Journal of Machine Learning Research 7 (2006) 2651-2667 Submitted 7/06; Revised 10/06; Published 12/06

Universal Kernels

Charles A. Micchelli CAM@MATH.ALBANY.EDU
Department of Mathematics and Statistics
State University of New York
The University at Albany
Albany, New York 12222, USA

Yuesheng Xu YXU06@SYR.EDU
Haizhang Zhang HZHANG12@SYR.EDU
Department of Mathematics
Syracuse University
Syracuse, NY 13244, USA

Editor: Gabor Lugosi

Abstract

In this paper we investigate conditions on the features of a continuous kernel so that it may approximate an arbitrary continuous target function uniformly on any compact subset of the input space. A number of concrete examples are given of kernels with this universal approximating property.

Keywords: density, translation invariant kernels, radial kernels

1. Introduction

Let $\mathcal{X}$ be a prescribed input space and set $\mathbb{N}_n := \{1,2,\ldots,n\}$. We shall call a function K from $\mathcal{X}\times\mathcal{X}$ to $\mathbb{C}$ a *kernel* on $\mathcal{X}$ provided that for *any* finite sequence of inputs $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq \mathcal{X}$ the matrix

$$K_{\mathbf{x}} := (K(x_j, x_k) : j,k \in \mathbb{N}_n) \tag{1}$$

is Hermitian and positive semi-definite. Kernels are an essential component in a multitude of novel algorithms for pattern analysis (Bishop, 1995; Hastie et al., 2001; Schölkopf and Smola, 2002). Besides their superior performance on a wide spectrum of learning tasks from data, they have a substantial theoretical basis, as they are reproducing kernels of Hilbert spaces of functions on $\mathcal{X}$ for which point evaluation is always continuous (Aronszajn, 1950). Such spaces are called *Reproducing Kernel Hilbert Spaces* (RKHS) and an important reason for the interest in kernels is the (essentially) unique correspondence between them and RKHS. This relationship leads, by means of the regularization approach to learning, functions having the representation

$$f := \sum_{j\in\mathbb{N}_n} c_j K(\cdot, x_j) \tag{2}$$

where $\{c_j : j \in \mathbb{N}_n\} \subseteq \mathbb{C}$ are parameters typically obtained from training data (Bishop, 1995; Evgeniou et al., 2000; Hastie et al., 2001; Schölkopf and Smola, 2002). This useful fact is known as the *Representer Theorem* and has wide applicability (Schölkopf et al., 1999; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Wahba, 1990). We shall refer to the function in the sum on the right hand side of (2) as *sections* of the kernel K.

Certainly, the choice of the kernel in (2) affects the performance of kernel based learning algorithms and so, is important. For recent work in this direction, see Argyriou et al. (2005, 2006), Bach et al. (2004), Lanckriet et al. (2004), Micchelli and Pontil (2005), Micchelli et al. (2006), Neumann et al. (2004), Ong et al. (2005), Sonnenburg et al. (2006) and references therein. Following Poggio et al. (2002), we ask a conceptually simpler, but very basic question about choosing the kernel : can the function representation (2), as the number of summands increases without bound, approximate any target function arbitrarily close? In the study of this question it is important which norm is used to compute the error between the function appearing in (2) and a given target function. Indeed, it is well-known that if we use the norm in the RKHS whose kernel is K then all members of this Hilbert space are approximable arbitrarily by functions of the type appearing in (2). Actually, this is the way the Hilbert space associated with a kernel is constructed from the kernel itself (Aronszajn, 1950).

Our concern here is with the *uniform norm*. To this end, we assume that the input space $\mathcal{X}$ is a Hausdorff topological space and that all kernels to be considered are *continuous* on $\mathcal{X} \times \mathcal{X}$. To begin to address the problem which interests us here, we let $\mathcal{Z}$ be a fixed but arbitrary compact subset of $\mathcal{X}$ and, as usual, let $C(\mathcal{Z})$ be the space of all continuous complex-valued functions from $\mathcal{Z}$ to $\mathbb{C}$ equipped with maximum norm $\|\cdot\|_{\mathcal{Z}}$. Our hypothesis that the input space is Hausdorff ensures that it has an abundance of compact subsets. We shall always enforce this hypothesis throughout and for simplicity of presentation we do not mention it again.

Given a kernel K we form the space of *kernel sections*

$$K(\mathcal{Z}) := \overline{\text{span}}\{K_y : y \in \mathcal{Z}\},$$

where $K_y : \mathcal{X} \to \mathbb{C}$ is the function defined at every $x \in \mathcal{X}$ by the equation $K_y(x) := K(x,y)$. The set $K(\mathcal{Z})$ consists of all functions in $C(\mathcal{Z})$ which are uniform limits of functions of the form (2) where $\{x_j : j \in \mathbb{N}_n\} \subseteq \mathcal{Z}$.

We want to identify kernels with the following *universal approximating property*: given *any* prescribed compact subset $\mathcal{Z}$ of $\mathcal{X}$, any positive number ε and any function $f \in C(\mathcal{Z})$ there is a function $g \in K(\mathcal{Z})$ such that $\|f - g\|_{\mathcal{Z}} \leq \varepsilon$. That is, for any choice of compact subset $\mathcal{Z}$ of the input space $\mathcal{X}$, the set $K(\mathcal{Z})$ is *dense* in $C(\mathcal{Z})$ in the maximum norm. When a kernel has this property we call it a *universal kernel*. In other words, a universal kernel K has the property that $K(\mathcal{Z}) = C(\mathcal{Z})$. It is this question of characterizing universal kernels that we address here. We shall demonstrate that it has a satisfactory resolution in terms of any feature map representation of the kernel K. Indeed, we provide a necessary and sufficient condition for K to have the universal approximating property in terms of its features, thereby completing preliminary remarks made about this problem in Micchelli and Pontil (2004) and Micchelli et al. (2003).

Concrete examples of kernels with their feature maps and observations about the associated density problem are investigated in Section 3. In Section 4, we stress translation invariant kernels on $\mathbb{R}^d$ and give several useful sufficient conditions for K to be a universal translation invariant kernel. This discussion includes the popular choice of the Gaussian kernel. We end the paper with a remark about issues for further investigation.

2. Kernels Defined by Feature Maps

We start from a Hilbert space $\mathcal{W}$ over $\mathbb{C}$ and a continuous kernel K on $\mathcal{X}\times\mathcal{X}$. A *feature map* for the kernel K is any *continuous* function $\Phi:\mathcal{X}\to\mathcal{W}$ such that for each $(x,y)\in\mathcal{X}\times\mathcal{X}$

$$K(x,y)=(\Phi(x),\Phi(y))_{\mathcal{W}} \tag{3}$$

where $(\cdot,\cdot)_{\mathcal{W}}$ is the inner product on $\mathcal{W}$. Every kernel has such a representation and conversely whenever it does then it is a kernel. However, a feature space representation is *not* unique. Let us elaborate on these well-known facts. First, it is straightforward to see that any function K which has the representation (3) is a kernel. The reason for this fact is that the input matrix appearing in (1) is formed by the mutual inner products of the set of vectors $\{\Phi(x_j):j\in\mathbb{N}_n\}$ and such a matrix is certainly Hermitian and positive semi-definite. To establish the converse, we first construct the Hilbert space $\mathcal{H}$ associated with the continuous kernel K and then observe for all $x,y\in\mathcal{X}$, by the reproducing kernel property, that

$$K(x,y):=(K_x,K_y)_{\mathcal{H}}.$$

Hence, we may choose $\mathcal{W}=\mathcal{H}$ and for any $x\in\mathcal{X}$ we let $\Phi(x)=K_x$. This feature space representation is continuous because for all $x,y\in\mathcal{X}$ we have that

$$\|\Phi(x)-\Phi(y)\|_{\mathcal{W}}^2=K(x,x)+K(y,y)-K(x,y)-K(y,x)$$

where $\|\cdot\|_{\mathcal{W}}$ denotes the norm on $\mathcal{W}$.

