# Lecture-05: PDS Kernels

## 1 PDS Kernels

**Definition 1.1 (Normalized kernels).** To any kernel $K$, we can associate a **normalized kernel** $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined for all $x, y \in \mathcal{X}$ by

$$K'(x,y) = \begin{cases} \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}, & K(x,x)K(y,y) \neq 0, \\ 0, & K(x,x)K(y,y) = 0. \end{cases}$$

*Remark* 1. For any $x \in \mathcal{X}$ such that $K(x,x) \neq 0$, we have $K'(x,x) = 1$.

**Example 1.2 (Gaussian kernel).** For $\sigma > 0$, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as $K(x,y) = \exp\left(\frac{\langle x,y \rangle}{\sigma^2}\right)$. The normalized kernel associated with this kernel is the **Gaussian kernel** $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with parameter $\sigma > 0$, defined for all $x, y \in \mathcal{X}$ as

$$K'(x,y) = \exp\left(\frac{1}{2\sigma^2}(2\langle x,y \rangle - \|x\|^2 - \|y\|^2)\right) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right).$$

### 1.1 Properties

**Lemma 1.3 (Normalized PDS kernels).** *Let $K$ be a PDS kernel. Then, the normalized kernel $K'$ associated to $K$ is PDS.*

*Proof.* Consider an $m$-sized sample $S = (x_1, \dots, x_m) \in \mathcal{X}^m$. We will show that the gram matrix $\mathbf{K}'$ generated by the sample $S$ and kernel $K'$ is SPSD. Symmetry of $K'$ follows from the symmetry of $K$, and hence the gram matrix $\mathbf{K}'$ is symmetric.

To see the positive semi-definiteness of the gram matrix $\mathbf{K}'$, we note that its $(i,j)$-th entry $\mathbf{K}'(x_i, x_j) = \frac{\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathbb{H}}}{\|\Phi(x_i)\|_{\mathbb{H}} \|\Phi(x_j)\|_{\mathbb{H}}}$. Hence, for any arbitrary vector $c \in \mathbb{R}^m$, we have

$$\sum_{i,j=1}^{m} c_i K'(x_i, x_j) c_j = \sum_{i,j=1}^{m} c_i \frac{K(x_i, x_j)}{\sqrt{K(x_i, x_i) K(x_j, x_j)}} c_j = \sum_{i,j=1}^{m} c_i \frac{\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathbb{H}}}{\|\Phi(x_i)\|_{\mathbb{H}} \|\Phi(x_j)\|_{\mathbb{H}}} c_j = \left\| \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_{\mathbb{H}}} \right\|_{\mathbb{H}}^2 \geqslant 0.$$

□

Advantages of working with kernel is that no explicit definition of a feature map $\Phi$ is needed.

Following are the advantages of working with explicit feature map $\Phi$.

(i) For primal method in various optimization problems.

(ii) To derive an approximation based on $\Phi$.

(iii) Theoretical analysis where $\Phi$ is more convenient.

**Definition 1.4 (Empirical kernel map).** Given a sample $S = (x_1, \dots, x_m) \in \mathcal{X}^m$ and a PDS kernel $K$, the associated **empirical kernel map** $\Phi$ is a feature mapping defined for all $x \in \mathcal{X}$ by

$$\Phi(x) = \begin{bmatrix} K(x, x_1) \\ \vdots \\ K(x, x_m) \end{bmatrix}.$$

*Remark* 2. The empirical kernel map evaluated at a point $x$ is the vector of $K$-similarity measure of $x$ with each of the training points.

*Remark* 3. For any $i \in [m]$, we have $\Phi(x_i) = \mathbf{K}e_i$, where $e_i$ is the $i$-th unit vector. Hence,

$$\left\langle \mathbf{K}e_i, \mathbf{K}e_j \right\rangle = \left\langle e_i, \mathbf{K}^2 e_j \right\rangle.$$

That is, the kernel matrix associated with the empirical kernel map $\Phi$ is $\mathbf{K}^2$.

**Definition 1.5.** Let $\mathbf{K}^\dagger$ denote the pseudo-inverse of the gram matrix $\mathbf{K}$ and let $(\mathbf{K}^\dagger)^{\frac{1}{2}}$ denote the SPSD matrix whose square is $\mathbf{K}^\dagger$. We define a feature map $\Psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ using the empirical kernel map $\Phi$ and the matrix $(\mathbf{K}^\dagger)^{\frac{1}{2}}$ as

$$\Psi(x) = (\mathbf{K}^\dagger)^{\frac{1}{2}}, \text{ for all } x \in \mathcal{X}.$$

*Remark* 4. Using the identity $\mathbf{K}\mathbf{K}^\dagger\mathbf{K} = \mathbf{K}$, we see that

$$\left\langle \Psi(x_i), \Psi(x_j) \right\rangle = \left\langle (\mathbf{K}^\dagger)^{\frac{1}{2}}\Phi(x_i), (\mathbf{K}^\dagger)^{\frac{1}{2}}\Phi(x_j) \right\rangle = \left\langle \mathbf{K}e_i, \mathbf{K}^\dagger\mathbf{K}e_j \right\rangle = \left\langle e_i, \mathbf{K}e_j \right\rangle.$$

Thus, the kernel matrix associated to map $\Psi$ is $\mathbf{K}$.

