Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

MAX-PLANCK-GESELLSCHAFT

Technical Report No. 127

Kernels, Associated Structures and Generalizations

Matthias Hein[1], Olivier Bousquet[1]

July 2004

[1] Department Schölkopf, email: matthias.hein;olivier.bousquet@tuebingen.mpg.de

Kernels, Associated Structures and Generalizations

Matthias Hein and Olivier Bousquet

Abstract. This paper gives a survey of results in the mathematical literature on positive definite kernels and their associated structures. We concentrate on properties which seem potentially relevant for Machine Learning and try to clarify some results that have been misused in the literature. Moreover we consider different lines of generalizations of positive definite kernels. Namely we deal with operator-valued kernels and present the general framework of Hilbertian subspaces of Schwartz which we use to introduce kernels which are distributions. Finally indefinite kernels and their associated reproducing kernel spaces are considered.

1 Introduction

Positive definite kernels are extremely powerful and versatile tools. They allow to construct spaces of functions on an arbitrary set with the convenient structure of a Hilbert space. Methods based on such kernels are usually very tractable because of the particular structure (reproducing property) of the space of functions. This has a large number of applications, in particular for statistical learning, approximation or interpolation where one has to manipulate functions defined on various types of data, see e.g. [1, 2, 3].

Our goal is to survey some of the results relevant for machine learning. Since the literature is scattered among various fields of mathematics we believe that the learning community would benefit from a unified exposition of the results and relationships between them. This work is a first attempt to go into that direction. Although the theory can be quite technical, we want to shed light on its essence and convey several important messages that anyone working with kernels and associated spaces should have in mind.

A first message is that there is an equivalence (in a strong) sense between several objects: positive definite kernels (which are specific functions of two variables), Hilbert spaces of functions with a certain topological property, Gaussian processes and a class of positive operators. A second message is that the mysterious ”feature maps” associated to kernels are not related to the Mercer property and they exist and can be defined in many different ways as soon as the kernel is positive definite. A third message is that the integral operator associated to a kernel has nice properties even if the kernel is not continuous. In particular it is tightly related to the covariance operator (i.e. the population limit of a covariance matrix) as they have the same spectrum. A fourth message is that most attempts to generalize kernels (e.g. to operator-valued or generalized functions) end up being special cases. This may seem surprising but it easily seen by changing the point of view one adopts, going from sets to functions on these sets. Finally, we recall that there exists a well-developed theory of indefinite kernels (i.e. kernels that are not positive definite) and their associated structures, based on the notion of reproducing kernel Krein spaces.

2 Positive Definite Kernels and Associated Structures

We restrict ourselves to the real-valued case and denote by $\mathbb{R}^{\mathcal{X}}$ the vector space of functions from $\mathcal{X}$ to $\mathbb{R}$ where $\mathcal{X}$ is an arbitrary *index* set[1] and by $\mathbb{R}^{[\mathcal{X}]}$ the vector space of finite linear combinations of evaluation functionals (i.e. of the form $\sum_{i=1}^n a_i \delta_{x_i}$). We define a bilinear map from $\mathbb{R}^{[\mathcal{X}]} \times \mathbb{R}^{\mathcal{X}}$ to $\mathbb{R}$ as $\langle \sum_{i=1}^n \alpha_i \delta_{x_i}, f \rangle_{\mathbb{R}^{[\mathcal{X}]}, \mathbb{R}^{\mathcal{X}}} := \sum_{i=1}^n \alpha_i f(x_i)$ where $x_1, \ldots, x_n \in \mathcal{X}$.

In this first section we shortly review the notion of positive definite (PD) kernels and its associated structures. Indeed such a kernel can be associated to a space of functions, called reproducing kernel Hilbert space (RKHS), to a linear operator called positive symmetric kernel (PSK) operator and to a Gaussian process in a natural way. The following diagram illustrates the fact that all these notions are tightly related.

$$\begin{tikzcd}
\boxed{\text{PD kernel}} \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & \boxed{\text{RKHS}} \arrow[d, leftrightarrow] \\
\boxed{\text{PSK operator}} \arrow[r, leftrightarrow] & \boxed{\text{Gaussian Process}}
\end{tikzcd}$$

[1] or also called *input space*.

2.1 Definitions

We now give the definitions of the four objects in the preceding diagram.

Definition 1 *A real-valued symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* **positive definite (PD) kernel** *if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathbb{R}$*

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \tag{1}$$

The set of all real-valued positive definite kernels on $\mathcal{X}$ is denoted $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$.

Definition 2 *A* **positive symmetric kernel (PSK) operator** *K is a linear operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$ which is symmetric*

$$\forall v', w' \in \mathbb{R}^{[\mathcal{X}]}, \quad \langle v', Kw' \rangle_{\mathbb{R}^{[\mathcal{X}]},\mathbb{R}^{\mathcal{X}}} = \langle w', Kv' \rangle_{\mathbb{R}^{[\mathcal{X}]},\mathbb{R}^{\mathcal{X}}} \,,$$

and positive: $\forall v' \in \mathbb{R}^{[\mathcal{X}]},, \quad \langle v', Kv' \rangle_{\mathbb{R}^{[\mathcal{X}]},\mathbb{R}^{\mathcal{X}}} \geq 0$.
The set of all such operators is denoted $L_+(\mathbb{R}^{\mathcal{X}})$.

Definition 3 *A* **reproducing kernel Hilbert space (RKHS)** *$\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ where all evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x(f) = f(x)$ are continuous[2], equivalently for all $x \in \mathcal{X}$, there exists a $M_x < \infty$ such that*

$$\forall\, f \in \mathcal{H},\ |f(x)| \leq M_x \, \|f\|_{\mathcal{H}} \ .$$

The set of all such spaces is denoted $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$.

This definition stresses the fact, that an RKHS is a Hilbert space of pointwise defined functions, where norm convergence implies pointwise convergence.

Definition 4 *A* **centered Gaussian process** *indexed by $\mathcal{X}$ is a family X_x, $x \in \mathcal{X}$, of jointly normal random variables, that is for each finite set $x_1, \ldots, x_n \in \mathcal{X}$, the vector $(X_{x_1}, \ldots, X_{x_n})$ is centered Gaussian[3].*
The set of all such processes is denoted $G(\mathcal{X})$.

Note that we restrict ourselves to centered Gaussian random variables. In principle the results can be transferred to the non-centered case.

2.2 Properties and Connections

The fundamental and most important property of PD kernels is the relationship with inner product spaces. Often the use of kernel methods is justified by the implicit mapping of the input space $\mathcal{X}$ into a 'high-dimensional' feature space. As the next proposition shows, such a mapping exists as soon as the kernel is positive definite and actually characterizes such kernels.

Proposition 1 *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PD kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}$, $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.*

Note that this result has nothing to do with Mercer's theorem (we will come back to this issue in section 3.1). There exist many proofs of the above proposition and we will give one later.
We will now establish the connections between the four objects we have introduced in the previous section. It is well known (see e.g. [4]) that $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ is invariant under addition, multiplication by a non-negative number and point-wise limits and has an order relationship ($k_1 \succeq k_2$ if $k_1 - k_2$ is PD). It is less known that all the other sets introduced above ($L_+(\mathbb{R}^{\mathcal{X}})$, $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$ and $G(\mathcal{X})$) have a similar structure. Actually, the following strong equivalence between these spaces and their structures holds.

