}_{x' \sim \mu}[k(\cdot, x') e_j(x')] = \lambda_j e_j$.
- $\langle e_i, e_j \rangle_{L^2(\mu)} = \mathbb{E}_{x \sim \mu}[e_j(x) e_i(x)] = \delta_{i,j}$.

The feature map. One significant consequence of the existence of spectral decomposition is that it provides a feature map:

$$\varphi : \mathcal{X} \mapsto \ell^2, \qquad \varphi(x) = \left(\sqrt{\lambda_1} e_1(x), \sqrt{\lambda_2} e_2(x), \ldots, \sqrt{\lambda_j} e_j(x), \ldots \right)^{\top}, \tag{7}$$

for which

$$k(x, x') = \sum_{j=1}^{\infty} \sqrt{\lambda_j} e_j(x) \sqrt{\lambda_j} e_j(x') = \langle \varphi(x), \varphi(x') \rangle_{\ell^2}. \tag{8}$$

Combining the above feature map with Theorem 4.8, we can obtain the following result:

Theorem 5.2 (Spectral representation of RKHS)**.** *Let k be a continuous kernel on a compact set $\mathcal{X}$, and $\{e_j\}$ be the orthonormal basis given in Mercer's theorem. Define*

$$\mathcal{H} = \left\{ f = \sum_j a_j e_j : \sum_j \frac{a_j^2}{\lambda_j} < \infty \right\},$$

with the inner product

$$\left\langle \sum_j a_j e_j, \sum_j b_j e_j \right\rangle_{\mathcal{H}} = \sum_j \frac{a_j b_j}{\lambda_j}.$$

Then, $\mathcal{H}$ is the RKHS $\mathcal{H}_k$.

Proof. Consider the feature map given by Eq. (8). Then, by Theorem 4.8, for any $f \in \mathcal{H}_k$, there must exist a $a_f \in \ell^2$ such that

$$f(x) = \sum_{j=1}^{\infty} (a_f)_j \varphi_j(x) = \sum_{j=1}^{\infty} (a_f)_j \lambda_j^{1/2} e_j(x).$$

Thus, $(a_f)_j = \langle f, e_j \rangle_{L^2(\mu)} / \lambda_j^{1/2}$. By Theorem 4.8, for any $f, g \in \mathcal{H}_k$, we have

$$\langle f, g \rangle_{\mathcal{H}_k} = \langle a_f, a_g \rangle_{\ell^2} = \sum_{j=1}^{\infty} \frac{\langle f, e_j \rangle_{L^2(\mu)} \langle g, e_j \rangle_{L^2(\mu)}}{\lambda_j}.$$

Thus, we complete the proof. □

Weighted L^2 space. In this way, RKHS can be viewed as a L^2 space weighted by the eigenvalues.

- L^2 space: $\|f\|_{L^2(\mu)}^2 = \sum_{j=1}^{\infty} a_j^2$.
- RKHS/Weighted L^2 space: $\|f\|_{\mathcal{H}_k}^2 = \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j}$.

Hence, the faster the eigenvalue decay is, the smaller the RKHS is. Consider $\lambda_j = \frac{1}{j^s}$. Then,

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} j^s a_j^2 < \infty \quad \overset{roughly}{\Longrightarrow} \quad a_j^2 = O\left(\frac{1}{j^{s+1+\delta}}\right) \quad \text{for some } \delta > 0,$$

A larger s leads to a faster the decay of the coefficients.

6 Examples of RKHSs

In this section, we provide some concrete examples of RKHS. We will use the following convention of Fourier transform and its inverse:

$$\mathcal{F}[f](\omega) = \hat{f}(\omega) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i \, \omega \cdot x} \, \mathrm{d}x$$
$$\mathcal{F}^{-1}[g](x) = \check{g}(x) = \int_{\mathbb{R}^d} h(\omega) e^{2\pi i \, \omega \cdot x} \, \mathrm{d}\omega.$$

