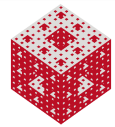
MATH 590: Meshfree Methods

Chapter 2 — Part 1: Positive Definite Kernels and Reproducing Kernel Hilbert Spaces

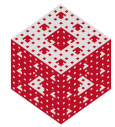
Greg Fasshauer

Department of Applied Mathematics
Illinois Institute of Technology

Fall 2014



Outline

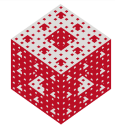


- We know that the solution of the scattered data interpolation problem with RBFs or kernels amounts to solving a linear system

$$\mathrm{K}\boldsymbol{c} = \boldsymbol{y},$$

where $\mathrm{K}_{ij} = \kappa(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2)$ or $\mathrm{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, N$.

- Linear algebra tells us that this system will have a unique solution whenever K is non-singular.
- Necessary and sufficient conditions to characterize this general non-singular case are still open.
- We will focus mostly on kernels K that generate positive definite matrices.



Positive Definite Matrices and Kernels

Definition

A real symmetric $N \times N$ matrix K is called positive definite if its associated quadratic form is positive for any nonzero $\boldsymbol{c} = [c_1, \ldots, c_N]^T \in \mathbb{R}^N$, i.e.,

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j \mathrm{K}_{ij} > 0,$$

or more compactly $\boldsymbol{c}^T \mathrm{K} \boldsymbol{c} > 0$. The matrix is called positive semi-definite if the quadratic form is allowed to be nonnegative.

Definition

A symmetric kernel K is called positive definite on $\Omega \times \Omega$ if its associated kernel matrix $\mathrm{K} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi-definite for any $N \in \mathbb{N}$ and any set of distinct points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \Omega$.

Positive Operators and Integrally PD Kernels

In analogy to the positive (semi-)definiteness of a symmetric matrix we consider this notion for a self-adjoint operator [Hoc73, Section 3.5]:

Definition

A self-adjoint operator $\mathcal{K}$ acting on a Hilbert space $\mathcal{H}$ is called positive if $\langle \mathcal{K}f, f \rangle_{\mathcal{H}} \geq 0$ for all $f \in \mathcal{H}$.

This leads to another generalization [Mer09] of positive (semi-)definite matrices:

Definition

A symmetric kernel K is called integrally positive definite on $\Omega \times \Omega$ if

$$\int_{\Omega} \int_{\Omega} K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z} \geq 0$$

for all $f \in L_2(\Omega)$.

If the operator $\mathcal{K}$ is defined as an integral operator, i.e.,

$$(\mathcal{K}f)(\boldsymbol{x}) = \int_\Omega K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

and

$$\langle f, g \rangle_{\mathcal{H}} = \int_\Omega f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

then the quadratic form

$$\int_\Omega \int_\Omega K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z}$$

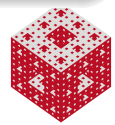
is just $\langle \mathcal{K}f, f \rangle_{\mathcal{H}}$, so that an integrally positive definite kernel is the kernel of a positive integral operator.

Remark

Bochner [Boc33] showed that the two notions of positive definiteness are equivalent for continuous kernels.

Remark

- *Unfortunately, analysts in the early 20^{th} century (such as [Mer09, Mat23, Boc32]) defined positive definite functions/kernels in analogy to positive semi-definite matrices.*
- *This concept is not strong enough to guarantee a non-singular interpolation matrix and so [Mic86] later defined strictly positive definite functions (see also [Fas07]).*
- *This leads to rather unfortunate differences in terminology used in the context of matrices, kernels and functions.*
- *Sometimes the literature is confusing about this and you might find papers that use the term positive definite function/kernel in the strict sense, and others that use it in the way we defined it above. Hopefully, the authors are clear about their use of terminology.*



Example

The (complex-valued) kernel

$$K_e(\boldsymbol{x}, \boldsymbol{z}) = e^{i(\boldsymbol{x}-\boldsymbol{z})\cdot \boldsymbol{t}}, \qquad \boldsymbol{t} \in \mathbb{R}^d \text{ fixed},$$

is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$ since its quadratic form is

$$\begin{aligned}
\sum_{j=1}^{N}\sum_{k=1}^{N} c_j c_k K_e(\boldsymbol{x}_j, \boldsymbol{x}_k) &= \sum_{j=1}^{N}\sum_{k=1}^{N} c_j c_k e^{i(\boldsymbol{x}_j-\boldsymbol{x}_k)\cdot \boldsymbol{t}} \\
&= \sum_{j=1}^{N} c_j e^{i\boldsymbol{x}_j\cdot \boldsymbol{t}} \sum_{k=1}^{N} c_k e^{-i\boldsymbol{x}_k\cdot \boldsymbol{t}} \\
&= \left| \sum_{j=1}^{N} c_j e^{i\boldsymbol{x}_j\cdot \boldsymbol{t}} \right|^2 \geq 0.
\end{aligned}$$

Example

The cosine function gives rise to a positive definite kernel on $\mathbb{R} \times \mathbb{R}$.