*Remark* 5. For the feature mapping $\Omega : \mathcal{X} \to \mathbb{R}^m$ defined by $\Omega(x) = \mathbf{K}^\dagger\Phi(x)$ for all $x \in \mathcal{X}$, we check that the

$$\left\langle \Omega(x_i), \Omega(x_j) \right\rangle = \left\langle \mathbf{K}^\dagger\Phi(x_i), \mathbf{K}^\dagger\Phi(x_j) \right\rangle = \left\langle \mathbf{K}e_i, \mathbf{K}^\dagger e_j \right\rangle = \left\langle e_i, \mathbf{K}\mathbf{K}^\dagger e_j \right\rangle.$$

Thus, the kernel matrix associated to map $\Omega$ is $\mathbf{K}\mathbf{K}^\dagger$.

**Definition 1.6 (Tensor product).** The **tensor product** of two kernels $K_1, K_2$ is denoted by $K_1 \otimes K_2 : \mathcal{X}^4 \to \mathbb{R}$ and defined for all $x_1, y_1, x_2, y_2 \in \mathcal{X}$ as

$$(K_1 \otimes K_2)(x_1, x_2, y_1, y_2) = K_1(x_1, y_1)K_2(x_2, y_2).$$

**Theorem 1.7 (Closure properties of PDS kernels).** *PDS kernels are closed under sum, product, tensor product, point-wise limit, and composition with a power series $\sum_{n=0}^\infty a_n x^n$ with $a_n \geqslant 0$ for all $n \in \mathbb{N}$.*

*Proof.* Let $(K_n : n\ in \mathbb{N})$ be a sequence of PDS kernels on $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, and let $\mathbf{K}_n$ be the gram matrix generated by a sample $S = (x_1, \dots, x_m) \in \mathcal{X}^m$ for the kernel $K_n$ for each $n \in \mathbb{N}$.

(i) It suffices to show that $\mathbf{K}_1 + \mathbf{K}_2$ is SPSD. Since $\mathbf{K}_1, \mathbf{K}_2$ are SPSD, it follows that $\mathbf{K}_1 + \mathbf{K}_2$ is symmetric. From the linearity of inner products and positive semi definiteness of $\mathbf{K}_1, \mathbf{K}_2$, we have $\langle c, (\mathbf{K}_1 + \mathbf{K}_2)c \rangle = \langle c, \mathbf{K}_1, c \rangle + \langle c, \mathbf{K}_2, c \rangle \geqslant 0$ for any $c \in \mathbb{R}^m$.

(ii) It suffices to show that the matrix $\mathbf{K}_{ij} = [(\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij}]$ is SPSD. Symmetry follows from the symmetry of SPSD matrices $\mathbf{K}_1$ and $\mathbf{K}_2$.

Since $\mathbf{K}_1$ is SPSD, we have $\mathbf{K}_1 = \mathbf{M}\mathbf{M}^T$ by singular value decomposition or Cholesky decomposition. Therefore, $(\mathbf{K}_1)_{ij}(\mathbf{K}_2)_{ij} = \sum_{k=1}^m \mathbf{M}_{ik}\mathbf{M}_{jk}(\mathbf{K}_2)_{ij}$ and hence for any $c \in \mathbb{R}^m$, we can write

$$\sum_{i,j=1}^m c_i c_j (\sum_{k=1}^m \mathbf{M}_{ik}\mathbf{M}_{jk})(\mathbf{K}_2)_{ij} = \sum_{k=1}^m \sum_{i,j=1}^m (c_i\mathbf{M}_{ik})(\mathbf{K}_2)_{ij}(c_j\mathbf{M}_{jk}).$$

Defining $z_k = (c_i\mathbf{M}_{ik} : i \in [m])$, we see that $c^T\mathbf{K}c = \sum_{k=1}^m z_k^T\mathbf{K}_2 z_k \geqslant 0$.

(iii) The tensor product of two kernels $K_1, K_2$ can be thought of as the product of two PDS kernels

$$(x_1, x_2, y_1, y_2) \mapsto K_1(x_1, y_1), \qquad (x_1, x_2, y_1, y_2) \mapsto K_2(x_2, y_2).$$

(iv) Let $K$ be the point-wise limit of the sequence of PDS kernels $(K_n : n \in \mathbb{N})$. Let $\mathbf{K}$ be the gram matrix generated by the map $K$ and the sample $S = (x_1, \dots, x_m) \in \mathcal{X}^m$. Symmetry of $\mathbf{K}$ follows from the symmetry of each $\mathbf{K}_n$. From the continuity of inner products, we have for any $c \in \mathbb{R}^m$

$$0 \leqslant \langle c, \mathbf{K}_n c \rangle = \langle c, \mathbf{K}c \rangle.$$