Theorem 1 *[5] There exist bijections which preserve the structure of ordered, closed convex cones between each two of the following sets*

$$\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}} \ , \ L_+(\mathbb{R}^{\mathcal{X}}) \ , \ \mathrm{Hilb}(\mathbb{R}^{\mathcal{X}}) \ , \ G(\mathcal{X}) \,.$$

An example how the order is transferred from $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ to $\mathrm{Hilb}(\mathbb{R}^{\mathcal{X}})$ is the following.

Theorem 2 *[4] Let $k_1, k_2 \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ and $\mathcal{H}_1, \mathcal{H}_2$ their associated RKHS. Then $\mathcal{H}_1 \subset \mathcal{H}_2$, and $\|f_1\|_{\mathcal{H}_1} \geq \|f_1\|_{\mathcal{H}_2}$, $\forall\, f_1 \in \mathcal{H}_1$ if and only if $k_1 \preceq k_2$.*

[2] with respect to the topology induced by the norm of $\mathcal{H}$

[3] equivalently, all linear combinations $\sum \alpha_i X_{x_i}$ are real Gaussian random variables with zero mean.

The remaining part of this section will show several of these bijections, but due to space limitations we are not able to show all of them explicitly. Additionally we introduce in the appendix several objects associated to a Gaussian Process. These objects become interesting if one is interested for example in sample path properties of a Gaussian Process.

2.2.1 PD Kernels and PSK Operators

The bijection between kernels and kernel operators is made explicit in the following lemma.

Lemma 1 *[6] Let $k \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$. The linear operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$ defined by $K(\delta_x) = k(x,\cdot)$, is a PSK operator. Conversely, given $K \in L_+(\mathcal{X})$, the function k defined as $k(x,y) = \langle \delta_x, K\delta_y \rangle_{\mathbb{R}[\mathcal{X}],\mathbb{R}^{\mathcal{X}}}$ is a PD kernel.*

The above lemma indicates the close correspondence between the kernel function and its associated operator. In particular, symmetry of one corresponds to symmetry of the other, while positive definiteness of the former one corresponds to positivity of the latter.

2.2.2 PD Kernels and RKHS

The following fundamental theorems illustrate the link between RKHS and PD kernels.

Theorem 3 *[4] Let $\mathcal{H}$ be a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$, $\mathcal{H}$ is a RKHS if and only if there exists a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

$$\begin{aligned} &\forall\, x \in \mathcal{X}, \quad k(x,\cdot) \in \mathcal{H}, \\ &\forall\, f \in \mathcal{H}, \quad \langle f(\cdot), k(x,\cdot) \rangle_{\mathcal{H}} = f(x)\,. \end{aligned}$$

If such a k exists, it is unique and it is a PD kernel.

The second property is called the *reproducing property* of the RKHS and k is called the (reproducing) kernel of $\mathcal{H}$.

Theorem 4 *(Moore) If k is a positive definite kernel then there exists a unique reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is k.*

Proof: We give a sketch of the proof (of both theorems above) which involves an important construction. The proof proceeds in three steps. The first step is to consider the set of all finite linear combinations of the kernel: $\mathcal{G} = \text{Span}\{k(x,.) : x \in \mathcal{X}\}$ and to endow it with the following inner product

$$\left\langle \sum_i a_i k(x_i,.), \sum_j b_j k(x_j,.) \right\rangle_{\mathcal{G}} = \sum_{i,j} a_i b_j k(x_i,x_j)\,. \tag{2}$$

It can be shown that this is indeed a well-defined inner product. At this point we already have the reproducing property on $\mathcal{G}$. The second step is to construct the semi-norm associated to this inner product and to show (thanks to the Cauchy-Schwarz inequality) that it is actually a norm. Hence, and this is the third step, $\mathcal{G}$ is a pre-Hilbert space which can be completed[4] into a Hilbert space $\mathcal{H}$ of functions. Finally, one has to check that the reproducing property carries over to the completion. It is then easy to show that any other Hilbert space with the same reproducing kernel has to be isometric isomorphic. Namely let $\mathcal{K}$ be another RKHS with reproducing kernel k. It is obvious that $\mathcal{H}$ has to be a closed subspace of $\mathcal{K}$. Then $\mathcal{K}$ can be decomposed into $\mathcal{K} = \mathcal{H} \oplus \mathcal{H}^{\perp}$. Now let $f \in \mathcal{K}$, but $f \notin \mathcal{H}$. Then for all $x \in \mathcal{X}$

$$f(x) = \langle f, k(x,\cdot) \rangle_{\mathcal{K}} = \left\langle f^{\|} + f^{\perp}, k(x,\cdot) \right\rangle_{\mathcal{K}} = f^{\|}(x)$$

Therefore $f \equiv f^{\|}$, which is a contradiction and we get $\mathcal{K} = \mathcal{H}$. □

Hence $\mathcal{H}$ is simply the completion of the linear span (i.e. finite linear combinations) of the functions $k(x,\cdot)$ endowed with the inner product (2).

2.2.3 PD Kernels and Gaussian Processes

It is well-known that a centered Gaussian process $(X_x)_{x\in\mathcal{X}}$ is uniquely determined by its covariance function $\mathbb{E}\left[X_s X_t\right]$, which is a positive definite kernel. Conversely any positive definite kernel defines a covariance function and therefore a unique Gaussian process by Theorem 14.

[4] i.e. we add to $\mathcal{G}$ the pointwise limits of all Cauchy sequences of elements of $\mathcal{G}$

3 Useful Properties

A quit useful relationship between $k \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ and the set $\mathcal{X}$ is that k induces a semi-metric on $\mathcal{X}$ by $d_k(x,y) = \|k(x,\cdot) - k(y,\cdot)\|_{\mathcal{H}}$. Many properties of the RKHS can be stated in terms of this (semi)-metric space $(\mathcal{X}, d_k)$ as we will later see in the study of the separability of the RKHS.

3.1 Feature Maps

Often Mercer's theorem is mentioned as a necessary condition to have a feature map. The goal of this section is to show, that it is a sufficient condition but it requires additional assumptions on $\mathcal{X}$ and k. As we have seen in Proposition 1 a necessary and sufficient condition that such a feature map into a Hilbert space exists is that the kernel is positive definite. Two questions can then be raised: Can such a map be constructed explicitly ? What is the induced representation for the kernel ? Both questions have an affirmative answer without any further assumptions on k as the following feature maps $\Phi : X \to \mathcal{H}$ show.