First we remember that for any $x, z \in \mathbb{R}$

$$\cos(x - z) = \frac{1}{2}\left(e^{i(x-z)} + e^{-i(x-z)}\right).$$

By our earlier example the kernel $K_e(\boldsymbol{x}, \boldsymbol{z}) = e^{i(\boldsymbol{x}-\boldsymbol{z})\cdot\boldsymbol{t}}$ is positive definite on $\mathbb{R}^d \times \mathbb{R}^d$ for any fixed $\boldsymbol{t} \in \mathbb{R}^d$.
Therefore, $K_1(x, z) = e^{i(x-z)}$ and $K_2(x - z) = e^{-i(x-z)}$ are positive definite on $\mathbb{R} \times \mathbb{R}$.
We will show below that the sum of two positive definite kernels is positive definite. Thus,

$$K(x, z) = \cos(x - z)$$

is positive definite.

Hilbert–Schmidt Operators

Definition

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ a bounded linear operator. The operator $\mathcal{T}$ is called a Hilbert–Schmidt operator if there is an orthonormal basis $\{e_n\}$ in $\mathcal{H}$ such that

$$\sum_{n=1}^{\infty} \|\mathcal{T} e_n\|^2 < \infty,$$

where $\|\cdot\|$ is the norm in $\mathcal{H}$ induced by its inner product $\langle \cdot, \cdot \rangle$.

$\|\mathcal{T}\|_{HS} = \sqrt{\sum_{n=1}^{\infty} \|\mathcal{T} e_n\|^2}$ is called the Hilbert–Schmidt norm of $\mathcal{T}$.

Remark

If $\mathcal{H}$ is finite-dimensional, i.e., $\mathcal{T}$ is a matrix, then the HS-norm turns into the Frobenius norm.

Trace of a bounded linear operator

The trace of an $N \times N$ matrix T is defined as

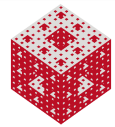
$$\operatorname{trace} \mathsf{T} = \sum_{i=1}^{N} \mathsf{T}_{ii} = \sum_{i=1}^{N} \boldsymbol{e}_i^T \mathsf{T} \boldsymbol{e}_i.$$

Analogously, for a bounded linear operator $\mathcal{T}$ we define

$$\operatorname{trace} \mathcal{T} = \sum_{n=1}^{\infty} \langle \mathcal{T} e_n, e_n \rangle.$$

Sometimes the HS-norm is expressed in terms of the trace.
If $\mathcal{T}^*$ is the adjoint of $\mathcal{T}$ (so that $\langle \mathcal{T} f, g \rangle = \langle f, \mathcal{T}^* g \rangle$) then

$$\|\mathcal{T}\|_{HS} = \sqrt{\operatorname{trace}(\mathcal{T}^* \mathcal{T})}.$$



Hilbert–Schmidt integral operators and their kernels

Theorem

Let $\mathcal{H} = L_2(\Omega, \rho)$ be a Hilbert space on $\Omega \subseteq \mathbb{R}^d$ and ρ a weight function such that $\int_\Omega \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. Further, let the kernel $K : (\boldsymbol{x}, \boldsymbol{z}) \mapsto K(\boldsymbol{x}, \boldsymbol{z})$ be in $L_2(\Omega \times \Omega, \rho \times \rho)$, i.e., assume that

$$\int_\Omega \int_\Omega |K(\boldsymbol{x}, \boldsymbol{z})|^2 \rho(\boldsymbol{x})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} < \infty.$$

Then the operator $\mathcal{K}$ defined by

$$(\mathcal{K}f)(\boldsymbol{x}) = \int_\Omega K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{z})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \qquad f \in L_2(\Omega), \qquad (\star)$$

is a Hilbert–Schmidt operator.
Conversely, every Hilbert–Schmidt operator on $L_2(\Omega, \rho)$ is of the form $(\star)$ for some unique kernel $K : (\boldsymbol{x}, \boldsymbol{z}) \mapsto K(\boldsymbol{x}, \boldsymbol{z})$ in $L_2(\Omega \times \Omega, \rho \times \rho)$.

The Hilbert–Schmidt eigenvalue problem on $L_2(\Omega, \rho)$

This eigenvalue problem can be viewed as a homogeneous Fredholm integral equation of the second kind, i.e., for appropriate eigenvalues λ and eigenfunctions φ we have

$$\int_\Omega K(\boldsymbol{x}, \boldsymbol{z})\varphi(\boldsymbol{z})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \lambda\varphi(\boldsymbol{x}) \quad \Longleftrightarrow \quad (\mathcal{K}\varphi)(\boldsymbol{x}) = \lambda\varphi(\boldsymbol{x}).$$

Using the ρ-weighted L_2 inner product

$$\langle f, g\rangle_{L_2(\Omega,\rho)} = \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x})\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

we can also write this as

$$\langle K(\boldsymbol{x}, \cdot), \varphi\rangle_{L_2(\Omega,\rho)} = \lambda\varphi(\boldsymbol{x}),$$

which is reminiscent of the reproducing property, but of course applies only to the eigenfunctions of $\mathcal{K}$.

