
Peter Bartlett
Overview

- Kernel methods
  - Kernels
  - Reproducing kernel Hilbert spaces
  - Mercer’s theorem
  - Constructing kernels
Recall: Inner Products

For the perceptron algorithm and its analysis, all we needed was an inner product on some vector space:

$$\hat{y} = \text{sign} \left( \sum_j \alpha_j \langle \Phi(x_j), \Phi(x) \rangle \right),$$

$$\Phi : \mathcal{X} \mapsto \mathcal{V}.$$  

We don’t need to explicitly evaluate $\Phi(x)$, as long as we can evaluate the inner products (which might be much cheaper).
Kernels and inner product spaces

**Definition:** \( k : \mathcal{X}^2 \to \mathbb{R} \) is **positive semidefinite** if, for all \( n \) and all \( x_1, \ldots, x_n \in \mathcal{X} \), the Gram matrix \( K \in \mathbb{R}^{n \times n} \)—defined by \( K_{ij} = k(x_i, x_j) \)—is positive semidefinite.

**Definition:** \( k : \mathcal{X}^2 \to \mathbb{R} \) is a **kernel** if it is

1. Symmetric: \( k(u, v) = k(v, u) \), and
2. Positive semidefinite: every Gram matrix \( K_{ij} = k(x_i, x_j) \) is positive semidefinite.
Kernels and inner product spaces

**Theorem:** If $k$ is a kernel, then there is an inner product space $\mathcal{F}$ and a feature map $\Phi$ such that $k(u, v) = \langle \Phi(u), \Phi(v) \rangle$.

Consider:

$\Phi(x) := k(\cdot, x),
\mathcal{F} := \text{span} \{ \Phi(x) : x \in X \},
\left\langle \sum_i \alpha_i \Phi(u_i), \sum_j \beta_j \Phi(v_j) \right\rangle := \sum_{i,j} \alpha_i \beta_j k(u_i, v_j).$
Kernels and inner product spaces

We can augment this inner product space a little, by including all the
limit points, i.e., making it complete (wrt the metric
\[\|f - g\| = \sqrt{\langle f - g, f - g \rangle}\]):

**Definition:** A metric space \( \mathcal{F} \) is complete if every Cauchy
sequence (ie: elements approach each other) converges to an
\( f \in \mathcal{F} \).

A *Hilbert space* is an inner product space that is a complete
metric space wrt the norm induced by the inner product.
**Kernels and reproducing kernel Hilbert spaces**

**Definition:** A reproducing kernel Hilbert space is a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, with a reproducing kernel $k : \mathcal{X}^2 \to \mathbb{R}$, that is, the span of $\{k(\cdot, x) : x \in \mathcal{X}\}$ is dense in $\mathcal{H}$, and $k(x, \cdot) \in \mathcal{H}$ is the point evaluation function for $\mathcal{H}$: $f(x) = \langle k(x, \cdot), f \rangle$. 
Kernels and reproducing kernel Hilbert spaces

- For our construction of a Hilbert space $\mathcal{H}$ from a kernel $k$, it’s easy to check that $k$ is the reproducing kernel of the Hilbert space, and that $\mathcal{H}$ is unique.
- There are alternative (equivalent) ways of define an RKHS.
- Not all Hilbert spaces have a reproducing kernel.
Fix a symmetric function $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ on a compact set $\mathcal{X} \subset \mathbb{R}^d$, and consider the integral operator $T_k : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{X})$ defined as

$$T_k f(\cdot) = \int_{\mathcal{X}} k(\cdot, x) f(x) \, dx.$$ 

We say $T_k$ is positive semidefinite if, for all $f \in L_2(\mathcal{X})$,

$$\langle f, T_k f \rangle_{L_2(\mathcal{X})} \geq 0,$$

that is,

$$\int_{\mathcal{X}^2} k(u, v) f(u) f(v) \, du \, dv \geq 0.$$
Mercer’s Theorem

**Theorem:** If $k$ is continuous and $T_k$ is positive semidefinite, then $T_k$ has eigenfunctions $\psi_i \in L_2(\mathcal{X})$ (say $\|\psi_i\|_{L_2} = 1$) with eigenvalues $\lambda_i \geq 0$, and for all $u, v \in \mathcal{X}$, we can write

$$k(u, v) = \sum_{i=1}^{\infty} \lambda_i \psi_i(u) \psi_i(v).$$

Furthermore, this series converges uniformly.
Mercer’s Theorem: finite-dimensional analog

Consider the finite-dimensional analog: Write $K_{i,j} = k(x_i, x_j)$; identify $f \in \mathbb{R}^X$ with a vector $f = (f_1, \ldots, f_n) \in \mathbb{R}^n$. Then

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

so for all $f \in \mathbb{R}^n$,

$$f^T K f \geq 0.$$
Mercer’s Theorem: finite-dimensional analog

That is, $K$ is positive semidefinite, so we can write it as

$$K = \sum_{i=1}^{n} \lambda_i v_i v_i^T,$$

with $\lambda_i \geq 0$. Then we have

$$k(x_i, x_j) = K_{ij} = (V \Lambda V^T)_{ij}$$

$$= \sum_{t=1}^{n} \lambda_t v_{ti} v_{tj}$$

$$= \sum_{t=1}^{n} \lambda_t \psi_t(x_i) \psi_t(x_j),$$

where $\psi_t : \mathcal{X} \rightarrow \mathbb{R}$ is given by $\psi_t(x_i) = v_{t,i}$. 


