Injective Hilbert Space Embeddings of Probability Measures

Bharath K. Sriperumbudur[1]*, Arthur Gretton[2], Kenji Fukumizu[3], Gert Lanckriet[1] and Bernhard Schölkopf[2]

[1]Department of ECE, UC San Diego, La Jolla, CA 92093, USA.
[2]MPI for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany.
[3]Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.
bharathsv@ucsd.edu, {arthur,bernhard.schoelkopf}@tuebingen.mpg.de
fukumizu@ism.ac.jp, gert@ece.ucsd.edu

Abstract

A Hilbert space embedding for probability measures has recently been proposed, with applications including dimensionality reduction, homogeneity testing and independence testing. This embedding represents any probability measure as a mean element in a reproducing kernel Hilbert space (RKHS). The embedding function has been proven to be injective when the reproducing kernel is universal. In this case, the embedding induces a metric on the space of probability distributions defined on compact metric spaces.

In the present work, we consider more broadly the problem of specifying characteristic kernels, defined as kernels for which the RKHS embedding of probability measures is injective. In particular, characteristic kernels can include non-universal kernels. We restrict ourselves to translation-invariant kernels on Euclidean space, and define the associated metric on probability measures in terms of the Fourier spectrum of the kernel and characteristic functions of these measures. The support of the kernel spectrum is important in finding whether a kernel is characteristic: in particular, the embedding is injective if and only if the kernel spectrum has the entire domain as its support. Characteristic kernels may nonetheless have difficulty in distinguishing certain distributions on the basis of finite samples, again due to the interaction of the kernel spectrum and the characteristic functions of the measures.

1 Introduction

The concept of distance between probability measures is a fundamental one and has many applications in probability theory and statistics. In probability theory, this notion is used to metrize the weak convergence (convergence in distribution) of probability measures defined on a metric space. Formally, let $\mathfrak{S}$ be the set of all Borel probability measures defined on a metric measurable space $(M, \rho, \mathcal{M}_\rho)$ and let γ be its metric, i.e., $(\mathfrak{S}, \gamma)$ is a metric space. Then P_n is said to converge weakly to P if and only if $\gamma(P_n, P) \overset{n\to\infty}{\longrightarrow} 0$, where $P, \{P_n\}_{n\geq 1} \in \mathfrak{S}$. When M is separable, examples for γ include the *Lévy-Prohorov distance* and the *dual-bounded Lipschitz distance* (*Dudley metric*) [Dud02, Chapter 11]. Other popular examples for γ include the *Monge-Wasserstein distance*, *total variation distance* and the *Hellinger distance*, which yield a stronger notion of convergence of probability measures [Sho00, Chapter 19].

In statistics, the notion of distance between probability measures is used in a variety of applications, including homogeneity tests (the two-sample problem), independence tests, and goodness-of-fit tests. The two-sample problem involves testing the null hypothesis $H_0 : P = Q$ versus the alternative $H_1 : P \neq Q$, using random samples $\{X_l\}_{l=1}^m$ and $\{Y_l\}_{l=1}^n$ drawn i.i.d. from distributions P and Q on a measurable space $(M, \mathcal{M})$. If γ is a metric (or more generally a semi-metric[1]) on $\mathfrak{S}$, then $\gamma(P, Q)$ can be used as a test statistic to address the two-sample problem. This is because $\gamma(P, Q)$ takes the unique and distinctive value of zero only when $P = Q$. Thus, the two-sample problem can be reduced to testing $H_0 : \gamma(P, Q) = 0$ versus $H_1 : \gamma(P, Q) > 0$. The problems of testing independence and goodness-of-fit can be posed in an analogous form.

Several recent studies on kernel methods have focused on applications in distribution comparison: the advantage being that kernels represent a linear way of dealing with higher order statistics. For instance, in homogeneity testing, differences in higher order moments are encoded in mean differences computed in the right reproducing kernel Hilbert space (RKHS) [GBR+07]; in kernel ICA [BJ02, GHS+05], general nonlinear dependencies show up as linear correlations once they are computed in a suitable RKHS. Instrumental to these studies is the notion of a Hilbert space embedding for probability measures [SGSS07], which involves representing any probability measure as a mean element in an RKHS $(\mathcal{H}, k)$, where k is the reproducing kernel [Aro50,

*The author wishes to acknowledge the support from the Max Planck Institute (MPI) for Biological Cybernetics, 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 the author was an intern at MPI.
The authors thank anonymous reviewers for their comments to improve the paper.

[1]Given a set M, a *metric* for M is a function $\rho : M \times M \to \mathbb{R}_+$ such that *(i)* $\forall\, x,\ \rho(x, x) = 0$, *(ii)* $\forall\, x, y,\ \rho(x, y) = \rho(y, x)$, *(iii)* $\forall\, x, y, z,\ \rho(x, z) \leq \rho(x, y) + \rho(y, z)$, and *(iv)* $\rho(x, y) = 0 \Rightarrow x = y$ [Dud02, Chapter 2]. A semi-metric only satisfies *(i)*, *(ii)* and *(iv)*.

SS02]. For this reason, the RKHSs used have to be "sufficiently large" to capture all nonlinearities that are relevant to the problem at hand, so that differences in embeddings correspond to differences of interest in the distributions. The question of how to choose such RKHSs is the central focus of the present paper.

Recently, Fukumizu *et al.* [FGSS08] introduced the concept of a *characteristic kernel*, this being an RKHS kernel for which the mapping $\Pi : \mathfrak{S} \to \mathcal{H}$ from the space of Borel probability measures $\mathfrak{S}$ to the associated RKHS $\mathcal{H}$ is injective ($\mathcal{H}$ is denoted as a characteristic RKHS). Clearly, a characteristic RKHS is sufficiently large in the sense we have described: in this case $\gamma(P, Q) = 0$ implies $P = Q$, where γ is the induced metric on $\mathfrak{S}$ by Π, defined as the RKHS distance between the mappings of P and Q. Under what conditions, then, is Π injective? As discussed in [GBR+07, SGSS07], when M is compact, the RKHS is characteristic when its kernel is universal in the sense of Steinwart [Ste02, Definition 4]: the induced RKHS should be dense in the Banach space of bounded continuous functions with respect to the supremum norm (examples include the Gaussian and Laplacian kernels). Fukumizu *et al.* [FGSS08, Lemma 1] considered injectivity for non-compact M, and showed Π to be injective if the direct sum of $\mathcal{H}$ and $\mathbb{R}$ is dense in the Banach space of p-power ($p \geq 1$) integrable functions (we denote RKHSs satisfying this criterion as *F*-characteristic). In addition, for $M = \mathbb{R}^d$, Fukumizu *et al.* provide sufficient conditions on the Fourier spectrum of a translation-invariant kernel for it to be characteristic [FGSS08, Theorem 2]. Using this result, popular kernels like Gaussian and Laplacian can be shown to be characteristic on all of $\mathbb{R}^d$.

In the present study, we provide an alternative means of determining whether kernels are characteristic, for the case of translation-invariant kernels on $\mathbb{R}^d$. This addresses several limitations of the previous work: in particular, it can be difficult to verify the conditions that a universal or *F*-characteristic kernel must satisfy; and universality is in any case an overly restrictive condition because universal kernels assume M to be compact. In other words, they induce a metric only on the space of probability measures that are compactly supported on M. In addition, there are compactly supported kernels which are not universal, e.g. B_{2n+1}-splines, which can be shown to be characteristic. We provide simple verifiable rules in terms of the Fourier spectrum of the kernel that characterize the injective behavior of Π, and derive a relationship between the family of kernels and the family of probability measures for which $\gamma(P, Q) = 0$ implies $P = Q$. In particular, we show that a translation-invariant kernel on $\mathbb{R}^d$ is characteristic if and only if its Fourier spectrum has the entire domain as its support.

We begin our presentation in §2 with an overview of terminology and notation. In §3, we briefly describe the approach of Hilbert space embedding of probability measures. Assuming the kernel to be translation-invariant in $\mathbb{R}^d$, in §4, we deduce conditions on the kernel and the set of probability measures for which the RKHS is characteristic. We show that the support of the kernel spectrum is crucial: $\mathcal{H}$ is characteristic if and only if the kernel spectrum has the entire domain as its support. We note, however, that even using such a kernel does not guarantee that one can easily distinguish distributions based on finite samples. In particular, we provide two illustrations in §5 where interactions between the kernel spectrum and the characteristic functions of the probability measures can result in an arbitrarily small $\gamma(P, Q) = \epsilon > 0$ for non-trivial differences in distributions $P \neq Q$. Proofs of the main theorems and related lemmas are provided in §6. The results presented in this paper use tools from *distribution theory* and Fourier analysis: the related technical results are collected in Appendix A.

2 Notation

For $M \subset \mathbb{R}^d$ and μ a Borel measure on M, $L^p(M, \mu)$ denotes the Banach space of p-power ($p \geq 1$) μ-integrable functions. We will also use $L^p(M)$ for $L^p(M, \mu)$ and dx for $d\mu(x)$ if μ is the Lebesgue measure on M. $C_b(M)$ denotes the space of all bounded, continuous functions on M. The space of all q-continuously differentiable functions on M is denoted by $C^q(M)$, $0 \leq q \leq \infty$. For $x \in \mathbb{C}$, $\overline{x}$ represents the complex conjugate of x. We denote as i the complex number $\sqrt{-1}$.

The set of all compactly supported functions in $C^\infty(\mathbb{R}^d)$ is denoted by $\mathscr{D}_d$ and the space of rapidly decreasing functions in $\mathbb{R}^d$ is denoted by $\mathscr{S}_d$. For an open set $U \subset \mathbb{R}^d$, $\mathscr{D}_d(U)$ denotes the subspace of $\mathscr{D}_d$ consisting of the functions with support contained in U. The space of linear continuous functionals on $\mathscr{D}_d$ (resp. $\mathscr{S}_d$) is denoted by $\mathscr{D}'_d$ (resp. $\mathscr{S}'_d$) and an element of such a space is called as a *distribution* (resp. *tempered distribution*). m_d denotes the normalized Lebesgue measure defined by $dm_d(x) = (2\pi)^{-\frac{d}{2}}\, dx$. $\hat{f}$ and $\check{f}$ represent the Fourier transform and inverse Fourier transform of f respectively.

For a measurable function f and a signed measure P, $Pf := \int f\, dP = \int_M f(x)\, dP(x)$. δ_x represents the Dirac measure at x. The symbol δ is overloaded to represent the Dirac measure, the Dirac-delta function, and the Kronocker-delta, which should be distinguishable from the context.

3 Maximum Mean Discrepancy

We briefly review the theory of RKHS embedding of probability measures proposed by Smola *et al.* [SGSS07]. We lead to these embeddings by first introducing the maximum mean discrepancy (MMD), which is based on the following result [Dud02, Lemma 9.3.2], related to the weak convergence of probability measures on metric spaces.

Lemma 1 ([Dud02]) *Let (M, ρ) be a metric space with Borel probability measures P and Q defined on M. Then $P = Q$ if and only if $Pf = Qf$, $\forall\, f \in C_b(M)$.*

Originally, Gretton *et al.* [GBR+07] defined the maximum mean discrepancy as follows.