L_2-orthonormality of the eigenfunctions will play an important role in Mercer's theorem, i.e.,

$$\langle \varphi_m, \varphi_n \rangle_{L_2(\Omega,\rho)} = \int_\Omega \varphi_m(\boldsymbol{x})\varphi_n(\boldsymbol{x})\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \delta_{mn}.$$

If we assume that K is a reproducing kernel (more details later), then we can use its reproducing property to compute

$$\begin{aligned}
\langle \mathcal{K}f, g \rangle_{\mathcal{H}_K(\Omega)} &= \left\langle \int_\Omega K(\cdot, \boldsymbol{z}) f(\boldsymbol{z})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, g \right\rangle_{\mathcal{H}_K(\Omega)} \\
&= \int_\Omega \langle K(\cdot, \boldsymbol{z}), g \rangle_{\mathcal{H}_K(\Omega)} f(\boldsymbol{z})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \int_\Omega g(\boldsymbol{z}) f(\boldsymbol{z})\rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \langle g, f \rangle_{L_2(\Omega,\rho)}.
\end{aligned}$$

Thus $\mathcal{K}$ can be interpreted as the adjoint of the operator that continuously embeds $\mathcal{H}_K(\Omega)$ into $L_2(\Omega, \rho)$ (and so $\mathcal{H}_K(\Omega) \subset L_2(\Omega, \rho)$). Since $\mathcal{K}$ is self-adjoint we know that this embedding operator is $\mathcal{K}$ itself.

If we let

$$f = \varphi_m \quad \text{and} \quad g = \varphi_n$$

and employ the eigenvalue relation

$$\mathcal{K}\varphi_m = \lambda_m \varphi_m$$

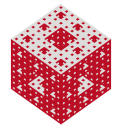
we see that

$$\begin{aligned} & \langle \mathcal{K}f, g \rangle_{\mathcal{H}_K(\Omega)} &&= \langle g, f \rangle_{L_2(\Omega,\rho)} \\ \iff \quad & \langle \mathcal{K}\varphi_m, \varphi_n \rangle_{\mathcal{H}_K(\Omega)} &&= \langle \varphi_n, \varphi_m \rangle_{L_2(\Omega,\rho)} = \delta_{nm}, \end{aligned}$$

so that

$$\langle \varphi_m, \varphi_n \rangle_{\mathcal{H}_K(\Omega)} = \begin{cases} 0, & m \neq n, \\ \frac{1}{\lambda_m}, & m = n, \end{cases}$$

i.e., the eigenfunctions are also orthogonal in the RKHS $\mathcal{H}_K(\Omega)$.



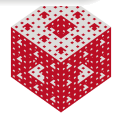
The spectral theory is similar as in the familiar finite-dimensional case:

- eigenvalues of a compact self-adjoint operator $\mathcal{K}$ are real
- eigenfunctions associated with different eigenvalues are orthogonal
- The spectral theorem implies

$$\mathcal{K}f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \lambda_n \langle f, \varphi_n \rangle \varphi_n(\boldsymbol{x}), \qquad f \in L_2(\Omega, \rho).$$

Remark

Since this identity holds for arbitrary L_2 functions f, one might hope for a series representation of the kernel K of $\mathcal{K}$ itself in terms of the eigenvalues and eigenfunctions $\rightarrow$ *Mercer's theorem*



Integral Eigenvalue Problem Example

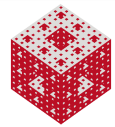
Example (Brownian motion kernel)

Consider the domain $\Omega = [0, 1]$, and let

$$K(x, z) = \min(x, z) = \begin{cases} x, & x \leq z, \\ z, & x > z. \end{cases}$$

Another way to write this kernel is

$$K(x, z) = \frac{1}{2}(x + z - |x - z|) = \begin{cases} \frac{1}{2}(x + z - (z - x)) = x, & x \leq z, \\ \frac{1}{2}(x + z - (x - z)) = z, & x > z. \end{cases}$$



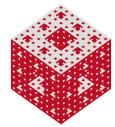
Example (cont.)

Start with the generic Hilbert–Schmidt integral eigenvalue problem

$$\mathcal{K}\varphi = \lambda\varphi \quad \Longleftrightarrow \quad \int_\Omega K(x,z)\varphi(z)\rho(z)\mathrm{d}z = \lambda\varphi(x)$$

and take $\Omega = [0,1]$, $\rho(x) \equiv 1$ and $K(x,z) = \min(x,z)$, i.e.,

$$\begin{aligned} \int_0^1 \min(x,z)\varphi(z)\mathrm{d}z &= \lambda\varphi(x) \\ \Longleftrightarrow \quad \int_0^x z\varphi(z)\mathrm{d}z + \int_x^1 x\varphi(z)\mathrm{d}z &= \lambda\varphi(x) \end{aligned}$$



Example (cont.)

