

# Hilbert space methods

Risi Kondor

June 22, 2008

## 1 Positive definite functions

**Positive (semi-)definite function.** A function  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is said to be positive (semi-)definite if (a)  $k$  is symmetric ( $k(x, x') = k(x', x)$ ), (b) for any  $m \in \{1, 2, \dots\}$ , any  $x_1, x_2, \dots, x_m \in \mathcal{X}$  and any  $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$ ,

$$\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) \geq 0. \quad (1)$$

A function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be positive (semi-)definite if (a)  $k(x) = k(-x)$ , (b) for any  $m \in \{1, 2, \dots\}$ , any  $x_1, x_2, \dots, x_m \in \mathcal{X}$  and any  $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$ ,

$$\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j g(x_i - x_j) \geq 0. \quad (2)$$

**Conditionally positive (semi-)definite function.** A function  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is said to be conditionally positive (semi-)definite if (a)  $k$  is symmetric ( $k(x, x') = k(x', x)$ ), (b) for any  $m \in \{1, 2, \dots\}$ , any  $x_1, x_2, \dots, x_m \in \mathcal{X}$  and any  $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$  satisfying  $\sum_{i=1}^m \alpha_i = 0$ ,

$$\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) \geq 0. \quad (3)$$

A function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be positive (semi-)definite if (a)  $k(x) = k(-x)$ , (b) for any  $m \in \{1, 2, \dots\}$ , any  $x_1, x_2, \dots, x_m \in \mathcal{X}$  and any  $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$  satisfying  $\sum_{i=1}^m \alpha_i = 0$ ,

$$\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j g(x_i - x_j) \geq 0. \quad (4)$$

**Bochner/Mathias theorems** Mathias' theorem (1923) states that if  $f: \mathcal{R} \rightarrow \mathbb{R}$  is a positive semi-definite with Fourier transform  $\hat{f}$ , then  $\hat{f}(\omega) \geq 0$  for

all  $\omega \in \mathbb{R}$ . Bochner generalized this in 1933 to functions which do not have Fourier transforms by proving that if  $f$  is continuous and positive definite, then there is a bounded monotone increasing function  $V$  such that  $f(x) = \int e^{ikx} dV(k)$ . Both theorems have generalizations to  $\mathbb{R}^n$ , and compact groups in general.

## 2 Kernels

**Kernel.** In machine learning a kernel on a space  $\mathcal{X}$  is a function  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  which is positive semi-definite.

**Translation invariant kernel.** A kernel  $k: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  is called translation invariant if  $k(x, x') = k(x - x')$  for some function  $k: \mathbb{R}^n \rightarrow \mathbb{R}$ .

**Kernel vector.** Given a kernel  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  and a point  $x \in \mathcal{X}$ , the corresponding kernel vector is the function  $k_x: \mathcal{X} \rightarrow \mathbb{R}$ ,  $k_x(x') = k(x, x')$ , viewed as a vector in a space of functions.

**Kernel map.** The kernel map is a function from  $\mathcal{X}$  to a Hilbert space  $\mathcal{H}$  such that  $k(x, x') = \langle \phi(x), \phi(x') \rangle$ .

**Kernel operator.** The kernel operator  $\mathcal{K}$  is the linear operator

$$[\mathcal{K}f](x) = \int k(x, x') f(x') dx'.$$

**Gram matrix.** The Gram matrix of a kernel  $k$  and a set of points  $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_m \in \mathcal{X}$  is the  $m \times m$  matrix  $K$  with elements  $K_{i,j} = k(x_i, x_j)$ .

## 3 Hilbert spaces

**Inner product.** An inner product on a space  $V$  is a function  $\langle \cdot, \cdot \rangle: V \times V \rightarrow \mathbb{R}$  satisfying

$$\begin{aligned} \langle u+v, w \rangle &= \langle u, w \rangle + \langle v, w \rangle, \\ \langle \alpha u, v \rangle &= \alpha \langle u, v \rangle, \\ \langle u, v \rangle &= \langle v, u \rangle, \\ \langle u, u \rangle &\geq 0 \text{ with equality only when } u = 0, \end{aligned}$$

for all  $u, v, w \in V$  and  $\alpha \in \mathbb{R}$ . The norm induced by the inner product is  $\|u\| = \sqrt{\langle u, u \rangle}$ .