1. *Aronszajn map*
   $\phi : x \mapsto k(x,\cdot)$, $\mathcal{H}$ is the associated RKHS, $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle$

2. *Kolmogorov map*
   $\phi : x \mapsto X_x$, $\mathcal{H} = L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ where μ is a Gaussian measure[5], $k(x,y) = \mathbb{E}\left[X_x X_y\right]$

3. *Integral map*
   There exists a set T and a measure μ on T such that one has $\phi : x \mapsto (\Gamma_x(t))_{t\in T}$, $\mathcal{H} = L_2(T,\mu)$[6], $k(x,y) = \int \Gamma(x,t)\Gamma(y,t)d\mu(t)$

4. *Basis map*
   given any orthornormal basis[7] $(f_\alpha)_{\alpha\in I}$ of the RKHS associated to $\mathcal{H}$, one has $\phi : x \mapsto (f_\alpha(x))_{\alpha\in I}$, $\mathcal{H} = \ell_2(I)$[8] and $k(x,y) = \sum_{\alpha\in I} f_\alpha(x) f_\alpha(y)$.

When infinite sums are involved like in the last case, it is important to specify in which sense the sum converges. In general the convergence occurs for each pair (x,y). However, [7] shows one has stronger convergence, namely uniform on every set $A \times B \subset \mathcal{X} \times \mathcal{X}$, with A bounded and B compact (w.r.t. the topology induced by d_k).
Given additional structure of the kernel resp. the corresponding RKHS there exist other feature space interpretations. Mercer's theorem is a special case of the basis map. It gives stronger convergence properties of the kernel representation but needs additional assumption, namely $\mathcal{X}$ has to be compact and the kernel k continuous.

3.2 Boundedness and Continuity

Because of the PD property and Cauchy-Schwarz inequality, there are relationships between the function $x \mapsto k(x,x)$ and $(x,y) \mapsto k(x,y)$ when one considers boundedness or continuity properties of the kernel.

Lemma 2 *For a PD kernel k the following two statements are equivalent*

(i) $x \mapsto k(x,x)$ *is bounded;*

(ii) $(x,y) \mapsto k(x,y)$ *is bounded.*

Lemma 3 *[8] A PD kernel k is continuous on $\mathcal{X} \times \mathcal{X}$ if and only if the following two conditions are fulfilled*

(i) $x \mapsto k(x,x)$ *is continuous;*

(ii) for any fixed x the function $y \mapsto k(x,y)$ is continuous at $y = x$.

These conditions are equivalent to the continuity of the function $(x,y) \mapsto k(x,y)$ at every point of the diagonal $\{(x,y) : x = y\}$.

Corollary 1 *If k is continuous on $\mathcal{X} \times \mathcal{X}$ then the identity map $(\mathcal{X}, d) \to (\mathcal{X}, d_k)$ is continuous.*

[5] see Appendix A for details.

[6] The Kolmogorov map shows that such a set T and a measure μ always exist.

[7] such a basis always exists but may be uncountable, in which case, only a countable subset of the coordinates of any vector are non-zero.

[8] space of square summable functions on I with countable support

Proof: Follows directly from $d_k^2(x, x_n) = k(x, x) - 2k(x_n, x) + k(x_n, x_n)$. □

A related question is: when does the RKHS consist of continuous functions ? Since $k(x, \cdot)$ belongs to the associated RKHS, this means that k has to be at least separately continuous. The following theorem provides necessary and sufficient conditions in a rather general setting.

Theorem 5 *[6] Let $\mathcal{X}$ be a locally compact space and $C(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$ with the topology of uniform convergence on compact subsets. The canonical injection $i : \mathcal{H}_k \to C(\mathcal{X})$ is continuous if and only if $k(x, y)$ is separately continuous on $\mathcal{X} \times \mathcal{X}$ and locally bounded.*

3.3 When is a Function in a RKHS ?

Let us suppose we are given a function f and want to know if it is contained in the RKHS associated to a PD kernel k. Some mistakes have been made concerning this question in the Machine Learning literature. We give a general result.

Lemma 4 *[8] The function f belongs to the RKHS $\mathcal{H}$ associated to k if and only if there exists $\epsilon > 0$ such that*

$$R_\epsilon(x, y) = k(x, y) - \epsilon f(x) f(y)\,,$$

is a positive definite kernel. Equivalently this corresponds to the condition

$$\sup_{|I|<\infty,\,(a_i)_{i\in I}\in\mathbb{R},\,(x_i)_{i\in I}\in\mathcal{X}} \frac{\sum_{i\in I} a_i f(x_i)}{\Big(\sum_{i,j\in I} a_i a_j k(x_i, x_j)\Big)^{1/2}} < \infty.$$

If this is satisfied, one can compute the norm of f as the value of the above supremum, or as $\|f\|_{\mathcal{H}} = \inf\{1/\sqrt{\epsilon}\,|\,\epsilon > 0, R_\epsilon \succeq 0\}$.

A simple consequence of this lemma is that the RKHS associated to any bounded kernel cannot contain unbounded functions.

3.4 Separability of the RKHS

Some convergence proofs of iterative algorithms require the separability of the RKHS. However, this is seldom made explicit in the Machine Learning literature. The first result gives a necessary and sufficient condition for separability.

Theorem 6 *[9] $\mathcal{H}_k$ is separable if and only if $(\mathcal{X}, d_k)$ is separable.*

Proof: Let $\mathcal{H}_k$ be separable, then $\mathcal{H}_k$ and every subset of $\mathcal{H}_k$ is second countable. Particularly the set $k(\mathcal{X}, \cdot) := \{k(x, \cdot)\,|\,x \in \mathcal{X}\}$ is second countable and therefore separable. Since $(\mathcal{X}, d_k)$ is isometric to the set $k(\mathcal{X}, \cdot)$, $(\mathcal{X}, d_k)$ is separable.

We sketch the proof of the other direction. Since $(\mathcal{X}, d_k)$ is separable, $k(\mathcal{X}, \cdot)$ is separable. Then it is easy to show that the span of $k(\mathcal{X}, \cdot)$ with rational numbers is dense in Span $k(\mathcal{X}, \cdot)$ and since $\mathcal{H}_k = \overline{\text{Span}\, k(\mathcal{X}, \cdot)}$ we are done. □

In the case of continuous kernels we get the following consequence

Theorem 7 *[8] Let $\mathcal{X}$ be a topological space, k a PD kernel which is continuous on $\mathcal{X} \times \mathcal{X}$, and $\mathcal{H}$ its associated RKHS. If $\mathcal{X}$ is separable, then $\mathcal{H}$ is separable.*

As a result any continuous kernel on $\mathbb{R}^n$ induces a separable RKHS e.g. the RKHS associated to the RBF kernel $k(x, y) = \exp(-\|x - y\|/\sigma^2)$ is separable. In the case, where $\mathcal{H}_k$ is separable, the basis feature map can be written with a countable sum. Again, this does not require anything like Mercer's theorem.