6.1 Brownian kernel

Let $\mathcal{X} = [0, 1]$ and consider the *Brownian kernel* (also known as the *Wiener kernel*):

$$k_B(x, y) = \min(x, y).$$

To characterize the RKHS associated with this kernel, we aim to construct its feature map representation. Specifically, let

$$\phi(t; x) = \mathbf{1}_{[0,x]}(t).$$

We can verify that the map $x \mapsto \phi(\cdot; x) \in L^2([0,1])$ generates the Brownian kernel as follows:

$$\min(x, y) \;=\; \int_0^1 \varphi(t; x)\, \varphi(t; y)\, \mathrm{d}t.$$

Hence, by Theorem 4.8, any function f in the RKHS admits the representation

$$f(x) = \int_0^1 a(t)\, \varphi(t; x)\, \mathrm{d}t = \int_0^1 a(t)\, \mathbf{1}_{[0,x]}(t)\, \mathrm{d}t = \int_0^x a(t)\, \mathrm{d}t,$$

for some coefficient function $a \in L^2([0,1])$. Consequently,

$$\mathcal{H}_{k_B} = \left\{ f(x) = \int_0^x a(t)\, \mathrm{d}t \;:\; a \in L^2([0,1]) \right\}.$$

Since f can be written as an integral of $a \in L^2([0,1])$, it follows that f is absolutely continuous, and moreover, $f'(x) = a(x)$ almost everywhere and $f(0) = 0$.

To see how the inner product is realized, we use the fact that $a_f = f'$ and $a_g = g'$. Then, the RKHS inner product is

$$\langle f, g \rangle_{\mathcal{H}_{k_B}} = \int_0^1 a_f(t)\, a_g(t)\, \mathrm{d}t = \int_0^1 f'(t)\, g'(t)\, \mathrm{d}t.$$

Thus, the RKHS norm of f is given by $\|f\|^2_{\mathcal{H}_{k_B}} = \int_0^1 |f'(t)|^2\, \mathrm{d}t$.

Putting these observations together, the RKHS associated with k_B is the Sobolev space:

$$\big\{ f : f(0) = 0,\ f \text{ is absolutely continuous},\ f' \in L^2([0,1]) \big\}.$$

Note that functions in this Sobolev space vanish only at $x = 0$. One can easily verify that the classical Sobolev space $H^1_0([0,1])$, whose functions vanish at both $x = 0$ and $x = 1$, is the RKHS generated by the Brownian bridge kernel

$$k(x, y) = \min(x, y) - xy.$$

6.2 Bandlimited functions

Consider bandlimited functions given by

$$\mathcal{H} = \{ f : \mathbb{R} \mapsto \mathbb{R} : \hat{f}(\omega) = 0 \text{ for } |\omega| \geq 1/2 \}$$

equipped with the L^2 inner product $\langle f, g \rangle_{\mathcal{H}} = \int_{\mathbb{R}} f(x) g(x)\, \mathrm{d}x$. Then, we claim that $\mathcal{H}$ is a RKHS with the kernel given by the sinc function: $k(x, y) = \frac{\sin \pi(x-y)}{\pi(x-y)} = \mathrm{sinc}(x - y)$.

One can easily verify that $\mathcal{H}$ is indeed a Hilbert space and here, we only verify the reproducing property. Note that

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = \int f(y) \mathrm{sinc}(x - y)\, \mathrm{d}y = (f * \mathrm{sinc})(x). \tag{9}$$

To evaluate this convolution, we leverage the Fourier transform, which simplifies convolution operations by transforming them into multiplications in the frequency domain. The Fourier transform of the sinc function is well-known:

$$\widehat{\operatorname{sinc}}(\omega) = \operatorname{rect}(\omega) = \begin{cases} 0, & \text{if } |\omega| \geq 1/2 \\ 1, & \text{if } |\omega| < 1/2 \end{cases}$$

Since $\hat{f}(\omega) = 0$ for any $|\omega| > 1/2$, for any $\omega \in \mathbb{R}$ it holds: $\hat{f}(\omega)\widehat{\operatorname{sinc}}(\omega) = \hat{f}(\omega)$. Applying the inverse Fourier transform, we get

$$\langle f, k(x, \cdot)\rangle_{\mathcal{H}} = \mathcal{F}^{-1}[\hat{f} \cdot \widehat{\operatorname{sinc}}](x) = f(x),$$

which verifies the reproducing property.

