The Representer Theorem in Reproducing Kernel Hilbert Spaces

The **Representer Theorem** is a fundamental result in kernel methods, particularly in the context of Reproducing Kernel Hilbert Spaces (RKHS). It states that the solution to a wide class of regularized empirical risk minimization problems in an RKHS can always be expressed as a finite linear combination of kernel functions centered at the training data points.

1 Theorem Statement

Theorem 1 (Representer Theorem):

- **Context:** Consider a set of training data points $X = \{x_i\}_{i=1}^n$. We are working in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, which is associated with a kernel function $k(\cdot, \cdot)$.
- **Optimization Problem:** We aim to find a function f^* that minimizes a regularized empirical risk:

$$f^* \in \arg\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} \ell(f(x_i), y_i) + \eta \Omega(\|f\|_k) \right)$$

  where:

  - $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a loss function.
  - $\eta \geq 0$ is a regularization parameter.
  - $\Omega(\|f\|_k)$ is a non-decreasing penalty term dependent on the RKHS norm of f (typically $\Omega(\|f\|_k) = \|f\|_k^2$).
- **Conclusion:** The theorem states that the optimal solution f^* can always be written in the form:

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$$

  for some coefficients $\alpha_i \in \mathbb{R}$.

2 Proof Explanation

The proof relies on the decomposition of functions in a Hilbert space and the Reproducing Property of the RKHS.

2.1 1. Decomposition of f

Any function $f \in \mathcal{H}$ can be uniquely decomposed into two orthogonal components with respect to the finite-dimensional subspace S spanned by the kernel functions evaluated at the training data points:

$$S = \text{span}\{k(x_1, \cdot), \ldots, k(x_n, \cdot)\}$$

Thus, we write $f = f_\parallel + f_\perp$, where $f_\parallel \in S$ and $f_\perp \in S^\perp$.

2.2 2. Norm and Orthogonality

Due to the orthogonality, the square of the RKHS norm satisfies the Pythagorean theorem:

$$\|f\|_k^2 = \|f_\parallel\|_k^2 + \|f_\perp\|_k^2$$

From this, we see that $\|f\|_k^2 \geq \|f_\parallel\|_k^2$.

2.3 3. Applying the Reproducing Property

The **reproducing property** states that for any $f \in \mathcal{H}$ and any $x \in \mathcal{X}$:

$$f(x) = \langle f, k(x, \cdot) \rangle_k$$

Applying this to a training point x_i and using the decomposition $f = f_{\|} + f_{\perp}$:

$$\begin{aligned} f(x_i) &= \langle f, k(x_i, \cdot) \rangle_k \\ &= \langle f_{\|} + f_{\perp}, k(x_i, \cdot) \rangle_k \\ &= \langle f_{\|}, k(x_i, \cdot) \rangle_k + \langle f_{\perp}, k(x_i, \cdot) \rangle_k \end{aligned}$$

2.4 4. Consequence of Orthogonality

Since $f_{\perp}$ is orthogonal to S, it is orthogonal to every basis function $k(x_i, \cdot)$. Thus, the second term is zero:

$$\langle f_{\perp}, k(x_i, \cdot) \rangle_k = 0$$

The expression simplifies to:

$$f(x_i) = \langle f_{\|}, k(x_i, \cdot) \rangle_k$$

By the reproducing property applied to $f_{\|}$, we also have $f_{\|}(x_i) = \langle f_{\|}, k(x_i, \cdot) \rangle_k$. Therefore, we conclude the critical result:

$$f(x_i) = f_{\|}(x_i)$$

The orthogonal component $f_{\perp}$ does not affect the function's value at any training data point x_i.

2.5 5. Minimization Argument

Substituting $f(x_i) = f_{\|}(x_i)$ into the optimization objective, and assuming the common case where $\Omega(\|f\|_k) = \|f\|_k^2$:

$$\min_{f \in \mathcal{H}} \left(\sum_{i=1}^{n} \ell(f_{\|}(x_i), y_i) + \eta(\|f_{\|}\|_k^2 + \|f_{\perp}\|_k^2) \right)$$

- The loss term $\sum_{i=1}^{n} \ell(f_{\|}(x_i), y_i)$ depends only on $f_{\|}$.
- The regularization term $\eta\|f\|_k^2$ contains the strictly non-negative term $\eta\|f_{\perp}\|_k^2$.

To minimize the overall objective, since the loss is fixed by $f_{\|}$, we must minimize the remainder of the penalty, which requires setting $\|f_{\perp}\|_k^2 = 0$. In a Hilbert space, this implies that the optimal solution f^* must have $f_{\perp} = 0$.

2.6 6. Conclusion on the Form of f^*

Since $f_{\perp} = 0$, the optimal function f^* must satisfy $f^* = f_{\|}$. By definition, $f_{\|} \in S$, which is the span of the kernel functions at the training points. Therefore, f^* must be expressible as:

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$$

This concludes the proof.

3 Significance

The Representer Theorem is incredibly powerful because it simplifies the search for an optimal function from an **infinite-dimensional RKHS** to a **finite-dimensional problem** of finding the coefficients α_i. This is the theoretical justification for the **kernel trick** in algorithms like Support Vector Machines (SVMs) and Kernel Principal Component Analysis (KPCA), allowing us to work implicitly in high-dimensional feature spaces while keeping the computation tractable.