4 Integral and Covariance Operators

In general we assume in statistical learning theory that the space $\mathcal{X}$ is endowed with a probability measure P. Then samples X_i are drawn according to this probability measure P. These define then the empirical measure $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

In kernel-algorithms one uses the so-called *kernel matrix* $K_n : L_2(\mathcal{X}, P_n) \to L_2(\mathcal{X}, P_n)$ defined as $K_n = \frac{1}{n}(k(X_i, X_j))_{i,j=1,\ldots,n}$ and the *empirical covariance operator* $C_n : \mathcal{H}_k \to \mathcal{H}_k$ defined as $C_n = \frac{1}{n}\sum_{k=1}^n \Phi(X_k) \otimes \Phi(X_k)$. These are under some conditions finite sample approximations of operators $K : L_2(\mathcal{X}, P) \to L_2(\mathcal{X}, P)$

resp. $C : \mathcal{H}_k \to \mathcal{H}_k$ defined for the whole probability measure P.
We will study the properties of the operators K and C and the convergence of the empirical counterparts to the true operators under the following assumptions on the kernel.

- $k(x, y)$ is measurable,
- $k(x, y)$ is a positive definite kernel,
- $\int_{\mathcal{X}} k(x, x)dP(x) < \infty$.

Note that the second assumption implies $k \in L_2(\mathcal{X} \times \mathcal{X}, P \otimes P)$ by the Cauchy-Schwarz inequality. Also note that in our setting we have no assumptions on the separability of $\mathcal{H}$ or $L_2(\mathcal{X}, P)$.

Theorem 8 *Let $i : \mathcal{H} \to L_2(\mathcal{X}, P)$ be the canonical injection. Then under the stated assumptions i is continuous. Moreover i is a Hilbert-Schmidt operator with $\|i\|^2_{HS} \leq \int_{\mathcal{X}} k(x, x)dP(x)$.*
Proof: Let i be the canonical injection $i : \mathcal{H} \to L_2(\mathcal{X}, P)$. Then for all $f \in \mathcal{H}$,

$$\|if\|^2_{L_2(\mathcal{X},P)} = \int |f(x)|^2 dP(x) = \int \langle f, k(x, \cdot)\rangle^2_{\mathcal{H}} \, dP(x) \leq \|f\|^2_{\mathcal{H}} \int k(x, x)dP(x) \, .$$

Therefore i is a bounded operator.
Denote by $\{e_\alpha, \, \alpha \in A\}$ an orthonormal basis (possibly uncountable) of $L_2(\mathcal{X}, P)$. i is Hilbert-Schmidt if and only if $\sum_{\alpha\in A} \|i \, e_\alpha\|^2_{L_2(\mathcal{X},P)} < \infty$. For all finite sets $F \subset A$ we have

$$\begin{aligned}
\sum_{\alpha\in F} \|ie_\alpha\|^2_{L_2(\mathcal{X},P)} &= \int_{\mathcal{X}} \sum_{\alpha\in F} |e_\alpha(x)|^2 dP(x) = \int_{\mathcal{X}} \sum_{\alpha\in F} \left|\langle e_\alpha, k(x, \cdot)\rangle_{\mathcal{H}}\right|^2 dP(x) \\
&\leq \int_{\mathcal{X}} \|k(x, \cdot)\|^2_{\mathcal{H}} \, dP(x) = \int_{\mathcal{X}} k(x, x)dP(x)
\end{aligned}$$

where we have used Bessel's inequality. Let now $S_{fin}(A) = \{P \subset A \,|\, P \text{ finite}\}$ be the directed set of finite subsets of A with the set inclusion as partial order. Since all summands are positive, the limit of the net of partial sums can be computed as follows

$$\sum_{\alpha\in A} \|i \, e_\alpha\|^2_{L_2(\mathcal{X},P)} = \sup\{\sum_{\alpha\in F} \|i \, e_\alpha\|^2_{L_2(\mathcal{X},P)} \, , \; F \in S_{fin}(A)\} \leq \int_{\mathcal{X}} k(x, x)dP(x).$$

□

The next proposition connects the canonical injection i with the integral and the covariance operator:
Proposition 2 *The integral operator K*

$$K : L_2(\mathcal{X}, P) \to L_2(\mathcal{X}, P), \; (Kf)(x) = \int_{\mathcal{X}} k(x, y)f(y)dP(y) \, . \tag{3}$$

and the covariance operator C

$$C : \mathcal{H} \to \mathcal{H}, \; \langle f, Cg\rangle = \int_{\mathcal{X}} f(x)g(x)dP(x) \, . \tag{4}$$

are both positive, self-adjoint, Hilbert-Schmidt and trace-class. Moreover they can be decomposed as $K = ii^*$ and $C = i^*i$ and have the same spectrum, which implies that $\operatorname{tr} K = \operatorname{tr} C$ and $\|C\|_{HS} = \|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X},P\otimes P)}$.
Proof: We showed in theorem 8 that i is continuous. Therefore the adjoint $i^* : L_2(\mathcal{X}, P) \to \mathcal{H}$ exists and is defined for $g \in L_2(\mathcal{X}, P)$ and $f \in \mathcal{H}$ as $\langle i^*g, f\rangle_{\mathcal{H}} = \langle g, if\rangle_{L_2(\mathcal{X},P)}$. In particular, choosing $f = k(x, \cdot) \in \mathcal{H}$ we see that $(i^*g)(x) = \langle k(x, \cdot), i^*g\rangle_{\mathcal{H}} = \langle ik(x, \cdot), g\rangle = \int_{\mathcal{X}} k(x, y)g(y)dP(y)$, so that $K = ii^*$. As a consequence, K is positive and self-adjoint. Moreover it is trace-class since

$$\operatorname{tr} K = \sum_{\alpha\in A} \langle e_\alpha, K \, e_\alpha\rangle_{L_2(\mathcal{X},P)} = \sum_{\alpha\in A} \|i^*e_\alpha\|_{\mathcal{H}} = \|i^*\|^2_{HS} \leq \int_{\mathcal{X}} k(x, x)dP(x),$$

where we use the fact $\|i\|_{HS} = \|i^*\|_{HS}$.
Moreover, for $f, g \in \mathcal{H}$, $\langle f, i^* i g\rangle_{\mathcal{H}} = \langle if, ig\rangle_{L_2(\mathcal{X},P)} = \mathbb{E}\left[f(X)g(X)\right]$ so that C is positive, self-adjoint and $C = i^* i$. It follows easily that C is trace-class with

$$\operatorname{tr} C = \sum_{\alpha\in A} \langle e_\alpha, C\, e_\alpha\rangle_{\mathcal{H}} = \sum_{\alpha\in A} \|i\, e_\alpha\|^2_{L_2(\mathcal{X},P)} = \|i\|^2_{HS}\,.$$

Both C and K are trace-class and therefore compact, which implies that they only have a discrete spectrum. Moreover they have the same spectrum and all non-zero eigenvalues have the same multiplicity. Let λ_n be an eigenvalue of K and denote by Λ_n the corresponding finite-dimensional eigenspace. Then

$$ii^*\Lambda_n = \lambda_n \Lambda_n \Rightarrow (i^* i) i^* \Lambda_n = \lambda_n i^* \Lambda_n \tag{5}$$

that is $i^*\Lambda_n$ is an eigenspace of C to the corresponding eigenvalue λ_n and the same argumentation holds in the other direction. Also $\dim \Lambda_n = \dim i^*(\Lambda_n)$ since it follows from (5) that $\Lambda_n \not\subseteq Ker(i^*)$ and $i^*(\Lambda_n) \not\subseteq Ker(i)$.
It is a classical result that $k \in L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$ implies that K is Hilbert-Schmidt and $\|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X},P\otimes P)}$, see [10] (note that this is true, even if $L_2(\mathcal{X}, P)$ is not separable). Since a compact self-adjoint operator is Hilbert-Schmidt if and only if $\sum_i \lambda_i^2 < \infty$ it follows directly from the equality of the spectra that C is Hilbert-Schmidt with $\|C\|_{HS} = \|K\|_{HS} = \|k\|_{L_2(\mathcal{X}\times\mathcal{X},P\otimes P)}$. □