6.3 Sobolev spaces in torus

The periodic (fractional) Sobolev space $H^s(\mathbb{T})$ is defined over the torus $\mathbb{T} = [0, 1)$. Here, $s \geq 0$ is the smoothness parameter that controls the regularity of the functions in the space. Given a $f : \mathbb{T} \to \mathbb{C}$, let its Fourier series is given by

$$f(x) = \sum_{n\in\mathbb{Z}} \hat{f}(n) e^{2\pi i n x}.$$

The Sobolev norm can be defined using the Fourier transform as follows

$$\|f\|_{H^s}^2 = \sum_{n\in\mathbb{Z}} (1 + |n|^2)^s |\hat{f}(n)|^2.$$

The term $(1+|n|^2)^s$ acts as a weight that increases with $|n|$, penalizing higher-frequency components more heavily as s grows. When s is positive integer, $f^{(s)}(x) = \sum_{n\in\mathbb{Z}} (2\pi i n)^s \hat{f}(n) e^{2\pi i n x}$ and its L^2 norm satisfies $\int_{\mathbb{T}} |f^{(s)}(x)|^2 \, \mathrm{d}x \asymp \sum_{n\in\mathbb{Z}} n^{2s} \hat{f}(n)^2$, where $\asymp$ holds because $(2\pi)^{2s}$ is constant. Thus, we have

$$\|f\|_{H^s}^2 \asymp \sum_{n=0}^{s} \int_{\mathbb{T}} |f^{(s)}(x)|^2 \, \mathrm{d}x,$$

where $f^{(0)} = f$. Thus, $H^s(\mathbb{T})$ consists of functions whose derivatives up to order s have finite L^2 norm, reflecting their smoothness properties.

Lemma 6.1. *Consider a periodic kernel $k : \mathbb{T} \times \mathbb{T} \mapsto \mathbb{R}$ given by $k(x, x') = \kappa(x - x')$. Suppose the Fourier series of κ decay as $\hat{\kappa}(n) \asymp (1 + |n|^2)^{-s}$ for some $s > 1/2$. Then, $\mathcal{H}_k$ is equivalent to $H^s(\mathbb{T})$.*

The condition $s > 1/2$ ensures that $\sum_{n\in\mathbb{Z}} (1 + |n|)^s < \infty$, which is necessary for the kernel to be a well-defined PSD kernel.

Let $e_n(x) = e^{2\pi i n x}$ be the Fourier basis functions and the underlying distribution to be $\mu = \operatorname{Unif}(\mathbb{T})$. Then, the kernel has the following expansion:

$$k(x, x') = \kappa(x - x') = \sum_{n\in\mathbb{Z}} \hat{\kappa}(n) e^{2\pi i n (x-y)} = \sum_{n\in\mathbb{Z}} \hat{\kappa}(n) e_n(x) \overline{e_n(x')},$$

suggesting that the eigenfunctions of periodic kernels are the Fourier basis and $\{\hat{\kappa}(n)\}_{n\in\mathbb{Z}}$ are the corresponding eigenvalues. By the spectral representation of RKHS, we have

$$\|f\|_{\mathcal{H}_k}^2 = \sum_{n\in\mathbb{Z}} \mu_n^{-1} \langle f, e_n\rangle_{L^2(\mathbb{T})} = \sum_{n\in\mathbb{Z}} \frac{\hat{f}(n)^2}{\hat{\kappa}(n)}. \tag{10}$$

Plugging $\hat{\kappa}(n) \asymp (1+|n|^2)^{-s}$, we have

$$\|f\|_{\mathcal{H}_k}^2 \asymp \sum_{n\in\mathbb{Z}} (1+|n|^2)^s \hat{f}(n)^2 \asymp \|f\|_{H^s}^2.$$

This establishes the equivalence.