Definition 2 (Maximum Mean Discrepancy) *Let $\mathcal{F} = \{f \mid f : M \to \mathbb{R}\}$ and let P, Q be Borel probability measures defined on (M, ρ). Then the maximum mean discrepancy is defined as*

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |Pf - Qf| \,. \tag{1}$$

With this definition, one can derive various metrics (mentioned in §1) that are used to define the weak convergence of probability measures on metric spaces. To start with, it is easy to verify that, independent of $\mathcal{F}$, $\gamma_{\mathcal{F}}$ in Eq. (1) is a pseudometric[2] on $\mathfrak{S}$. Therefore, the choice of $\mathcal{F}$ determines whether or not $\gamma_{\mathcal{F}}(P,Q) = 0$ implies $P = Q$. In other words, $\mathcal{F}$ determines the metric property of $\gamma_{\mathcal{F}}$ on $\mathfrak{S}$. By Lemma 1, $\gamma_{\mathcal{F}}$ is a metric on $\mathfrak{S}$ when $\mathcal{F} = C_b(M)$. When $\mathcal{F}$ is the set of bounded, ρ-uniformly continuous functions on M, by the Portmanteau theorem [Sho00, Chapter 19, Theorem 1.1], $\gamma_{\mathcal{F}}$ is not only a metric on $\mathfrak{S}$ but also metrizes the weak topology on $\mathfrak{S}$. $\gamma_{\mathcal{F}}$ is a *Dudley metric* [Sho00, Chapter 19, Definition 2.2] when $\mathcal{F} = \{f : \|f\|_{BL} \leq 1\}$ where $\|f\|_{BL} = \|f\|_\infty + \|f\|_L$ with $\|f\|_\infty := \sup\{|f(x)| : x \in M\}$ and $\|f\|_L := \sup\{|f(x) - f(y)|/\rho(x,y) : x \neq y \text{ in } M\}$. $\|f\|_L$ is called the Lipschitz seminorm of a real-valued function f on M. By the Kantorovich-Rubinstein theorem [Dud02, Theorem 11.8.2], when (M,ρ) is separable, $\gamma_{\mathcal{F}}$ equals the *Monge-Wasserstein distance* for $\mathcal{F} = \{f : \|f\|_L \leq 1\}$. $\gamma_{\mathcal{F}}$ is the *total variation metric* when $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$. If $\mathcal{F} = \{e^{i\langle\omega,\cdot\rangle} : \omega \in \mathbb{R}^d\}$, then $\gamma_{\mathcal{F}}(P,Q)$ reduces to finding the maximal difference between the characteristic functions of P and Q. By the uniqueness theorem for characteristic functions [Dud02, Theorem 9.5.1], we have $\gamma_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow \phi_P = \phi_Q \Leftrightarrow P = Q$, where ϕ_P and ϕ_Q represent the characteristic functions of P and Q, respectively.[3] Therefore, the function class $\mathcal{F} = \{e^{i\langle\omega,\cdot\rangle} : \omega \in \mathbb{R}^d\}$ induces a metric on $\mathfrak{S}$. Gretton *et al.* [GBR+07, Theorem 3] showed $\gamma_{\mathcal{F}}$ to be a metric on $\mathfrak{S}$ when $\mathcal{F}$ is chosen to be a unit ball in a universal RKHS $\mathcal{H}$. This choice of $\mathcal{F}$ yields an injective map, $\Pi : \mathfrak{S} \to \mathcal{H}$, as proposed by Smola *et al.* [SGSS07]. A similar injective map can also be obtained by choosing $\mathcal{F}$ to be a unit ball in an RKHS induced by kernels satisfying the criteria in [FGSS08, Lemma 1, Theorem 2] (which we denote *F*-characteristic kernels).

We henceforth assume $\mathcal{F}$ to be a unit ball in an RKHS $(\mathcal{H},k)$ (not necessarily universal or *F*-characteristic) defined on $(M,\mathcal{M})$ with $k : M \times M \to \mathbb{R}$, i.e., $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$. The following result provides a different representation for $\gamma_{\mathcal{F}}$ defined in Eq. (1) by exploiting the reproducing property of $\mathcal{H}$, and will be used later in deriving our main results.

Theorem 3 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H},k)$ defined on a measurable space $(M,\mathcal{M})$ with k measurable and bounded. Then*

$$\gamma_{\mathcal{F}}(P,Q) = \|Pk - Qk\|_{\mathcal{H}}, \tag{2}$$

where $\|.\|_{\mathcal{H}}$ represents the RKHS norm.

Proof: Let $T_P : \mathcal{H} \to \mathbb{R}$ be a linear functional defined as $T_P[f] := \int_M f(x)\,dP(x)$ with $\|T_P\| := \sup_{f\in\mathcal{H}} \frac{|T_P[f]|}{\|f\|_{\mathcal{H}}}$. Consider

$$\begin{aligned} |T_P[f]| &= \left|\int_M f(x)\,dP(x)\right| \leq \int_M |f(x)|\,dP(x) \\ &= \int_M |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}|\,dP(x) \leq \sqrt{C}\|f\|_{\mathcal{H}}, \end{aligned}$$

where we have exploited the reproducing property and boundedness of the kernel to show T_P is a bounded linear functional on $\mathcal{H}$. Here, $C > 0$ is the bound on k, i.e., $|k(x,y)| \leq C < \infty$, $\forall\, x,y \in M$. Therefore, by the Riesz representation theorem [RS72, Theorem II.4], there exists a unique $\lambda_P \in \mathcal{H}$ such that $T_P[f] = \langle f, \lambda_P\rangle_{\mathcal{H}}$, $\forall\, f \in \mathcal{H}$. Let $f = k(\cdot,u)$ for some $u \in M$. Then, $T_P[k(\cdot,u)] = \langle k(\cdot,u), \lambda_P\rangle_{\mathcal{H}} = \lambda_P(u)$, which implies $\lambda_P = T_P[k] = Pk = \int_M k(\cdot,x)\,dP(x)$. Therefore, with $|Pf - Qf| = |\langle f, \lambda_P - \lambda_Q\rangle_{\mathcal{H}}|$, we have $\gamma_{\mathcal{F}}(P,Q) = \sup_{\|f\|_{\mathcal{H}}\leq 1} |Pf - Qf| = \|\lambda_P - \lambda_Q\|_{\mathcal{H}} = \|Pk - Qk\|_{\mathcal{H}}$. ■

The representation of $\gamma_{\mathcal{F}}$ in Eq. (2) yields the embedding, $\Pi[P] = \int_M k(\cdot,x)\,dP(x)$ as proposed in [SGSS07, FGSS08], which is injective when k is characteristic. While the representation of $\gamma_{\mathcal{F}}$ in Eq. (2) holds irrespective of the characteristic property of k , it need not be a metric on $\mathfrak{S}$, as Π is not guaranteed to be injective. The obvious question to ask is "For what class of kernels is Π injective?". To understand this in detail, we are interested in the following questions which we address in this paper.

Q1. Let $\mathfrak{D} \subsetneq \mathfrak{S}$ be a set of Borel probability measures defined on $(M,\mathcal{M})$. Let $\mathcal{K}$ be a family of positive definite kernels defined on M. What are the conditions on $\mathfrak{D}$ and $\mathcal{K}$ for which $\Pi : \mathfrak{D} \to \mathcal{H}_k$, $P \mapsto \int_M k(\cdot,x)\,dP(x)$ is injective, i.e., $\gamma_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow P = Q$ for $P,Q \in \mathfrak{D}$? Here, $\mathcal{H}_k$ represents the RKHS induced by $k \in \mathcal{K}$.

Q2. What are the conditions on $\mathcal{K}$ so that Π is injective on $\mathfrak{S}$?

Note that Q1 is a restriction of Q2 to $\mathfrak{D}$. The idea is that the kernels that do not make $\gamma_{\mathcal{F}}$ as a metric on $\mathfrak{S}$ may make it as a metric on some restricted class of probability measures, $\mathfrak{D} \subsetneq \mathfrak{S}$. Our next step, therefore, is to characterize the relationship between classes of kernels and probability measures, which is addressed in the following section.

4 Characteristic Kernels & Main Theorems

In this section, we present main results related to the behavior of MMD. We start with the following definition of characteristic kernels, which was recently introduced by Fukumizu *et al.* [FGSS08] in the context of measuring conditional (in)dependence using positive definite kernels.

Definition 4 (Characteristic kernel) *A positive definite kernel k is characteristic to a set $\mathfrak{D}$ of probability measures defined on $(M,\mathcal{M})$ if $\gamma_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow P = Q$ for $P,Q \in \mathfrak{D}$.*

Remark 5 *Equivalently, k is said to be characteristic to $\mathfrak{D}$ if the map, $\Pi : \mathfrak{D} \to \mathcal{H}$, $P \mapsto \int_M k(\cdot,x)\,dP(x)$, is injective. When $M = \mathbb{R}^d$, the notion of characteristic kernel is a generalization of the characteristic function, $\phi_P(\omega) = \int_{\mathbb{R}^d} e^{i\omega^T x}\,dP(x)$, $\forall\,\omega \in \mathbb{R}^d$, which is the expectation of the*

[2] A pseudometric only satisfies *(i)-(iii)* of the properties of a metric (see footnote 1). Unlike a metric space (M,ρ), points in a pseudometric space need not be distinguishable: one may have $\rho(x,y) = 0$ for $x \neq y$ [Dud02, Chapter 2].

[3] The characteristic function of a probability measure, P on $\mathbb{R}^d$ is defined as $\phi(\omega) := \int_{\mathbb{R}^d} e^{i\omega^T x}\,dP(x)$, $\forall\,\omega \in \mathbb{R}^d$.

complex-valued positive definite kernel, $k(\omega, x) = e^{i\omega^T x}$. Thus, the definition of a characteristic kernel generalizes the well-known property of the characteristic function that ϕ_P uniquely determines a Borel probability measure P on $\mathbb{R}^d$. See [FGSS08] for more details.

It is obvious from Definition 4 that universal kernels defined on a compact M and F-characteristic kernels on M are characteristic to the family of all probability measures defined on $(M, \mathcal{M})$. The characteristic property of the kernel relates the family of positive definite kernels and the family of probability measures. We would like to characterize the positive definite kernels that are characteristic to $\mathfrak{S}$. Among the kernels that are not characteristic to $\mathfrak{S}$, we would like to determine those kernels that are characteristic to some appropriately chosen subset $\mathfrak{D}$, of $\mathfrak{S}$. Intuitively, the smaller the set $\mathfrak{D}$, larger is the family of kernels that are characteristic to $\mathfrak{D}$. To this end, we make the following assumption.

Assumption 1 *$k(x, y) = \psi(x - y)$ where ψ is a bounded continuous real-valued positive definite function[4] on $M = \mathbb{R}^d$.*

The above assumption means that k is translation-invariant in $\mathbb{R}^d$. A whole family of such kernels can be generated as the Fourier transform of a finite non-negative Borel measure, given by the following result due to Bochner, which we quote from [Wen05, Theorem 6.6].

Theorem 6 (Bochner) *A continuous function $\psi : \mathbb{R}^d \to \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$, i.e.*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\Lambda(\omega), \quad \forall\, x \in \mathbb{R}^d. \tag{3}$$

Since the translation-invariant kernels in $\mathbb{R}^d$ are characterized by the Bochner's theorem, it is theoretically interesting to ask which subset in the Fourier images gives characteristic kernels. Before we describe such kernels k that are characteristic to $\mathfrak{S}$, in the following example, we show that there exist kernels that are not characteristic to $\mathfrak{S}$. Here, $\mathfrak{S}$ represents the family of all Borel probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ represents the Borel σ-algebra defined by open sets in $\mathbb{R}^d$ (see Assumption 1).