Corollary 2 *If $Ker(i) = 0$ then $\mathcal{H} = \overline{i^*(L_2(X,P))}$ and $\mathcal{H}$ is separable.*

Proof: If $Ker(i) = 0$ then $\overline{Ran(i^*)} = Ker(i)^\perp = \mathcal{H}$. Since i^* is compact, $\overline{Ran(i^*)}$ is separable and therefore $\mathcal{H}$ is separable. □

In other words if the zero function is the only function in the RKHS $\mathcal{H}$ which is zero P-almost everywhere then the image of the integral operator K is dense in the RKHS and the RKHS is automatically separable.

Corollary 3 *If $\mathcal{H}$ is separable then $\|i\|^2_{HS} = \operatorname{tr} C = \operatorname{tr} K = \int_{\mathcal{X}} k(x,x)dP(x)$.*

Proof: Let $\{e_n\}_{n=1}^\infty$ be a complete orthonormal basis of $\mathcal{H}$. Then

$$\begin{aligned}\|i\|^2_{HS} &= \lim_{N\to\infty}\sum_{n=1}^N \|i\, e_n\|^2_{L_2(\mathcal{X},P)} = \lim_{N\to\infty}\sum_{n=1}^N \int_{\mathcal{X}} |e_n(x)|^2 dP(x) = \lim_{N\to\infty}\sum_{n=1}^N \int_{\mathcal{X}} |\left\langle k(x,\cdot), e_n\right\rangle_{\mathcal{H}}|^2 dP(x)\\ &= \int_{\mathcal{X}} \lim_{N\to\infty}\sum_{n=1}^N |\left\langle k(x,\cdot), e_n\right\rangle_{\mathcal{H}}|^2 dP(x) = \int_{\mathcal{X}} \|k(x,\cdot)\|^2_{\mathcal{H}}\, dP(x) = \int_{\mathcal{X}} k(x,x)dP(x) < \infty\end{aligned}$$

where the fourth step follows from the monotone convergence theorem and fifth step is Parseval's identity. □

The next corollary establishes a feature map in $L_2(\mathcal{X}, P)$.

Corollary 4 *If $k \in L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$, then there exists an orthonormal system (ϕ_n) in $L_2(P)$ such that*

$$k(x,y) = \sum_{n\in\mathbb{N}} \lambda_n \phi_n(x)\phi_n(y)\,, \tag{6}$$

where $\lambda_n \geq 0$ and the convergence of the sum occurs in $L_2(\mathcal{X}\times\mathcal{X}, P\otimes P)$. The associated feature map is thus

$$\Phi(x) = (\sqrt{\lambda_n}\phi_n(x))_{n\in\mathbb{N}}\,.$$

Proof: That is a classical result in functional analysis, see e.g. [11]. □

The remaining question is how the empirical counterparts K_n and C_n are related to the operators K and C.

Proposition 3 *Let K be the integral operator defined in* (3) *and X_i an i.i.d. set of random variables drawn from P. For all $f \in L_2(\mathcal{X}, P)$ we have:*

$$\begin{aligned}\lim_{n\to\infty}\langle f, Kf\rangle_{L_2(\mathcal{X},P_n)} &= \lim_{n\to\infty} n^{-2}\sum_{i,j=1}^n f(X_i)f(X_j)k(X_i,X_j) = \int_{\mathcal{X}^2} f(x)f(y)k(x,y)dP(x)dP(y)\\ &= \langle f, Kf\rangle_{L_2(\mathcal{X},P)} \textit{ a.s.}\end{aligned}$$

Proof: The proof is essentially an application of a result in [12]. Given an i.i.d. set of random variables $X_i \in \mathcal{X}$ drawn from P and a measurable symmetric function $g(x,y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ it states that $\lim_{n\to\infty} n^{-2} \sum_{i,j=1}^{n} g(X_i, X_j) = Eg(X,Y)$ almost surely if $E|g(X,Y)| < \infty$ and $E\sqrt{|g(X,X)|} < \infty$. Let now $g(x,y) = f(x)f(y)k(x,y)$, then the conditions require that $\int_{\mathcal{X}^2} f(x)f(y)k(x,y)dP(x)dP(y) < \infty$ and $\int_{\mathcal{X}} |f(x)|\sqrt{k(x,x)}dP(x) < \infty$. The second condition implies the first one and we have

$$\int_{\mathcal{X}} |f(x)|\sqrt{k(x,x)}dP(x) \leq \int_{\mathcal{X}} |f(x)|^2 dP(x) \int_{X} k(x,x)dP(x) < \infty,$$

since $\|f\|_{L_2(\mathcal{X},P)} \leq \|f\|_{\mathcal{H}} \int_{\mathcal{X}} k(x,x)dP(x)$. □

The next statement relates C_n and C:

Proposition 4

$$\langle f, C_n g\rangle_{\mathcal{H}_k} \xrightarrow{a.s.} \langle f, Cg\rangle_{\mathcal{H}_k}, \ \forall f, g \in \mathcal{H}_k.$$

Proof: The proof is a simple application of the strong law of large numbers. □

As a final remark we would like to note that if k is bounded then all the assumptions are fulfilled and the theorems of this section apply for any probability measure P.

5 Generalizations

Now that we have the general picture in mind, we investigate possible generalizations of the presented notions. We consider the generalization of kernel functions to operator-valued functions and of the RKHS to Hilbertian subspaces. We will show that they are both special cases of the general theory above.

5.1 Operator-Valued Kernels

Recently there was interest in the machine learning community to extend real-valued kernels to operator-valued kernels in order to learn vector-valued functions [13]. This concept is not new in the mathematics literature. It can at least traced back to the paper of [14].
Let $\mathcal{X}$ be a set and $\mathcal{G}$ a Hilbert space[9]. The goal is to generate a (generalized) RKHS whose functions are from $\mathcal{X}$ to $\mathcal{G}$ (instead of $\mathcal{X} \to \mathbb{R}$). We define a (generalized) notion of positive definite kernel:

Definition 5 *A function $k : \mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$[10] such that $k(x,y) = k(y,x)^*$ is called a* **positive definite operator-valued kernel function** *if for all $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathcal{G}$, $\sum_{i,j=1}^{n} \langle c_i\, k(x_i, x_j), c_j\rangle \geq 0$*

This seems to generalize the PD kernels we introduced before, and indeed, several papers deal with the notion of operator-valued kernels. However, a slight change of point of view allows to recast operator-valued kernels in the standard setting of real-valued ones, showing their great generality. We have the following result.

