On the relation between universality, characteristic kernels and RKHS embedding of measures

Bharath K. Sriperumbudur
Dept. of ECE, UC San Diego
La Jolla, USA.
bharathsv@ucsd.edu

Kenji Fukumizu
The Institute of Statistical
Mathematics, Tokyo, Japan.
fukumizu@ism.ac.jp

Gert R. G. Lanckriet
Dept. of ECE, UC San Diego
La Jolla, USA.
gert@ece.ucsd.edu

Abstract

Universal kernels have been shown to play an important role in the achievability of the *Bayes risk* by many kernel-based algorithms that include binary classification, regression, etc. In this paper, we propose a notion of universality that *generalizes* the notions introduced by Steinwart and Micchelli et al. and study the necessary and sufficient conditions for a kernel to be universal. We show that all these notions of universality are closely linked to the injective embedding of a certain class of Borel measures into a reproducing kernel Hilbert space (RKHS). By exploiting this relation between universality and the embedding of Borel measures into an RKHS, we establish the relation between universal and *characteristic* kernels. The latter have been proposed in the context of the RKHS embedding of probability measures, used in statistical applications like homogeneity testing, independence testing, etc.

1 INTRODUCTION

Kernel methods have been popular in machine learning and pattern analysis for their superior performance on a wide spectrum of learning tasks. They are broadly established as an easy way of constructing nonlinear algorithms from linear ones, by embedding points into higher dimensional reproducing kernel Hilbert spaces (RKHSs) (Schölkopf & Smola, 2002). In the regularization approach to learning, these algorithms generally invoke the representer theorem and learn a function in a RKHS that has the representation,

$$f := \sum_{j\in\mathbb{N}_n} c_j k(\cdot, x_j), \tag{1}$$

where $\mathbb{N}_n := \{1, 2, \ldots, n\}$ and $k : X \times X \to \mathbb{R}$ is a positive definite (pd) kernel on some arbitrary space, X. $\{c_j : j \in \mathbb{N}_n\} \subset \mathbb{R}$ are parameters typically obtained from training data, $\{x_j : j \in \mathbb{N}_n\} \subset X$. As noted in Micchelli et al. (2006), one can ask whether the function representation in (1) approximates any real-valued target function arbitrarily *well* as the number of summands increases without bound. This is an important question to consider because if the answer is affirmative, then the kernel-based learning algorithm is *consistent* in the sense that for any target function, $f^\star$, the discrepancy between f (which is learned from the training data) and $f^\star$ goes to zero (in some sense) as the sample size goes to infinity. Since

$$\Big\{\sum_{j\in\mathbb{N}_n} c_j k(\cdot, x_j) : n \in \mathbb{N}, \{c_j\} \subset \mathbb{R}, \{x_j\} \subset X\Big\}$$

is dense in the RKHS, $\mathcal{H}$ associated with k (Aronszajn, 1950), and assuming that the kernel-based algorithm makes f "converge to an appropriate function" in $\mathcal{H}$ as $n \to \infty$, the above question is equivalent to the question of whether $\mathcal{H}$ is rich enough to approximate any $f^\star$ arbitrarily *well*. This paper deals with the characterization of such RKHSs, where the *wellness* of approximation is measured in terms of the *uniform norm*. Below, we first present the prior work that dealt with the above approximation problem and then briefly discuss our contribution, which generalizes these earlier works.

Steinwart (2001) considered the above approximation problem when X is a compact metric space and defined a continuous kernel, k as *universal* (in this paper, we refer to it as *c-universal*) if its associated RKHS, $\mathcal{H}$ is dense in the Banach space, $C(X)$ of real-valued continuous functions (on X) w.r.t. the uniform norm, i.e., for any $f^\star \in C(X)$, there exists a $g \in \mathcal{H}$ that uniformly approximates $f^\star$. In the context of learning,

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

this indicates that if a kernel is *c-universal*, then the corresponding kernel-based learning algorithm could be consistent in the sense that any target function, $f^\star \in C(X)$ could be approximated arbitrarily well in the uniform norm by f in (1) as $n \to \infty$ (see Corollary 5.29 in Steinwart and Christmann (2008) for a rigorous result). By applying the Stone-Weierstraß theorem, Steinwart (2001) then provided sufficient conditions for a kernel to be *c-universal*. However, one limitation in the setup considered by Steinwart (2001) is that X is assumed to be compact, which excludes many interesting spaces, such as $\mathbb{R}^d$ and *infinite* discrete sets.

To overcome the limitation of compact X in *c-universality*, Micchelli et al. (2006) proposed a notion of universality, wherein a continuous k is said to be *universal* (in this paper, we refer to it as *cc-universal*), if for any choice of compact set Z of a Hausdorff topological space, X, the set $K(Z) := \overline{\text{span}}\{k(\cdot, y) : y \in Z\}$ is dense in $C(Z)$ in the uniform norm. It can be shown that k is *cc-universal* if and only if for any compact set $Z \subset X$, for any $f^\star \in C(Z)$, there exists a $g \in \mathcal{H}_{|Z}$ that uniformly approximates $f^\star$, i.e., $\mathcal{H}$ is dense in $C(X)$ with the *topology of compact convergence.* Here, $\mathcal{H}_{|Z} := \{f_{|Z} : f \in \mathcal{H}\}$ is the restriction of $\mathcal{H}$ to Z and $f_{|Z}$ is the restriction of f to Z. Although *cc-universality* can handle non-compact domains unlike *c-universality*, the topology of compact convergence is weaker than the topology of *uniform convergence*, i.e., a sequence of functions, $\{f_n\} \subset C(X)$ converging to $f \in C(X)$ in the topology of uniform convergence ensure that they converge in the topology of compact convergence but not vice-versa. So, the natural question to ask is whether we can characterize $\mathcal{H}$ that are rich enough to approximate any $f^\star$ on non-compact X in a stronger sense, i.e., uniformly, by some $g \in \mathcal{H}$.