Example 1 (Trivial kernel) *Let $k(x, y) = \psi(x - y) = C$, $\forall\, x, y \in \mathbb{R}^d$ with $C > 0$. It can be shown that ψ is the Fourier transform of $\Lambda = C\delta_0$ with support $\{0\}$.*

Consider $Pk = \int_{\mathbb{R}^d} k(\cdot, x)\, dP(x) = C \int_{\mathbb{R}^d} dP(x) = C$. Since $Pk = C$ irrespective of $P \in \mathfrak{S}$, the map Π is not injective. In addition, $\gamma_{\mathcal{F}}(P, Q) = 0$ for any $P, Q \in \mathfrak{S}$. Therefore, the trivial kernel, k is not characteristic to $\mathfrak{S}$.

4.1 Main theorems

The following theorem characterizes all translation-invariant kernels in $\mathbb{R}^d$ that are characteristic to $\mathfrak{S}$.

[4]Let M be a nonempty set. A function $\psi : M \to \mathbb{R}$ is called positive definite if and only if $\sum_{j,l=1}^{n} c_j c_l \psi(x_j - x_l) \geq 0$, $\forall\, x_j \in M$, $\forall\, c_j \in \mathbb{R}$, $\forall\, n \in \mathbb{N}$.

Theorem 7 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Suppose k satisfies Assumption 1. Then k is a characteristic kernel to the family, $\mathfrak{S}$, of all probability measures defined on $\mathbb{R}^d$ if and only if supp$(\Lambda) = \mathbb{R}^d$.*

We provide a sketch of the proof of the above theorem, which is proved in §6.2.1 using a number of intermediate lemmas. The first step is to derive an alternate representation for $\gamma_{\mathcal{F}}$ in Eq. (2) under Assumption 1. Lemma 13 provides the Fourier representation of $\gamma_{\mathcal{F}}$ in terms of the kernel spectrum, Λ and the characteristic functions of P and Q. The advantage of this representation over the one in Eq. (2) is that it is easy to obtain necessary and sufficient conditions for the existence of $P \neq Q$, $P, Q \in \mathfrak{S}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, which are captured in Lemma 15. We then show that if supp$(\Lambda) = \mathbb{R}^d$, the conditions mentioned in Lemma 15 are violated, meaning $\nexists\, P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, thereby proving the sufficient condition in Theorem 7. Proving the converse is equivalent to proving that k is not characteristic to $\mathfrak{S}$ when supp$(\Lambda) \subsetneq \mathbb{R}^d$. So, when supp$(\Lambda) \subsetneq \mathbb{R}^d$, the result is proved using Lemma 19, which shows the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

Theorem 7 shows that the embedding function Π, associated with a positive definite translation-invariant kernel in $\mathbb{R}^d$ is injective if and only if the kernel spectrum has the entire domain as its support. Therefore, this result provides a simple verifiable rule for Π to be injective, unlike the results in [SGSS07, FGSS08] where the universality and F-characteristic properties of a given kernel are not easy to verify. In addition, the universality and F-characteristic properties are sufficient conditions for a kernel to induce an injective map Π, whereas Theorem 7 provides supp$(\Lambda) = \mathbb{R}^d$ as the necessary and sufficient condition. Therefore, we have answered question Q2 posed in §3. Examples of kernels that are characteristic to $\mathfrak{S}$ include the Gaussian, Laplacian and B_{2n+1}-splines. In fact, the whole family of compactly supported translation-invariant kernels on $\mathbb{R}^d$ are characteristic to $\mathfrak{S}$, as shown by the following corollary of Theorem 7.

Corollary 8 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Suppose k satisfies Assumption 1 and supp(ψ) is compact. Then k is a characteristic kernel to $\mathfrak{S}$.*

Proof: Since supp(ψ) is compact in $\mathbb{R}^d$, by Lemma 25, which is a corollary of the Paley-Wiener theorem (see also [GW99, Theorem 31.5.2, Proposition 31.5.4]), we deduce that supp$(\Lambda) = \mathbb{R}^d$. Therefore, the result follows from Theorem 7. ■

The above result is interesting in practice because of the computational advantage in dealing with compactly supported kernels. By Theorem 7, it is clear that kernels with supp$(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $\mathfrak{S}$. However, they can be characteristic to some $\mathfrak{D} \subsetneq \mathfrak{S}$ (see Q1 in §3). The following result addresses this setting.

Theorem 9 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of all compactly supported probability measures on $\mathbb{R}^d$ with characteristic functions in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Suppose k satisfies Assumption 1 and supp$(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior. Then k is a characteristic kernel to $\mathfrak{D}$.*

$\psi(x)$, $\Omega = \text{supp}(\Lambda)$	$\mathfrak{D}$	Characteristic	$\gamma_{\mathcal{F}}$	Reference
$\Omega = \mathbb{R}^d$	$\mathfrak{S}$	Yes	Metric	Theorem 7
$\text{supp}(\psi)$ is compact	$\mathfrak{S}$	Yes	Metric	Corollary 8
$\Omega \subsetneq \mathbb{R}^d$ has a non-empty interior	$\{P : \text{supp}(P)$ is compact, $\phi_P \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\}$	Yes	Metric	Theorem 9
$\Omega \subsetneq \mathbb{R}^d$	$\mathfrak{S}$	No	Pseudometric	Theorem 7

Table 1: k satisfies Assumption 1 and is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$. $\mathfrak{S}$ is the set of all probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. P represents a probability measure in $\mathbb{R}^d$ and ϕ_P is its characteristic function. If k is characteristic to $\mathfrak{S}$, then $(\mathfrak{S}, \gamma_{\mathcal{F}})$ is a metric space, where $\mathcal{F}$ is a unit ball in an RKHS $(\mathcal{H}, k)$.

The proof is given in §6.2.2 and the strategy is similar to that of Theorem 7, where the Fourier representation of $\gamma_{\mathcal{F}}$ (see Lemma 13) is used to derive necessary and sufficient conditions for the existence of $P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$ (see Lemma 17). We then show that if $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior, the conditions mentioned in Lemma 17 are violated, which means $\nexists\, P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, thereby proving the result.

Although, by Theorem 7, the kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $\mathfrak{S}$, Theorem 9 shows that there exists $\mathfrak{D} \subsetneq \mathfrak{S}$ to which a subset of these kernels are characteristic. This type of result is not available for the methods studied in [SGSS07, FGSS08]. An example of a kernel that satisfies the conditions in Theorem 9 is the Sinc kernel, $\psi(x) = \frac{\sin(\sigma x)}{x}$ which has $\text{supp}(\Lambda) = [-\sigma, \sigma]$. The condition that $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior is important for Theorem 9 to hold. If $\text{supp}(\Lambda)$ has an empty interior (examples include periodic kernels), then one can construct $P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. See §6.2.2 for the related discussion and an example.

We have shown that the support of the Fourier spectrum of a positive definite translation-invariant kernel in $\mathbb{R}^d$ characterizes the injective or non-injective behavior of Π. In particular, $\text{supp}(\Lambda) = \mathbb{R}^d$ is the necessary and sufficient condition for the map Π to be injective on $\mathfrak{S}$, which answers question Q2 posed in §3. We also showed that kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ can be characteristic to some $\mathfrak{D} \subsetneq \mathfrak{S}$ even though they are not characteristic to $\mathfrak{S}$, which in turn answers question Q1 in §3. A summary of these results is given in Table 1.

4.2 A result on periodic kernels and discrete probability measures

Proposition 10 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$ where k satisfies Assumption 1. Let $\mathfrak{D} = \{P : P = \sum_{n=1}^{\infty} \beta_n \delta_{x_n}, \sum_{n=1}^{\infty} \beta_n = 1, \beta_n \geq 0, \forall\, n\}$ be the set of probability measures defined on $M' = \{x_1, x_2 \ldots\} \subsetneq \mathbb{R}^d$. Then $\exists\, P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$ if the following conditions hold:*

(i) ψ is τ-periodic[5] in $\mathbb{R}^d$, i.e., $\psi(x) = \psi(x + \eta \bullet \tau)$, $\eta \in \mathbb{Z}^d$, $\tau \in \mathbb{R}^d_+$,

(ii) $x_s - x_t = l_{st} \bullet \tau$, $l_{st} \in \mathbb{Z}^d$, $\forall\, s, t$,

where $\bullet$ represents the Hadamard multiplication.

Proof: Let ψ be τ-periodic in $\mathbb{R}^d$ and $x_s - x_t = l_{st} \bullet \tau$, $l_{st} \in \mathbb{Z}^d$, $\forall\, s, t$. Consider $P, Q \in \mathfrak{D}$ given by $P = \sum_{n=1}^{\infty} \tilde{p}_n \delta_{x_n}$ and $Q = \sum_{n=1}^{\infty} \tilde{q}_n \delta_{x_n}$ such that $\tilde{p}_n, \tilde{q}_n \geq 0, \forall\, n$; $\sum_{n=1}^{\infty} \tilde{p}_n = 1$, $\sum_{n=1}^{\infty} \tilde{q}_n = 1$. Then $\gamma_{\mathcal{F}}(P, Q) = \|Pk - Qk\|_{\mathcal{H}} = \| \int_{\mathbb{R}^d} \psi(. - x)\, d(P - Q)(x) \|_{\mathcal{H}} = \| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_n) \|_{\mathcal{H}} = \| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_1 - l_{n1} \bullet \tau) \|_{\mathcal{H}} = \| \psi(. - x_1) \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \|_{\mathcal{H}} = 0$. This holds for any $P, Q \in \mathfrak{D}$. ■

The converse of Proposition 10, if true, would make the result more interesting. This is because any non-periodic translation invariant kernel on $\mathbb{R}^d$ would then be characteristic to the set of discrete probability measures on $\mathbb{R}^d$. In order to prove the converse, we would need to show that *(i)* and *(ii)* in Proposition 10 hold when $\gamma_{\mathcal{F}}(P, Q) = 0$ for $P \neq Q$, $P, Q \in \mathfrak{D}$. However, this is not true as the trivial kernel yields $\gamma_{\mathcal{F}}(P, Q) = 0$ for any $P, Q \in \mathfrak{S}$ and not just $P, Q \in \mathfrak{D}$.

Let us consider $\gamma_{\mathcal{F}}(P, Q) = 0$ for $P, Q \in \mathfrak{D}$. This is equivalent to $\| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_n) \|_{\mathcal{H}} = 0$. Squaring on both sides and using the reproducing property of k, we get $\sum_{s,t=1}^{\infty} \tilde{r}_t \tilde{r}_s \psi(x_s - x_t) = 0$ where $\{\tilde{r}_n = \tilde{p}_n - \tilde{q}_n\}_{n=1}^{\infty}$ satisfy $\sum_{s=1}^{\infty} \tilde{r}_s = 0$ and $\{\tilde{r}_s\}_{s=1}^{\infty} \in [-1, 1]$. So, to prove the converse, we need to characterize all ψ, $\{\tilde{r}_n\}_{n=1}^{\infty}$ and $\{x_n\}_{n=1}^{\infty}$ that satisfy $\mathcal{R} = \{\sum_{s,t=1}^{\infty} \tilde{r}_t \tilde{r}_s \psi(x_s - x_t) = 0 : \sum_{s=1}^{\infty} \tilde{r}_s = 0, \{\tilde{r}_s\}_{s=1}^{\infty} \in [-1, 1]\}$, which is not easy. However, choosing some ψ, $\{\tilde{r}_n\}_{n=1}^{\infty}$ and $\{x_n\}_{n=1}^{\infty}$ is easy, as shown in Proposition 10. Suppose there exists a class, $\mathcal{K}$ of positive definite translation-invariant kernels in $\mathbb{R}^d$ with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ and a class, $\mathfrak{E} \subset \mathfrak{D}$ of probability measures that jointly violate $\mathcal{R}$, then any $k \in \mathcal{K}$ is characteristic to $\mathfrak{E}$.