Apply the differential operator $\frac{d^2}{dx^2}$ to the integral equation, i.e.,

$$\frac{d^2}{dx^2}\left[\int_0^x z\varphi(z)dz + \int_x^1 x\varphi(z)dz\right] = \frac{d^2}{dx^2}\left[\lambda\varphi(x)\right]$$

$$\frac{d^2}{dx^2}\left[\int_0^x z\varphi(z)dz - x\int_1^x \varphi(z)dz\right] = \lambda\varphi''(x)$$

$$\frac{d}{dx}\left[x\varphi(x) - \int_1^x \varphi(z)dz - x\varphi(x)\right] = \lambda\varphi''(x)$$

$$-\frac{d}{dx}\left[\int_1^x \varphi(z)dz\right] = \lambda\varphi''(x)$$

$$-\varphi(x) = \lambda\varphi''(x) \quad \Longleftrightarrow \quad -\varphi''(x) = \frac{1}{\lambda}\varphi(x)$$

Therefore, the eigenvalues of the integral operator $\mathcal{K}$ correspond to reciprocals of eigenvalues of the differential operator $\mathcal{L} = -\frac{d^2}{dx^2}$. The eigenfunctions are the same. We will solve this ODE eigenvalue problem later.

Differential and Integral Operators

Remark

- *Later it will be important to have an inverse relation between the integral operator $\mathcal{K}$ and the differential operator $\mathcal{L}$ (as just illustrated), which has K as its Green's kernel:*
  - *All regular ordinary differential operators have compact inverse integral operators and vice versa.*
  - *Moreover, if the differential operator is self-adjoint, so is the inverse integral operator (see, e.g., [CH53, Chapter V], [Hoc73, Chapter 3], [AG93, Appendix II]).*
  - *As a special case, Sturm–Liouville eigenvalue problems are inverse to integral eigenvalue problems for compact integral operators.*
- *As a consequence, we will see — as in the example — that both operators have the same eigenfunctions, and that the eigenvalues of the differential operator are reciprocals of the eigenvalues of the integral operator.*

Theorem (Mercer's theorem [Mer09])

Let $\Omega \subset \mathbb{R}^d$, let ρ be a weight function and $K \in L_2(\Omega \times \Omega, \rho \times \rho)$ be a kernel with positive integral operator

$$(\mathcal{K}f)(\boldsymbol{x}) = \int_\Omega K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{z}) \rho(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}.$$

Let $\varphi_n \in L_2(\Omega, \rho)$, $n = 1, 2, \ldots,$ be the $L_2(\Omega, \rho)$ orthonormal eigenfunctions of $\mathcal{K}$ associated with the eigenvalues $\lambda_n > 0$. Then the following are true:

(1) *The eigenvalues $\{\lambda_n\}_{n=1}^\infty$ are absolutely summable, and so $\mathcal{K}$ has finite trace.*

(2) *The kernel has a Mercer expansion*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{n=1}^\infty \lambda_n \varphi_n(\boldsymbol{x}) \varphi_n(\boldsymbol{z})$$

which converges absolutely and uniformly on Ω.

Remark

- *Mercer's theorem provides an infinite series representation (or an eigenfunction expansion) of a positive definite kernel.*
- *A transparent proof for a continuous kernel K in the case $\Omega = [0, 1]$ and $\rho(z) = z$ can be found in [Hoc73, pg. 90].*
- *A general proof is given in [Kön86, Chapter 3].*
- *Mercer's theorem guarantees uniform convergence of the series whenever $\mathcal{K}$ is a positive operator. A related result by Schmidt [Sch07] establishes only L_2 convergence of the series, but for arbitrary compact self-adjoint operators $\mathcal{K}$.*
- *A modern discussion of Mercer's theorem with a number of generalizations can be found in [SS12, Sun05].*

The Karhunen-Loève expansion theorem can be viewed as a corollary to Mercer's theorem (but in the stochastic process setting).
It plays an important role in polynomial chaos approximations and uncertainty quantification.

Theorem (Karhunen–Loève expansion)

A centered mean-square continuous stochastic process Y with continuous covariance kernel $K : \Omega \times \Omega \to \mathbb{R}$ has an orthogonal expansion of the form

$$Y_{\boldsymbol{x}}(\omega) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} Z_n(\omega) \varphi_n(\boldsymbol{x}),$$

where $\lambda_n > 0$ are the eigenvalues and $\varphi_n \in \mathcal{H}_K(\Omega)$ are the associated eigenfunctions of $\mathcal{K}$, and Z_n are "orthonormal" random variables

$$Z_n(\omega) = \frac{1}{\sqrt{\lambda_n}} \int_{\Omega} Y_{\boldsymbol{x}}(\omega) \varphi_n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

such that $\boldsymbol{E}[Z_m Z_n] = \delta_{mn}$.

Reproducing kernel Hilbert spaces (RKHSs)

Reproducing kernels were introduced by Aronszajn and Bergman in the first half of the 20th century (see [Aro50, Ber50]).