Proposition 5 *Let k be a PD operator-valued kernel $\mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$. Define ℓ as the function on $(\mathcal{X} \times \mathcal{G})$ such that $\ell((x,f),(y,g)) = \langle f, k(x,y)g\rangle_{\mathcal{G}}$. The map $k \mapsto \ell$ thus defined, is a bijection between PD operator-valued kernels $\mathcal{X} \times \mathcal{X} \to L(\mathcal{G})$ and real-valued PD kernels $(\mathcal{X},\mathcal{G}) \times (\mathcal{X},\mathcal{G}) \to \mathbb{R}$ which are bilinear on $\mathcal{G} \times \mathcal{G}$[11]. If $\mathcal{G}$ is finite dimensional, dim $\mathcal{G} = d$, one can also define, (e_i) being an orthonormal basis of $\mathcal{G}$, $\ell((x,i),(y,j)) = \langle e_i, k(x,y)e_j\rangle$, such that $k \mapsto \ell$ is a bijection to real-valued PD kernels on $(\mathcal{X}, \{1,\ldots,d\})$.*

Proof: We prove the above proposition in the finite dimensional case (the general case has a similar proof). Let $\ell((x,i),(y,j)$ be a PD kernel on $(\mathcal{X}, \{1,\ldots,d\})$. Define a bilinear form on $\mathbb{R}^d$ by defining the matrix $k(x,y) : \mathbb{R}^d \to \mathbb{R}^d$ as

$$k_{ij}(x,y) = \langle e_i, k(x,y)e_j\rangle_{\mathbb{R}^d} = \ell((x,i),(y,j))$$

[9] the same theory can be developed for Banach spaces or locally convex spaces.

[10] set of bounded linear operators on $\mathcal{G}$

[11] i.e. $k((x,g_1),(y,g_2))$ is bilinear in g_1, g_2.

where e_i denotes a basis in $\mathbb{R}^d$. Conversely given the PD operator valued kernel $k(x,y)$, define by the above expression the kernel function $\ell((x,i),(y,j))$. Then we have with $v_m \in \{1,\ldots,d\}$

$$\begin{aligned}\sum_{i,j=1}^{n}\sum_{m,n=1}^{d}\alpha_{im}\alpha_{jn}\ell((x_i,v_m),(x_j,v_n)) &= \sum_{i,j=1}^{n}\sum_{m,n=1}^{d}\alpha_{im}\alpha_{jn}k_{v_m v_n}(x_i,x_j)\\ &= \sum_{i,j=1}^{n}\left\langle \sum_{m=1}^{d}\alpha_{im}e_{v_m}, k(x_i,x_j)\sum_{n=1}^{d}\alpha_{jn}e_{v_n}\right\rangle\\ &= \sum_{i,j=1}^{n}\langle c_i, k(x_i,x_j)c_j\rangle\end{aligned}$$

with $c_i = \sum_{m=1}^{d}\alpha_{im}e_{v_m}$. Now if ℓ is positive definite then consider the index set of size nd given by $z_{im} = (x_i, v_m)$ which gives the above expression and implies that $k(x,y)$ is a PD operator-valued kernel, since we can express any vector $c \in \mathbb{R}^d$ in the form $\sum_{m=1}^{d}\alpha_m e_{v_m}$. Conversely let $k(x,y)$ be a PD operator-valued kernel and take as vectors $c_i = \alpha_i e_{v_i}$, then $\ell((x,i),(y,j))$ is a PD kernel function since we can express all index sets in the form $z_i = (x_i, v_i)$. □

The meaning of the above proposition is that at the price of changing the index set, one can simply work with real-valued kernels, and the positive definiteness of these kernels implies the positivity of the corresponding operator valued kernels. Moreover one can use the properties of the real-valued kernels to derive the properties of the operator-valued one.

5.2 Hilbertian Subspaces

Instead of trying to generalize the PD kernels, one may, as in the work of Schwartz [6] generalize the notion of RKHS and kernel operator. The idea is to consider instead of Hilbert spaces of real-valued functions, that is a Hilbertian subspace of $\mathbb{R}^{\mathcal{X}}$, subspaces of quit general spaces equipped with the structure of a Hilbert space that may not even contain functions. The framework of Schwartz is formulated in the very general setting of locally convex topological vector spaces (l.c.s.), see [11, 15] for an introduction. Note that $\mathbb{R}^{\mathcal{X}}$ with the topology of pointwise convergence is a complete l.c.s.. This topology is equivalent to the weak topology induced by the duality map $\langle\cdot,\cdot\rangle_{\mathbb{R}^{[\mathcal{X}]},\mathbb{R}^{\mathcal{X}}}$ defined above. In the following E denotes a complete l.c.s..

Definition 6 *A linear subspace $\mathcal{H} \subset E$ is called a* **Hilbertian subspace** *if*

(i) it is provided with $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\mathcal{H}$ is a Hilbert space.

(ii) The injection of $\mathcal{H}$ into E is continuous; that is convergence in $\mathcal{H}$ implies convergence in E.

Definition 7 *A* **kernel operator** *K is a linear, symmetric map*[12] *from E'*[13] *into E. K is said to be* **positive** *if for all $e' \in E'$, $\langle e', Ke'\rangle_{E',E} \geq 0$.*

The following theorem gives the analogue of the bijection between positive definite kernels and RKHS.

Theorem 9 *[6] There is a one-to-one correspondence between the closed convex cone of Hilbertian subspaces $\mathcal{H}$ and the positive kernel operators K. To $\mathcal{H}$ corresponds the kernel operator $K = j \circ \theta \circ j'$, where $j : \mathcal{H} \to E$ is the natural injection, $j' : E' \to \mathcal{H}'$ its adjoint and $\theta : \mathcal{H}' \to \mathcal{H}$ the canonical isomorphism. Moreover given a positive kernel operator K, the Hilbert space is given by $\mathcal{H} = \overline{KE'}$ with the inner product on KE' defined as*

$$\langle Ke', Kf'\rangle_{\mathcal{H}} = \langle e', Kf'\rangle_{E',E}\,.$$

The inner product in $\mathcal{H}$ defined in the above way 'reproduces' the value of e' on any element of E contained in $\mathcal{H}$. **Example:** [Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$] We have defined in a previous section a positive symmetric kernel operator $K : \mathbb{R}^{[\mathcal{X}]} \to \mathbb{R}^{\mathcal{X}}$. Since $\mathbb{R}^{\mathcal{X}}$ is a complete l.c.s., it is also a positive kernel operator in the sense of Schwartz. Additionally by Theorem 4, the associated reproducing kernel Hilbert spaces are Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$. So we do recover the standard RKHS as a special case of Schwartz's theory. The setting of Schwartz seems at first

[12] Note that a linear, symmetric map is weakly continuous.