(v) Let's assume that $K$ is a PDS kernel with $|K(x,y)| < \rho$ for all $x, y \in \mathcal{X}$, and let $f : x \mapsto \sum_{n=0}^\infty a_n x^n$, be a power series with $a_n \geqslant 0$ and radius of convergence $\rho$. Then, for any $n \in \mathbb{N}$, both $K^n$ and thus $a_n K^n$ are PDS by closure under product. For any $N \in \mathbb{N}$, the sum $\sum_{n=0}^N a_n K^n$ is PDS by closure under sum of PDS kernels $(a_n K^n : n \geqslant 0)$ and $f \circ K$ is PDS by closure under the limit of $\sum_{n=0}^N a_n K^n$ as $N \to \infty$.

□

**Example 1.8 (Gaussian kernels).** For any PDS kernel $K$, the kernel $\exp(K)$ is also PDS since it can be written as a power series with an infinite radius of convergence. We can check that a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $K(x,y) = \langle x,y \rangle$ for all $x,y \in \mathcal{X}$ is PDS kernel, and hence $K' = \exp(K)$ defined by $K'(x,y) = \exp\left(\frac{\langle x,y \rangle}{\sigma^2}\right)$ for all $x,y \in \mathcal{X}$ is PDS kernel. Therefore, the Gaussian kernel is PDS since it is normalized kernel of $K'$.

## 1.2 Kernel-based algorithms

We can generalize SVMs in the input space $\mathcal{X}$ to the SVMs in the feature space $\mathbb{H}$ mapped by the feature mapping $\Phi$. Recall that $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$ for all $x,y \in \mathcal{X}$, and hence the gram matrix $\mathbf{K}$ generated by the kernel map $K$ and the training sample $S = (x_1, \dots, x_m)$ suffices to describe the SVM solution completely.

Defining Hadamard product of two vectors $x,y \in \mathbb{R}^m$ as $x \circ y \in \mathbb{R}^m$ such that $(x \circ y)_i = x_i y_i$, we can write the dual problem for non-separable training data in this high dimensional space $\mathbb{H}$ as

$$\max_{\alpha} \mathbf{1}^T \alpha - \frac{1}{2}(\alpha \circ y)^T \mathbf{K} (\alpha \circ y)$$
$$\text{subject to: } 0 \leqslant \alpha \leqslant C \text{ and } \alpha^T y = 0.$$

The solution hypothesis $h$ can be written as

$$h(x) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b\right),$$

where $b = y_i - (\alpha \circ y)^T \mathbf{K} e_i$ for all $x_i$ such that $0 < \alpha_i < C$.

## 1.3 Representer theorem

Observe that modulo the offset $b$, the hypothesis solution of SVMs can be written as a linear combination of the functions $K(x_i, \cdot)$, where $x_i$ is a sample point. The following theorem known as the representer theorem shows that this is in fact a general property that holds for a broad class of optimization problems, including that of SVMs with no offset.

**Theorem 1.9 (Representer theorem).** *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel and $\mathbb{H}$ its corresponding RKHS. Then for any non decreasing function $G : \mathbb{R} \to \mathbb{R}$ and any loss function $L : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, the optimization problem*

$$\arg\min_{h \in \mathbb{H}} F(h) = \arg\min_{h \in \mathbb{H}} G(\|h\|_{\mathbb{H}}) + L(h(x_1), \dots, h(x_m)),$$

*has a solution of the form $h^* = \sum_{i=1}^{m} \alpha_i K(x_i, \cdot)$. If $G$ is strictly increasing, then any solution has this form.*

*Proof.* Let $\mathbb{H}_1 = \operatorname{span}(K(x_i, \cdot) : i \in [m])$. We can write the RKHS $\mathbb{H}$ as the direct sum of span of $(K(x_i, \cdot) : i \in [m])$ and the orthogonal space $\mathbb{H}^{\perp}$, i.e. $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}^{\perp}$. Hence, any hypothesis $h \in \mathbb{H}$, can be written as $h = h_1 + h^{\perp}$. Since $G$ is non-decreasing

$$G(\|h_1\|_{\mathbb{H}}) \leqslant G\left(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h^{\perp}\|_{\mathbb{H}}^2}\right) = G(\|h\|_{\mathbb{H}}).$$

By the reproducing property, we have for all $i \in [m]$

$$h(x_i) = \langle h, K(x_i, \cdot) \rangle = \langle h_1, K(x_i, \cdot) \rangle = h_1(x_i).$$

Therefore, $L(h(x_1), \dots, h(x_m)) = L(h_1(x_1), \dots, h_1(x_m))$, and hence $F(h_1) \leqslant F(h)$. If $G$ is strictly increasing, then $F(h_1) < F(h)$ when $\|h^{\perp}\|_{\mathbb{H}} > 0$ and any solution of the optimization problem must be in $\mathbb{H}_1$. $\square$