CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: RKHS and the Representer Theorem

Date: October 22nd, 2025 *Author: Surbhi Goel*

1 Recap and Preview

Last time, we saw that pointwise evaluation is ill-behaved in standard function spaces like L^2. This motivated the need for **kernels**, the tools required to build special Hilbert spaces where evaluation is well-behaved.

Today, we construct these spaces—called **Reproducing Kernel Hilbert Spaces (RKHS)**—and prove the powerful **Representer Theorem**, which makes infinite-dimensional optimization practical.

Roadmap:

1. Reproducing Kernel Hilbert Spaces (RKHS)—special function spaces with a "magic" property
2. The Representer Theorem—why infinite-dimensional optimization is tractable
3. Applications to kernel ridge regression and SVMs

2 Reproducing Kernel Hilbert Spaces

2.1 Definition and Properties

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. For any point $x \in \mathcal{X}$, define the function $K_x : \mathcal{X} \to \mathbb{R}$ as $K_x(\cdot) = K(x, \cdot)$. We will build our Hilbert space from these functions.

- Start with the set of all finite linear combinations of these functions:

$$\mathcal{H}_0 = \left\{ f : f(\cdot) = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot) \text{ for some } m, \alpha_i, x_i \right\}$$

- Define an inner product on $\mathcal{H}_0$: For $f = \sum_i \alpha_i K_{x_i}$ and $g = \sum_j \beta_j K_{y_j}$,

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, y_j)$$

- The induced norm is $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.
- **Definition:** The **Reproducing Kernel Hilbert Space (RKHS)** $\mathcal{H}_K$ is the completion of $\mathcal{H}_0$ with respect to this norm.
- "Completion" means: add all limit points to make it a complete Hilbert space.

2.2 The Reproducing Property

This is the key property that makes RKHS special.

Theorem 1 (Reproducing Property)**.** *For any $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$:*

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

- **Proof for $f \in \mathcal{H}_0$:** If $f = \sum_i \alpha_i K_{x_i}$, then:

$$\begin{aligned} \langle f, K_x \rangle_{\mathcal{H}} &= \left\langle \sum_i \alpha_i K_{x_i}, K_x \right\rangle_{\mathcal{H}} \\ &= \sum_i \alpha_i \langle K_{x_i}, K_x \rangle_{\mathcal{H}} \\ &= \sum_i \alpha_i K(x_i, x) \\ &= f(x) \end{aligned}$$

- For general $f \in \mathcal{H}_K$: use continuity and limits.
- **Special case:** $K_x(y) = \langle K_x, K_y \rangle_{\mathcal{H}} = K(x, y)$
- **Intuition:** Think of K_x as the "x-th basis vector" in function space. The reproducing property says: to evaluate f at x, just take the inner product with K_x.
- This is like $e_i^T x = x_i$ in $\mathbb{R}^n$—the i-th standard basis vector "extracts" the i-th coordinate.

2.3 Key Consequences

- **Evaluation is continuous:** If $f_n \to f$ in $\mathcal{H}_K$, then $f_n(x) \to f(x)$ for all x.

  Proof: By Cauchy-Schwarz,

$$\begin{aligned} |f_n(x) - f(x)| &= |\langle f_n - f, K_x \rangle_{\mathcal{H}}| \\ &\leq \|f_n - f\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} \\ &= \|f_n - f\|_{\mathcal{H}} \sqrt{K(x, x)} \to 0 \end{aligned}$$

- In general Hilbert spaces like L^2, convergence in norm doesn't imply pointwise convergence. RKHS are special!
- **Theorem (Moore-Aronszajn):** Every kernel K defines a unique RKHS $\mathcal{H}_K$ with K as its reproducing kernel. Conversely, every RKHS has a unique reproducing kernel.
- There's a one-to-one correspondence: {kernels} $\leftrightarrow$ {RKHS}.

2.4 Examples

- **Linear kernel:** $K(x, y) = x^T y$ gives $\mathcal{H}_K = \{f(x) = w^T x : w \in \mathbb{R}^d\}$ (finite-dimensional)
- **Polynomial kernel:** $K(x, y) = (x^T y + 1)^2$ gives finite-dimensional RKHS of polynomials
- **Gaussian kernel:** $K(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ gives infinite-dimensional RKHS of smooth functions

3 The Representer Theorem

Now for the main result: optimization over infinite-dimensional RKHS reduces to finite dimensions.

3.1 The Setup

Consider the regularized empirical risk minimization problem:

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} \ell(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \right]$$

where:

- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function (e.g., squared loss, hinge loss)
- $(x_1, y_1), \ldots, (x_N, y_N)$ are training data
- $\lambda > 0$ is a regularization parameter
- We optimize over *all functions* in the (possibly infinite-dimensional) RKHS $\mathcal{H}_K$

Question: How can we solve this? There are infinitely many functions to search over!