[5] A τ-periodic ψ in $\mathbb{R}$ is the Fourier transform of $\Lambda = \sum_{n=-\infty}^{\infty} \alpha_n \delta_{\frac{2\pi n}{\tau}}$, where $\delta_{\frac{2\pi n}{\tau}}$ is the Dirac measure at $\frac{2\pi n}{\tau}$, $n \in \mathbb{Z}$ with $\alpha_n \geq 0$ and $\sum_{n=-\infty}^{\infty} \alpha_n < \infty$. Thus, $\text{supp}(\Lambda) = \{\frac{2\pi n}{\tau} : \alpha_n > 0, n \in \mathbb{Z}\} \subsetneq \mathbb{R}$. $\{\alpha_n\}_{-\infty}^{\infty}$ are called the Fourier series coefficients of ψ.

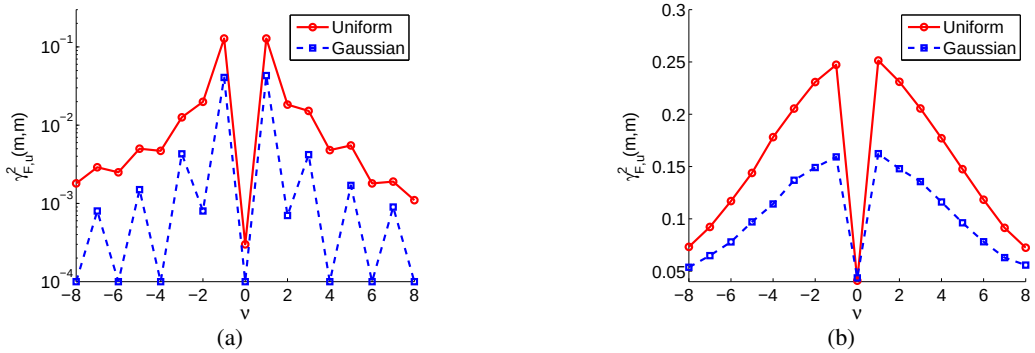

Figure 1: Behavior of the empirical estimate of $\gamma^2_{\mathcal{F}}(P,Q)$ w.r.t. ν for the (a) B_1-spline kernel and (b) Gaussian kernel. P is constructed from Q as defined in Eq. (4). "Uniform" corresponds to $Q = \mathcal{U}[-1,1]$ and "Gaussian" corresponds to $Q = \mathcal{N}(0,2)$. $m = 1000$ samples are generated from P and Q to estimate $\gamma^2_{\mathcal{F}}(P,Q)$ through $\gamma^2_{\mathcal{F},u}(m,m)$. See Example 2 for details.

5 Dissimilar Distributions with Small Mean Discrepancy

So far, we have studied the behavior of $\gamma_{\mathcal{F}}$ and have shown that it depends on the support of the spectrum of the kernel. As mentioned in §1, applications like homogeneity testing exploit the metric property of $\gamma_{\mathcal{F}}$ to distinguish between probability distributions. Since the metric nature of $\gamma_{\mathcal{F}}$ is guaranteed only for kernels with $\text{supp}(\Lambda) = \mathbb{R}^d$, tests based on other kernels can fail to distinguish between different probability distributions. However, in the following, we show that the characteristic kernels, while guaranteeing $\gamma_{\mathcal{F}}$ to be a metric on $\mathfrak{S}$, may nonetheless have difficulty in distinguishing certain distributions on the basis of finite samples. Before proving the result, we motivate it through the following example.

Example 2 *Let P be defined as*

$$p(x) = q(x) + \alpha q(x)\sin(\nu\pi x), \tag{4}$$

where q is a symmetric probability density function with $\alpha \in \mathbb{R}$, $\nu \in \mathbb{R}\backslash\{0\}$. Consider a B_1-spline kernel on $\mathbb{R}$ given by $k(x,y) = \psi(x-y)$ where

$$\psi(x) = \begin{cases} 1-|x|, & |x| \le 1 \\ 0, & \textit{otherwise} \end{cases}, \tag{5}$$

with its Fourier transform given by $\Psi(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{\sin^2\frac{\omega}{2}}{\omega^2}$ (see footnote 10 for the definition of Ψ). Since ψ is characteristic to $\mathfrak{S}$, $\gamma_{\mathcal{F}}(P,Q) > 0$ (see Theorem 7). However, it would be of interest to study the behavior of $\gamma_{\mathcal{F}}(P,Q)$ as a function of ν. We do this through an unbiased, consistent estimator[6] of $\gamma^2_{\mathcal{F}}(P,Q)$ as proposed by Gretton et al. [GBR+07, Lemma 7].

Figure 1(a) shows the behavior of the empirical estimate of $\gamma^2_{\mathcal{F}}(P,Q)$ as a function of ν for $q = \mathcal{U}[-1,1]$ and $q = \mathcal{N}(0,2)$ using the B_1-spline kernel in Eq. (5). Since the Gaussian kernel, $k(x,y) = e^{-(x-y)^2}$ is also a characteristic kernel, its effect on the behavior of $\gamma^2_{\mathcal{F},u}(m,m)$ is shown in Figure 1(b) in comparison to that of the B_1-spline kernel.

From Figure 1, we observe two circumstances under which the mean discrepancy may be small. First, $\gamma^2_{\mathcal{F},u}(m,m)$ decays with increasing $|\nu|$, and can be made as small as desired by choosing a sufficiently large $|\nu|$. Second, in Figure 1(a), $\gamma^2_{\mathcal{F},u}(m,m)$ has troughs at $\nu = \frac{\omega_0}{\pi}$ where $\omega_0 = \{\omega : \Psi(\omega) = 0\}$. Since $\gamma^2_{\mathcal{F},u}(m,m)$ is a consistent estimate of $\gamma^2_{\mathcal{F}}(P,Q)$, one would expect similar behavior from $\gamma^2_{\mathcal{F}}(P,Q)$. This means that though the B_1-spline kernel is characteristic to $\mathfrak{S}$, in practice, it becomes harder to distinguish between P and Q with finite samples, when P is constructed as in Eq. (4) with $\nu = \frac{\omega_0}{\pi}$. In fact, one can observe from a straightforward spectral argument that the troughs in $\gamma^2_{\mathcal{F}}(P,Q)$ can be made arbitrarily deep by widening q, when q is Gaussian.

For characteristic kernels, although $\gamma_{\mathcal{F}}(P,Q) > 0$ when $P \neq Q$, Example 2 demonstrates that one can construct distributions such that $\gamma^2_{\mathcal{F},u}(m,m)$ is indistinguishable from zero with high probability, for a given sample size m. Below, in Theorem 12, we investigate the decay mode of MMD for large $|\nu|$ (see Example 2) by explicitly constructing $P \neq Q$ such that $|P\varphi_l - Q\varphi_l|$ is large for some large l, but $\gamma_{\mathcal{F}}(P,Q)$ is arbitrarily small, making it hard to detect a non-zero value of the *population* MMD on the basis of a finite sample. Here, $\varphi_l \in L^2(M)$ represents the bounded orthonormal eigenfunctions of a positive definite integral operator[7] associated with k.

Consider the formulation of MMD in Eq. (1). The construction of P for a given Q such that $\gamma_{\mathcal{F}}(P,Q)$ is small, though not zero, can be intuitively seen by re-writing Eq. (1) as

$$\gamma_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{H}} \frac{|Pf - Qf|}{\|f\|_{\mathcal{H}}}. \tag{6}$$

[6] Starting from the expression for $\gamma_{\mathcal{F}}$ in Eq. (2), we get $\gamma^2_{\mathcal{F}}(P,Q) = \mathbb{E}_{X,X'\sim P}k(X,X') - 2\mathbb{E}_{X\sim P,Y\sim Q}k(X,Y) + \mathbb{E}_{Y,Y'\sim Q}k(Y,Y')$, where X, X' are independent random variables with distribution P and Y, Y' are independent random variables with distribution Q. An unbiased *empirical estimate* of $\gamma^2_{\mathcal{F}}$, denoted as $\gamma^2_{\mathcal{F},u}(m,m)$ is given by $\gamma^2_{\mathcal{F},u}(m,m) = \frac{1}{m(m-1)}\sum_{l\neq j}^{m} h(Z_l,Z_j)$, which is a one-sample U-statistic with $h(Z_l,Z_j) := k(X_l,X_j) + k(Y_l,Y_j) - k(X_l,Y_j) - k(X_j,Y_l)$, where $Z_1,\ldots,Z_m$ are m i.i.d. random variables with $Z_j := (X_j,Y_j)$ (see [GBR+07, Lemma 7]).

[7] See [SS02, Theorem 2.10] for definition of positive definite integral operator and its corresponding eigenfunctions.

When $P \neq Q$, $|Pf - Qf|$ can be large for some $f \in \mathcal{H}$. However, $\gamma_{\mathcal{F}}(P, Q)$ can be made small by selecting P such that the maximization of $\frac{|Pf-Qf|}{\|f\|_{\mathcal{H}}}$ over $\mathcal{H}$ requires an f with large $\|f\|_{\mathcal{H}}$. More specifically, higher order eigenfunctions of the kernel (φ_l for large l) have large RKHS norms, and so if they are prominent in P, Q (i.e., highly non-smooth distributions), one can expect $\gamma_{\mathcal{F}}(P, Q)$ to be small even when there exists an l for which $|P\varphi_l - Q\varphi_l|$ is large. To this end, we need the following lemma, which we quote from [GSB$^+$04, Lemma 6].