**Cauchy sequence.** A sequence of points  $a_1, a_2, \dots$  in a normed space is called a Cauchy sequence if for any  $\epsilon > 0$ , there is an  $n \in \mathbb{N}$  such that  $\|a_i - a_j\| < \epsilon$  for all  $i, j \geq n$ .

**Complete normed space.** A normed space  $V$  is said to be complete if for any Cauchy sequence  $a_1, a_2, \dots \in V$ , there is an  $a \in V$  such that  $\lim_{i \rightarrow \infty} a_i = a$ .

**Hilbert space.** A Hilbert space is a vector space  $V$  endowed with an inner product which is complete with respect to the norm induced by the inner product.

**Reproducing kernel Hilbert space (RKHS).** The RKHS  $\mathcal{F}_k$  associated to the kernel  $k$  is the space of finite linear combinations of the functions  $\{k_x \mid x \in \mathcal{X}\}$  endowed with an inner product satisfying  $\langle k_x, k_{x'} \rangle = k(x, x')$ , and completed with respect to the norm induced by the inner product. Clearly, in this case the kernel map is  $\phi(x) = k_x$ .

**Reproducing property.** The reproducing property of the RKHS  $\mathcal{F}_k$  is that for any  $f \in \mathcal{F}$ , and any  $x \in \mathcal{X}$ ,  $f(x) = \langle f, k_x \rangle$ . In particular,  $k_{x'}(x) = \langle k_{x'}, k_x \rangle = k(x, x')$ .

**Representer theorem.** The representer theorem states that in an RKHS  $\mathcal{F}$  induced by a kernel  $k$ , for any increasing function  $\Omega$ , the solution of the regularized risk minimization problem

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \left[ \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i) + \Omega(\langle f, f \rangle) \right]$$

may be expressed in the form

$$f(x) = \sum_{i=1}^m \alpha_i k(x_i, x)$$

for some  $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{R}$ .

**Mercer's theorem.** If  $\mathcal{X}$  is compact and the kernel  $k$  is continuous, then there is a basis  $\{\phi_i\}_i$  of  $L_2(\mathcal{X})$  such that

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x'),$$

where the  $\phi_i$  are eigenvectors of  $\mathcal{K}$ , the  $\lambda_i$  are the corresponding eigenvalues and the convergence is absolute and uniform on  $\mathcal{X} \times \mathcal{X}$ . This addresses Yee Whye's question: if  $\mathcal{X}$  is compact and  $k$  is continuous, then  $k$  can be represented in a Hilbert space of countable dimensionality. However, not all kernels satisfy these conditions, and not all Hilbert spaces are of countable dimension.

## 4 Gaussian processes

Recall that the general form of ( $L_2$  regularized) Hilbert space learning algorithms is

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \left[ \frac{1}{m} \sum_{i=1}^m L(f(x_i), y_i) + \lambda \langle f, f \rangle \right]. \quad (5)$$

A (centered) Gaussian process on  $\mathcal{X}$  with covariance function  $k$  is a distribution over functions  $f: \mathcal{X} \rightarrow \mathbb{R}$  such that for any  $x_1, x_2, \dots, x_m \in \mathcal{X}$ ,

$$p(f(x_1) = t_1, \dots, f(x_m) = t_m) \sim e^{-t^\top K^{-1} t}, \quad (6)$$

where  $C$  is the Gram matrix with elements  $K_{i,j} = K(x_i, x_j)$ . We show that this is equivalent to the distribution on the RKHS  $\mathcal{F}$  induced by  $k$  given by

$$p(f) \propto e^{-\langle f, f \rangle / 2} \quad (7)$$

by decomposing  $\mathcal{F}$  into  $V = \text{span}\{k_{x_1}, k_{x_2}, \dots, k_{x_m}\}$  and its orthogonal complement  $V^\perp$  and noting that any  $f$  obeying  $f(x_1) = t_1, \dots, f(x_m) = t_m$  may be written as

$$f = f_V + f_\perp = \left( \sum_{i=1}^m \alpha_i k_{x_i} \right) + f_\perp$$

with  $f_\perp \in V^\perp$ . The vector of coefficients  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)^\top$  can be found from