Definition

Let $\Omega \subseteq \mathbb{R}^d$ and let $\mathcal{H}_K(\Omega)$ be a real Hilbert space of functions $f : \Omega \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$. A kernel $K : \Omega \times \Omega \to \mathbb{R}$ is called reproducing kernel for $\mathcal{H}_K(\Omega)$ if

(1) $K(\cdot, \boldsymbol{x}) \in \mathcal{H}_K(\Omega)$ for all $\boldsymbol{x} \in \Omega$,

(2) $\langle f, K(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_K(\Omega)} = f(\boldsymbol{x})$ for all $f \in \mathcal{H}_K(\Omega)$ and all $\boldsymbol{x} \in \Omega$.

The name *reproducing kernel* is motivated by the reproducing property (2), which states that the reproducing kernel at $\boldsymbol{x}$, $K(\cdot, \boldsymbol{x})$, is the Riesz representer of function evaluation at $\boldsymbol{x}$.

Properties of reproducing kernels

1. For all $\boldsymbol{x}, \boldsymbol{z} \in \Omega$

$$\langle K(\cdot, \boldsymbol{x}), K(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_K(\Omega)} = K(\boldsymbol{x}, \boldsymbol{z}) = K(\boldsymbol{z}, \boldsymbol{x}).$$

2. For every $f \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$ we have

$$|f(\boldsymbol{x})| \leq \sqrt{K(\boldsymbol{x}, \boldsymbol{x})} \|f\|_{\mathcal{H}_K(\Omega)}.$$

$$K(\boldsymbol{x}, \boldsymbol{x}) = \langle K(\cdot, \boldsymbol{x}), K(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_K(\Omega)} = \|K(\cdot, \boldsymbol{x})\|^2_{\mathcal{H}_K(\Omega)} \geq 0.$$

Remark

(2) shows that reproducing kernel Hilbert spaces are special "smooth" Hilbert spaces in which point evaluation is bounded, i.e., continuous. So, in an RKHS, values of functions at nearby points are closely related to each other (since this is just what the concept of continuity implies).

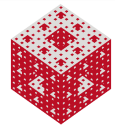
Properties of reproducing kernels (cont.)

3. If $\mathcal{H}_K$ is a reproducing kernel Hilbert space with reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}$, then K is positive definite.

4. Moreover, K is strictly positive definite if and only if the point evaluation functionals $\delta_{\boldsymbol{x}}$ are linearly independent in $\mathcal{H}_K^*$.

Remark

Here the space of bounded linear functionals on $\mathcal{H}_K$ is known as its dual, and denoted by $\mathcal{H}_K^*$.



Proof.

We analyze the quadratic form of K.
For distinct points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and arbitrary $\boldsymbol{c} \in \mathbb{R}^N$ we have

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j \langle K(\cdot, \boldsymbol{x}_i), K(\cdot, \boldsymbol{x}_j)\rangle_{\mathcal{H}_K(\Omega)} \\
&= \left\langle \sum_{i=1}^{N} c_i K(\cdot, \boldsymbol{x}_i), \sum_{j=1}^{N} c_j K(\cdot, \boldsymbol{x}_j) \right\rangle_{\mathcal{H}_K(\Omega)} \\
&= \left\| \sum_{i=1}^{N} c_i K(\cdot, \boldsymbol{x}_i) \right\|^2_{\mathcal{H}_K(\Omega)} \geq 0.
\end{aligned}$$

Thus K is positive definite.

Proof (cont.).

To establish the second claim we assume K is not strictly positive definite and show that the point evaluation functionals must be linearly dependent.

If K is not strictly positive definite then there exist distinct points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and nonzero coefficients c_i such that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.$$

The same manipulation of the quadratic form as above therefore implies

$$\sum_{i=1}^{N} c_i K(\cdot, \boldsymbol{x}_i) = 0.$$

Proof (cont.).

Now we take the Hilbert space inner product with an arbitrary function $f \in \mathcal{H}_K$ and use the reproducing property of K to obtain

$$\begin{aligned} 0 &= \left\langle f, \sum_{i=1}^{N} c_i K(\cdot, \boldsymbol{x}_i) \right\rangle_{\mathcal{H}_K(\Omega)} \\ &= \sum_{i=1}^{N} c_i \langle f, K(\cdot, \boldsymbol{x}_i) \rangle_{\mathcal{H}_K(\Omega)} \\ &= \sum_{i=1}^{N} c_i f(\boldsymbol{x}_i) = \sum_{i=1}^{N} c_i \delta_{\boldsymbol{x}_i}(f). \end{aligned}$$

This, however, implies the linear dependence of the point evaluation functionals $\delta_{\boldsymbol{x}_i}(f) = f(\boldsymbol{x}_i)$, $i = 1, \ldots, N$, since the coefficients c_i were assumed to be not all zero.

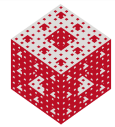
An analogous argument can be used to establish the converse. □

Remark

(3) and (4) provide one direction of the connection between strictly positive definite kernels and reproducing kernels.

However, we are also interested in the other direction.

Since our kernel-based approximation methods generally use strictly positive definite kernels, we want to know how to construct an associated reproducing kernel Hilbert space. We provide that discussion in Chapter 2 — Part 3.