Lemma 11 ([GSB$^+$04]) *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on compact M. Let $\varphi_l \in L^2(M)$ be orthonormal eigenfunctions (assumed to be absolutely bounded), and λ_l be the corresponding eigenvalues (arranged in a decreasing order for increasing l) of a positive definite integral operator associated with k. Assume λ_l^{-1} increases superlinearly with l. Then for $f \in \mathcal{F}$ where $f(x) := \sum_{j=1}^{\infty} \tilde{f}_j \varphi_j(x)$, we have $\{|\tilde{f}_j|\}_{j=1}^{\infty} \in \ell_1$ and for every $\epsilon > 0$, $\exists\, l_0 \in \mathbb{N}$ such that $|\tilde{f}_l| < \epsilon$ if $l > l_0$.*

Theorem 12 ($P \neq Q$ can give small MMD) *Assume the conditions in Lemma 11 hold. Then there exists a probability distribution $P \neq Q$ defined on M for which $|P\varphi_l - Q\varphi_l| > \beta - \epsilon$ for some non-trivial β and arbitrarily small $\epsilon > 0$, yet for which $\gamma_{\mathcal{F}}(P, Q) < \eta$ for an arbitrarily small $\eta > 0$.*

Proof: Let us construct $p(x) = q(x) + \alpha_l e(x) + \beta\varphi_l(x)$ where $e(x) = \mathbb{1}_M(x)$. For P to be a probability distribution, the following conditions need to be satisfied:

$$\int_M [\alpha_l e(x) + \beta\varphi_l(x)]\, dx = 0, \tag{7}$$

$$\min_{x \in M} [q(x) + \alpha_l e(x) + \beta\varphi_l(x)] \geq 0. \tag{8}$$

Expanding $e(x)$ and $f(x)$ in the orthonormal basis $\{\varphi_l\}_{l=1}^{\infty}$, we get $e(x) = \sum_{l=1}^{\infty} \tilde{e}_l \varphi_l(x)$ and $f(x) = \sum_{l=1}^{\infty} \tilde{f}_l \varphi_l(x)$, where $\tilde{e}_l := \langle e, \varphi_l \rangle_{L^2(M)}$ and $\tilde{f}_l := \langle f, \varphi_l \rangle_{L^2(M)}$. Therefore, $Pf - Qf = \int_M f(x)[\alpha_l e(x) + \beta\varphi_l(x)]\, dx$ reduces to

$$Pf - Qf = \alpha_l \sum_{j=1}^{\infty} \tilde{e}_j \tilde{f}_j + \beta \tilde{f}_l, \tag{9}$$

where we used the fact that[8] $\langle \varphi_j, \varphi_t \rangle_{L^2(M)} = \delta_{jt}$. Rewriting Eq. (7) and substituting for $e(x)$ gives $\int_M [\alpha_l e(x) + \beta\varphi_l(x)]\, dx = \int_M e(x)[\alpha_l e(x) + \beta\varphi_l(x)]\, dx = \alpha_l \sum_{j=1}^{\infty} \tilde{e}_j^2 + \beta\tilde{e}_l = 0$, which implies

$$\alpha_l = -\frac{\beta\tilde{e}_l}{\sum_{j=1}^{\infty} \tilde{e}_j^2}. \tag{10}$$

Now, let us consider $P\varphi_t - Q\varphi_t = \alpha_l \tilde{e}_t + \beta\delta_{tl}$. Substituting for α_l gives

$$P\varphi_t - Q\varphi_t = \beta\delta_{tl} - \beta\frac{\tilde{e}_t\tilde{e}_l}{\sum_{j=1}^{\infty} \tilde{e}_j^2} = \beta\delta_{tl} - \beta\tau_{tl}, \tag{11}$$

where $\tau_{tl} := \frac{\tilde{e}_t\tilde{e}_l}{\sum_{j=1}^{\infty} \tilde{e}_j^2}$. By Lemma 11, $\{|\tilde{e}_l|\}_{l=1}^{\infty} \in \ell_1 \Rightarrow \sum_{j=1}^{\infty} \tilde{e}_j^2 < \infty$, and choosing large enough l gives $|\tau_{tl}| < \epsilon$, $\forall\, t$, for any arbitrary $\epsilon > 0$. Therefore, $|P\varphi_t - Q\varphi_t| > \beta - \epsilon$ for $t = l$ and $|P\varphi_t - Q\varphi_t| < \epsilon$ for $t \neq l$. By appealing to Lemma 1, we therefore establish that $P \neq Q$. In the following we prove that $\gamma_{\mathcal{F}}(P, Q)$ can be arbitrarily small, though non-zero.

Recall that $\gamma_{\mathcal{F}}(P, Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |Pf - Qf|$. Substituting for α_l in Eq. (9), we have

$$\gamma_{\mathcal{F}}(P, Q) = \sup \left\{ \beta \sum_{j=1}^{\infty} \nu_{jl} \tilde{f}_j \;:\; \sum_{j=1}^{\infty} \frac{\tilde{f}_j^2}{\lambda_j} \leq 1 \right\}, \tag{12}$$

where we used the definition of RKHS norm as $\|f\|_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\tilde{f}_j^2}{\lambda_j}$ and $\nu_{jl} := \delta_{jl} - \tau_{jl}$. Eq. (12) is a convex quadratic program in $\{\tilde{f}_j\}_{j=1}^{\infty}$. Solving the Lagrangian yields $\tilde{f}_j = \frac{\nu_{jl}\lambda_j}{\sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j}}$. Therefore, $\gamma_{\mathcal{F}}(P, Q) = \beta\sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j} = \beta\sqrt{\lambda_l - 2\tau_{ll}\lambda_l + \sum_{j=1}^{\infty} \tau_{jl}^2 \lambda_j} \to 0$ as $l \to \infty$ because
(i) by choosing sufficiently large l, $|\tau_{jl}| < \epsilon$, $\forall\, j$, for any arbitrary $\epsilon > 0$,
(ii) $\lambda_l \to 0$ as $l \to \infty$ [SS02, Theorem 2.10]. ■

6 Proofs of the Main Theorems

In this section, we prove the main theorems in Section 4.

6.1 Preliminary lemmas

Using the Fourier characterization of ψ given by Eq. (3), under Assumption 1, we derive the following result that provides the Fourier representation of MMD. This result requires tools from *distribution theory* related to the Fourier transforms of distributions.[9] We refer the reader to [Rud91, Chapters 6,7] for the detailed treatment of distribution theory. Another good and basic reference on distribution theory is [Str03].

Lemma 13 (Fourier representation of MMD) *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$ with k satisfying Assumption 1. Let ϕ_P and ϕ_Q be the characteristic functions of probability measures P and Q defined on $\mathbb{R}^d$. Then*

$$\gamma_{\mathcal{F}}(P, Q) = \|[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^{\vee}\|_{\mathcal{H}}, \tag{13}$$

where $-$ represents complex conjugation, $\vee$ represents the inverse Fourier transform and Λ represents the finite non-negative Borel measure on $\mathbb{R}^d$ as defined in Eq. (3). $(\overline{\phi}_P - \overline{\phi}_Q)\Lambda$ represents a finite Borel measure defined by Eq. (26).

Proof: From Theorem 3, we have $\gamma_{\mathcal{F}}(P, Q) = \|Pk - Qk\|_{\mathcal{H}}$. Consider $Pk = \int_{\mathbb{R}^d} k(\cdot, x)\, dP(x) = \int_{\mathbb{R}^d} \psi(\cdot - x)\, dP(x)$. By Eq. (23), $\int_{\mathbb{R}^d} \psi(\cdot - x)\, dP(x)$ represents the convolution of ψ and P, denoted as $\psi * P$. By appealing to the convolution theorem (Theorem 22), we have $(\psi * P)^{\wedge} = \hat{P}\Lambda$, where $\hat{P}(\omega) =$

[8] Here δ is used in the Kronecker sense.

[9] Here, the term *distribution* should not be confused with probability distributions. In short, distributions refer to generalized functions which cannot be treated as functions in the Lebesgue sense. Classical examples of distributions are the Dirac-delta function and Heaviside's function, for which derivatives and Fourier transforms do not exist in the usual sense.

$\int_{\mathbb{R}^d} e^{-i\omega^T x}\, dP(x),\ \forall\, \omega \in \mathbb{R}^d$ (by Lemma 20). Note that $\hat{P} = \overline{\phi}_P$. Therefore, $\gamma_{\mathcal{F}}(P,Q) = \|\psi * P - \psi * Q\|_{\mathcal{H}} = \left\|(\overline{\phi}_P\Lambda)^\vee - (\overline{\phi}_Q\Lambda)^\vee\right\|_{\mathcal{H}}$. Using the linearity of the Fourier inverse, we get the desired result. ■

Remark 14 *(a) If Ψ is the distributional derivative[10] of Λ, then Eq. (13) can also be written as*

$$\gamma_{\mathcal{F}}(P,Q) = \|[(\overline{\phi}_P - \overline{\phi}_Q)\Psi]^\vee\|_{\mathcal{H}}, \tag{14}$$

where the term inside the RKHS norm is the Fourier inverse of a tempered distribution.

(b) By Assumption 1, ψ is real-valued and symmetric in $\mathbb{R}^d$. Therefore, by (ii) in Lemma 20, Λ and Ψ are real-valued, symmetric tempered distributions.

The representation of MMD in terms of the kernel spectrum as in Eq. (13) will be central to deriving our main theorems. It is easy to see that characteristic kernels can be described indirectly by deriving conditions for the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P,Q) = 0$. Using the Fourier representation of $\gamma_{\mathcal{F}}$, the following result provides necessary and sufficient conditions for the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P,Q) = 0$.

Lemma 15 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$, and let P, Q be probability distributions on $\mathbb{R}^d$ such that $P \neq Q$. Suppose that k satisfies Assumption 1 and supp$(\Lambda) \subset \mathbb{R}^d$. Then $\gamma_{\mathcal{F}}(P,Q) = 0$ if and only if there exists $\theta \in \mathscr{S}'_d$ that satisfies the following conditions:*

(i) $p - q = \check{\theta}$,

(ii) $\theta\Lambda = 0$,

where p and q represent the distributional derivatives of P and Q respectively, and $\theta\Lambda$ represents a finite Borel measure defined by Eq. (26).

Proof: The proof follows directly from the formulation of $\gamma_{\mathcal{F}}$ in Eq. (13).

($\Rightarrow$) Let $\theta \in \mathscr{S}'_d$ satisfy *(i)* and *(ii)*. Since $\theta \in \mathscr{S}'_d$, we have $\theta = \hat{\check{\theta}} = (p-q)^\wedge = \hat{p} - \hat{q} = \overline{\phi}_P - \overline{\phi}_Q$. Therefore, by *(ii)*, we have $\gamma_{\mathcal{F}}(P,Q) = \|[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^\vee\|_{\mathcal{H}} = \|[\theta\Lambda]^\vee\|_{\mathcal{H}} = 0$.

($\Leftarrow$) Let $\gamma_{\mathcal{F}}(P,Q) = \|[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^\vee\|_{\mathcal{H}} = 0$, which implies $[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^\vee = 0$. Since $(\overline{\phi}_P - \overline{\phi}_Q)\Lambda$ is a finite Borel measure as defined by Eq. (26), it is therefore a tempered distribution and so $(\overline{\phi}_P - \overline{\phi}_Q)\Lambda = [[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^\vee]^\wedge = 0$. Let $\theta := \overline{\phi}_P - \overline{\phi}_Q$. Clearly $\theta \in \mathscr{S}'_d$ as by Lemma 20, $\overline{\phi}_P, \overline{\phi}_Q \in \mathscr{S}'_d$. So, $p - q = (\overline{\phi}_P)^\vee - (\overline{\phi}_Q)^\vee = (\overline{\phi}_P - \overline{\phi}_Q)^\vee = \check{\theta}$. ■

$\theta = 0$ trivially satisfies *(ii)* in Lemma 15. However, it violates our assumption of $P \neq Q$ when it is used in condition *(i)*. If we relax this assumption, then the result is trivial as $P = Q \Rightarrow \gamma_{\mathcal{F}}(P,Q) = 0$. For the results we derive later, it is important to understand the properties of θ, which we present in the following proposition.

Proposition 16 (Properties of θ) *θ in Lemma 15 satisfies the following properties:*

(a) θ is a conjugate symmetric, bounded and uniformly continuous function on $\mathbb{R}^d$.

(b) $\theta(0) = 0$.

(c) supp$(\theta) \subset \overline{\mathbb{R}^d \backslash \Omega}$ where $\Omega :=$ supp(Λ). In addition, if $\Omega = \{a_1, a_2, \ldots\}$, then $\theta(a_j) = 0,\ \forall\, a_j \in \Omega$.

Proof: *(a)* From Lemma 15, we have $\theta = \overline{\phi}_P - \overline{\phi}_Q$. Therefore, the result in *(a)* follows from Lemma 20, which shows that ϕ_P, ϕ_Q are conjugate symmetric, bounded, and uniformly continuous functions on $\mathbb{R}^d$.