To answer this question, we propose a notion of universality that can handle non-compact X while uniformly approximating any $f^\star$ by some $g \in \mathcal{H}$. However, instead of approximating any $f^\star \in C(X)$ for non-compact X, as is the case with *cc-universality*, in this notion, only a subset of $C(X)$ (defined below) is approximated. Note that this notion addresses limitations associated with both *c-* and *cc-universality.* To formalize this, we define k to be c_0*-universal* if k is bounded, $k(\cdot, x) \in C_0(X)$, $\forall\, x \in X$ and its corresponding RKHS, $\mathcal{H}$ is dense in $C_0(X)$ w.r.t. the uniform norm, where X is a locally compact Hausdorff (LCH) space (also see Carmeli et al. (2009)) and $C_0(X)$ is the Banach space of bounded continuous functions vanishing at infinity, endowed with the uniform norm (see Section 2 for the definition of $C_0(X)$). The necessary and sufficient condition for a kernel to be c_0*-universal* is derived in Section 3 (see Theorem 3), and is shown to be related to the injective embedding of a *certain class* of Borel measures into $\mathcal{H}$. Using this result, simple necessary and sufficient conditions are derived for translation invariant kernels on $\mathbb{R}^d$, Fourier kernels on the d-Torus, $\mathbb{T}^d$, and radial kernels on $\mathbb{R}^d$ to be c_0*-universal.* Examples of c_0*-universal* kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, B_{2l+1}-spline, inverse multiquadratics, Matérn class, etc. In addition, by providing a novel characterization for *c-* and *cc-universality*, which is also related to the injective embedding of *certain* class of Borel measures into $\mathcal{H}$, in Sections 3.1 and 3.2, we relate *c-* and *cc-universality* to c_0*-universality.* We show that all these three notions of equivalent when X is compact (which also trivially follows from their definitions), while c_0*-universality* is stronger than *cc-universality* when X is not compact, i.e., if a kernel is c_0*-universal*, then it is *cc-universal* but not vice-versa.

Recently, the RKHS embedding of probability measures,

$$\mathbb{P} \mapsto \int_X k(\cdot, x)\, d\mathbb{P}(x), \tag{2}$$

has been studied (Sriperumbudur et al., 2008; Fukumizu et al., 2009b; Sriperumbudur et al., 2009) and has been used in many statistical applications like homogeneity testing (Gretton et al., 2007), independence testing (Gretton et al., 2008; Fukumizu et al., 2008), dimensionality reduction (Fukumizu et al., 2009a), etc. Here, X is a topological space, k is a measurable, bounded kernel and $\mathbb{P}$ is a Borel probability measure on X. The motivation to consider such an embedding is that it provides a powerful and straightforward method of dealing with higher-order statistics of random variables. In all the above mentioned applications, it is critical that the embedding in (2) is injective so that probability measures can be distinguished by their images in $\mathcal{H}$. A bounded, measurable k is said to be *characteristic* if and only if (2) is injective. Gretton et al. (2007) related characteristic and universal kernels by showing that if k is *c-universal*, then it is *characteristic.* Besides this result, not much is known or understood about the relation between universal and characteristic kernels. In Section 4, we relate universality and *characteristic* kernels by using the results in Section 3 that relate universality and the RKHS embedding of measures. In particular, we show that a translation invariant kernel on $\mathbb{R}^d$ (in general, any locally compact Abelian group) is c_0*-universal* if and only if it is *characteristic.* We also show that the converse to the result by Gretton et al. (2007) is not true, i.e., if a kernel is *characteristic*, it need not be *c-universal.*

In Section 5, we briefly discuss a generalization of c_0*-universality* and issues associated with it. The appendix contains a supplementary result used in proofs.

To summarize, we have proposed a stronger notion of universality that generalizes the earlier notions (Steinwart, 2001; Micchelli et al., 2006) and presented a unified approach to understand these notions by relating them to the RKHS embedding of measures. By exploiting this connection between universal kernels and the RKHS embedding of measures, we also clarified the relationship between universal and characteristic kernels.

2 DEFINITIONS & NOTATION

Let X be a topological space. $C(X)$ (*resp.* $C_b(X)$) denotes the space of all continuous (*resp.* bounded continuous) functions on X. For an LCH space, X, $f \in C(X)$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \geq \epsilon\}$ is compact. The class of all continuous f on X which vanish at infinity is denoted as $C_0(X)$. The spaces $C_b(X)$ and $C_0(X)$ are endowed with the uniform norm, $\|\cdot\|_u$ defined as $\|f\|_u := \sup_{x\in X}|f(x)|$ for $f \in C_0(X) \subset C_b(X)$.

If X denotes a topological vector space, we denote by X' the vector space of continuous linear functionals on X, and X' is called the *topological dual space* (in this paper, we simply refer to it as the *dual*).

Radon measures: A *Radon measure* μ on a Hausdorff space X is a Borel measure on X satisfying (i) $\mu(C) < \infty$ for each compact subset $C \subset X$ and (ii) $\mu(B) = \sup\{\mu(C)|C \subset B, C\,\text{compact}\}$ for each B in the Borel σ-algebra of X. μ is said to be finite if $\|\mu\| := |\mu|(X) < \infty$, where $|\mu|$ is the total-variation of μ. $M_b(X)$ denotes the space of all finite signed Radon measures on X, while $M^1_+(X)$ denotes the space of all Radon probability measures. $M_{bc}(X)$ denotes the space of all compactly supported finite signed Radon measures on X. For $\mu \in M_b(X)$, the support of μ is defined as $\text{supp}(\mu) = \{x \in X\,|\,\text{for any open set } U \text{ such that } x \in U, |\mu|(U) \neq 0\}$. We refer the reader to Berg et al., (1984, Chapter 2) for a general reference on the theory of Radon measures.