There are alternate means to construct a feature space representation for a continuous kernel K which has the advantage that the Hilbert space $\mathcal{W}$ can be chosen to be *separable*. To construct such a representation we must *choose* a compact subset $\mathcal{Z}$ of $\mathcal{X}$ and a finite Borel measure μ on $\mathcal{Z}$ with $\operatorname{supp}(\mu)=\mathcal{Z}$ (see (15) for the definition of the support of a Borel measure). This measure yields a linear operator $T:L^2(\mathcal{Z},\mu)\to L^2(\mathcal{Z},\mu)$ defined for $g\in L^2(\mathcal{Z},\mu)$ by the equation

$$Tg:=\int_{\mathcal{Z}}K(\cdot,y)g(y)d\mu(y). \tag{4}$$

Following the ideas of Mercer (1909), T has countably many nonnegative Eigenvalues (each of finite multiplicities with zero as the only accumulation point of nonnegative Eigenvalues) $\{\lambda_i:i\in\mathbb{N}\}$ and corresponding orthonormal Eigenfunctions $\{\phi_i:i\in\mathbb{N}\}\subseteq L^2(\mathcal{Z},\mu)$ such that

$$K(x,y):=\sum_{i\in\mathbb{N}}\lambda_i\phi_i(x)\overline{\phi_i(y)},\quad (x,y)\in\mathcal{Z}\times\mathcal{Z} \tag{5}$$

where the series above converges absolutely and uniformly on $\mathcal{Z}\times\mathcal{Z}$, see also Lax (2002).

To write K in the form (3), we let $\ell^2(\mathbb{N})$ be the Hilbert space of square summable sequences on $\mathbb{N}$ and define a feature map $\Psi:\mathcal{Z}\to\ell^2(\mathbb{N})$ at each $x\in\mathcal{Z}$ and $j\in\mathbb{N}$ as

$$\Psi(x)(j):=\sqrt{\lambda_j}\phi_j(x).$$

Therefore, the Mercer representation in Equation (5) establishes for each $x,y\in\mathcal{Z}$ that

$$K(x,y)=(\Psi(x),\Psi(y))_{\ell^2(\mathbb{N})}.$$

Since we have for all $x, y \in \mathcal{Z}$ that

$$\|\Psi(x) - \Psi(y)\|^2_{\ell^2(\mathbb{N})} = K(x,x) + K(y,y) - K(x,y) - K(y,x),$$

the continuity of the K implies that the feature map $\Psi : \mathcal{Z} \to \ell^2(\mathbb{N})$ is also continuous.

Of course, to find an Eigenfunction feature representation of a kernel, except in special circumstances, is a serious challenge both analytically and computationally. Moreover, it should be observed that this feature space representation depends on the measure μ. Recent extensions of the Mercer theorem can be found in Sun (2005) and the reference therein.

Let us now return to the general case of formula (3). To this end, we need to recall facts about the dual space of $C(\mathcal{Z})$, that is, the space of all continuous linear functionals on $C(\mathcal{Z})$. By the Riesz representation theorem the linear functionals in dual space of $C(\mathcal{Z})$ are identified as regular complex-valued measures on $\mathcal{Z}$ (see, for example, Lax, 2002; Royden, 1988). The norm of a complex-valued measure, which is inherited from the norm on $C(\mathcal{Z})$, is its *total variation* and is defined as

$$\mathrm{TV}(\nu) := \sup\{|\int_{\mathcal{Z}} g(x) d\nu(x)| : \|g\|_{\mathcal{Z}} \le 1, g \in C(\mathcal{Z})\}.$$

We denote the space of all regular complex-valued measures on $\mathcal{Z}$ with this norm by $B(\mathcal{Z})$. For any $\nu \in B(\mathcal{Z})$, we wish to define the integral $\int_{\mathcal{Z}} \Phi(x) d\nu(x)$ as an element of $\mathcal{W}$. This is done by noting that the *conjugate linear functional* L defined on $\mathcal{W}$ for each $u \in \mathcal{W}$ by the equation

$$L(u) := \int_{\mathcal{Z}} (\Phi(x), u)_{\mathcal{W}} d\nu(x) \tag{6}$$

has a norm satisfying the inequality

$$\|L\| \le \mathrm{TV}(\nu)\|\Phi\|_\infty < \infty,$$

where $\|\Phi\|_\infty := \max\{\|\Phi(x)\|_{\mathcal{W}} : x \in \mathcal{Z}\}$. Therefore, by the Riesz representation theorem, for the Hilbert space $\mathcal{W}$ (Lax, 2002; Rudin, 1991) there exists a *unique* element $w \in \mathcal{W}$ such that for each $u \in \mathcal{W}$ that

$$L(u) = (w, u)_{\mathcal{W}}.$$

It is this vector w which we shall denote by $\int_{\mathcal{Z}} \Phi(x) d\nu(x)$. Consequently, we have the useful formula

$$(\int_{\mathcal{Z}} \Phi(x) d\nu(x), u)_{\mathcal{W}} = \int_{\mathcal{Z}} (\Phi(x), u)_{\mathcal{W}} d\nu(x) \tag{7}$$

valid for all $u \in \mathcal{W}$.

Next, we introduce a map $U : B(\mathcal{Z}) \to \mathcal{W}$ by letting for each $\nu \in B(\mathcal{Z})$

$$U(\nu) := \int_{\mathcal{Z}} \Phi(x) d\nu(x) \tag{8}$$

and so the formula (7) becomes

$$(U(\nu), u)_{\mathcal{W}} = \int_{\mathcal{Z}} (\Phi(x), u)_{\mathcal{W}} d\nu(x). \tag{9}$$

For any $y \in \mathcal{Z}$ we set $u = \Phi(y)$ in formula (9) above to obtain by the definition of the kernel (3) that

$$(U(\nu), \Phi(y))_{\mathcal{W}} = \int_{\mathcal{Z}} K(x,y) d\nu(x). \tag{10}$$

We now conjugate both sides of this equation, integrate both sides of the resulting equation with respect to the complex-valued measure ν and then simplify the resulting left side by another application of Equation (7) with the choice $u = U(\nu)$. Next, we use the feature space representation of the kernel in (3) on the right hand side of the equation to obtain the equation

$$\|U(\nu)\|_{\mathcal{W}}^2 = \int_{\mathcal{Z}}\int_{\mathcal{Z}} K(x,y)d\overline{\nu}(y)d\nu(x) \tag{11}$$

where $\overline{\nu}$ is the conjugate of the complex-valued measure ν defined for each Borel set $\mathcal{S} \subseteq \mathcal{Z}$ by $\overline{\nu}(\mathcal{S}) := \overline{\nu(\mathcal{S})}$. We remark here that since the kernel K is continuous and $\mathcal{Z}$ is compact, Fubini's theorem assures that we can interchange the order in the integral on the right hand side of the above equation (see, for example, Royden, 1988). Moreover, from this formula it follows that the linear operator U is continuous. Indeed, its norm satisfies the inequality

$$\|U\| \leq \sqrt{\|K\|_\infty}$$

where $\|K\|_\infty$ denotes the maxium norm of K on $C(\mathcal{Z}\times\mathcal{Z})$.

