1 Mercer's Theorem and Feature Map

1.1 Mercer's Theorem

Definition 1.1 (Definite Kernel). *The function $k : [a, b] \times [a, b] \to \mathbb{R}$ is a definite kernel where the following double integral:*

$$J(f) = \int_a^b \int_a^b k(x, y) f(x) f(y) dx dy, \tag{1}$$

satisfies $J(f) > 0$ for all $f(x) \neq 0$.

Mercer improved over Hilbert's work to propose his theorem, the Mercer's theorem, introduced in the following.

Theorem 1.1 (Mercer's Theorem). *Suppose $k : [a, b] \times [a, b] \to \mathbb{R}$ is a continuous symmetric positive semi-definite kernel which is bounded:*

$$\sup_{x,y} k(x, y) < \infty. \tag{2}$$

Assume the operator T_k takes a function $f(x)$ as its argument and outputs a new function as:

$$T_k f(x) := \int_a^b k(x, y) f(y) dy, \tag{3}$$

which is a Fredholm integral equation. The operator T_k is called the Hilbert–Schmidt integral operator. This output function is positive semi-definite:

$$\iint k(x, y) f(y) dx dy \geq 0. \tag{4}$$

Then, there is a set of orthonormal bases $\{\psi_i(.)\}_{i=1}^{\infty}$ of $L^2(a, b)$ consisting of eigenfunctions of T_K such that the corresponding sequence of eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are non-negative:

$$\int k(x, y) \psi_i(y) dy = \lambda_i \psi_i(x). \tag{5}$$

The eigenfunctions corresponding to the non-zero eigenvalues are continuous on $[a, b]$ and k can be represented as:

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y), \tag{6}$$

where the convergence is absolute and uniform.

Proof. A roughly high-level proof for the Mercer's theorem is as follows.

Step 1 of proof: According to assumptions of theorem, the Hilbert-Schmidt integral operator T_k is a symmetric operator on $L^2(a, b)$ space. Consider a unit ball in $L^2(a, b)$ as input to the operator. As the kernel is bounded, $\sup_{x,y} k(x, y) < \infty$, the sequence $f_1, f_2, \dots$ converges in norm, i.e. $\|f_n - f\| \to 0$ as $n \to 0$. Therefore, according to the Arzelà-Ascoli theorem, the image of the unit ball after applying the operator is compact. In other words, the operator T_k is compact.

Step 2 of proof: According to the spectral theorem, there exist several orthonormal bases $\{\psi_i(.)\}_{i=1}^{\infty}$ in $L^2(a, b)$ for the compact operator T_k. This provides a spectral (or eigenvalue) decomposition for the operator T_k:

$$T_k \psi_i(x) = \lambda_i \psi_i(x), \tag{7}$$

where $\{\psi_i(.)\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvectors and eigenvalues of the operator T_k, respectively. Noticing the defined Eq. (3) and the eigenvalue decomposition, Eq. (7), we have:

$$\int k(x,y)\psi_i(y)dy \overset{(3)}{=} T_k\psi_i(x) \overset{(7)}{=} \lambda_i\psi_i(x). \tag{8}$$

This proves the Eq. (5) which is the eigenfunction decomposition of the operator T_k. Note that the eigenvectors $\{\psi_i(.)\}_{i=1}^{\infty}$ are referred to as the *eigenfunctions* because the decomposition is applied on a function or operator rather than a matrix. Note that eigenfunctions will be explained more in Section **??**.

Step 3 of proof: According to Parseval's theorem, the Bessel's inequality can be converted to equality. For the orthonormal bases $\{\psi_i(.)\}_{i=1}^{\infty}$ in the Hilbert space $\mathcal{H}$ associated with kernel k, we have for any function $f \in L^2(a,b)$:

$$f = \sum_{i=1}^{\infty} \langle f, \psi_i\rangle_k \psi_i. \tag{9}$$

If we replace ψ_i with f in Eq. (7) and consider Eq. (9), we will have:

$$T_k f = \sum_{i=1}^{\infty} \lambda_i \langle f, \psi_i\rangle_k \psi_i. \tag{10}$$

One can consider Eq. (3) as $T_k f = kf$. Noticing this and Eq. (10) results in:

$$kf = \sum_{i=1}^{\infty} \lambda_i \langle f, \psi_i\rangle_k \psi_i. \tag{11}$$

Ignoring f from Eq. (11) gives:

$$k(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y), \tag{12}$$

which is Eq. (6); hence, that is proved.

Step 4 of proof: We define the truncated kernel r_n (with parameter n) as:

$$\begin{aligned} r_n(x,y) &:= k(x,y) - \sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(y) \\ &= \sum_{i=n+1}^{\infty} \lambda_i \psi_i(x)\psi_i(y). \end{aligned} \tag{13}$$

As T_k is an integral operator, this truncated kernel has positive kernel, i.e., for every $x \in [a,b]$, we have:

$$r_n(x,x) = k(x,x) - \sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(x) \geq 0$$

which implies

$$\sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(x) \leq k(x,x) \leq \sup_{x\in[a,b]} k(x,x). \tag{14}$$

By Cauchy-Schwartz inequality, we have:

$$\begin{aligned} \Big|\sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(y)\Big|^2 &\leq \Big(\sum_{i=1}^{n} \lambda_i \psi_i(x)\psi_i(x)\Big)\Big(\sum_{i=1}^{n} \lambda_i \psi_i(y)\psi_i(y)\Big) \\ &\overset{(14)}{\leq} \Big(\sup_{x\in[a,b]} k(x,x)\Big)^2. \end{aligned}$$

Taking second root from the sides of inequality gives:

$$\sum_{i=1}^{n} \lambda_i \psi_i(x) \psi_i(x) \leq \sup_{x \in [a,b]} |k(x,x)| \overset{(2)}{\leq} \infty. \tag{15}$$

This shows that the sequence $\sum_{i=1}^{n} \lambda_i \psi_i(x) \psi_i(x)$ converges absolutely and uniformly. $\square$