6.4 Sobolev spaces in $\mathbb{R}^d$

For **Sobolev spaces** $H^s(\mathbb{R}^d)$, which consist of functions whose weak derivatives up to order s are in $L^2(\mathbb{R}^d)$, this property depends on the parameters s (the smoothness order) and d (the dimension). The Sobolev norm is typically defined as:

$$\|f\|_{H^s}^2 = \int_{\mathbb{R}^d} (1+\|\omega\|^2)^s |\hat{f}(\omega)|^2 \, \mathrm{d}\omega,$$

where $\hat{f}$ is the Fourier transform of f. The key insight comes from the **Sobolev embedding theorem**: when $s > \frac{d}{2}$, functions in $H^s(\mathbb{R}^d)$ are continuous. This is because the embedding $H^s(\mathbb{R}^d) \hookrightarrow C^0(\mathbb{R}^d)$ (the space of continuous functions vanishing at infinity) holds, and there exists a constant C such that:

$$\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)| \le C\|f\|_{H^s}.$$

Since $|f(x)| \le \|f\|_\infty$, it follows that:

$$|\delta_x(f)| = |f(x)| \le C\|f\|_{H^s},$$

proving that δ_x is a continuous linear functional on $H^s(\mathbb{R}^d)$. Thus, when $s > \frac{d}{2}$, $H^s(\mathbb{R}^d)$ is an RKHS. Moreover, the corresponding kernel is **Matérn Kernels**(widely used in applications like Gaussian processes and machine learning), defined as:

$$k_{\nu,d}(x,y) = \kappa_{\nu,d}(x-y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|x-y\|}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{\|x-y\|}{\rho}\right),$$

where:

- $\nu > 0$ is the smoothness parameter,
- $\rho > 0$ is the length scale,
- K_ν is the modified Bessel function of the second kind,
- Γ is the Gamma function.

To justify that the RKHS generated by the Matérn kernel is precisely $H^s(\mathbb{R}^d)$, we can follow an argument based on the feature map perspective.

Lemma 6.2. *Let $k(x,y) := \kappa(x-y)$ be a translation invariant kernel. Then, we have*

$$\|f\|_{\mathcal{H}_k}^2 = \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)^2}{\hat{\kappa}(\omega)} \,\mathrm{d}\omega. \tag{11}$$

Proof. By the inverse Fourier transform, we have

$$\kappa(x-y) = \int_{\mathbb{R}^d} \hat{\kappa}(\omega) e^{2\pi i \omega^\top (x-y)} \,\mathrm{d}\omega = \langle \varphi(\cdot;x), \varphi(\cdot;y) \rangle_{L^2(\mathbb{R}^d)},$$

where $\varphi(\omega;x) := \hat{\kappa}^{1/2}(\omega) e^{2\pi i \omega^\top x} \in L^2(\mathbb{R}^d)$. Therefore, by Theorem 4.8, the RKHS functions must admit the following representation

$$f(x) = \int_{\mathbb{R}^d} a_f(\omega) \hat{\kappa}^{1/2}(\omega) e^{2\pi i \omega^\top x} \,\mathrm{d}\omega,$$

for some $a_f \in L^2(\mathbb{R}^d)$. By the inverse Fourier transform, we have

$$a_f(\omega) = \hat{f}(\omega) \hat{\kappa}^{-1/2}(\omega) \quad a.e.$$

Thus, the RKHS norm of f can be written as

$$\|f\|_{\mathcal{H}_k}^2 = \int_{\mathbb{R}^d} a_f(\omega)^2 \,\mathrm{d}\omega = \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)^2}{\kappa(\omega)} \,\mathrm{d}\omega.$$

□

For the Fourier transform of Matérn kernel, up to constants, it is approximately:

$$\hat{\kappa}_{\nu,d}(\omega) \asymp \left(1 + \|\omega\|^2\right)^{-(\nu + \frac{d}{2})}.$$

Thus, by (11), it holds that

$$\|f\|_{\mathcal{H}_{k_{\nu,d}}}^2 \asymp \int_{\mathbb{R}^d} (1 + \|\omega\|^2)^{\nu + d/2} |\hat{f}(\omega)|^2 \,\mathrm{d}\omega = \|f\|_{H^{\nu+d/2}}^2.$$

This establishes that the RKHS generated by Matérn kernel is exactly the Sobolev space.