To continue, we recall the notion of *annihilator* of a subset $\mathcal{V}$ of $C(\mathcal{Z})$. This consists of all elements in $B(\mathcal{Z})$ which are zero on all functions in $\mathcal{V}$. In other words, we have that

$$\mathcal{V}^\perp := \{\nu : \nu \in B(\mathcal{Z}), \int_{\mathcal{Z}} f(x)d\nu(x) = 0, f \in \mathcal{V}\}.$$

Note that the annihilator of a subset of $C(\mathcal{Z})$ is a subspace of $B(\mathcal{Z})$. Furthermore, the *closed linear span* of subset $\mathcal{V}$ of $C(\mathcal{Z})$, denoted by $\overline{\text{span}}\,\mathcal{V}$, has the same annihilator as the set $\mathcal{V}$ itself. Moreover, two subsets $\mathcal{V}_1$ and $\mathcal{V}_2$ of $C(\mathcal{Z})$ have the same annihilator if and only if $\overline{\text{span}}\,\mathcal{V}_1 = \overline{\text{span}}\,\mathcal{V}_2$. Also, recall that two closed subspaces are equal if and only if their annihilators are the same (see, for example, Lax, 2002; Royden, 1988; Rudin, 1991, for these facts).

We shall denote the null space of U by $\mathcal{N}(U)$, that is, the subspace of all elements ν in $B(\mathcal{Z})$ for which $U(\nu)$ is zero, given by

$$\mathcal{N}(U) := \{\nu : \nu \in B(\mathcal{Z}), U(\nu) = 0\}.$$

We remark here that the operator U depends on the set $\mathcal{Z}$.

Proposition 1 *If $\mathcal{Z}$ is a compact subset of the input space $\mathcal{X}$ then*

$$K(\mathcal{Z})^\perp = \mathcal{N}(U). \tag{12}$$

Consequently, $K(\mathcal{Z}) = C(\mathcal{Z})$ if and only if U is injective.

Proof By the Hahn Banach theorem, the linear span of a subset $\mathcal{V}$ is dense in $C(\mathcal{Z})$, that is, $\overline{\text{span}}\,\mathcal{V} = C(\mathcal{Z})$ if and only if $\mathcal{V}^\perp = \{0\}$ (Lax, 2002; Royden, 1988; Rudin, 1991). Therefore, the second claim follows from (12). We now turn to the proof of this equation. If $\nu \in K(\mathcal{Z})^\perp$ then by definition, for all $y \in \mathcal{Z}$ we have that

$$\int_{\mathcal{Z}} K(x,y)d\nu(x) = 0$$

and so by (11) we get that $\nu \in \mathcal{N}(U)$. Therefore, we have established that $K(\mathcal{Z})^\perp \subseteq \mathcal{N}(U)$. To prove the opposite inclusion we suppose that $\nu \in \mathcal{N}(U)$, then appeal to (10) and conclude that $\nu \in K(\mathcal{Z})^\perp$, thereby proving the theorem.

■

Equation (7) has another consequence. To this end, we introduce a subspace of $\mathcal{W}$ defined as

$$\Phi(\mathcal{Z}) := \overline{\operatorname{span}}\{\Phi(x) : x \in \mathcal{Z}\}.$$

If Q is a linear mapping between two linear spaces and $\mathcal{S}$ is a subset of its domain we use the standard notation $Q(\mathcal{S})$ for its image under Q. When the set $\mathcal{S}$ is the *domain* of Q then its image is the range of Q and is denoted by $\mathcal{R}(Q)$. From these definitions it follows that $Q(\operatorname{span}\mathcal{S}) = \operatorname{span} Q(\mathcal{S})$.

Proposition 2

$$\overline{\mathcal{R}(U)} = \Phi(\mathcal{Z}). \tag{13}$$

Proof We shall prove the proposition by showing that

$$\mathcal{R}(U)^\perp = \Phi(\mathcal{Z})^\perp.$$

If $u \in \Phi(\mathcal{Z})^\perp$ then (7) implies for any $\nu \in B(\mathcal{Z})$ that $u \in \mathcal{R}(U)^\perp$. Conversely, if $u \in \mathcal{R}(U)^\perp$ then again we get for any $\nu \in B(\mathcal{Z})$ from (7) that $\int_{\mathcal{Z}} (\Phi(x), u)_{\mathcal{W}} d\nu(x) = 0$. In particular, choosing ν to be the point evaluation at an arbitrary $x \in \mathcal{Z}$ we obtain that $(\Phi(x), u)_{\mathcal{W}} = 0$ and so we conclude that $u \in \Phi(\mathcal{Z})^\perp$, thereby establishing (13). ■

Let us now introduce another linear operator $V : \mathcal{W} \to C(\mathcal{Z})$ defined for any $u \in \mathcal{W}$ and $x \in \mathcal{Z}$ as $(V(u))(x) := (\Phi(x), u)_{\mathcal{W}}$. Certainly, V is bounded, as its norm satisfies the inequality $\|V\| \leq \|\Phi\|_\infty$. Moreover, according to (7) we have that

$$(U(\nu), u)_{\mathcal{W}} = \int_{\mathcal{Z}} (V(u))(x) d\nu(x) \tag{14}$$

which means that the *adjoint* of the operator V denoted by $V^* : B(\mathcal{Z}) \to \mathcal{W}$ is U, that is, $U = V^*$. Next, we point out a consequence of this fact and Proposition 1.

Corollary 3 $K(\mathcal{Z}) = \overline{\mathcal{R}(V)}$.