Mercer’s theorem gives another representation of $k$ as an inner product, this time with feature map

$$\Psi(x) = \begin{bmatrix}
\psi_1(x) \\
\vdots \\
\psi_n(x)
\end{bmatrix}.$$ 

Notice that $T_k$ is positive semidefinite iff for all $x_1, \ldots, x_n \in \mathcal{X}$ the Gram matrix $K$ is positive semidefinite. So we have another characterization.
**Theorem:** For $X \subset \mathbb{R}^d$ compact and $k : X^2 \to \mathbb{R}$ continuous and symmetric, the following are equivalent:

1. Every Gram matrix is positive semidefinite.
2. The integral operator $T_k$ is positive semidefinite.
3. We can express $k$ as

   $$k(u, v) = \sum_i \lambda_i \psi_i(u)\psi_i(v)$$

   for fixed $\lambda_i \geq 0$ and $\psi_i : X \to \mathbb{R}$.
4. $k$ is the reproducing kernel of an RKHS on $X$. 


Mercer’s Theorem

Notes:

- We have seen two representations of $k(u, v)$ as an inner product $k(u, v) = \langle \Phi(u), \Phi(v) \rangle$:

  $\Phi_1(u) = k(\cdot, u)$
  $\langle k(\cdot, u), k(\cdot, v) \rangle = k(u, v)$

  $\Phi_2(u) = \begin{pmatrix}
    \sqrt{\lambda_1} \psi_1(u) \\
    \sqrt{\lambda_2} \psi_2(u) \\
    \vdots
  \end{pmatrix}$
  $\langle \Phi_2(u), \Phi_2(v) \rangle = \sum_i \lambda_i \psi_i(u) \psi_i(v)$.

  So they are not unique.

- Computing a kernel $k$ is equivalent to computing inner products, in what might be an infinite-dimensional space.
Mercer’s Theorem

- An infinite-dimensional RKHS is approximated by a finite-dimensional subspace, since we have uniform absolute convergence:

\[
\lim_{n \to \infty} \sup_{u,v \in \mathcal{X}} \left| k(u,v) - \sum_{i=1}^{n} \lambda_i \psi_i(u) \psi_i(v) \right| = 0.
\]
Constructing Kernels

If $k_1$ and $k_2$ are kernels on $\mathcal{X}$, then the following are also kernels:

1. $k(u, v) = a_1 k_1(u, v) + a_2 k_2(u, v)$ (for $a_1, a_2 \geq 0$).
2. $k(u, v) = k_1(u, v) k_2(u, v)$
3. $k(u, v) = k_1(f(u), f(v))$, where $f : \mathcal{V} \to \mathcal{X}$. 
4. $k(u, v) = g(u)g(v)$, where $g : \mathcal{X} \to \mathbb{R}$.

5. $k(u, v) = p(k_1(u, v))$, where $p$ is a polynomial with positive coefficients.

6. $k(u, v) = \exp(k_1(u, v))$.

7. $k(u, v) = \exp(-||u - v||^2/2)$.
The gaussian kernel is an example of a translation-invariant kernel: $k(u, v) = f(u - v)$, where $f : [-\pi, \pi] \rightarrow \mathbb{R}$ is a continuous, even function. Then we can write

$$f(x) = \sum_{n=0}^{\infty} a_n \cos(nx) \quad (a_n \geq 0)$$

$$k(u, v) = \sum_{n=0}^{\infty} a_n \left( \sin(nu) \sin_nv + \cos(nu) \cos(nv) \right)$$

$$= \sum_{n=0}^{\infty} \lambda_n \psi_n(u) \psi_n(v),$$

where $\{\psi_i(u)\} = \{1, \sin(u), \cos(u), \sin(2u), \cos(2u), \ldots\}$. 
Marginalized kernels

Given a probability distribution $P$ on $\mathcal{X} \times \mathcal{H}$, and a kernel $k$ defined on $(x, h)$ pairs, we can define

$$k_M(x, x') = \sum_{h, h'} k((x, h), (x', h')) P(h|x) P(h'|x').$$

For example, if $x$ is a graph, and $h$ is a random walk on the graph, and $k$ reflects the similarity of the nodes on the two random walks, this gives a useful (and efficiently computable) approach to computing an inner product between two graphs.