Positive definite and strictly positive definite: A function $k : X \times X \to \mathbb{R}$ is called *positive definite* (pd) if, for all $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and all $x_1, \dots, x_n \in X$, we have

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

Furthermore, k is said to be *strictly pd* if, for mutually distinct $x_1, \dots, x_n \in X$, equality in (3) only holds for $\alpha_1 = \cdots = \alpha_n = 0$. ψ is said to be a positive definite function on $\mathbb{R}^d$ if $k(x, y) = \psi(x-y)$ is positive definite.

Fourier transform in $\mathbb{R}^d$: For $X \subset \mathbb{R}^d$, let $L^p(X)$ denote the Banach space of p-power ($p \geq 1$) integrable functions w.r.t. the Lebesgue measure. For $f \in L^1(\mathbb{R}^d)$, $\hat{f}$ represents the Fourier transform of f given by

$$\hat{f}(y) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-iy^Tx} f(x)\, dx,\ y \in \mathbb{R}^d,$$

where i denotes the imaginary unit $\sqrt{-1}$. For a finite Borel measure, μ on $\mathbb{R}^d$, the Fourier transform of μ is given by

$$\hat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^Tx}\, d\mu(x),\ \omega \in \mathbb{R}^d,$$

which is a bounded, uniformly continuous function on $\mathbb{R}^d$.

3 UNIVERSAL KERNELS

In Section 1, we have briefly discussed the notions of *c-* and *cc-universality* along with some of their limitations. To address these limitations, in this paper, we consider the following notion of universality.

Definition 1 (c_0-universal)**.** *Let X be an LCH space with the kernel, k being bounded and $k(\cdot, x) \in C_0(X), \forall\, x \in X$. k is said to be c_0-universal if the RKHS, $\mathcal{H}$ induced by k is dense in $C_0(X)$ w.r.t. the uniform norm, i.e., for every function $g \in C_0(X)$ and all $\epsilon > 0$, there exists an $f \in \mathcal{H}$ such that $\|f-g\|_u \leq \epsilon$.*

Note that the above definition of universality can handle non-compact X and ensures uniform convergence over entire X, therefore removing the limitations associated with *c-universality* and *cc-universality.* Since $C_0(X) = C(X)$ when X is compact, it is easy to see that the notions of *c-universal*, *cc-universal* and c_0-*universal* are the same. However, when X is not compact, it is not straightforward to see how the notions of c_0-*universal* and *cc-universal* are related. Before we discuss this relation (see Section 3.2), we are primarily interested in the characterization of c_0-*universal* kernels. To obtain a characterization for c_0-*universal* kernels, we need the following result, usually called under the name of Hahn-Banach theorem, which we quote from Rudin, (1991, Theorem 3.5).

Theorem 2 (Hahn-Banach)**.** *Suppose A be a subspace of a locally convex topological vector space Y. Then A is dense in Y if and only if $A^\perp = \{0\}$, where*

$$A^\perp := \{T \in Y' : \forall x \in A,\ T(x) = 0\}.$$

The main results in this paper hinge on the above theorem, where we choose A to be the RKHS, $\mathcal{H}$ and Y to be $C_0(X)$ or $C(X)$ (depending on the notion of universality) for which Y' is known through the Riesz representation theorem. Using Theorem 2 with $A := \mathcal{H}$ and $Y := C_0(X)$, the following result presents a necessary and sufficient condition for k to be c_0-*universal.*

Theorem 3 (Characterization of c_0-*universality*)**.** *k is c_0-universal if and only if the embedding,*

$$\mu \mapsto \int_X k(\cdot,x)\,d\mu(x),\ \mu \in M_b(X), \qquad (4)$$

is injective.

Proof. By Definition 1, k is c_0*-universal* if $\mathcal{H}$ is dense in $C_0(X)$. We now invoke Theorem 2 to characterize the denseness of $\mathcal{H}$ in $C_0(X)$, which means we need to consider the dual $C_0'(X) := (C_0(X))'$ of $C_0(X)$. By the Riesz representation theorem (Folland, 1999, Theorem 7.17), $C_0'(X) = M_b(X)$ in the sense that there is a bijective linear isometry $\mu \mapsto T_\mu$ from $M_b(X)$ onto $C_0'(X)$, given by the natural mapping, $T_\mu(f) = \int_X f\,d\mu,\ f \in C_0(X)$. Therefore, by Theorem 2, $\mathcal{H}$ is dense in $C_0(X)$ if and only if $\mathcal{H}^\perp := \{\mu \in M_b(X) : \forall\, f \in \mathcal{H},\ \int_X f\,d\mu = 0\} = \{0\}$.

$(\Leftarrow)$ If (4) is injective, i.e., for $\mu \in M_b(X)$, $\int_X k(\cdot,x)\,d\mu(x) = 0 \Rightarrow \mu = 0$, then by Lemma 20 (see Appendix), we have $\int_X f\,d\mu = \langle f, \int_X k(\cdot,x)\,d\mu(x)\rangle_{\mathcal{H}} = 0, \forall\, f \in \mathcal{H}$, which means $\mathcal{H}$ is dense in $C_0(X)$ and therefore k is c_0*-universal.*

$(\Rightarrow)$ Suppose $(\int_X k(\cdot,x)\,d\mu(x) = 0 \Rightarrow \mu = 0)$ does not hold, i.e., $\exists\,\mu \in M_b(X)$, $\mu \neq 0$ such that $\int_X k(\cdot,x)\,d\mu(x) = 0$, which means $\exists\,\mu \in M_b(X)$, $\mu \neq 0$ such that $\int_X f\,d\mu = 0$ for every $f \in \mathcal{H}$ and therefore $\mathcal{H}$ is not dense in $C_0(X)$. □

Theorem 3 shows that the c_0*-universality* of k is related to the injective embedding of $\mu \in M_b(X)$ into $\mathcal{H}$. Recently, such injective embeddings have been considered when μ is a Borel probability measure on a measurable space, X, which as mentioned in Section 1 is related to characteristic kernels. Before we relate these to c_0*-universality* (see Section 4), we obtain an alternate and equivalent characterization of c_0*-universality*, which resembles the condition for k to be strictly pd, though not equivalent (see Proposition 5).