Proof It is generally true for any bounded linear operator that $\mathcal{N}(Q^*) = \mathcal{R}(Q)^\perp$ (Lax, 2002; Rudin, 1991) and, in particular, $\mathcal{N}(U) = \mathcal{R}(V)^\perp$ so the result follows directly from Proposition 1. ■

We recall that the linear span of a subset $\mathcal{S}$ is dense in $\mathcal{W}$, that is, $\overline{\operatorname{span}}S = \mathcal{W}$, if and only if the only $u \in \mathcal{W}$ with $(u, v) = 0$ for all $v \in \mathcal{S}$ is $u = 0$ (We already use a similar fact for the space $C(\mathcal{Z})$). It follows directly from Equation (14), for any subset $\mathcal{S}$ of $\mathcal{W}$ such that $\operatorname{span}\mathcal{S}$ is dense in $\mathcal{W}$, that $V(\mathcal{S})^\perp = \mathcal{N}(U)$ and so with this remark and Proposition 1 we conclude that

$$K(\mathcal{Z}) = \overline{\operatorname{span}} V(\mathcal{S}).$$

We use this equation in the following manner. Recall that a subset $\mathcal{Y}$ of $\mathcal{W}$ is *orthonormal* if for every distinct elements $u,v \in \mathcal{Y}$ we have that $(u,v) = 0$ and also $(u,u) = 1$. Every Hilbert space has an orthonormal basis $\mathcal{Y}$ (which may not be countable) such that for every $u \in \mathcal{W}$ the set $\{y : y \in \mathcal{Y}, (u,y) \neq 0\}$ is countable. Moreover, we have for $u \in \mathcal{W}$ the decomposition

$$u = \sum_{y \in \mathcal{Y}} (u,y)y,$$

where the sum on the right hand side of this equation converges in $\mathcal{W}$ for *any* ordering of elements of $\mathcal{Y}$ (Lax, 2002; Rudin, 1991). Corresponding to each element y in an orthonormal basis $\mathcal{Y}$ of $\mathcal{W}$ we define the function $F_y \in C(\mathcal{Z})$ at $x \in \mathcal{Z}$ by the equation $F_y(x) = (\Phi(x),y)_{\mathcal{W}}$ and introduce the corresponding subspace of $C(\mathcal{Z})$

$$\Phi(\mathcal{Y}) := \overline{\text{span}}\{F_y : y \in \mathcal{Y}\}.$$

Note the important difference between the sets $\Phi(\mathcal{Z})$ and $\Phi(\mathcal{Y})$. The first is in the Hilbert space $\mathcal{W}$ and the second is in $C(\mathcal{Z})$. Combining the above remarks we obtain the following equivalence between density of kernel representation and feature function density in $C(\mathcal{Z})$.

Theorem 4 *If $\mathcal{Z}$ is a compact subset of the input space $\mathcal{X}$, K a kernel with feature space representation (3) and $\mathcal{Y}$ an orthornormal basis for $\mathcal{W}$ then $K(\mathcal{Z}) = \Phi(\mathcal{Y})$.*

Parallel to our notion that a kernel is universal, we say a *feature map* Φ is *universal* provided that given any compact subset $\mathcal{Z}$ of the input space $\mathcal{X}$, any positive number ε and any function $f \in C(\mathcal{Z})$ there is a function $g \in \Phi(\mathcal{Y})$ such that $\|f - g\|_{\mathcal{Z}} \leq \varepsilon$. That is, for *any* choice of compact subset $\mathcal{Z}$ of the input space $\mathcal{X}$, the set $\{F_y : y \in \mathcal{Y}\}$ is *dense* in $C(\mathcal{Z})$. In other words, we have that $\Phi(\mathcal{Y}) = C(\mathcal{Z})$. Therefore, with this terminology we can succinctly summarize our conclusion in Theorem 4 by saying that *a kernel K expressed in feature space form (3) is universal if and only if its features are universal*!

We now consider an alternate way to express the universality of a kernel K in terms of the operator T and the corresponding measure μ defined in Equation (4) which determines it. To this end, we recall the that the *support* of a Borel measure ν on $\mathcal{Z}$ is defined to be the closed set

$$\text{supp}(\nu) := \bigcap \left\{ S \subseteq \mathcal{Z} : S \text{ is closed, } \nu(S^C) = 0 \right\}. \tag{15}$$

Consequently, if $\int_{\mathcal{Z}} f(x)d\nu(x) = 0$, ν a Borel measure with $\text{supp}(\nu) = \mathcal{Z}$ and f is a nonnegative and continuous function on $\mathcal{Z}$ then $f = 0$.

The first statement we make is about the mapping T defined in Equation (4) which is an immediate consequence of Theorem 4 concerning its Eigenfunctions.

Corollary 5 *If K is a kernel on an input space $\mathcal{X}$, $\mathcal{Z}$ a compact subset of $\mathcal{X}$, $\{\lambda_i : i \in \mathbb{N}\} \subseteq \mathbb{R}_+ \backslash \{0\}$ and $\{\phi_i : i \in \mathbb{N}\} \subseteq L^2(\mathcal{Z},\mu)$ are the nonzero Eigenvalues and corresponding orthonormal Eigenfunctions of the compact operator T where $\text{supp}(\mu) = \mathcal{Z}$ then $K(\mathcal{Z}) = C(\mathcal{Z})$ if and only if $\overline{\text{span}}\{\phi_i : i \in \mathbb{N}\} = C(\mathcal{Z})$.*

The next comment concerns the range of the operator T.

Theorem 6 *If supp(μ)= $\mathcal{Z}$ for the measure appearing in Equation (4) then $K(\mathcal{Z}) = \overline{\mathcal{R}(T)}$.*

Proof It suffices to show that $K(Z)^\perp = \mathcal{R}(T)^\perp$. If $\nu \in K(Z)^\perp$ then for each $y \in Z$

$$\int_Z K(x,y)d\nu(x) = 0.$$

By Fubini's theorem, we observe for each $g \in L^2(Z,\mu)$ that

$$\int_Z (Tg)(x)d\nu(x) = \int_Z g(y)\{\int_Z K(x,y)d\nu(x)\}d\mu(y) = 0$$

and so we conclude that $\nu \in \mathcal{R}(T)^\perp$. Conversely, $\nu \in \mathcal{R}(T)^\perp$ then by the above equation we obtain for any $g \in C(Z)$ that

$$\int_Z g(y)\{\int_Z K(x,y)d\nu(x)\}d\mu(y) = 0.$$

We now choose $g = \overline{\int_Z K(x,\cdot)d\nu(x)}$ in this equation and conclude that

$$\int_Z |g(y)|^2 d\mu(y) = 0.$$

Since $\operatorname{supp}(\mu) = Z$, we obtain that $g = 0$, that is, $\nu \in K(Z)^\perp$. ■

As a consequence of Theorem 6, we observe that $K(Z) = C(Z)$ if and only if $\overline{\mathcal{R}(T)} = C(Z)$.

We end this section by remarking that the results presented here may be extended from $C(Z)$ to L^p- spaces where $p \in [1,\infty)$. However, as we remarked in the introduction our focus here is on the maximum norm and so we do not go into this matter here.

3. Examples of Universal Kernels

In this section, we give examples of kernels defined by feature maps and study the corresponding density problem. We begin with a set $\{\phi_j : j \in I\}$ of continuous complex-valued functions on X where I is a countable set of indices and define the kernel K by

$$K(x,y) := \sum_{j\in I} \phi_j(x)\overline{\phi_j(y)}, \quad (x,y) \in X \times X, \tag{16}$$

where we assume that the series converges uniformly on $Z \times Z$ for every compact subset Z of X. To make use of the presentation in Section 2 we set $\mathcal{W} = \ell^2(\mathbb{N})$, choose the standard orthonormal basis $\mathcal{Y}$ for $\mathcal{W}$ in Theorem 4 and obtain the following result.