[13] E' denotes the topological dual space E.

much to general for machine learning tasks. However as we will see soon it provides us with the right setting to deal with distribution valued kernels, which is a generalization of the usual kernel function. One could ask at this point why it is a good idea to consider kernels on functions instead of kernels on points. One can argue that because of the limited precision of the measurement device measurements of real-valued physical quantities can never be made with arbitrary precision. This measurement error can be modelled by considering, instead of points, functions with compact support which are concentrated on the measured points. The width of the function then models the uncertainty in the measurement. This means we smear the points before we compare them with the kernel function. The following famous theorem characterizes the form of the kernel operator when one considers Hilbertian subspace of distributions.

Theorem 10 (Schwartz kernel theorem) *The topological vector space of continuous linear maps $D(\mathbb{R}^n) \to D'(\mathbb{R}^n)$*[14]*, with the strong topology, is canonically isomorphic to the topological vector space $D'(\mathbb{R}^n \times \mathbb{R}^n)$.*

This theorem guarantees that we have again a unique correspondence between the kernel operator and a generalized kernel function as in the case of usual positive definite kernels. Indeed, in the abstract framework of Hilbertian subspaces, it is not clear that a function of two variables is naturally associated to a subspace. However, thanks to this result, it is true in the case of Hilbertian subspaces of distributions: they are naturally associated to a (generalized) kernel function which is actually a distribution on $\mathbb{R}^n \times \mathbb{R}^n$. We give a simple yet illustrative example of this phenomenon.

Example: [$L_2(\mathbb{R}^n)$ as a Hilbertian subspace of $D'(\mathbb{R}^n)$] Let $K = \delta(x-y) \in D'(\mathbb{R}^n \times \mathbb{R}^n)$. Then we have for all $f \in D(\mathbb{R}^n)$, $(Kf)(x) = \int_{\mathbb{R}^n} \delta(x-y) f(y) = f(x)$ and the inner product on $KD(\mathbb{R}^n)$ is defined as:

$$\langle Kf, Kg \rangle := \langle Kf, g \rangle_{D'(\mathbb{R}^n), D(\mathbb{R}^n)} = \int_{\mathbb{R}^n} f(x) g(x) dx.$$

Since $D(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$ and the above inner product induces an isometry between $KD(\mathbb{R}^n)$ and $L^2(\mathbb{R}^n)$ restricted to $D(\mathbb{R}^n)$ we get the desired result that $L^2(\mathbb{R}^n)$ is isometrically isomorphic to the Hilbertian subspace $\overline{KD(\mathbb{R}^n)} \subset D'(\mathbb{R}^n)$.

Remark: The example on Hilbertian subspaces of $\mathbb{R}^{\mathcal{X}}$ suggests that the framework of Hilbertian subspaces is a generalization of the Aronszajn framework of RKHS. But one can always see the elements of the Hilbertian subspace $H \subset E$ as linear functions on the dual E' acting via $h(e') = \langle e', h \rangle_{E',E}$. So $\mathcal{H}$ can be considered as a Hilbertian subspace of $\mathbb{R}^{E'}$. Since E' must have a special structure, whereas the Aronszajn approach works for any set $\mathcal{X}$, from this point of view Hilbertian subspaces are actually less general. For example the framework of distributions can be seen as a RKHS on $\mathbb{R}^{\mathcal{X}}$. The problem of the Aronszajn approach is that the special properties of the underlying set $\mathcal{X}$ play no role and are 'forgotten'. In general it seems that from the structural point of view the framework of Schwartz is better, from the practical point of view the framework of Aronszajn is maybe easier to handle.

5.3 The General Indefinite Case

In general it is not easy to check if a given symmetric function is a positive definite kernel. In some cases like $k(x,y) = \tanh(\alpha \langle x, y \rangle + \beta)$ it is even known that the associated kernel matrix can have negative eigenvalues. Nevertheless it is sometimes used in support vector machines. Naturally the question arises if there still exists something like reproducing kernel spaces, such that we can interpret this non-positive definite kernel as an indefinite inner product in these space. The theory of reproducing kernel spaces with indefinite inner products was to our knowledge first explored by Schwartz [6] in the framework of hermitian subspaces. A more explicit treatment following Aronszajn was done by Sorjonen [16].

5.3.1 Reproducing Kernel Pontryagin Spaces

Definition 8 *A symmetric kernel function $K(s,t) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to have κ* ***negative squares****, κ a nonnegative integer, if $\forall\, n \geq 1$, and all $x_1, \dots, x_n \in \mathcal{X}$ the matrix $(k(x_i, x_j)_{i,j=1,\dots,n})$ has at most κ negative eigenvalues and at least one such matrix has exactly κ negative eigenvalues.*

Now we define a generalization of Hilbert spaces.

[14] $D'(\mathbb{R}^n)$ denotes the distributions on $\mathbb{R}^n$ and $D(\mathbb{R}^n)$ the space of smooth functions on $\mathbb{R}^n$ with compact support with the strict inductive limit topology.

Definition 9 *A **Krein space** is an inner product space $\mathcal{H}$, which can be written as the orthogonal sum $\mathcal{H} = \mathcal{H}_+ \oplus \mathcal{H}_-$ of a Hilbert space $\mathcal{H}_+$ and the antispace*[15] *$\mathcal{H}_-$ of a Hilbert space. If the antispace $\mathcal{H}_-$ is finite dimensional then $\mathcal{H}$ is called **Pontryagin space**.*

This decomposition is not unique, but the resulting spaces are all isomorphic. The dimensions of $H_\pm$ are independent of the choice of the decomposition and are called positive and negative indices of $\mathcal{H}$.

Definition 10 *A **reproducing kernel Pontryagin space (RKPS)** $\mathcal{H}$ on $\mathcal{X}$ is a Pontryagin space of functions from $\mathcal{X}$ to $\mathbb{R}$ with a reproducing kernel $k(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that*

$$\begin{aligned} &\forall\, x \in \mathcal{X}, \quad k(x, \cdot) \in \mathcal{H} \\ &\forall\, f \in \mathcal{H}, \quad \langle f(\cdot), k(x, \cdot)\rangle_{\mathcal{H}} = f(x). \end{aligned}$$

The RKPS are very similar in their structure as the following two theorems show.

Theorem 11 *[16] A Pontryagin space $\mathcal{H}$ of real-valued functions on Ω admits a reproducing kernel $K(s, t)$ if and only if all evaluation functionals are continuous. In this case, $K(s, t)$ is unique, and it is a hermitian kernel having κ negative squares, where k is the negative index of $\mathcal{H}$.*

Theorem 12 *[16] If $K(s, t)$ is a hermitian kernel on $\mathcal{X} \times \mathcal{X}$ having κ negative squares, then there is a unique Pontryagin space $\mathcal{H}$ of functions on $\mathcal{X}$ with $dim\, \mathcal{H}^- = \kappa$ having $K(s, t)$ as reproducing kernel.*

5.3.2 Reproducing Kernel Krein Spaces

The following theorem gives necessary and sufficient conditions for a symmetric function to be a reproducing kernel of a Krein space.