Proposition 4. *k is c_0-universal if and only if*

$$\iint_X k(x,y)\,d\mu(x)\,d\mu(y) > 0, \forall\, 0 \neq \mu \in M_b(X). \quad (5)$$

Proof. $(\Leftarrow)$ Suppose k is not c_0*-universal.* By Theorem 3, $\exists\,\mu \in M_b(X), \mu \neq 0$ such that $\int_X k(\cdot,x)\,d\mu(x) = 0 \Rightarrow \|\int_X k(\cdot,x)\,d\mu(x)\|_{\mathcal{H}} = 0$. This means

$$\begin{aligned} 0 &= \Big\langle \int_X k(\cdot,x)\,d\mu(x), \int_X k(\cdot,x)\,d\mu(x) \Big\rangle_{\mathcal{H}} \\ &\overset{(a)}{=} \iint_X k(x,y)\,d\mu(x)\,d\mu(y), \end{aligned}$$

where (a) follows from Lemma 20 (see Appendix). By our assumption in (5), this leads to a contradiction. Therefore, if (5) holds, then k is c_0*-universal.*

$(\Rightarrow)$ Suppose $\exists\,\mu \in M_b(X), \mu \neq 0$ such that $\iint_X k(x,y)\,d\mu(x)\,d\mu(y) = 0 \Rightarrow \|\int_X k(\cdot,x)\,d\mu(x)\|_{\mathcal{H}} = 0$, which implies $\int_X k(\cdot,x)\,d\mu(x) = 0$. This means, the embedding in (4) is not injective, which by Theorem 3 implies that k is not c_0*-universal.* Therefore, if k is c_0*-universal*, then k satisfies (5). □

The following result shows that k being strictly pd is a necessary condition for k to be c_0*-universal.*

Proposition 5 (c_0*-universal* kernels are strictly pd)**.** *If k is c_0-universal, then it is strictly pd.*

Proof. Suppose k is not strictly pd. This means for some $n \in \mathbb{N}$ and for mutually distinct $x_1, \ldots, x_n \in X$, there exists $\mathbb{R} \ni \alpha_j \neq 0$ for some $j \in \{1,\ldots,n\}$ such that $\sum_{l,j=1}^n \alpha_l \alpha_j k(x_l,x_j) = 0$. Define $\mu := \sum_{j=1}^n \alpha_j \delta_{x_j}$, where δ_x represents the Dirac measure at x. Clearly $\mu \neq 0$ and $\mu \in M_b(X)$. Therefore, there exists $0 \neq \mu \in M_b(X)$ such that $\iint_X k(x,y)\,d\mu(x)\,d\mu(y) = 0$, which by Proposition 4 implies k is not c_0*-universal.* □

Remark 6. *By combining Propositions 4 and 5, it is easy see that if k satisfies (5), then k is strictly pd. However, the converse is not true. See Steinwart and Christmann, (2008, Proposition 4.60, Theorem 4.62) for the related discussion.*

Although the condition in (5) for c_0*-universality* is easy to interpret, it is not always easy to check. To this end, we present easily checkable characterizations for the following classes of kernels:

(A_1) k is translation invariant on $\mathbb{R}^d \times \mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, where $0 \neq \psi \in C_b(\mathbb{R}^d)$ is a pd function on $\mathbb{R}^d$.

(A_2) k is a radial kernel on $\mathbb{R}^d \times \mathbb{R}^d$, i.e., $k(x,y) = \varphi(\|x-y\|_2^2)$, $x,y \in \mathbb{R}^d$, where $\varphi \in C_0(\mathbb{R})$ is *completely monotone* (Wendland, 2005, Chapter 7) on $[0,\infty)$.

(A_3) X is an LCH space with bounded k. Let $k(x,y) = \sum_{j\in I} \phi_j(x)\phi_j(y)$, $(x,y) \in X \times X$, where we assume that series converges uniformly on $X \times X$. $\{\phi_j : j \in I\}$ is a set of continuous real-valued functions on X where I is a countable index set.

Translation invariant kernels on $\mathbb{R}^d$: (A_1)

The following result provides a necessary and sufficient condition (which is easily checkable) for k to be c_0*-universal* when k is translation invariant, i.e., when it satisfies (A_1). Before we present the result, we need a theorem due to Bochner that characterizes translation invariant kernels on $\mathbb{R}^d$, which is quoted from Wendland, (2005, Theorem 6.6).

Theorem 7 (Bochner)**.** *$\psi \in C_b(\mathbb{R}^d)$ is pd on $\mathbb{R}^d$ if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$, i.e.,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\,d\Lambda(\omega),\ x \in \mathbb{R}^d. \qquad (6)$$

Proposition 8 (Translation invariant kernels on $\mathbb{R}^d$)**.** *Suppose (A_1) holds and $\psi \in C_0(\mathbb{R}^d)$. Then k is c_0-universal if and only if* $\text{supp}(\Lambda) = \mathbb{R}^d$.