Theorem 7 *The kernel K defined by Equation (16) is universal if and only if the set of features $\{\phi_j : j \in I\}$ is universal.*

We shall now apply Theorem 7 to *dot product kernels* on various domains of $\mathbb{R}^d$ and $\mathbb{C}^d$. To this end, we start with an entire function G defined at any $z \in \mathbb{C}$ by the equation

$$G(z) := \sum_{n\in\mathbb{Z}_+} a_n z^n \tag{17}$$

where the coefficients $\{a_n : n \in \mathbb{Z}_+\}$ are assumed to be all *positive*. The function G induces the dot product kernel K defined at $x,y \in \mathbb{C}^d$ by the equation

$$K(x,y) := G((x,y)) \tag{18}$$

where we shall always use (x,y) for the standard inner product between the vectors x and y. For an extensive discussion of dot product kernels see FitzGerald et al. (1995).

Corollary 8 *The dot product kernel defined in Equation (18) is universal on $\mathbb{C}^d$ and $\mathbb{R}^d$.*

Proof For any lattice vector $\alpha := (\alpha_j : j \in \mathbb{N}_d) \in \mathbb{Z}_+^d$ we set $|\alpha| := \sum_{j\in\mathbb{N}_d} \alpha_j$. Using the multinomial expansion we conclude that the dot product kernel defined in Equation (18) can be expressed in the form

$$K(x,y) := \sum_{\alpha\in\mathbb{Z}_+^d} \phi_\alpha(x)\overline{\phi_\alpha(y)}, \;\; x,y \in \mathbb{C}^d$$

where the features are defined for $\alpha \in \mathbb{Z}_+^d$ at $x \in \mathbb{C}^d$ as

$$\phi_\alpha(x) := \sqrt{a_{|\alpha|}\binom{|\alpha|}{\alpha}} \;\; x^\alpha.$$

As is well-known, for example as a special case of the Stone-Weierstrass approximation theorem (Rudin, 1991, page 122), these features are universal on $\mathbb{C}^d$ and $\mathbb{R}^d$ so the result follows from Theorem 7. ■

The next result we present is a version of the above remark appropriate for the unit ball $\mathbb{B}^d := \{x : (x,x) < 1\}$ in $\mathbb{R}^d$. We have in mind the following fact. Again, we start with the function G defined above in (17) but in the next result we only assume that it is analytic in the unit disc $\Delta := \{z : |z| < 1, z \in \mathbb{C}\}$

Corollary 9 *If G is analytic in Δ and has all positive coefficients then K is universal on $\mathbb{B}^d$.*

The proof is identical to the proof of Corollary 8 and therefore is omitted.

We end our discussion of dot product kernels by considering the case of the unit sphere $\mathbb{S}^d$ in $\mathbb{R}^{d+1}$. To this end, we review the construction of *Schoenberg kernels* on $\mathbb{S}^d$ (Schoenberg, 1942). Let P_k^d, $k \in \mathbb{Z}_+$ be the k-th degree ultraspherical polynomial. When $d = 1$, P_k^1 is the k-th degree Chebyshev polynomial (Rivlin, 1990) and for $d > 1$, P_k^d is determined by the generating function

$$\frac{1}{(1-2zt+z^2)^{(d-1)/2}} = \sum_{k\in\mathbb{Z}_+} P_k^d(t) z^k, \;\; z \in \Delta, \, t \in [-1,1].$$

We assume that we have a sequence of *nonnegative* numbers $\{a_k : k \in \mathbb{Z}_+\}$ such that

$$\sum_{k\in\mathbb{Z}_+} a_k P_k^d(1) < \infty. \tag{19}$$

Let the function $g : [0,\pi] \to \mathbb{R}$ be given at $t \in [0,\pi]$ by the equation

$$g(t) := \sum_{k\in\mathbb{Z}_+} a_k P_k(\cos t). \tag{20}$$

The condition (19) ensures that the series in (20) converges uniformly on $[0,\pi]$, since P_k^d achieves its maximum in absolute value on the interval $[-1,1]$ at 1 (Szegö, 1959, page 166).

The geodesic distance between $x, y \in \mathbb{S}^d$ is given by

$$D_d(x,y) := \arccos(x,y)$$

and Schoenberg proved in Schoenberg (1942) that K is kernel on $\mathbb{S}^d$ if and only if it has this form

$$K(x,y) := g(D_d(x,y)), \quad x,y \in \mathbb{S}^d. \tag{21}$$

Theorem 10 *The kernel given by Equation (21) is universal on $\mathbb{S}^d$ if and only if for all $k \in \mathbb{Z}_+$, a_k is positive.*

Proof We write the kernel in (21) in the feature form. For this purpose, we recall some basic facts about *spherical harmonics* which can be found in Stein and Weiss (1971). Let $\mathcal{H}_k$ be the set of all homogeneous harmonic polynomials of total degree k on $\mathbb{R}^{d+1}$ restricted to $\mathbb{S}^d$ and set $h_k := \dim \mathcal{H}_k$. We view $\mathcal{H}_k$ as a subspace of the $L^2(\mathbb{S}^d, \omega_d)$ where ω_d is the Lebesgue measure on $\mathbb{S}^d$. Let $\{Y_j^k : j \in \mathbb{N}_{h_k}\}$ be an orthonormal basis for $\mathcal{H}_k$ and recall that $\mathcal{H}_k$ is orthogonal to $\mathcal{H}_k'$ if $k \neq k'$ (Stein and Weiss, 1971). For each $k \in \mathbb{Z}_+$, there exists a positive constant c_k such that for all $x, y \in \mathbb{S}^d$

$$P_k((x,y)) = c_k \sum_{j \in \mathbb{N}_{h_k}} Y_j^k(x) Y_j^k(y). \tag{22}$$

Therefore, by Equations (20) and (22), we have that

$$K(x,y) = \sum_{k \in \mathbb{Z}_+} a_k c_k \sum_{j \in \mathbb{N}_{h_k}} Y_j^k(x) Y_j^k(y), \;\; x,y \in \mathbb{S}^d.$$

We let $I := \{(k,j) : k \in \mathbb{Z}_+, j \in \mathbb{N}_{h_k}\}$ and introduce for each $\ell = (k,j) \in I$ the feature

$$\phi_\ell := \sqrt{a_k c_k} Y_j^k.$$

Now, if all the a_k, $k \in \mathbb{Z}_+$ are positive we conclude that $\operatorname{span}\{\phi_\ell : \ell \in I\}$ is the linear space of *all* polynomials and, in particular, is universal. However, if there exists a $m \in \mathbb{Z}_+$ such that $a_m = 0$ then $\operatorname{span}\{\phi_\ell : \ell \in I\}$ is orthogonal to $\mathcal{H}_m$ and hence is not universal. ■