Theorem 13 *[6] If $k(x, y)$, $x, y \in \mathcal{X}$, is a symmetric function with values in $\mathbb{R}$, the following assertions are equivalent*

(i) *k is the reproducing kernel of a Krein space $\mathcal{H}_k$ of functions on $\mathcal{X}$.*

(ii) *There exists an $\ell \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ such that $-\ell \preceq k \preceq \ell$.*

(iii) *$k = k_+ - k_-$ for some $k_+, k_- \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$.*

Unfortunately there exist counterexamples of symmetric functions which do not fulfill these conditions, but when the above conditions are satisfied, the reproducing kernel Krein space (RKKS) is characterized in the following way.

Proposition 6 *[6] If $k = k_+ - k_-$ with $k_+, k_- \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$, then one can choose k_+ and k_- such that the associated RKHS of k_+ and k_-, $\mathcal{H}_+$ respectively $\mathcal{H}_-$, fulfill $\mathcal{H}_+ \bigcap \mathcal{H}_- = \{0\}$. In this case the RKKS associated to k consists of the functions $f = f_+ + f_-$, $f_+ \in \mathcal{H}_+$, $f_- \in \mathcal{H}_-$ with the indefinite inner product $[f, g] = \langle f_+, g_+\rangle_{\mathcal{H}_+} - \langle f_-, g_-\rangle_{\mathcal{H}_-}$.*

6 Conclusion

We have tried to extract, from the huge and scattered mathematical literature on kernels, the basic facts that are relevant to the researchers in Machine Learning working with kernel methods. The motivation for such a work came from noticing that these concepts were sometimes misused or ignored by the community. In particular, if one wants to develop generalizations of these concepts, it should be clear that there already exist several points of view for such generalizations and that, changing the point of view they can be cast in the same framework.

Finally, we have to say that this work is far from being complete and there exist many other notions to be explored (and made accessible to the community). We hope to be able to provide an extended and more comprehensive account (covering for example Gaussian measures, generalized stochastic processes, group representations in RKHS, spectral decompositions of kernels, regularization theory and various results of applications in approximation, interpolation etc.) in the near future.

Acknowledgements

The authors are grateful to Gilles Blanchard, Bernhard Schölkopf and Ulrike von Luxburg for helpful comments.

[15] An antispace of a Hilbert space is $(\mathcal{H}, \langle\cdot,\cdot\rangle)_{\mathcal{H}}$ is given by $(\mathcal{H}, -\langle\cdot,\cdot\rangle_{\mathcal{H}})$.

A Structures Associated to a Gaussian Process

In this section we introduce extra objects that are naturally associated to a Gaussian process (hence to a PD kernel). We refer to [17] for additional details.
We denote by E an arbitrary locally convex space.

Definition 11 *A Borel probability measure μ on E is a* **Gaussian measure** *if each $e' \in E'$, regarded as a random variable defined on the probability space (E, μ) is Gaussian.*

Definition 12 *A random variable X with values in E is a* **Gaussian vector** *if the real-valued random variable $\langle e', X\rangle_{E',E}$ is Gaussian for every $e' \in E'$, or equivalently, if the distribution of X is a Gaussian measure on E.*

Theorem 14 (Kolmogorov extension theorem) *Let $\Omega = \mathbb{R}^{\mathcal{X}}$, where $\mathcal{X}$ is an arbitrary index set, and let $\mathcal{F}$ be the product σ-field $\mathcal{B}^{\mathcal{X}}$ on Ω. Suppose that for every finite subset $\mathcal{Y} \subseteq \mathcal{X}$, we are given a (consistent) probability measure $P_{\mathcal{Y}}$ on $\mathbb{R}^{\mathcal{Y}}$; then there exists a unique probability measure on $\mathbb{R}^{\mathcal{X}}$ such that the projection onto $\mathbb{R}^{\mathcal{Y}}$ induces $P_{\mathcal{Y}}$ for every finite $\mathcal{Y}$.*

It follows from this theorem that all the objects introduced before are tightly related.

Proposition 7 *Every Gaussian process $(X_x)_{x\in\mathcal{X}}$ defines a unique Gaussian measure on $\mathbb{R}^{\mathcal{X}}$ and a unique random vector X with values in $\mathbb{R}^{\mathcal{X}}$.*

We now give the construction of a feature map via the Kolmogorov theorem [18]. Given a PD kernel k on $\mathcal{X}$ define for any finite subset $\mathcal{Y} = x_1, \dots, x_n$ a probability measure which is centered Gaussian and has covariance matrix $(k(x_i, x_j))_{i,j}$. By Theorem 14 there exists a measure μ on $\mathbb{R}^{\mathcal{X}}$ and it is Gaussian. If we consider the Hilbert space $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ and define $X_x := f(x),\ f \in \mathbb{R}^{\mathcal{X}}$ (where f has the distribution μ), then X_x is an element of $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ and $\mathbb{E}\left[X_x X_y\right] = \int f(x)f(y)d\mu(f) = k(x, y)$. Moreover one can check that the completion of the subspace of Gaussian random variables X_x in $L_2(\mathbb{R}^{\mathcal{X}}, \mu)$ still consists only of Gaussian random variables. Therefore it is called **Gaussian Hilbert space**. It is shown in Janson [17] that the Gaussian Hilbert space is isometric isomorphic to the RKHS associated to the PD kernel k.

References

[1] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1990.

[2] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[3] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] M. Atteia. *Hilbertian Kernels and Spline Functions*. North Holland, Amsterdam, 1992.

[6] L. Schwartz. Sous-espaces hilbertiens et noyaux associés. *Journal d'Analyse, Jerusalem*, XIII:115–256, 1964.

[7] H. Niemi. Stochastic processes as Fourier transforms of stochastic measures. *Ann. Acad. Sci. Fenn. Ser. A I*, 591, 1975.

[8] M. Krein. Hermitian-positive kernels on homogeneous spaces I and II. *Amer. Math. Soc. Translations Ser. 2*, 34:69–164, 1963. Original: Ukrain. Mat. Z., 1, 64-98,(1949), and 2, 10-59,(1950).

[9] M. N. Lukić and J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Am. Math. Soc.*, 353:3945–3969, 2001.

[10] J. B. Conway. *A course in Functional Analysis*. Springer, New York, 1985.

[11] M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis*. Academic Press, San Diego, 1980.

[12] E. Giné and J. Zinn. Marcinkiewicz type laws of large numbers and convergence of moments for U-statistics. In *Probability in Banach spaces, 8, Brunswick, 1991*, pages 273–291. Birkhäuser Boston, 1992.

[13] C. A. Micchelli and M. Pontil. On learning vector-valued functions. Technical Report RN/03/08, Dept. of Computer Science, University College London, 2003.

[14] A. Devinatz. On measurable positive definite operator functions. *J. London Math. Soc.*, 35:417–424, 1960.

[15] H. H. Schaefer and M. P. Wolff. *Topological Vector Spaces*. Springer, New York, 1999. Second edition.

[16] P. Sorjonen. Pontrjaginräume mit einem reproduzierenden Kern. *Ann. Acad. Sci. Fenn. Ser. A I*, 594, 1975.

[17] S. Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, Cambridge, 1997.

[18] K. R. Parthasarathy and K. Schmidt. *Positive definite kernels, continuous tensor products, and central limit theorems of probability theory*, volume 272 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1972.