Proof. $(\Leftarrow)$ Consider $B := \iint_{\mathbb{R}^d} k(x,y)\, d\mu(x)\, d\mu(y)$ for any $0 \neq \mu \in M_b(\mathbb{R}^d)$ with $k(x,y) = \psi(x-y)$.

$$\begin{aligned} B &= \iint_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) \\ &\overset{(a)}{=} \iiint_{\mathbb{R}^d} e^{-i(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\ &\overset{(b)}{=} \iint_{\mathbb{R}^d} e^{-ix^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{iy^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\ &= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) = \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega), \quad (7) \end{aligned}$$

where Theorem 7 is invoked in (a), while Fubini's theorem (Folland, 1999, Theorem 2.37) is invoked in (b). If $\text{supp}(\Lambda) = \mathbb{R}^d$, then it is clear that $B > 0$. Therefore, by Proposition 4, k is c_0-*universal.*

$(\Rightarrow)$ Suppose k is c_0-*universal*, which by Theorem 3 means that $\mu \mapsto \int k(\cdot,x)\, d\mu(x)$ is injective for $\mu \in M_b(\mathbb{R}^d)$. This means $\mu \mapsto \int k(\cdot,x)\, d\mu(x)$ is injective for $\mu \in M^1_+(\mathbb{R}^d)$ and therefore Theorem 7 in Sriperumbudur et al. (2008) yields $\text{supp}(\Lambda) = \mathbb{R}^d$. □

The above result shows that a translation invariant kernel on $\mathbb{R}^d$ is c_0-*universal* if and only if the support of its Fourier transform is entire $\mathbb{R}^d$. Examples of c_0-*universal* translation invariant kernels on $\mathbb{R}^d$ include the Gaussian [$k(x,y) = \exp(-\alpha\|x-y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\alpha > 0$], Laplacian [$k(x,y) = \exp(-\alpha\|x-y\|_1)$, $x, y \in \mathbb{R}^d$, $\alpha > 0$], B_{2l+1}-spline [$k(x,y) = (1-|x-y|)\mathbb{1}_{[-1,1]}(x-y)$, $x, y \in \mathbb{R}$], inverse multiquadratics [$k(x,y) = (\beta + \|x-y\|_2^2)^{-\alpha}$, $x, y \in \mathbb{R}^d$, $\alpha > 0$, $\beta > 0$], etc., as it can be shown that all these kernels have Fourier transforms supported on entire $\mathbb{R}^d$ (Wendland, 2005, Chapter 6). An example of a translation invariant kernel on $\mathbb{R}$ that is not c_0-*universal* is $k(x,y) = \psi(x-y) = \frac{\sin^2((x-y)/2)}{(x-y)^2}$ (called the *sinc-squared* kernel) as $\text{supp}(\Lambda) = [-1,1] \subsetneq \mathbb{R}$. However, since the *sinc-squared* kernel is strictly pd (Wendland, 2005, Theorem 6.11), the converse to Proposition 5 is not true.

As a corollary to Proposition 8, the following result shows that all compactly supported translation invariant kernels on $\mathbb{R}^d$ are c_0-*universal.*

Corollary 9. *Suppose (A_1) holds. If* $\text{supp}(\psi)$ *is compact, then k is c_0-universal.*

Proof. The proof is same as that of Corollary 8 in Sriperumbudur et al. (2008). □

Radial kernels on $\mathbb{R}^d$: (A_2)

Proposition 11 provides a necessary and sufficient condition for k to be c_0-*universal* when k satisfies (A_2). Before that, we present Schoenberg's characterization of radial kernels, which we quote from Wendland, (2005, Corollary 7.12 and Theorem 7.13).

Theorem 10 (Schoenberg)**.** *$k(x,y) = \varphi(\|x-y\|_2^2)$ is a kernel on $\mathbb{R}^d \times \mathbb{R}^d$ if and only if there exists a finite nonnegative Borel measure, ν on $[0,\infty)$ such that for all $r > 0$,*

$$\varphi(r) = \int_0^\infty e^{-rt}\, d\nu(t). \quad (8)$$

Proposition 11 (Radial kernels on $\mathbb{R}^d$)**.** *Suppose (A_2) holds. Then k is c_0-universal if and only if* $\text{supp}(\nu) \neq \{0\}$.

Proof. $(\Leftarrow)$ Consider $B := \iint k(x,y)\, d\mu(x)\, d\mu(y)$ with $k(x,y) = \varphi(\|x-y\|_2^2)$. Using the representation for φ in (8), we have

$$\begin{aligned} B &= \iint_{\mathbb{R}^d} \int_0^\infty e^{-t\|x-y\|_2^2}\, d\nu(t)\, d\mu(x)\, d\mu(y) \\ &\overset{(a)}{=} \int_0^\infty \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2 \hat{g}_t(\omega)\, d\omega\, d\nu(t) \\ &\overset{(b)}{=} \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2 \left[\int_0^\infty \hat{g}_t(\omega)\, d\nu(t)\right] d\omega, \quad (9) \end{aligned}$$

where $\hat{g}_t(\omega) := (2t)^{-d/2} e^{-\|\omega\|_2^2/4t}$ is the Fourier transform of $e^{-t\|x\|_2^2}$. Here Fubini's theorem and (7) is invoked in (a) while Fubini's theorem is invoked in (b). Since $\text{supp}(\nu) \neq \{0\}$, the inner integral in (9) is positive for every $\omega \in \mathbb{R}^d$ and so $B > 0$. Therefore k is c_0-*universal* by Proposition 4.

$(\Rightarrow)$ If k is c_0-*universal*, then it is strictly pd (by Proposition 5). The result therefore follows from Wendland, (2005, Theorem 7.14) which says that if k is strictly pd, then $\text{supp}(\nu) \neq \{0\}$. □

Examples of radial kernels that are c_0-*universal* include the Gaussian, inverse multiquadratics etc.