4. Translation Invariant Kernels on $\mathbb{R}^d$

In the remaining part of this paper we shall focus on *translation invariant* kernels on $\mathbb{R}^d$ which have the form

$$K(x,y) = k(x-y), \quad x,y \in \mathbb{R}^d$$

for some function k which is continuous on $\mathbb{R}^d$. Recall that Bochner (1959) proved that K is a kernel if and only if there is a unique finite Borel measure μ on $\mathbb{R}^d$ such that k at any $x \in \mathbb{R}^d$ has the form

$$k(x) := \int_{\mathbb{R}^d} e^{i(x,y)} d\mu(y). \tag{23}$$

We shall study the question of the universality of the kernel K in terms of the properties of its corresponding finite Borel measure μ. We start by identifying the input space as $\mathcal{X} := \mathbb{R}^d$ and then introduce our Hilbert space $\mathcal{W}$ of all complex-valued functions on $\operatorname{supp}(\mu)$ with inner product

$$(f,g)_{\mathcal{W}} := \int_{\mathrm{supp}(\mu)} f(x)\overline{g(x)}d\mu(x).$$

Next, we introduce the feature map $\Phi : \mathbb{R}^d \to \mathcal{W}$ which is defined by setting for each $x, y \in \mathbb{R}^d$

$$\Phi(x)(y) := e^{i(x,y)} \tag{24}$$

so that $K(x,y) = (\Phi(x), \Phi(y))_{\mathcal{W}}$, $x, y \in \mathbb{R}^d$. Note that in this case the Hilbert space $\mathcal{W}$ is the usual $L^2(\mathrm{supp}(\mu), \mu)$ space of all square integrable complex-valued functions relative to the measure μ on $\mathrm{supp}(\mu)$. Since μ is a finite Borel measure every bounded continuous function on $\mathrm{supp}(\mu)$ is contained in $\mathcal{W}$.

Next, we introduce the set of exponentials

$$\mathcal{E}(\mu) := \{\Phi(x) : x \in \mathrm{supp}(\mu)\}.$$

We say $\mathcal{E}(\mu)$ is universal provided that $\mathcal{E}(\mu)$ is dense in $C(\mathcal{Z})$ for every compact subset $\mathcal{Z}$ of $\mathbb{R}^d$.

Lemma 11 *For each compact subset $\mathcal{Z}$ of $\mathbb{R}^d$, $K(\mathcal{Z}) = C(\mathcal{Z})$ if and only if $\overline{\mathrm{span}}\,\mathcal{E}(\mu) = C(\mathcal{Z})$.*

Proof We observe by Fubini's theorem that the map $U : B(\mathcal{Z}) \to \mathcal{W}$ corresponding to Φ in Equation (24) is identified by formula (7) for each $\nu \in B(\mathcal{Z})$ and $y \in \mathrm{supp}(\mu)$ to be

$$U(\nu)(y) := \int_{\mathcal{Z}} e^{i(x,y)} d\nu(x).$$

Hence, we see that $\mathcal{N}(U) = \mathcal{E}(\mu)^{\perp}$ and so U is injective if and only if $\mathrm{span}\,\mathcal{E}(\mu)$ is dense in $C(\mathcal{Z})$. Therefore, the result follows from Proposition 1. ■

As a consequence of Lemma 11, we find that the universality of K depends on the density of the set $\mathcal{E}(\mu)$ of complex exponentials.

Theorem 12 *The translation kernel K is universal if and only if the set of exponential features $\mathcal{E}(\mu)$ is universal.*

An interesting feature of this result is that whenever a translation kernel K is universal with corresponding measure μ then *any* kernel corresponding to *any* other measure ρ with the *same* support as μ is also universal! Another consequence of this result pertains to the integral operator T defined by Equation (4). Specifically, let $\mathcal{Z}$ be a compact subset of $\mathbb{R}^d$, ν a finite Borel measure on $\mathcal{Z}$ such that $\mathrm{supp}(\nu) = \mathcal{Z}$ and T the integral operator defined by (4) then Theorem 6 and Lemma 11 yield the following result.

Corollary 13 *If K is a translation kernel on $\mathbb{R}^d$ then $\overline{\mathcal{R}(T)} = C(\mathcal{Z})$ if and only if $\overline{\mathrm{span}}\,\mathcal{E}(\mu) = C(\mathcal{Z})$.*

We now turn our attention to describing various conditions on the support of the measure μ which ensures the corresponding set of exponential features $\mathcal{E}$ is universal. To this end, we say, as in Micchelli et al. (2003), that a subset $\mathcal{S}$ of $\mathbb{C}^d$ is a *uniqueness* set if an *entire function* on $\mathbb{C}^d$ vanishes on $\mathcal{S}$ then it is everywhere zero on $\mathbb{C}^d$. We recall the following result from Micchelli et al. (2003).

Proposition 14 *If* $\mathrm{supp}\,(\mu)$ *is a uniqueness subset of* $\mathbb{C}^d$ *then the translation kernel K is universal.*

Proof By Theorem 12, it suffices to show that for each compact set $\mathcal{Z}$ of $\mathbb{R}^d$ there does not exist nontrivial $\nu \in B(\mathcal{Z})$ satisfying for each $y \in \mathrm{supp}\,(\mu)$ that

$$\int_{\mathcal{Z}} e^{i(x,y)} d\nu(x) = 0. \tag{25}$$

Suppose there exists $\nu \in B(\mathcal{Z})$ that satisfies (25) for all $y \in \mathrm{supp}\,(\mu)$. Then the entire function F defined for each $z \in \mathbb{C}^d$ as

$$F(z) := \int_{\mathcal{Z}} e^{i(z,x)} d\nu(x)$$

vanishes on $\mathrm{supp}\,(\mu)$. Consequently, F must be everywhere zero and so $\nu = 0$. ■

We note here that the proof above adapts to show that for each finite Borel measure ω on $\mathcal{Z}$ and $p \in [1,\infty)$, $K(\mathcal{Z}) = L^p(\mathcal{Z},\omega)$ if and only if $\overline{\mathrm{span}}\,\mathcal{E}(\mu) = L^p(\mathcal{Z},\omega)$. This fact, together with the remark at the end of Section 2, implies that $\overline{\mathcal{R}(T)} = L^p(\mathcal{Z},\omega)$ if and only if $\overline{\mathrm{span}}\,\mathcal{E}(\mu) = L^p(\mathcal{Z},\omega)$.

Proposition 15 *If* $\mathrm{supp}\,(\mu)$ *has positive Lebesgue measure on* $\mathbb{R}^d$ *then the translation kernel K is universal.*

Proof By Proposition 14, we suffice to point out the well-known fact that the real zeros of any nontrivial entire function on $\mathbb{C}^d$ form a set of Lebesgue measure zero on $\mathbb{R}^d$. ■

By Proposition 15, the uniqueness condition is satisfied by a large class of finite Borel measures on $\mathbb{R}^d$. To elaborate on this point further, we apply the Lebesgue decomposition theorem to μ and write it uniquely as

$$\mu = \mu_c + \mu_s$$

where μ_c is the continuous part of μ (Royden, 1988), in other words, there is a nonnegative function $g \in L^1(\mathbb{R}^d)$, such that for all Borel sets $S \subseteq \mathbb{R}^d$

$$\mu_c(S) = \int_S g(x)dx \tag{26}$$

and μ_s is the singular part of μ so that the Lebesgue measure of its support is zero. Our next result makes use of this decomposition.