Answer (Representer Theorem): The optimal solution has the form:

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

for some coefficients $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$.

So even though we optimize over an infinite-dimensional space, the solution only has N degrees of freedom!

3.2 Statement and Proof

Theorem 2 (Representer Theorem)**.** *Let $\mathcal{H}_K$ be an RKHS with kernel K, let $\ell : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$ be any function, and let $\Omega : [0, \infty) \to \mathbb{R}$ be strictly increasing. Consider:*

$$f^* = \arg\min_{f \in \mathcal{H}_K} \left[\ell(f(x_1), \ldots, f(x_N); y_1, \ldots, y_N) + \Omega(\|f\|_{\mathcal{H}}^2) \right]$$

Then f^* has the form:

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

for some $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$.

Proof idea: Use projection.

Proof. **Step 1: Decompose f.**

Let $U = \text{span}\{K_{x_1}, \ldots, K_{x_N}\}$ be the span of kernel functions at training points. By the projection theorem for Hilbert spaces, any $f \in \mathcal{H}_K$ decomposes uniquely as:

$$f = f_U + f_{U^\perp}$$

where $f_U \in U$ and $f_{U^\perp} \in U^\perp$ (orthogonal complement).

Since these are orthogonal:

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

Since Ω is strictly increasing:

$$\Omega(\|f\|_{\mathcal{H}}^2) \geq \Omega(\|f_U\|_{\mathcal{H}}^2)$$

Step 2: Show $f(x_i)$ only depends on f_U.

By the reproducing property:

$$f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}} = \langle f_U + f_{U^\perp}, K_{x_i} \rangle_{\mathcal{H}}$$

But $K_{x_i} \in U$ and $f_{U^\perp} \in U^\perp$, so $\langle f_{U^\perp}, K_{x_i} \rangle_{\mathcal{H}} = 0$. Thus:

$$f(x_i) = \langle f_U, K_{x_i} \rangle_{\mathcal{H}} = f_U(x_i)$$

This means the predictions at training points only depend on f_U!

Step 3: Conclude.

The objective is:

$$\ell(f(x_1), \ldots, f(x_N); y_1, \ldots, y_N) + \Omega(\|f\|_{\mathcal{H}}^2)$$

From Step 2: the loss term only depends on f_U.

From Step 1: the regularization term is minimized when $f_{U^\perp} = 0$, i.e., when $f = f_U \in U$.

Therefore, $f^* \in U = \text{span}\{K_{x_1}, \ldots, K_{x_N}\}$, which means:

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

□

3.3 Interpretation and Consequences

- **Geometric view:** The optimal function lies in the N-dimensional subspace spanned by $\{K_{x_1}, \ldots, K_{x_N}\}$.
- Even though $\mathcal{H}_K$ might be infinite-dimensional (e.g., Gaussian kernel), the solution only uses N basis functions!
- **Computational view:** We only need to find N coefficients $\alpha_1, \ldots, \alpha_N$. This reduces to finite-dimensional optimization.
- This is a general projection result: we find the best solution in the subspace of functions "generated" by the training data.
- The Representer Theorem holds for *any* loss function and *any* increasing regularizer. Very general!

4 Application: Kernel Ridge Regression

Problem: Given data $(x_1, y_1), \ldots, (x_N, y_N)$, minimize:

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right]$$

Solution: By the Representer Theorem, $f^*(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$ for some α.

Substituting: predictions are $f^*(x_j) = \sum_i \alpha_i K(x_i, x_j)$, so in vector form $\hat{y} = K\alpha$ where $K_{ij} = K(x_i, x_j)$.

The norm is $\|f^*\|_{\mathcal{H}}^2 = \alpha^T K \alpha$ (by bilinearity of inner product).

Therefore we minimize:

$$\min_{\alpha \in \mathbb{R}^N} \left[\|K\alpha - y\|_2^2 + \lambda \alpha^T K \alpha \right]$$

Setting the gradient to zero gives:

$$\alpha = (K + \lambda I)^{-1} y$$

To predict on a new point: $f^*(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$.

The kernel trick in action. Compare to linear ridge regression: $\hat{\theta} = (X^T X + \lambda I)^{-1} X^T y$. The form is similar, but now we work with the $N \times N$ kernel matrix. With the Gaussian kernel, we get *infinite features* for the cost of $O(N^3)$ time!

5 Summary

The journey:

- We began with linear algebra in finite dimensions (vector spaces, bases, projections).
- We extended these concepts to infinite dimensions via Hilbert spaces and introduced kernels to solve the problem of pointwise evaluation.
- **Today:** RKHS and the Representer Theorem make infinite-dimensional optimization tractable.

Key insight: Even though we optimize over infinite-dimensional RKHS, the solution has only N parameters: $f^*(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x)$.

Applications: This framework underpins Support Vector Machines, Gaussian Processes, kernel ridge regression, and many other ML methods.