Notably, in one dimension ($d = 1$), for $H^1(\mathbb{R})$ ($s = 1$), the reproducing kernel is the exponential kernel

$$k_{1,1}(x,y) = \kappa_{1,1}(x-y) = \frac{1}{2} e^{-|x-y|},$$

whose Fourier transform is given by $\hat{\kappa}_{1,1}(\omega) \asymp 1/(1+|\omega|^2)$.

7 Proofs

7.1 Proof of Theorem 2.4

We show that (3) indeed defines a valid inner product. First,

$$\langle f, g \rangle_{\mathcal{H}^0} = \sum_{i=1}^n \alpha_i g(x_i) = \sum_{j=1}^n \beta_j f(x'_j).$$

It is implied that that the inner product is independent of the specific representation of f and g. The triangular inequality is easy to verify. Next, we show that $\|f\|_{\mathcal{H}^0} = 0$ if and only if $f = 0$. If there exist $x_0 \in \mathcal{X}$ such that $f(x_0) \neq 0$. Assume $f(x) = \sum_{j=1}^m a_j k(x_j, \cdot)$ and consider

$$0 \leq \|\lambda f + f(x_0)k(\cdot, x_0)\|_{\mathcal{H}^0}^2 = \lambda^2 \|f\|_{\mathcal{H}^0}^2 + 2\lambda f^2(x_0) + f^2(x_0)k(x_0, x_0).$$

Taking $\lambda \to -\infty$, the RHS will be negative and this causes contradictory.

What remains is to show that the convergence of Cauchy sequence [1]. We refer to Link for a complete proof.

What remains is to show that k is a reproducing kernel of $\mathcal{H}_k$. For $f \in \mathcal{H}^0$, we can write $f(x) = \sum_{j=1}^m a_j k(\cdot, x_j)$. By definition,

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = \sum_{j=1}^m a_j k(x, x_j) = f(x).$$

For any $f \in \mathcal{H}_k$, let $\lim_{n\to\infty} f_n(x) = f(x)$. Then,

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = \lim_{n\to\infty} \langle f_n, k(\cdot, x)\rangle_{\mathcal{H}_k} = \lim_{n\to\infty} f_n(x) = f(x).$$

□

7.2 Proof of Lemma 2.7

Proof. First, by Moore-Aronsajn theorem, there exists a RKHS with k being the reproducing kernel. Assume $\mathcal{H}_1$ and $\mathcal{H}_2$ be two RKHSs with k being the reproducing kernel. First, by definition, $k(\cdot, x) \in \mathcal{H}_1$ for any $x \in \mathcal{X}$. Hence, $\mathcal{H}^0 \subset \mathcal{H}_1$. Moreover, $\mathcal{H}^0$ is dense in $\mathcal{H}_1$ since if there exists $f \in \mathcal{H}$ such that $f \perp \mathcal{H}^0$, we must have

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_1} = f(x) = 0 \qquad \forall x \in \mathcal{X}.$$

For $f = \sum_{j=1}^m a_j k(\cdot, x_j)$,

$$\begin{aligned}\|f\|_{\mathcal{H}_1}^2 &= \left\langle \sum_i^n a_i k(\cdot, x_i), \sum_{j=1}^m a_j k(\cdot, x_j) \right\rangle_{\mathcal{H}_1} = \sum_{i,j=1}^n a_i a_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}_1} \\ &\overset{(i)}{=} \sum_{i,j=1}^n a_i a_j k(x_i, x_j) = \|f\|_{\mathcal{H}^0}^2.\end{aligned}$$

where (i) follows from the reproducing property. Hence, $\|f\|_{\mathcal{H}_1} = \|f\|_{\mathcal{H}^0}$ for $f \in \mathcal{H}_0$. By the same argument, the same results hold for $\mathcal{H}_2$. For any $f \in \mathcal{H}_1$, there must exits $(f_n) \subset \mathcal{H}^0$ such that $f(x) = \lim_{n\to\infty} f_n(x)$. This implies that $f \in \mathcal{H}_2$. Similarly, $\mathcal{H}_1$ and $\mathcal{H}_2$ contains the same functions. What remains is to check that the two norms coincide, which results from

$$\|f\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}^0} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_2} = \|f\|_{\mathcal{H}_2}.$$

□

[1] You can skip the verification of completeness.