Kernels of type in (A_3)

We now consider the characterization of c_0-*universality* for (A_3).

Proposition 12. *Suppose (A_3) holds. Let $k(\cdot,x) \in C_0(X), \forall\, x \in X$. Then k is c_0-universal if and only if for any $0 \neq \mu \in M_b(X)$, there exists some $j \in I$ for which $\int_X \phi_j\, d\mu \neq 0$.*

Proof. Using $k(x,y) = \sum_{j\in I} \phi_j(x)\phi_j(y)$, we have $\iint_X k(x,y)\, d\mu(x)\, d\mu(y) = \sum_{j\in I} \left|\int_X \phi_j(x)\, d\mu(x)\right|^2$.

$(\Leftarrow)$ Suppose for any $0 \neq \mu \in M_b(X)$, there exists some $j \in I$ for which $\int_X \phi_j\, d\mu \neq 0$. Then, it is clear that $\iint_X k(x,y)\, d\mu(x)\, d\mu(y) > 0, \forall\, 0 \neq \mu \in M_b(X)$ and therefore k is c_0-*universal*, which follows from Proposition 4.

$(\Rightarrow)$ Suppose there exists a non-zero measure, $\mu \in M_b(X)$ for which $\int_X \phi_j\, d\mu = 0$ for any $j \in I$. This means, there exists a $0 \neq \mu \in M_b(X)$ for which $\iint_X k(x,y)\,d\mu(x)\,d\mu(y) = 0$, i.e., k is not c_0-*universal* (by Proposition 4). □

3.1 RELATION BETWEEN c_0-*universality* AND *c-universality*

Let X be a compact metric space (and therefore a compact Hausdorff space). Then $C_0(X) = C(X)$. Using this, in Theorem 13, we provide a characterization for *c-universal* kernels, which is similar to that of Theorem 3. Unlike the characterization in Steinwart (2001), which only provides sufficient conditions for *c-universality*, the following result provides both necessary and sufficient conditions for k to be *c-universal.*

Theorem 13 (Characterization of *c-universality*)**.** *k is c-universal if and only if the embedding in (4) is injective.*

When X is compact, Proposition 12 can be used to study the universality of Taylor kernels, e.g., exponential kernel, binomial kernel, etc. See Corollary 4.57, Examples 4.9 and 4.11 in Steinwart and Christmann (2008) for the definition of these kernels and their corresponding $\{\phi_j\}_{j\in I}$. The sufficient condition for the *c-universality* of Taylor kernels can easily be obtained from Proposition 12, which coincides with the result in Corollary 4.57 of Steinwart and Christmann (2008).

We now consider $X = [0, 2\pi)^d =: \mathbb{T}^d$, called the d-Torus and present necessary and sufficient conditions for a translation invariant kernel on $\mathbb{T}^d$, i.e.,

(A_4) $k(x,y) = \psi((x-y)_{mod\,2\pi})$, where $\psi \in C(\mathbb{T}^d)$ is a pd function on $\mathbb{T}^d$,

to be *c-universal.* Steinwart and Christmann, (2008, Lemma 4.12) called these kernels as of *Fourier type* and presented sufficient conditions for them to be *c-universal.* Using the characterization in Theorem 13, we show that these conditions are also necessary. Before we present the result in Proposition 15, we now state Bochner's theorem on $\mathbb{T}^d$.

Theorem 14 (Bochner)**.** *$\psi \in C(\mathbb{T}^d)$ is pd if and only if*

$$\psi(x) = \sum_{n\in\mathbb{Z}^d} A_\psi(n) e^{ix^T n},\ x \in \mathbb{T}^d, \qquad (10)$$

where $A_\psi : \mathbb{Z}^d \to \mathbb{R}_+$, $A_\psi(-n) = A_\psi(n)$ and $\sum_{n\in\mathbb{Z}^d} A_\psi(n) < \infty$. A_ψ are called the Fourier series coefficients of ψ.

Proposition 15. *Suppose (A_4) holds. Then k is c-universal if and only if $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$.*

Proof. $(\Leftarrow)$ Consider $B := \iint_{\mathbb{T}^d} k(x,y)\,d\mu(x)\,d\mu(y)$ for $0 \neq \mu \in M_b(\mathbb{T}^d)$. Substituting for k as in (A_4) and for ψ as in (10), it can be shown that $B = (2\pi)^{2d} \sum_{n\in\mathbb{Z}^d} A_\psi(n)|A_\mu(n)|^2$, where $A_\mu(n) := (2\pi)^{-d} \int_{\mathbb{T}^d} e^{-in^T x}\,d\mu(x)$, $n \in \mathbb{Z}^d$, which is the Fourier transform of μ in $\mathbb{T}^d$. Since $A_\psi(n) > 0, \forall\, n \in \mathbb{Z}^d$, we have $B > 0$, which by Proposition 4 implies k is *c-universal.*

$(\Rightarrow)$ Proving necessity is equivalent to proving that if $A_\psi(n) = 0$ for some $n = n_0 \neq 0$, then there exists $0 \neq \mu \in M_b(\mathbb{T}^d)$ such that $\iint k(x,y)\,d\mu(x)\,d\mu(y) = 0$. Let $A_\psi(n) = 0$ for some $n = n_0$. Define $d\mu(x) = 2\alpha\cos(x^T n_0)\,dx$, $\alpha \in \mathbb{R}\backslash\{0\}$. It is easy to check that $0 \neq \mu \in M_b(\mathbb{T}^d)$ and $\iint k(x,y)\,d\mu(x)\,d\mu(y) = 0$. Therefore, k is not *c-universal.* □

3.2 RELATION BETWEEN c_0-*universality* AND *cc-universality*

In this section, we first present a novel characterization of *cc-universality*, which is related to the injective embedding of a certain class of Borel measures into $\mathcal{H}$. Using this result, we then relate the notions of c_0-*universality* and *cc-universality*: if k is c_0-*universal*, then it is *cc-universal* but not vice-versa.