Properties of reproducing kernels (cont.)

5. If K_1 and K_2 are reproducing kernels of spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, on the same domain Ω, then $K = K_1 + K_2$ is the reproducing kernel of the direct sum $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$.
   If $\mathcal{H}_1 \cap \mathcal{H}_2 = \{0\}$, then since $\mathcal{H}$ is a direct sum, $\mathcal{H}_2$ is the orthogonal complement of $\mathcal{H}_1$ in $\mathcal{H}$.

6. If $K : \Omega \times \Omega \to \mathbb{R}$ is the reproducing kernel of $\mathcal{H}$ and $\Omega_0 \subset \Omega$, then K_0, the restriction of K to $\Omega_0 \times \Omega_0$, is the reproducing kernel of $\mathcal{H}_0$, a space whose elements are restrictions of elements of $\mathcal{H}$ to Ω_0.

Remark

(5) can be generalized to non-negative linear combinations [SC08].
(6) can also be generalized and then shows that a complex-valued kernel when restricted from a complex domain to a real subdomain is not only a complex-valued kernel, but if the kernel is real-valued then it is also a kernel in the purely real sense [SC08, Lemmas 4.3 & 4.4].

Properties of reproducing kernels (cont.)

7. If K_1 and K_2 are reproducing kernels of spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ on domains Ω_1 and Ω_2, respectively, then the tensor product kernel

$$K((\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{z}_1, \boldsymbol{z}_2)) = K_1(\boldsymbol{x}_1, \boldsymbol{z}_1) K_2(\boldsymbol{x}_2, \boldsymbol{z}_2)$$

is the reproducing kernel of the tensor product space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$.

Remark

Tensor products are useful for "time-space" applications (possible PDE project?), where K_1 is a kernel in the spatial domain and K_2 in time, or to construct anisotropic kernels.

Examples

So far we have seen

- $K(\boldsymbol{x}, \boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{z}\|$ — norm (or distance) kernel
- $K(\boldsymbol{x}, \boldsymbol{z}) = e^{-\varepsilon^2 \|\boldsymbol{x}-\boldsymbol{z}\|^2}$ — Gaussian kernel
- $K(\boldsymbol{x}, \boldsymbol{z}) = e^{i(\boldsymbol{x}-\boldsymbol{z})\cdot \boldsymbol{t}}$ — special complex-valued kernel
- $K(x, z) = \cos(x - z)$ — cosine kernel
- $K(x, z) = \min(x, z)$ — Brownian motion kernel

We will give many more examples in Chapter 3.

Remark

Of these, only the Gaussian kernel is strictly positive definite. The last three are positive definite (i.e., have a zero eigenvalue), and the first is conditionally negative definite.

Remark

While the space $L_2(\Omega)$ is a Hilbert space, it is not a reproducing kernel Hilbert space.

Reasons:

- *$L_2(\Omega)$ is not a space of functions, but of equivalence classes of functions.*
- *While the delta functional acts as its "reproducing kernel", i.e.,*

$$\int_\Omega f(\boldsymbol{z})\delta(\boldsymbol{x}-\boldsymbol{z})\mathrm{d}\boldsymbol{z} = f(\boldsymbol{x}),$$

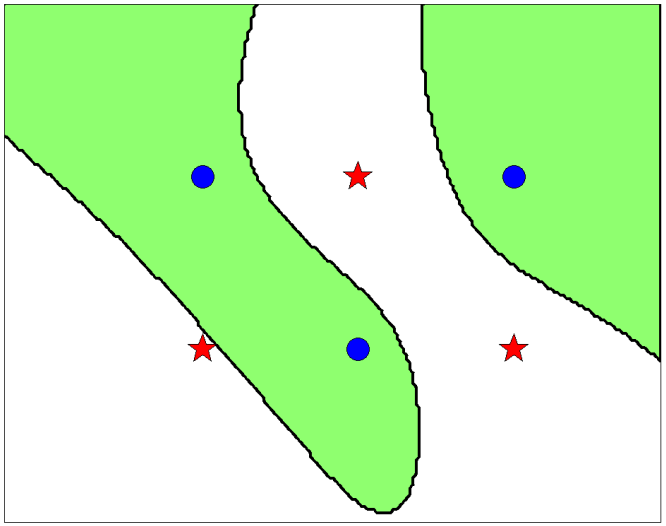
it does not belong to $L_2(\Omega)$. It belongs to the dual of the test function space, and is therefore technically a distribution.

Remark

On the other hand, the Kronecker delta symbol is a reproducing kernel for $\ell_2(\mathbb{R})$. However, $\ell_2(\mathbb{R})$ is not a Hilbert space of functions, but of real-valued sequences.

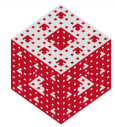
The Role of Kernels in Machine Learning

Support vector machines are used in machine learning to separate/classify data given in the input space.
Ideally, one wants to do this with a hyperplane. However, using a linear separation works only for very limited cases.