(b) By Lemma 20, $\phi_P(0) = \phi_Q(0) = 1$. Therefore, $\theta(0) = \overline{\phi}_P(0) - \overline{\phi}_Q(0) = 0$.

(c) Let $W := \{x \in \mathbb{R}^d \,|\, \theta(x) \neq 0\}$. It suffices to show that $W \subset \overline{\mathbb{R}^d \backslash \Omega}$. Suppose W is not contained in $\overline{\mathbb{R}^d \backslash \Omega}$. Then there is a non-empty open subset U such that $U \subset W \cap (\Omega \cup \partial\Omega)$. Fix further a non-empty open subset V with $\overline{V} \subset U$. Since $V \subset \Omega$, there is $\varphi \in \mathscr{D}_d(V)$ with $\Lambda(\varphi) \neq 0$. Take $h \in \mathscr{D}_d(U)$ such that $h = 1$ on $\overline{V}$, and define a continuous function $\varrho = \frac{h\varphi}{\theta}$ on $\mathbb{R}^d$, which is well-defined from supp$(h) \subset U$ and $\theta \neq 0$ on U. By *(ii)* of Lemma 15, $\theta\Lambda = 0$, where $\theta\Lambda$ is a finite Borel measure on $\mathbb{R}^d$ as defined by Eq. (26). Therefore,

$$\int_{\mathbb{R}^d} \varrho(x)\theta(x)\, d\Lambda(x) = 0. \tag{15}$$

The left hand side of Eq. (15) simplifies to

$$\begin{aligned}\int_{\mathbb{R}^d} \varrho(x)\theta(x)\, d\Lambda(x) &= \int_U \frac{h(x)\varphi(x)}{\theta(x)}\theta(x)\, d\Lambda(x) \\ &= \int_U \varphi(x)\, d\Lambda(x) = \Lambda(\varphi) \neq 0,\end{aligned}$$

resulting in a contradiction. So, supp$(\theta) \subset \overline{\mathbb{R}^d \backslash \Omega}$.

If $\Omega = \{a_1, a_2, \ldots\}$, then $\Lambda = \sum_{a_j \in \Omega} \beta_j \delta_{a_j}$, $\beta_j > 0$ and $\sum_j \beta_j < \infty$. $\theta\Lambda = 0$ implies $\int_{\mathbb{R}^d} \chi(x)\theta(x)\, d\Lambda(x) = \sum_j \beta_j \chi(a_j)\theta(a_j) = 0$ for any continuous function χ in $\mathbb{R}^d$. This implies $\theta(a_j) = 0,\ \forall\, a_j \in \Omega$. ■

Lemma 15 provides conditions under which $\gamma_{\mathcal{F}}(P,Q) = 0$ when $P \neq Q$. It shows that the kernel k cannot distinguish between P and Q if P is related to Q by condition *(i)*. Condition *(ii)* in Lemma 15 says that θ has to be chosen such that its support is disjoint with that of the kernel spectrum. This is what is precisely captured by *(c)* in Proposition 16. So, for a given Q, one can construct P such that $P \neq Q$ and $\gamma_{\mathcal{F}}(P,Q) = 0$ by choosing θ that satisfies the properties in Proposition 16. However, P should be a positive distribution so that it corresponds to a positive measure.[11] Therefore,

[10] If Λ is absolutely continuous w.r.t. the Lebesgue measure, then Ψ represents the Radon-Nikodym derivative of Λ w.r.t. the Lebesgue measure. In such a case, ψ is the Fourier transform of Ψ in the usual sense; i.e., $\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\Psi(\omega)\, dm_d(\omega)$. On the other hand, if Ψ is the distributional derivative of Λ, then Ψ is a symbolic representation of the derivative of Λ and will make sense only under the integral sign.

[11] A positive distribution is defined to be as the one that takes nonnegative values on nonnegative test functions. So, $D \in \mathscr{D}'_d(M)$

θ should also be such that $q + \check{\theta}$ is a positive distribution. Imposing such a constraint on θ is not straightforward, and therefore Lemma 15 does not provide a procedure to construct $P \neq Q$ given Q. However, by imposing some conditions on P and Q, we obtain the following result wherein the conditions on θ can be explicitly specified, yielding a procedure to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

Lemma 17 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of probability measures on $\mathbb{R}^d$ with characteristic functions either absolutely integrable or square integrable, i.e., for any $P \in \mathfrak{D}$, $\phi_P \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Suppose that k satisfies Assumption 1 and supp$(\Lambda) \subsetneq \mathbb{R}^d$. Then for any $Q \in \mathfrak{D}$, $\exists P \neq Q$, $P \in \mathfrak{D}$ given by*

$$p = q + \check{\theta} \tag{16}$$

such that $\gamma_{\mathcal{F}}(P, Q) = 0$ if and only if there exists a non-zero function $\theta : \mathbb{R}^d \to \mathbb{C}$ that satisfies the following conditions:

(i) $\theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ *is conjugate symmetric,*

(ii) $\check{\theta} \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$,

(iii) $\theta\Lambda = 0$,

(iv) $\theta(0) = 0$,

(v) $\inf_{x \in \mathbb{R}^d} \{\check{\theta}(x) + q(x)\} \geq 0$.

Proof: ($\Rightarrow$) Suppose there exists a non-zero function θ satisfying *(i) – (v)*. We need to show that $p = q + \check{\theta}$ is in $\mathfrak{D}$ for $q \in \mathfrak{D}$ and $\gamma_{\mathcal{F}}(P, Q) = 0$.

For any $Q \in \mathfrak{D}$, $\phi_Q \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$. When $\phi_Q \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$, the Riemann-Lebesgue lemma (Lemma 23) implies that $q = [\overline{\phi}_Q]^\vee \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$. When $\phi_Q \in L^2(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$, the Fourier transform in the L^2 sense[12] implies that $q = [\overline{\phi}_Q]^\vee \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Therefore, $q \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$. Define $p := q + \check{\theta}$. Clearly $p \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$. In addition, $\overline{\phi}_P = \hat{p} = \hat{q} + \hat{\check{\theta}} = \overline{\phi}_Q + \theta \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$. Since θ is conjugate symmetric, $\check{\theta}$ is real valued and so is p. Consider $\int_{\mathbb{R}^d} p(x)\, dx = \int_{\mathbb{R}^d} q(x)\, dx + \int_{\mathbb{R}^d} \check{\theta}(x)\, dx = 1 + \theta(0) = 1$. *(v)* implies that p is non-negative. Therefore, P represents a probability measure such that $P \neq Q$ and $P \in \mathfrak{D}$. Since P, Q are probability measures, $\gamma_{\mathcal{F}}(P, Q)$ is computed as $\gamma_{\mathcal{F}}(P, Q) = \|[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda]^\vee\|_{\mathcal{H}} = \|[\theta\Lambda]^\vee\|_{\mathcal{H}} = 0$.

($\Leftarrow$) Suppose that $P, Q \in \mathfrak{D}$ and $p = q + \check{\theta}$ gives $\gamma_{\mathcal{F}}(P, Q) = 0$. We need to show that θ satisfies *(i) – (v)*.

$P, Q \in \mathfrak{D}$ implies $\phi_P, \phi_Q \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ and $p, q \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$. Therefore, $\theta = \overline{\phi}_P - \overline{\phi}_Q \in (L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)) \cap C_b(\mathbb{R}^d)$ and $\check{\theta} = p - q \in L^1(\mathbb{R}^d) \cap (L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d))$. By Lemma 20, ϕ_P and ϕ_Q are conjugate symmetric and so is θ. Therefore θ satisfies *(i)* and $\check{\theta}$ satisfies *(ii)*. θ satisfies *(iv)* as $\theta(0) = \int_{\mathbb{R}^d} \check{\theta}(x)\, dx = \int_{\mathbb{R}^d} (p(x) - q(x))\, dx = 0$. Non-negativity of p yields *(v)*. $\gamma_{\mathcal{F}}(P, Q) = 0$ implies *(iii)*, with a proof similar to that of Lemma 15. ■

Remark 18 *Conditions (iii) and (iv) in Lemma 17 are the same as those of Proposition 16. Conditions (i) and (ii) are required to satisfy our assumption $P, Q \in \mathfrak{D}$ and Eq. (16). Condition (v) ensures that P is a positive measure, which was the condition difficult to impose in Lemma 15.*

In the above result, we restricted ourselves to probability measures P with characteristic functions ϕ_P in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. This ensures that the inverse Fourier transform of ϕ_P exists in the L^1 or L^2 sense. Without this assumption, ϕ_P is not guaranteed to have a Fourier transform in the L^1 or L^2 sense, and therefore has to be treated as a tempered distribution for the purpose of computing its Fourier transform. This implies $\theta = \overline{\phi}_P - \overline{\phi}_Q$ has to be treated as a tempered distribution, which is the setting in Lemma 15. Since we wanted to avoid dealing with distributions where the required positivity constraint is difficult to impose, we restricted ourselves to $\mathfrak{D}$.[13] Though this result explicitly captures the conditions on θ, it is a very restricted result as it only deals with continuous (a.e.) probability measures. However, we use this result in Lemma 19 to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

Lemmas 15 and 17 are the main results that provide conditions for the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. This means that if there exists a θ satisfying these conditions, then k cannot distinguish between P and Q where P is defined as in Eq. (16). Thus, the existence (resp. non-existence) of θ results in a non-injective (resp. injective) map Π. It is clear from Lemmas 15 and 17 that the dependence of $\gamma_{\mathcal{F}}$ on the kernel appears in the form of the support of the kernel spectrum. Therefore, two scenarios exist: *(a)* supp$(\Lambda) = \mathbb{R}^d$ and *(b)* supp$(\Lambda) \subsetneq \mathbb{R}^d$. The case of supp$(\Lambda) = \mathbb{R}^d$ is addressed by Theorem 7 while that of supp$(\Lambda) \subsetneq \mathbb{R}^d$ is addressed by Theorem 9. Using Lemma 17, the following result proves the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$ while using a kernel with supp$(\Lambda) \subsetneq \mathbb{R}^d$.

Lemma 19 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of all non-compactly supported probability measures on $\mathbb{R}^d$ with characteristic functions in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Suppose k satisfies Assumption 1 and supp$(\Lambda) \subsetneq \mathbb{R}^d$. Then $\exists P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.*

is a positive distribution if $D(\varphi) \geq 0$ for $0 \leq \varphi \in \mathscr{D}_d(M)$. If μ is a positive measure that is locally finite, then $D_\mu(\varphi) = \int_M \varphi\, d\mu$ defines a positive distribution. Conversely, every positive distribution comes from a locally finite positive measure [Str03, §6.4].

[12] If $f \in L^2(\mathbb{R}^d)$, the Fourier transform $\mathcal{F}[f] := \hat{f}$ of f is defined to be the limit, in the L^2-norm, of the sequence $\{\hat{f}_n\}$ of Fourier transforms of any sequence $\{f_n\}$ of functions belonging to $\mathscr{S}_d$, such that f_n converges in the L^2-norm to the given function $f \in L^2(\mathbb{R}^d)$, as $n \to \infty$. The function $\hat{f}$ is defined almost everywhere on $\mathbb{R}^d$ and belongs to $L^2(\mathbb{R}^d)$. Thus, $\mathcal{F}$ is a linear operator, mapping $L^2(\mathbb{R}^d)$ into $L^2(\mathbb{R}^d)$.