Proposition 16 *If the continuous part of* μ *in its Lebesgue decomposition is nonzero then the translation kernel K is universal.*

Proof We only need to show that $\mathrm{supp}\,(\mu)$ has positive Lebesgue measure if the continuous part μ_c of μ in its Lebesgue decomposition is nonzero. Let g be the nonnegative function in $L^1(\mathbb{R}^d)$ that determines μ_c by (26). The hypothesis that $\mu_c \neq 0$ implies $g \neq 0$. Since $\mathrm{supp}\,(\mu_c) = \mathrm{supp}\,(g) \subseteq \mathrm{supp}\,(\mu)$ it follows that $\mathrm{supp}\,(\mu)$ has positive Lebesgue measure. ■

We shall now turn our attention to the Schoenberg kernels on $\mathbb{R}^d \times \mathbb{R}^d$ (Schoenberg, 1938). A continuous function $g : \mathbb{R}_+ \to \mathbb{R}$ determines a *radial* kernel on $\mathbb{R}^d \times \mathbb{R}^d$ by the formula

$$K(x,y) := g(\|x-y\|^2), \;\; x,y \in \mathbb{R}^d \tag{27}$$

where $\|x\| := \sqrt{(x,x)}$ is the usual euclidean norm of $x \in \mathbb{R}^d$. It was proved in Schoenberg (1938) that K is a kernel on $\mathbb{R}^d \times \mathbb{R}^d$ for *all* $d \in \mathbb{N}$ if and only if there exists a finite Borel measure μ on $\mathbb{R}_+$ such that for all $t \in \mathbb{R}_+$

$$g(t) := \int_{\mathbb{R}_+} e^{-t\sigma} d\mu(\sigma). \tag{28}$$

All kernels of this type are *not* universal. Indeed, the choice of a measure concentrated only at $\sigma = 0$ gives a kernel K that is identically constant and therefore it is not universal. This is the only exceptional case as we shall explain in the next result.

Theorem 17 *If the measure μ in Equation (28) is not concentrated at zero then the radial kernel K in (27) is universal.*

Proof We first show how to prove the result using Proposition 16 when the measure μ has the additional property that for some $a > 0$ its support is contained in the ray $[a,\infty)$. In that case, we use the formula

$$e^{-\sigma\|x\|^2} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\frac{\pi}{\sigma}\right)^{d/2} e^{i(x,\xi)} e^{-\frac{\|\xi\|^2}{4\sigma}} d\xi$$

valid for all $x \in \mathbb{R}^d$ and $\sigma > 0$. Using Fubini's theorem, we express the function k in (23) for the kernel K in (27) at $x \in \mathbb{R}^d$ as

$$k(x) := \int_{\mathbb{R}^d} e^{i(x,y)} f(y) dy$$

where the function f is defined at $y \in \mathbb{R}^d$ by the equation

$$f(y) := \frac{1}{(2\pi)^d} \int_a^\infty \left(\frac{\pi}{\sigma}\right)^{d/2} e^{-\frac{\|y\|^2}{4\sigma}} d\mu(\sigma).$$

The function f is strictly positive and

$$\int_{\mathbb{R}^d} f(y) dy = \int_a^\infty d\mu(\sigma) > 0,$$

so the theorem is a consequence of Proposition 16. If the support of the measure is *not* a subset of the open ray $\mathbb{R}_+$ we proceed differently. A direct computation using the power series for the exponential function and the multinomial expansion yields the formula

$$K(x,y) = \sum_{\alpha \in \mathbb{Z}_+^d} \binom{|\alpha|}{\alpha} \frac{2^{|\alpha|}}{|\alpha|!} \int_{\mathrm{supp}\,(\mu)} \sigma^{|\alpha|} x^\alpha e^{-\sigma\|x\|^2} y^\alpha e^{-\sigma\|y\|^2} d\mu(\sigma).$$

This suggests that we introduce the Hilbert space $\mathcal{W}$ of real-valued functions on the set $\mathbb{M} := \mathbb{Z}_+^d \times \mathrm{supp}\,(\mu)$ with inner product

$$(F,G)_{\mathcal{W}} := \sum_{\alpha \in \mathbb{Z}_+^d} \binom{|\alpha|}{\alpha} \frac{2^{|\alpha|}}{|\alpha|!} \int_{\mathrm{supp}\,(\mu)} F(\alpha,\sigma) G(\alpha,\sigma) d\mu(\sigma)$$

and a feature map $\Phi : \mathbb{R}^d \to \mathcal{W}$ defined at $(\alpha,\sigma) \in \mathbb{M}$ and $x \in \mathbb{R}^d$ as

$$\Phi(x)(\alpha,\sigma) := \sigma^{|\alpha|/2} x^\alpha e^{-\sigma\|x\|^2}.$$

Hence we have the feature space representation for the kernel K given for each $x,y \in \mathbb{R}^d$ as

$$K(x,y) = (\Phi(x),\Phi(y))_{\mathcal{W}}.$$

Now, we let $\mathcal{Z}$ be some prescribed compact subset of $\mathbb{R}^d$. As in the proof of Theorem 11 we identify the operator $U : B(\mathcal{Z}) \to \mathcal{W}$ in (7) at $\nu \in B(\mathcal{Z})$ and $(\alpha,\sigma) \in \mathbb{M}$ as

$$U(\nu)(\alpha,\sigma) = \int_{\mathcal{Z}} \sigma^{|\alpha|/2} x^\alpha e^{-\sigma\|x\|^2} d\nu(x).$$

Therefore, if there is a *positive* $\rho \in \operatorname{supp}(\mu)$ and $\nu \in \mathcal{N}(U)$ then for any $\alpha \in \mathbb{Z}_+^d$ we have that

$$\int_{\mathcal{Z}} x^\alpha e^{-\rho\|x\|^2} d\nu(x) = 0.$$

This implies, by the density of all polynomials in $C(\mathcal{Z})$, that $\nu = 0$. In other words, U is injective and so the result follows from Proposition 1. ■

As a consequence of Theorem 17 we conclude that the following two classes of kernels are universal:

$$K(x,y) := e^{-\alpha\|x-y\|^2}, \;\; x,y \in \mathbb{R}^d$$

and

$$K(x,y) := (\beta + \|x-y\|^2)^{-\alpha}, \;\; x,y \in \mathbb{R}^d$$

where α and β are arbitrary positive numbers.

Next, we give a quite different condition on the support of the measure μ so that the corresponding translation kernel is universal. For this discussion we shall use the celebrated Stone-Weierstrass theorem (Rudin, 1991).