Therefore, we want a nonlinear separation.

It turns out that the setting of reproducing kernel Hilbert spaces provides a perfect framework to accomplish this while still applying linear techniques.



Feature Map

Lemma ([BTA04])

Consider a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a map $\Phi : \Omega \to \mathcal{H}$. The kernel $K : \Omega \times \Omega \to \mathbb{R}$ such that

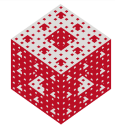
$$K(\boldsymbol{x}, \boldsymbol{z}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle_{\mathcal{H}}$$

is a positive definite kernel. The map Φ is called a feature map.

Thus, a kernel is just an inner product. The canonical map is given by $\Phi(\boldsymbol{x}) = K(\cdot, \boldsymbol{x})$, and we saw earlier that

$$\langle K(\cdot, \boldsymbol{x}), K(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}_K(\Omega)} = K(\boldsymbol{x}, \boldsymbol{z}).$$

A consequence of this lemma is the characterization of all possible reproducing kernels on $\Omega \times \Omega$ in terms of maps into the sequence space $\ell_2(\Omega)$.



Theorem

A function K defined on $\Omega \times \Omega$ is a reproducing kernel if and only if there exists a mapping $T : \Omega \to \ell_2(A)$, where A is an index set, such that for all $\boldsymbol{x}, \boldsymbol{z} \in \Omega$

$$K(\boldsymbol{x}, \boldsymbol{z}) = \langle T_{\boldsymbol{x}}, T_{\boldsymbol{z}} \rangle_{\ell_2(A)} = \sum_{\alpha \in A} (T_{\boldsymbol{x}})_\alpha (T_{\boldsymbol{z}})_\alpha.$$

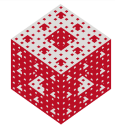
Proof.

[BTA04, Theorem 4]. □

This theorem is both

- a (non-unique) factorization theorem for the kernel K reminiscent of the spectral theorem for symmetric matrices,
- a way to construct reproducing kernels starting from a map T.

Given a feature map $\Phi : \Omega \to \mathcal{H}$, the map T can be viewed as the composition of Φ with an isometry that maps $\mathcal{H}$ to $\ell_2(\Omega)$.



Example (Brownian motion kernel)

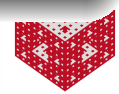
Take $\Omega = [0,1]$, $\mathcal{H} = \ell_2(\mathbb{N})$, and

$$T_x = (\sqrt{\lambda_1}\varphi_1(x), \sqrt{\lambda_2}\varphi_2(x), \ldots),$$

where $\lambda_n = \frac{4}{(2n-1)^2\pi^2}$ and $\varphi_n(x) = \sqrt{2}\sin\left((2n-1)\frac{\pi x}{2}\right)$ are eigenvalues and eigenfunctions of the Brownian motion kernel $K(x,z) = \min(x,z)$. Then

$$\begin{aligned} K(x,z) &= \langle T_x, T_z \rangle_{\ell_2(\mathbb{N})} \\ &= \sum_{n=1}^{\infty} \frac{8}{(2n-1)^2\pi^2} \sin\left((2n-1)\frac{\pi x}{2}\right) \sin\left((2n-1)\frac{\pi z}{2}\right) \\ &= \min(x,z). \end{aligned}$$

This is just the Mercer series for the Brownian motion kernel.



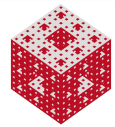
Closed form expression for the series is HW.

Remark

One can find the eigenvalues and eigenfunctions of the Brownian motion kernel

- *starting from the eigenvalue problem for the integral operator $\mathcal{K}$ with kernel $K(x,z) = \min(x,z)$,*
- *transforming it to its "inverse" differential eigenvalue problem,*
- *solving the resulting Sturm–Liouville eigenvalue problem.*

We will discuss this connection later.



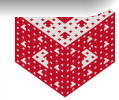
Example (Brownian motion kernel, Take 2)

Alternatively, we consider $T_x = \mathbf{1}_{[0,x]}$, with $\mathbf{1}_{[a,b]}$ the indicator function of $[a, b]$, i.e.,

$$\mathbf{1}_{[a,b]}(x) = \begin{cases} 1 & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

Also take $\mathcal{H} = L_2([0, 1])$, which is isomorphic to $\ell_2(\mathbb{N})$ via identification of an $f \in \mathcal{H}$ with its generalized Fourier coefficient. This gives us

$$\begin{aligned} K(x, z) &= \langle T_x, T_z \rangle_{L_2([0,1])} \\ &= \int_0^1 \mathbf{1}_{[0,x]}(t)\mathbf{1}_{[0,z]}(t)\mathrm{d}t \\ &= \int_0^{\min(x,z)} \mathrm{d}t = \min(x, z). \end{aligned}$$



Example (Brownian motion kernel, Take 3)