7.3 Proof of Proposition 4.2

We will need the following lemma, which establishes the *weak lower semicontinuity* of the norm in Hilbert spaces:

Lemma 7.1. *Let $\mathcal{H}$ be a Hilbert space, and let $\{x_n\}$ be a sequence in $\mathcal{H}$ such that x_n converges weakly to x. Then, $\|x\| \leq \liminf_{n\to\infty} \|x_n\|$.*

Proof. $x_n \rightharpoonup x$ implies $\langle x_n, y\rangle \to \langle x, y\rangle$ for every y. In particular, setting $y = x$ gives $\langle x_n, x\rangle \to \langle x, x\rangle = \|x\|^2$. By the Cauchy–Schwarz inequality, $\langle x_n, x\rangle \leq \|x_n\|\|x\|$. Taking the limit inferior as $n \to \infty$ on both sides yields

$$\|x\|^2 = \lim_{n\to\infty} \langle x_n, x\rangle = \liminf_{n\to\infty} \langle x_n, x\rangle \;\leq\; \liminf_{n\to\infty} \big(\|x_n\|\|x\|\big) = \|x\| \liminf_{n\to\infty} \|x_n\|.$$

If $\|x\| > 0$, dividing by $\|x\|$ completes the argument:

$$\|x\| \;\leq\; \liminf_{n\to\infty} \|x_n\|.$$

If $\|x\| = 0$, the inequality is trivially satisfied. □

Proof of Theorem 4.2. **Step 1: Extend $\theta^{(m)}$ to ℓ^2.** Since each $\theta^{(m)} \in \mathbb{R}^m$ has a dimension that increases with m, we define an extended sequence $\tilde{\theta}^{(m)} \in \ell^2$ by padding $\theta^{(m)}$ with zeros:

$$\tilde{\theta}^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_m^{(m)}, 0, 0, \ldots).$$

The ℓ^2 norm of this sequence is:

$$\|\tilde{\theta}^{(m)}\|_{\ell^2} = \sqrt{\sum_{j=1}^{\infty} |\tilde{\theta}_j^{(m)}|^2} = \sqrt{\sum_{j=1}^{m} |\theta_j^{(m)}|^2} = \|\theta^{(m)}\|_2 \leq B.$$

Thus, $\{\tilde{\theta}^{(m)}\}_{m=1}^{\infty}$ is a sequence in ℓ^2, uniformly bounded by B.

Step 2: Extract a weakly convergent subsequence. The Banach-Alaoglu theorem ensures that any bounded sequence in ℓ^2 has a weakly convergent subsequence. Therefore, there exists a subsequence $\{\tilde{\theta}^{(m_k)}\}_{k=1}^{\infty}$ and some $\theta^* \in \ell^2$ such that:

$$\tilde{\theta}^{(m_k)} \rightharpoonup \theta^* \quad \text{weakly in } \ell^2. \tag{12}$$

Then by Lemma 7.1, we have $\|\theta^*\|_{\ell^2} \leq \liminf_{k\to\infty} \|\tilde{\theta}^{(m_k)}\|_{\ell^2} \leq B$.

Step 3: Verifying the representation. First, $\sum_{j=1}^{\infty} \varphi_j(x)\theta_j^*$ is well defined for all $x \in \mathcal{X}$,

$$\sum_{j=1}^{\infty} |\varphi_j(x)\theta_j^*| \leq \left(\sum_{j=1}^{\infty} |\varphi_j(x)|^2\right)^{1/2} \left(\sum_{j=1}^{\infty} |\theta_j^*|^2\right)^{1/2} < \infty,$$

by Cauchy-Schwarz inequality. Moreover,

$$f^*(x) = \lim_{m\to\infty} \langle \varphi(x), \tilde{\theta}^{(m)}\rangle = \lim_{k\to\infty} \langle \varphi(x), \tilde{\theta}^{(m_k)}\rangle = \langle \varphi(x), \theta^*\rangle,$$

where the last step uses (12) and the definition of weak convergence. □

References

[Bach, 2024] Bach, F. (2024). *Learning theory from first principles*. MIT press.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.