Proposition 18 *If* $\operatorname{supp}(\mu)$ *is a subgroup of* $\mathbb{R}^d$ *such that for each* $x \in R^d \setminus \{0\}$ *the set* $\{(x,y) : y \in \operatorname{supp}(\mu)\} \not\subseteq \mathbb{Z}$ *then K is universal.*

Proof By Theorem 12, it suffices to show that $\operatorname{span}\mathcal{E}(\mu)$ is dense in $C(\mathcal{Z})$. Suppose all the hypotheses are satisfied and there exists some compact set $\mathcal{Z} \subseteq \mathbb{R}^d$ such that $K(\mathcal{Z})$ is not dense in $C(\mathcal{Z})$. Since $\operatorname{supp}(\mu)$ is a subgroup of $\mathbb{R}^d$ we see that $1 \in \operatorname{span}\mathcal{E}(\mu)$ and that for all $f,g \in \operatorname{span}\mathcal{E}(\mu)$, both fg and $\overline{f}$ belong to $\operatorname{span}\mathcal{E}(\mu)$. Therefore, by the Stone-Weierstrass theorem, there exist distinct points $x_1,x_2 \in \mathcal{Z}$ such that for all $f \in \operatorname{span}\mathcal{E}(\mu)$, $f(x_1) = f(x_2)$. That is, for each $y \in \operatorname{supp}(\mu)$ $e^{i(x_1-x_2,y)} = 1$ or in other words, $(x_1-x_2,y)/2\pi \in \mathbb{Z}$. This contradiction proves the proposition. ■

Corollary 19 *If* $d = 1$, $\operatorname{supp}(\mu)$ *is a subgroup of* $\mathbb{R}$ *and there exists* $y_1,y_2 \in \operatorname{supp}(\mu) \setminus \{0\}$ *such that* y_1/y_2 *is an irrational number then K is universal.*

Proof By the hypotheses of the corollary, it is clear that there does not exist $x \in \mathbb{R} \setminus \{0\}$ such that both xy_1 and xy_2 are integers. The result follows immediately from Proposition 18. ■

5. Conclusion

We have provided a variety of conditions for a kernel to be universal in terms of properties of its features. Several examples of universal dot product kernels are given. In the case of translation kernels we showed that universality depends on the density of a set of complex exponentials. This problem has attracted much interest in the literature. An extensive survey of existing results on the univariate case is given in Redheffer (1977) and additional information in Beurling and Malliavin (1967). With this available information a complete characterization of univariate translation kernels follows. We show that except in rare circumstances all Schoenberg radial kernels are universal.

Our study indicates that there is intimate relationship between uniformly approximating a prescribed target function by a kernel and approximating by its features. There is an important problem which is not treated here that deserves careful attention. Given a prescribed error $\varepsilon > 0$ and a prescribed target function f, what is the relationship between the number of features needed to represent f with error ε and the number of kernel sections needed for the same purpose. We intend to address this issue on another occasion.

Acknowledgments

The authors are indebted to one of the referees for bringing to their attention the relationship of Corollaries 3.2 and 3.3 with results in references Steinwart (2001) and Zhou (2003). This work was supported by the US National Science of Foundation under grant CCR-0407476, by the Natural Science Foundation of China under grant 10371122 and by the Ministry of Education of the People's Republic of China under the Changjiang Scholar Chair Professorship program. The corresponding author is Yuesheng Xu. He is also with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P. R. China.

References

A. Argyriou, C. A. Micchelli and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. *Proceeding of the 18th Annual Conference on Learning Theory (COLT'05)*, Bertinoro, Italy, 2005.

A. Argyriou, R. Hauser, C. A. Micchelli and M. Pontil. A DC-programming algorithm for kernel selection. *Proceeding of the 23rd International Conference on Machine Learning (ICML'06)*, forthcoming (see also Research Note RN/06/04, Department of Computer Science, UCL, 2006).

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68: 337–404, 1950.

F. R. Bach, G. R. G. Lanckriet and M. I. Jordan. Multiple kernel learning, conic duality and the SMO algorithm. *Proceeding of the 21st International Conference on Machine learning (ICML'04)*, 2004.

A. Beurling and P. Malliavin. On the closure of characters and the zeros of entire functions. *Acta. Math.*, 118: 79–93, 1967.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

S. Bochner. *Lectures on Fourier Integrals With an author's supplement on monotonic functions, Stieltjes integrals, and harmonic analysis*. Annals of Mathematics Studies, no. 42, Princeton University Press, New Jersey, 1959.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13: 1–50, 2000.

C. H. FitzGerald, C. A. Micchelli and A. Pinkus. Functions that preserve families of positive semidefinite matrices. *Linear Algebra Appl.*, 221: 83–102, 1995.

T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El-Ghaoui and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5: 27–72, 2004.

P. Lax. *Functional Analysis*. Wiley, New York, 2002.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Royal Soc. London*, 209: 415–446, 1909.

C. A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. *Proceeding of the 17th Annual Conference on Learning (COLT'04)*, 2004.

C. A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel. *Machine Learning*, forthcoming (see also: Research Note RN/05/11, Department of Computer Science, UCL, June, 2005).

C. A. Micchelli, M. Pontil, Q. Wu and D. X. Zhou. Error bounds for learning the kernel. Research Note RN/05/04, Department of Computer Science, UCL, 2006.

C. A. Micchelli, Y. Xu and P. Ye. Cucker Smale learning theory in Besov spaces. *Advances in Learning Theory: Methods, Models and Applications*. J. Suykens, G. Horvath, S. Basu, C. A. Micchelli and J. Vandewalle, editors. IOS Press, Amsterdam, The Netherlands, 2003, 47–68.

J. Neumann, C. Schnörr and G. Steidl. SVM-based feature selection by direct objective minimization. C.E. Rasmussen, H. H. Bülthoff, B. Schölkopf and M. A. Giese, editors. Lecture Notes in Computer Science, 3175: 212–219, *Proceeding of the 26th DAGM Symposium*, 2004.

C. S. Ong, A. J. Smola and R. C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6: 1043–1071, 2005.

T. Poggio, S. Mukherjee, R. Rifkin, A. Raklin and A. Verri. B. *Uncertainty in geometric computations*, J. Winkler and M. Niranjan, editors. Kluwer Academic Publishers, 22: 131–141, 2002.

R. M. Redheffer. Completeness of sets of complex exponentials. *Adv. Math.*, 24: 1–62, 1977.

T. J. Rivlin. *Chebyshev Polynomials*. 2nd Edition, John Wiley, New York, 1990.

H. Royden. *Real Analysis*. 3rd Edition, Macmillan Publishing Company, New York, 1988.

W. Rudin. *Functional Analysis*. 2nd Edition, McGraw Hill, New York, 1991.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math. (2)*, 39: 811–841, 1938.

I. J. Schoenberg. Positive definite functions on spheres. *Duke. Math. J.*, 9: 96–108, 1942.

B. Schölkopf, C. J. C. Burges and A. Smola. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, Mass, 1999.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, Mass, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

S. Sonnenburg, G. Rätsch and C. Schäfer. A general and efficient multiple kernel learning algorithm. Y. Weiss, B. Schölkopf and J. Platt, editors. *Advances in Neural Information Processing Systems*, 18. MIT Press, Cambridge, Mass, 2006.

E. M. Stein and G. Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, New Jersey, 1971.

I. Steinwart. On the influence of kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2: 67–93, 2001.

H. Sun. Mercer theorem for RKHS on noncompact sets. *J. Complexity*, 21: 337–349, 2005.

G. Szegö. *Orthogonal Polynomials*. American Mathematical Society Colloquium Publications 23. Revised Edition, Providence, RI, 1959.

G. Wahba. *Splines Models for Observational Data*. Series in Applied Mathematics 59. SIAM, Philadelphia, 1990.

D. X. Zhou. Density problem and approximation error in learning theory. *preprint*, 2003.