[13] Choosing $\mathfrak{D}$ to be the set of all probability measures with characteristic functions in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$ is the best possible restriction that avoids treating θ as a tempered distribution. The classical Fourier transforms on $\mathbb{R}^d$ are defined for functions in $L^p(\mathbb{R}^d)$, $1 < p \leq 2$. For $p > 2$, the only reasonable way to define Fourier transforms on $L^p(\mathbb{R}^d)$ is through distribution theory.

Proof: We claim that there exists a non-zero function, θ satisfying *(i) – (v)* in Lemma 17 which therefore proves the result. Consider the following function, $g_{\beta,\omega_0} \in C^\infty(\mathbb{R}^d)$ supported in $[\omega_0 - \beta, \omega_0 + \beta]$,

$$g_{\beta,\omega_0}(\omega) = \prod_{j=1}^{d} \mathbb{1}_{[-\beta_j,\beta_j]}(\omega_j - \omega_{0,j})\, e^{-\frac{\beta_j^2}{\beta_j^2 - (\omega_j - \omega_{0,j})^2}}, \quad (17)$$

where $\omega = (\omega_1, \ldots, \omega_d)$, $\omega_0 = (\omega_{0,1}, \ldots, \omega_{0,d})$ and $\beta = (\beta_1, \ldots, \beta_d)$. Since $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, there exists an open set $U \subset \mathbb{R}^d$ on which Λ is null. So, there exists β and $\omega_0 \neq 0$ with $\omega_0 > \beta$ such that $[\omega_0 - \beta, \omega_0 + \beta] \subset U$. Choose $\theta = \alpha(g_{\beta,\omega_0} + g_{\beta,-\omega_0})$, $\alpha \in \mathbb{R}\backslash\{0\}$, which implies $\text{supp}(\theta) = [-\omega_0 - \beta, -\omega_0 + \beta] \cup [\omega_0 - \beta, \omega_0 + \beta]$ is compact. Therefore, by the Paley-Wiener theorem (Theorem 24), $\check{\theta}$ is a rapidly decaying function, i.e., $\check{\theta} \in \mathscr{S}_d$. Since $\theta(0) = 0$ (by construction), $\check{\theta}$ will take negative values. However, $\check{\theta}$ decays faster than some $Q \in \mathfrak{D}$ of the form $q(x) \propto \prod_{j=1}^{d} \frac{1}{1+|x_j|^{l+\epsilon}}$, $\forall\, l \in \mathbb{N}$, $\epsilon > 0$ where $x = (x_1, \ldots, x_d)$. It can be verified that θ satisfies conditions *(i) – (v)* in Lemma 17. We conclude, there exists a non-zero θ as claimed earlier, which completes the proof. ■

The above result shows that k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ is not characteristic to the class of non-compactly supported probability measures on $\mathbb{R}^d$ with characteristic functions in either $L^1(\mathbb{R}^d)$ or $L^2(\mathbb{R}^d)$.

6.2 Main theorems: Proofs

We are now in a position to prove Theorems 7 and 9.

6.2.1 Proof of Theorem 7

($\Rightarrow$) Let $\text{supp}(\Lambda) = \mathbb{R}^d$. k is a characteristic kernel to $\mathfrak{S}$ if $\gamma_{\mathcal{F}}(P, Q) = 0 \Leftrightarrow P = Q$ for $P, Q \in \mathfrak{S}$. We only need to show the implication $\gamma_{\mathcal{F}}(P, Q) = 0 \Rightarrow P = Q$ as the other direction is trivial.

Assume that $\exists\, P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Then by Lemma 15, $\exists\, \theta$ satisfying *(i)* and *(ii)* given in Lemma 15. By Proposition 16, $\theta\Lambda = 0$ implies $\text{supp}(\theta) \subset \overline{\mathbb{R}^d\backslash\text{supp}(\Lambda)}$. Since $\text{supp}(\Lambda) = \mathbb{R}^d$ and θ is a uniformly continuous function in $\mathbb{R}^d$, we have $\text{supp}(\theta) = \emptyset$ which means $\theta = 0$ a.e. Therefore, by *(i)* of Theorem 15, we have $P = Q$, leading to a contradiction. Thus, $\nexists\, P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

($\Leftarrow$) Suppose k is characteristic to $\mathfrak{S}$. We then need to show that $\text{supp}(\Lambda) = \mathbb{R}^d$. This is equivalent to proving that k is not characteristic to $\mathfrak{S}$ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Let $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Choose $\mathfrak{D} \subsetneq \mathfrak{S}$ as the set of all non-compactly supported probability measures on $\mathbb{R}^d$ with characteristic functions in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. By Lemma 19, $\exists\, P \neq Q$, $P, Q \in \mathfrak{D} \subsetneq \mathfrak{S}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Therefore, k is not characteristic to $\mathfrak{S}$. ■

6.2.2 Proof of Theorem 9

Suppose $\exists P \neq Q$, $P, Q \in \mathfrak{D} \subsetneq \mathfrak{S}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Then by Lemma 15, there exists a $\theta \in \mathscr{S}'_d$ such that $\check{\theta} = p - q$ where p and q are the distributional derivatives of P and Q, respectively. Since $P, Q \in \mathfrak{D}$, we can apply Lemma 17 and so θ is a non-zero function that satisfies conditions *(i) – (v)* in Lemma 17. The condition $\theta\Lambda = 0$ implies $\text{supp}(\theta) \subset \overline{\mathbb{R}^d\backslash\text{supp}(\Lambda)}$. Since $\text{supp}(\Lambda)$ has a non-empty interior, we have $\text{supp}(\theta) \subsetneq \mathbb{R}^d$. Thus, there exists an open set, $U \subset \mathbb{R}^d$ such that $\theta(x) = 0$, $\forall\, x \in U$. By Lemma 25, this means that $\check{\theta}$ is not compactly supported in $\mathbb{R}^d$. Condition *(iv)* implies $\int_{\mathbb{R}^d} \check{\theta}(x)\, dx = 0$, which means that $\check{\theta}$ takes negative values. Since q is compactly supported in $\mathbb{R}^d$, $q(x) + \check{\theta}(x) < 0$ for some $x \in \mathbb{R}^d\backslash\text{supp}(Q)$, which violates condition *(v)* in Lemma 17. In other words, there does not exist a non-zero θ that satisfies conditions *(i) – (v)* in Lemma 17, thereby leading to a contradiction. ■

As discussed in §4.1, the condition that $\text{supp}(\Lambda)$ has a non-empty interior is important for Theorem 9 to hold. This is because if $\text{supp}(\Lambda)$ has an empty interior, then $\text{supp}(\theta) = \mathbb{R}^d$. In principle, one can construct such a θ by selecting $\theta \in \mathscr{S}_d$ so that it satisfies conditions *(i) – (iv)* of Lemma 17 while satisfying the decay conditions (Eq. (29) and Eq. (30)) given in the Paley-Wiener theorem (see Theorem 24). Therefore, by the Paley-Wiener theorem, $\check{\theta}$ is a C^∞ function with compact support. If θ is chosen such that $\text{supp}(\check{\theta}) \subset \text{supp}(Q)$, then condition *(v)* of Theorem 17 will be satisfied. Thus, one can construct $P \neq Q$, $P, Q \in \mathfrak{D}$ ($\mathfrak{D}$ being defined in Theorem 9) such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Note that conditions *(i)* and *(ii)* of Lemma 17 are automatically satisfied (except for conjugate symmetry) by choosing $\theta \in \mathscr{S}_d$. However, choosing θ such that it is also an entire function (so that the Paley-Wiener theorem can be applied) is not straightforward. In the following, we provide a simple example to show that $P \neq Q$, $P, Q \in \mathfrak{D}$ can be constructed such that $\gamma_{\mathcal{F}}(P, Q) = 0$, where $\mathcal{F}$ corresponds to a unit ball in an RKHS $(\mathcal{H}, k)$ induced by a periodic translation-invariant kernel for which $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has an empty interior.

Example 3 *Let Q be a uniform distribution on $[-\beta, \beta] \subset \mathbb{R}$, i.e., $q(x) = \frac{1}{2\beta}\mathbb{1}_{[-\beta,\beta]}(x)$ with its characteristic function, $\phi_Q(\omega) = \frac{1}{\beta\sqrt{2\pi}} \frac{\sin(\beta\omega)}{\omega}$ in $L^2(\mathbb{R})$. Let ψ be the Dirichlet kernel with period τ, where $\tau \leq \beta$, i.e., $\psi(x) = \frac{\sin\frac{(2l+1)\pi x}{\tau}}{\sin\frac{\pi x}{\tau}}$ and $\Psi(\omega) = \sum_{j=-l}^{l} \delta\left(\omega - \frac{2\pi j}{\tau}\right)$ with supp$(\Psi) = \{\frac{2\pi j}{\tau},\, j \in \{0, \pm 1, \ldots, \pm l\}\}$. Clearly, supp$(\Psi)$ has an empty interior. Let θ be*

$$\theta(\omega) = \frac{8\sqrt{2}\alpha}{i\sqrt{\pi}} \sin\left(\frac{\omega\tau}{2}\right) \frac{\sin^2\left(\frac{\omega\tau}{4}\right)}{\tau\omega^2}, \quad (18)$$

with $\alpha \leq \frac{1}{2\beta}$. It is easy to verify that $\theta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ and so θ satisfies (i) in Lemma 17. Since $\theta(\omega) = 0$ at $\omega = \frac{2\pi l}{\tau}$, $l \in \mathbb{Z}$, θ also satisfies (iii) and (iv) in Lemma 17. $\check{\theta}$ is given by

$$\check{\theta}(x) = \begin{cases} \frac{2\alpha\left|x + \frac{\tau}{2}\right|}{\tau} - \alpha, & -\tau \leq x \leq 0 \\ \alpha - \frac{2\alpha\left|x - \frac{\tau}{2}\right|}{\tau}, & 0 \leq x \leq \tau \\ 0, & \textit{otherwise,} \end{cases} \quad (19)$$

where $\check{\theta} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ satisfies (ii) in Lemma 17. Now, consider $p = q + \check{\theta}$ which is given as

$$p(x) = \begin{cases} \frac{1}{2\beta}, & x \in [-\beta, -\tau] \cup [\tau, \beta] \\ \frac{2\alpha\left|x + \frac{\tau}{2}\right|}{\tau} + \frac{1}{2\beta} - \alpha, & x \in [-\tau, 0] \\ \alpha + \frac{1}{2\beta} - \frac{2\alpha\left|x - \frac{\tau}{2}\right|}{\tau}, & x \in [0, \tau] \\ 0, & \textit{otherwise.} \end{cases}$$

Clearly, $p(x) \geq 0$, $\forall\, x$ and $\int_{\mathbb{R}} p(x)\, dx = 1$. $\phi_P = \phi_Q + \theta = \phi_Q + i\theta_I$ where $\theta_I = Im[\theta]$ and $\phi_P \in L^2(\mathbb{R})$. We have therefore constructed $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, where P and Q are compactly supported in $\mathbb{R}$ with characteristic functions in $L^2(\mathbb{R})$.

The condition of the compact support for probability measures mentioned in Theorem 9 is also critical for the result to hold. If this condition is relaxed, then k with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ is no longer characteristic to $\mathfrak{D}$, as shown in Lemma 19.