Theorem 16 (Characterization of *cc-universality*)**.** *Let X be an LCH space and k be continuous and bounded on $X \times X$. Then k is cc-universal if and only if the embedding,*

$$\mu \mapsto \int_X k(\cdot, x)\,d\mu(x),\ \mu \in M_{bc}(X), \qquad (11)$$

is injective.

Proof. As in the proof of Theorem 3, we need to consider the dual of $C(X)$ (endowed with the topology of compact convergence), which is $M_{bc}(X)$ (Hewitt, 1950). The rest of the proof follows the same idea as in the proof of Theorem 3. □

It is clear from Theorems 3 and 16 that any c_0-*universal* kernel is *cc-universal.* However, the converse is not true. To prove the converse is not true, we first re-derive a result due to Micchelli et al., (2006, Proposition 15), which provides a sufficient condition for k to be *cc-universal* when k is translation invariant on $\mathbb{R}^d$.

Proposition 17. *Suppose (A_1) holds. If* $\text{supp}(\Lambda)$ *has a non-empty interior, then k is cc-universal.*

Proof. Based on Theorem 16, it is easy show that if $\iint k(x,y)\,d\mu(x)\,d\mu(y) > 0, \forall\, 0 \neq \mu \in M_{bc}(X)$, then k is *cc-universal* (note that the proof of this claim is very similar to that of Proposition 4). Now, consider $B := \iint k(x,y)\,d\mu(x)\,d\mu(y)$ with $k(x,y) = \psi(x-y)$. As shown in the proof of Proposition 8, we have $B = \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\,d\Lambda(\omega)$, where Λ is defined in Theorem 7. Since $\mu \in M_{bc}(\mathbb{R}^d)$, by the Paley-Wiener theorem (Rudin, 1991, Theorem 7.23), we obtain $\text{supp}(\hat{\mu}) = \mathbb{R}^d$. Therefore if $\text{supp}(\Lambda)$ has a non-empty interior, $B > 0$ and therefore, k is *cc-universal.* □

An example of a *cc-universal* kernel is $k(x,y) = \frac{\sin^2((x-y)/2)}{(x-y)^2}$ as $\text{supp}(\Lambda) = [-1,1] \subsetneq \mathbb{R}$ has a non-empty interior. However, it is not c_0-*universal* as $\text{supp}(\Lambda) \neq \mathbb{R}$.

To summarize, so far, we have showed how the proposed notion of c_0-*universality* generalizes (i.e., stronger than) the notions of *c-universality* and *cc-universality* by relating them to the injective RKHS embeddings of certain class of Borel measures. We have also characterized the notion of c_0-*universality* for many interesting families of kernels. In the following section, we relate these various notions of *universality* to *characteristic kernels*, which are associated with the injective RKHS embedding of Borel probability measures.

4 CHARACTERISTIC KERNELS AND UNIVERSALITY

Recent studies in machine learning have considered the mapping of random variables into a suitable RKHS and showed that this provides a powerful and straightforward method of dealing with higher-order statistics of the variables. For sufficiently *rich* RKHSs, this notion is used to test for homogeneity (Gretton et al., 2007), independence (Gretton et al., 2008), conditional independence (Fukumizu et al., 2008), etc.

Key to the above applications is the notion of a *characteristic kernel*, which is defined as a kernel for which the embedding, $\mathbb{P} \mapsto \int_X k(\cdot,x)\,d\mathbb{P}(x)$ is injective. Here, $\mathbb{P}$ is a Borel probability measure defined on a topological space, X and k is a measurable, bounded kernel on X. In other words, a characteristic kernel induces an RKHS that is sufficiently rich in the sense that probability measures have unique images. From the point of view of applications, although the universality (which is motivated from the point of view of achieving consistency in kernel-based algorithms) and characteristic property may seem unrelated, in this paper, we show that they are closely related. The first result in this direction is by Gretton et al. (2007), wherein they showed that a *c-universal* kernel is characteristic. Besides this result, not much is known or understood about the relation between characteristic and universal kernels.

Based on the relation between universality and the RKHS embedding of measures which we established in Section 3, the following proposition presents the relation between universal and characteristic kernels.

Proposition 18 (Universal and characteristic kernels−I). *If k is*

(a) c_0- or c-universal, then it is characteristic to the set of probability measures contained in $M_b(X)$.

(b) cc-universal, then it is characteristic to the set of probability measures contained in $M_{bc}(X)$.

Proof. The proof follows from Theorems 3, 13, 16 and the definition of a characteristic kernel. □

Now, one can ask when is the converse true? The following result answers this question when k is translation invariant on $\mathbb{R}^d$ and $\mathbb{T}^d$, i.e., the kernels defined in (A_1) and (A_4).

Proposition 19 (Universal and characteristic kernels−II). *The following hold:*

(a) Suppose (A_1) holds with $\psi \in C_0(\mathbb{R}^d)$. Then k is c_0-universal if and only if it is characteristic to the set of all Borel probability measures on $\mathbb{R}^d$.

(b) Suppose (A_4) holds. Then k is c-universal if it is characteristic to the set of all Borel probability measures on $\mathbb{T}^d$ and $A_\psi(0) > 0$.

Proof. (a) Suppose k is c_0-*universal.* Then, by Proposition 18, k is characteristic to $M_+^1(\mathbb{R}^d)$. Conversely, if k is characteristic to $M_+^1(\mathbb{R}^d)$, we have $\text{supp}(\Lambda) = \mathbb{R}^d$ which follows from Theorem 7 in Sriperumbudur et al. (2008). The result therefore follows from Proposition 8.