Take the canonical feature map $T_x = K(\cdot, x) = \min(\cdot, x)$, and $\mathcal{H} = \mathcal{H}_K = \{f \in H^1([0,1]), f(0) = 0\}$ the RKHS with inner product $\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'(t)g'(t)\mathrm{d}t$. Then we have

$$\begin{aligned}\langle T_x, T_z\rangle_{\mathcal{H}} &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\min(t,x)\frac{\mathrm{d}}{\mathrm{d}t}\min(t,z)\mathrm{d}t \\ &= \int_0^1 \mathbf{1}_{[0,x]}(t)\mathbf{1}_{[0,z]}(t)\mathrm{d}t = \min(x,z).\end{aligned}$$

Remark

Note that we also know that $\langle K(\cdot,x), K(\cdot,z)\rangle_{\mathcal{H}_K} = K(x,z)$*, i.e., without having to integrate,*

$$\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\min(t,x)\frac{\mathrm{d}}{\mathrm{d}t}\min(t,z)\mathrm{d}t = \min(x,z).$$

The "Kernel Trick" in Machine Learning

The canonical feature map, i.e., $T_{\boldsymbol{x}} : \Omega \to \mathcal{H}_K$ such that

$$\boldsymbol{x} \mapsto T_{\boldsymbol{x}} = K(\cdot, \boldsymbol{x})$$

transforms a given problem from the input space Ω (think all sorts of stuff here) to the Hilbert space $\mathcal{H}_K$ (think "nice" math with mostly linear algorithms) via the reproducing kernel K.

Since Ω can be very general (e.g., texts, images, medical data, etc.) the feature map is at the heart of applications of kernel methods to problems in machine learning and its application is known there as the kernel trick (see, e.g., [SS02]).

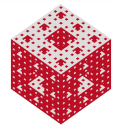
In essence, the feature map allows us to compute with all sorts of quantities that are not at all numerical by turning, e.g., a Shakespeare sonnet $\boldsymbol{x}$ into a function $T_{\boldsymbol{x}}$ that expresses — via the kernel K — the similarity of $\boldsymbol{x}$ to all other texts in the collection Ω.

Remark

A particularly attractive feature of the kernel trick is the fact that the actual (nonlinear) feature map need not be known; simply working with the kernel K is sufficient and easy.

In fact, all the techniques for determining optimal classifiers via structural risk minimization are essentially the same as our techniques for finding optimal kernel-based approximants.

While we have not investigated this yet, the kernel-based solution to the scattered data interpolation problem is such an optimal approximant. That's one reason that problem is so fundamental.



References I

[AG93] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, Dover Publications, 1993, Originally published in 1961.

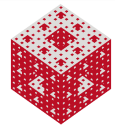
[Aro50] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society **68** (1950), no. 3, 337–404.

[Ber50] S. Bergman, *The Kernel Function and Conformal Mapping*, 2nd ed., American Mathematical Society, 1950.

[Boc32] S. Bochner, *Vorlesungen über Fouriersche Integrale*, Mathematik und ihre Anwendungen, vol. 12, Akademische Verlagsgesellschaft, Leipzig, 1932, Translated as: Lectures on Fourier Integrals. (AM-42) Princeton University Press, 1959.

[Boc33] _____, *Monotone Funktionen, Stieltjes Integrale und harmonische Analyse*, Math. Ann. **108** (1933), 378–410.

[Bre10] Haim Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2010.



References II

[BTA04] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Dordrecht, 2004.

[CH53] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol. 1*, Wiley, 1953, Reprinted 1989.

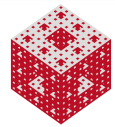
[Fas07] G. E. Fasshauer, *Meshfree Approximation Methods with* MATLAB, Interdisciplinary Mathematical Sciences, vol. 6, World Scientific Publishing Co., Singapore, 2007.

[Hoc73] Harry Hochstadt, *Integral Equations*, Wiley, 1973.

[Kön86] H König, *Eigenvalue Distribution of Compact Operators*, Birkhäuser Verlag, Basel; Boston, 1986.

[Mat23] M. Mathias, *Über positive Fourier-Integrale*, Math. Zeit. **16** (1923), 103–125.

[Men28] K. Menger, *Die Metrik des Hilbertschen Raumes*, Anzeiger der Akad. der Wissenschaften in Wien, Nat. Kl. **65** (1928), 159–160.



References III

[Mer09] J. Mercer, *Functions of positive and negative type, and their connection with the theory of integral equations*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **209** (1909), no. 441-458, 415–446.

[Mic86] C. A. Micchelli, *Interpolation of scattered data: Distance matrices and conditionally positive definite functions*, Constructive Approximation **2** (1986), no. 1, 11–22.

[SC08] I. Steinwart and A. Christmann, *Support Vector Machines*, Information Science and Statistics, Springer, New York, 2008.

[Sch07] E. Schmidt, *Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener*, Math. Ann. **63** (1907), 433–476.

[SS02] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, 2002.

References IV

[SS12] I. Steinwart and C. Scovel, *Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs*, Constructive Approximation **35** (2012), 363–417.

[Sun05] Hongwei Sun, *Mercer theorem for RKHS on noncompact sets*, Journal of Complexity **21** (2005), no. 3, 337–349.