7 Concluding Remarks

Previous works have studied the Hilbert space embedding for probability measures using universal kernels, which form a restricted family of positive definite kernels. These works showed that if the kernel is universal, then the embedding function from the space of probability measures to a reproducing kernel Hilbert space is injective. In this paper, we extended this approach to a larger family of kernels which are translation-invariant on $\mathbb{R}^d$. We showed that the support of the Fourier spectrum of the kernel determines whether the embedding is injective. In particular, the necessary and sufficient condition for the embedding to be injective is that the Fourier spectrum of the kernel should have the entire domain as its support. Our study in this paper was limited to kernels and probability measures that are defined on $\mathbb{R}^d$, and the results have been derived using Fourier analysis in $\mathbb{R}^d$. Since Fourier theory is available for more general groups apart from $\mathbb{R}^d$, one direction for future work is to extend the analysis to positive definite kernels defined on other groups.

Appendix A Supplementary Results

We show five supplementary results used to prove the results in §4 and §6. The first two are basic, and deal with the Fourier transform of a measure and the convolution theorem. The remaining three (the Riemann-Lebesgue lemma, the Paley-Wiener theorem, and its corollary) are stated without proof.

Lemma 20 (Fourier transform of a measure) *Let μ be a finite Borel measure on $\mathbb{R}^d$. The Fourier transform of μ is a tempered distribution given by*

$$\hat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^T x}\, d\mu(x), \quad \forall\, \omega \in \mathbb{R}^d \tag{20}$$

which is a bounded, uniformly continuous function on $\mathbb{R}^d$. In addition, $\hat{\mu}$ satisfies the following properties:

(i) $\overline{\hat{\mu}(\omega)} = \hat{\mu}(-\omega)$, $\forall\, \omega \in \mathbb{R}^d$,

(ii) $\hat{\mu}(\omega) = \hat{\mu}(-\omega)$, $\forall\, \omega \in \mathbb{R}^d$ *if and only if* $D_\mu(\varphi) = D_\mu(\tilde{\varphi})$, $\forall\, \varphi \in \mathscr{S}_d$ *where D_μ is the tempered distribution defined by μ and $\tilde{\varphi}(x) := \varphi(-x)$, $\forall\, x \in \mathbb{R}^d$.*

Proof: Let D_μ denote a tempered distribution defined by μ. For $\varphi \in \mathscr{S}_d$, we have $\widehat{D_\mu}(\varphi) = D_\mu(\hat{\varphi}) = \int_{\mathbb{R}^d} \hat{\varphi}(\omega)\, d\mu(\omega) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i\omega^T x} \varphi(x)\, dm_d(x)\, d\mu(\omega)$. From Fubini's theorem,

$$\widehat{D_\mu}(\varphi) = \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\mu(\omega) \right] \varphi(x)\, dm_d(x), \tag{21}$$

which proves Eq. (20). Clearly $\hat{\mu}$ is bounded as $|\hat{\mu}(\omega)| \leq 1$. By Lebesgue's dominated convergence theorem, $\hat{\mu}$ is uniformly continuous on $\mathbb{R}^d$ as $\lim_{h\to 0} |\hat{\mu}(\omega + h) - \hat{\mu}(\omega)| \leq \lim_{h\to 0} \int_{\mathbb{R}^d} |e^{-jh^T x} - 1|\, d\mu(x) = 0$, for any $\omega \in \mathbb{R}^d$.

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

(ii) ($\Rightarrow$) For $\varphi \in \mathscr{S}_d$, $\widehat{D_\mu}(\varphi) = D_\mu(\hat{\varphi}) = \int_{\mathbb{R}^d} \hat{\varphi}(x)\, d\mu(x) = \int_{\mathbb{R}^d} \hat{\mu}(x)\varphi(x)\, dm_d(x)$. Since $\hat{\varphi} \in \mathscr{S}_d$ and $D_\mu(\varphi) = D_\mu(\tilde{\varphi})$, $\forall\, \varphi \in \mathscr{S}_d$, we have $D_\mu(\hat{\varphi}) = D_\mu(\tilde{\hat{\varphi}}) = \int_{\mathbb{R}^d} \hat{\varphi}(-x)\, d\mu(x)$. Substituting for $\hat{\varphi}(-x)$, we get

$$D_\mu(\hat{\varphi}) = \int_{\mathbb{R}^d} \hat{\mu}(-x)\varphi(x)\, dm_d(x) = \int_{\mathbb{R}^d} \hat{\mu}(x)\varphi(x)\, dm_d(x),$$

for every $\varphi \in \mathscr{S}_d$, which implies $\hat{\mu}(x) = \hat{\mu}(-x)$, $\forall\, x \in \mathbb{R}^d$.

($\Leftarrow$) For $\varphi \in \mathscr{S}_d$, we have $D_\mu(\varphi) = (\widehat{D}_\mu)^\vee(\varphi) = \widehat{D}_\mu(\check{\varphi}) = \int_{\mathbb{R}^d} \hat{\mu}(x)\check{\varphi}(x)\, dm_d(x) = \int_{\mathbb{R}^d} \hat{\mu}(-x)\check{\varphi}(x)\, dm_d(x)$. Applying Fubini's theorem after substituting for $\hat{\mu}(-x)$ and $\check{\varphi}(x)$ gives

$$\begin{aligned} D_\mu(\varphi) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \delta(y + \omega)\varphi(y)\, dm_d(y)\, d\mu(\omega) \\ &= \int_{\mathbb{R}^d} \varphi(-\omega)\, d\mu(\omega) = D_\mu(\tilde{\varphi}), \end{aligned}$$

for every $\varphi \in \mathscr{S}_d$. ■

Remark 21 *(a) Property (i) in Lemma 20 shows that the Fourier transform of a finite Borel measure on $\mathbb{R}^d$ is "conjugate symmetric", which means that $Re[\hat{\mu}]$ is an even function and $Im[\hat{\mu}]$ is an odd function.*

(b) Property (ii) shows that real symmetric tempered distributions have real symmetric Fourier transforms. This can be easily understood when μ is absolutely continuous w.r.t. the Lebesgue measure. Suppose $d\mu = \Psi\, dm_d$. Then property (ii) implies that $\hat{\mu}$ is real and symmetric if and only if Ψ is real and symmetric.

The following result is popularly known as the *convolution theorem*. Before providing the result, we first define convolution: if f and g are complex functions in $\mathbb{R}^d$, their convolution $f * g$ is

$$(f * g)(x) = \int_{\mathbb{R}^d} f(y) g(x - y)\, dy, \tag{22}$$

provided that the integral exists for almost all $x \in \mathbb{R}^d$, in the Lebesgue sense. Let μ be a finite Borel measure on $\mathbb{R}^d$ and f be a bounded measurable function on $\mathbb{R}^d$. The convolution of f and μ, $f * \mu$, which is a bounded measurable function, is defined by

$$(f * \mu)(x) = \int_{\mathbb{R}^d} f(x - y)\, d\mu(y). \tag{23}$$

Theorem 22 (Convolution Theorem) *Let μ be a finite Borel measure and f be a bounded function on $\mathbb{R}^d$. Suppose f is written as*

$$f(x) = \int_{\mathbb{R}^d} e^{ix^T\omega}\, d\Lambda(\omega), \tag{24}$$

with a finite Borel measure Λ on $\mathbb{R}^d$. Then

$$(f * \mu)^\wedge = \hat{\mu}\Lambda, \tag{25}$$

where the right hand side is a finite Borel measure[14] *and the equality holds as a tempered distribution.*

Proof: Since the Fourier and inverse Fourier transform give one-to-one correspondence of $\mathscr{S}'_d$, it suffices to show

$$f * \mu = (\hat{\mu}\Lambda)^\vee. \tag{27}$$

For an arbitrary $\varphi \in \mathscr{S}_d$,

$$(\hat{\mu}\Lambda)^\vee(\varphi) = (\hat{\mu}\Lambda)(\check{\varphi}) = \int_{\mathbb{R}^d} \check{\varphi}(x)\hat{\mu}(x)\, d\Lambda(x). \tag{28}$$

Substituting for $\hat{\mu}$ in Eq. (28) and applying Fubini's theorem, we have $(\hat{\mu}\Lambda)^\vee(\varphi) =$

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} e^{i(\omega-y)^T x}\, d\Lambda(x)\right]\varphi(\omega)\, dm_d(\omega)\, d\mu(y),$$

which reduces to $\int_{\mathbb{R}^d}[\int_{\mathbb{R}^d} f(\omega - y)\, d\mu(y)]\varphi(\omega)\, dm_d(\omega) = (f * \mu)(\varphi)$ and therefore proves Eq. (27). ■

The following result, called the Riemann-Lebesgue lemma, is quoted from [Rud91, Theorem 7.5].

Lemma 23 (Riemann-Lebesgue) *If $f \in L^1(\mathbb{R}^d)$, then $\hat{f} \in C_b(\mathbb{R}^d)$, and $\|\hat{f}\|_\infty \leq \|f\|_1$.*

The following theorem is a version of the *Paley-Wiener theorem* for C^∞ functions, and is proved in [Str03, Theorem 7.2.2].

Theorem 24 (Paley-Wiener) *Let f be a C^∞ function supported in $[-\beta, \beta]$. Then $\hat{f}(\omega + i\sigma)$ is a entire function of exponential type β, i.e., $\exists\, C$ such that*

$$\left|\hat{f}(\omega + i\sigma)\right| \leq Ce^{\beta|\sigma|}, \tag{29}$$

and $\hat{f}(\omega)$ is rapidly decreasing, i.e., $\exists\, c_n$ such that

$$\left|\hat{f}(\omega)\right| \leq \frac{c_n}{(1 + |\omega|)^n}, \forall\, n \in \mathbb{N}. \tag{30}$$

Conversely, if $F(\omega + i\sigma)$ is an entire function of exponential type β, and $F(\omega)$ is rapidly decaying, then $F = \hat{f}$ for some such function f.

The following lemma is a corollary of the Paley-Wiener theorem, and is proved in [Mal98, Theorem 2.6].

Lemma 25 ([Mal98]) *If $g \neq 0$ has compact support, then its Fourier transform $\hat{g}$ cannot be zero on a whole interval. Similarly, if $\hat{g} \neq 0$ has compact support, then g cannot be zero on a whole interval.*

[14] Let μ be a finite Borel measure and f be a bounded measurable function on $\mathbb{R}^d$. We then define a finite Borel measure $f\mu$ by

$$(f\mu)(E) = \int_{\mathbb{R}^d} I_E(x) f(x)\, d\mu(x), \tag{26}$$

where E is an arbitrary Borel set and I_E is its indicator function.

References

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[BJ02] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK, 2002.

[FGSS08] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

[GBR+07] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[GHS+05] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, December 2005.

[GSB+04] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, B. Schölkopf, and N. Logothetis. Behaviour and convergence of the constrained covariance. Technical Report 130, MPI for Biological Cybernetics, 2004.

[GW99] C. Gasquet and P. Witomski. *Fourier Analysis and Applications*. Springer-Verlag, New York, 1999.

[Mal98] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.

[RS72] M. Reed and B. Simon. *Functional Analysis*. Academic Press, New York, 1972.

[Rud91] W. Rudin. *Functional Analysis*. McGraw-Hill, USA, 1991.

[SGSS07] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany, 2007.

[Sho00] G. R. Shorack. *Probability for Statisticians*. Springer-Verlag, New York, 2000.

[SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[Ste02] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

[Str03] R. S. Strichartz. *A Guide to Distribution Theory and Fourier Transforms*. World Scientific Publishing, Singapore, 2003.

[Wen05] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.