(b) Using the same idea as in the proof of the necessity part of Proposition 15, it can be shown that if k is characteristic, then $A_\psi(0) \geq 0$, $A_\psi(n) > 0, \forall\, n \neq 0$ (we skip the proof here). Therefore if k is characteristic with $A_\psi(0) > 0$, then it is *c-universal* by Proposition 15. □

The above result shows that the concepts of universality and characteristic property are equivalent (*resp.* almost equivalent) on the class of translation invariant kernels defined over $\mathbb{R}^d$ (*resp.* $\mathbb{T}^d$). This result can be easily extended to translation invariant kernels on locally compact Abelian groups.

Based on the discussion so far, one can summarize the similarity and difference between characteristic and universal kernels as follows: *(i)* Based on (2), (4) and (11), it is clear that characteristic and universal kernels are essentially the same except that universal kernels deal with some subset of $M_b(X)$ while characteristic kernels deal with probability measures. *(ii)* For the characteristic property, the constant function is not necessary in $\mathcal{H}$, which is clearly highlighted in Proposition 19(b).

5 CONCLUSIONS & DISCUSSION

In this work, we have generalized the notions of universality considered by Steinwart (2001) and Micchelli et al. (2006) by presenting a notion of universality that subsumes these other definitions. The properties of the proposed notion of universality are studied. It is also shown that all these notions of universality are closely linked to the injective RKHS embedding of a certain class of Borel measures, which therefore leads to the problem of understanding the relation between characteristic and universal kernels. This is fully settled in the case of translation invariant kernels on $\mathbb{R}^d$ and $\mathbb{T}^d$, where the equivalence between characteristic and universal kernels is established.

As an extension, one can further generalize the notion of universality that is considered in this paper. Suppose X is a topological space. A bounded continuous kernel, k can be defined to be c_b*-universal* if the RKHS, $\mathcal{H}$ induced by k is dense in $C_b(X)$ w.r.t. the uniform norm. Clearly, this concept of universality subsumes c_0*-universality* and addresses its limitation of approximating only a subset of $C(X)$. Following a technique similar to the proof of Theorem 3, it can be shown that k is c_b*-universal* if and only if $\mu \mapsto \int_X k(\cdot,x)\,d\mu(x)$ is injective, where X is a normal space and μ is a regular bounded finitely additive set function defined on the *field* (not a σ-field) generated by the closed sets (Dunford & Schwartz, 1958). Because of the technicalities involved in dealing with such a set function, we did not pursue it in this paper, though it will be of interest to deal with such a generalized notion.

Acknowledgements

The authors thank anonymous reviewers for their constructive comments and especially the reviewer who pointed out Carmeli et al. (2009). B. K. S. and G. R. G. L. wish to acknowledge support from the Institute of Statistical Mathematics (ISM), Tokyo, the National Science Foundation (grant DMS-MSPA 0625409), the Fair Isaac Corporation and the University of California MICRO program. Part of this work was done while B. K. S. was visiting ISM. K.F. was supported by JSPS KAKENHI 19500249.

APPENDIX

The following result is a simple generalization of the technique used in the proof of Sriperumbudur et al., (2008, Theorem 3).

Lemma 20. *Let k be a measurable and bounded kernel on a measurable space, X and let $\mathcal{H}$ be its associated RKHS. Then, for any $f \in \mathcal{H}$ and for any finite signed Borel measure, μ, $\int_X f(x)\,d\mu(x) = \langle f, \int_X k(\cdot,x)\,d\mu(x)\rangle_{\mathcal{H}}$.*

References

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

Berg, C., Christensen, J. P. R., & Ressel, P. (1984). *Harmonic analysis on semigroups.* New York: Spring Verlag.

Carmeli, C., Vito, E. D., Toigo, A., & Umanità, V. (2009). Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications.* To appear.

Dunford, N., & Schwartz, J. T. (1958). *Linear operators. I: General theory.* New York: Wiley-Interscience.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications.* New York: Wiley-Interscience.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2009a). Kernel dimension reduction in regression. *Annals of Statistics*, *37*, 1871–1905.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems 20* (pp. 489–496). Cambridge, MA: MIT Press.

Fukumizu, K., Sriperumbudur, B. K., Gretton, A., & Schölkopf, B. (2009b). Characteristic kernels on groups and semigroups. *Advances in Neural Information Processing Systems 21* (pp. 473–480).

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., & Smola, A. (2007). A kernel method for the two sample problem. *Advances in Neural Information Processing Systems 19* (pp. 513–520). MIT Press.

Gretton, A., Fukumizu, K., Teo, C.-H., Song, L., Schölkopf, B., & Smola, A. (2008). A kernel statistical test of independence. *Advances in Neural Information Processing Systems 20* (pp. 585–592). MIT Press.

Hewitt, E. (1950). Linear functionals on spaces of continuous functions. *Fundamenta Mathematicae*, *37*, 161–189.

Micchelli, C. A., Xu, Y., & Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, *7*, 2651–2667.

Rudin, W. (1991). *Functional analysis.* USA: McGraw-Hill.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels.* Cambridge, MA: MIT Press.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., & Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in Neural Information Processing Systems 22* (pp. 1750–1758). MIT Press.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., & Schölkopf, B. (2008). Injective Hilbert space embeddings of probability measures. *Proc. of the 21st Annual Conference on Learning Theory* (pp. 111–122).

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

Steinwart, I., & Christmann, A. (2008). *Support vector machines.* Springer.

Wendland, H. (2005). *Scattered data approximation.* Cambridge, UK: Cambridge University Press.