$$f(x_j) = \sum_{i=1}^m \alpha_i k_{x_i}(x_j) = \sum_{i=1}^m \alpha_i k(x_i, x_j) = t_j$$

leading to the matrix equation  $K\alpha = t$ . We can marginalize  $p$  to  $V$  just as in the finite dimensional case by

$$p_V(f_V) = p(f(x_1) = t_1, \dots) \propto \int_{V^\perp} p(f_V + f_\perp) df_\perp = e^{-\langle f_V, f_V \rangle / 2} \quad (8)$$

and expand

$$\langle f_V, f_V \rangle = \sum_{i=1}^m \sum_{j=1}^m [K^{-1}t]_i [K^{-1}t]_j \langle k_{x_i}, k_{x_j} \rangle = t^\top K^{-1} K K^{-1} t = t^\top K^{-1} t$$

showing the equivalence to (6).

Assuming a noise model

$$p(y | x, f) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-f(x))^2 / (2\sigma^2)},$$

given data  $D = (x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ , the negative log posterior of the GP is then

$$-\log p(f|D) = \frac{1}{2} \langle f|f \rangle + \frac{1}{2\sigma^2} \sum_{i=1}^m (f(x_i) - y_i)^2,$$

so the MAP problem (in this case also known as ridge regression) is indeed of the form (5) with  $\lambda = \sigma^2/m$  and  $L(f(x), y) = (f(x) - y)^2$ .

## 5 Regularization theory

For  $f, g \in L_2(\mathcal{X})$ , regularization theory contrasts

$$\langle f, g \rangle_{L_2} = \int f(x) g(x) dx$$

with the RKHS inner product  $\langle f, g \rangle_{\mathcal{F}}$ . In particular, the regularization operator  $\Upsilon$  is the operator which satisfies

$$\langle f, g \rangle_{\mathcal{F}} = \langle \Upsilon f, \Upsilon g \rangle_{L_2}.$$

As a continuous limit of the expansion  $f = \sum_{i=1}^m \alpha_i k_{x_i}$ , we write

$$f(x) = \int k(x, x') \alpha(x) dx = [\mathcal{K}\alpha](x), \quad (9)$$

for some (generalized) function  $\alpha$ , in terms of which (using the fact that  $\mathcal{K}$  is invertible and self-adjoint)

$$\begin{aligned} \langle f, f \rangle_{\mathcal{F}} &= \int \int \alpha(x) k(x, x') \alpha(x') dx' dx = \langle \alpha, \mathcal{K}\alpha \rangle_{L_2} = \\ &= \langle \mathcal{K}\alpha, \mathcal{K}^{-1}\mathcal{K}\alpha \rangle_{L_2} = \langle \mathcal{K}^{-1/2}f, \mathcal{K}^{-1/2}f \rangle_{L_2}, \end{aligned}$$

showing that  $\Upsilon = \mathcal{K}^{-1/2}$ .

For translation invariant kernels (9) is just convolution

$$f(x) = \int k(x - x') \alpha(x') dx,$$

so, by the convolution theorem, in Fourier space  $\widehat{f}(\omega) = \widehat{k}(\omega) \widehat{\alpha}(\omega)$ . As a particular example, for the Gaussian RBF kernel

$$k(x, x') = e^{-(x-x')^2/(2\sigma^2)},$$

$\widehat{k}(\omega) = e^{-2\omega^2\sigma^2}$ , so  $\widehat{f}(\omega) = e^{-2\omega^2\sigma^2} \widehat{\alpha}(\omega)$  and by Parseval's theorem

$$\langle f, f \rangle_{\mathcal{F}} = \int \alpha(\omega)^\dagger e^{-2\omega^2\sigma^2} \alpha(\omega) d\omega = \int e^{2\omega^2\sigma^2} |f(\omega)|^2 d\omega$$

showing that

$$\widehat{\Upsilon}f(\omega) = e^{\omega^2\sigma^2} \widehat{f}(\omega),$$

i.e., the Gaussian kernel biases learning algorithms towards smoothness by penalizing functions by the energy at each frequency weighted by a factor exponential in the frequency squared.