# Injective Hilbert Space Embeddings of Probability Measures

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

[1]Dept. of ECE, UC San Diego, La Jolla, CA 92093, USA.
bharathsv@ucsd.edu,gert@ece.uscd.edu

[2]MPI for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany.
{arthur,bernhard.schoelkopf}@tuebingen.mpg.de

[3]Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.
fukumizu@ism.ac.jp

February 20, 2008

**Abstract**

A Hilbert space embedding for probability measures has recently been proposed, which has been successfully applied in many statistical applications like dimensionality reduction, homogeneity testing, independence testing etc. This method involves embedding any probability measure as a mean element in some reproducing kernel Hilbert space (RKHS). The embedding function has been proven to be injective when the reproducing kernel is universal. The injective nature of the embedding function has been used to define a metric on the space of probability distributions defined on a compact metric space.

In the present work, we broaden this approach by determining explicitly the properties of the feature space for which the embedding function is injective, considering in particular non-universal kernels. We restrict ourselves to convolution kernels, and relate the metric (defined on the space of probability distributions) to the Fourier spectrum of the kernel and characteristic functions of probability measures. We show that the support of the kernel spectrum is an important quantity that determines the behavior of the embedding function and the embedding function is injective if and only if the kernel spectrum has the entire domain as its support. We also deduce conditions on the kernel and probability measures for which the associated embedding function is not injective.

## 1 Introduction

The concept of distance between probability measures is a fundamental one and has lot of 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, suppose $(M, \rho)$ be a metric space equipped with its Borel $\sigma$-field $\mathcal{M}_\rho$ and let $\mathfrak{S}$ be the set of all probability measures defined on $(M, \mathcal{M}_\rho)$ with $\gamma$ as its metric, i.e., $(\mathfrak{S}, \gamma)$ is a metric space. Then $P_n$ is said to converge in distribution to $P$ if and only if $\gamma(P_n, P) \overset{n\to\infty}{\longrightarrow} 0$, where $\{P_n : n \geq 1\}$, $P \in \mathfrak{S}$. When $M$ is separable, examples for $\gamma$ include the *Lévy-Prohorov distance* and the *dual-bounded Lipschitz distance* (*Dudley metric*), which metrize the weak convergence of probability measures [4, Chapter 11]. Other popular examples for $\gamma$ include the *total variation distance* and the *Hellinger distance*, which yield a stronger notion of convergence of probability measures than that of the previously mentioned metrics (see [19, Chapter 19, §2]).

In statistics, the notion of distance between probability measures is used in a variety of applications that include homogeneity test (two-sample problem), independence test, goodness-of-fit test, etc. The two-sample problem involves testing the null hypothesis $H_0 : P = Q$ versus the alternative $H_1 : P \neq Q$ using random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ drawn independently from distributions $P$ and $Q$ on

a measurable space $(M, \mathcal{M})$. The problems of testing independence and goodness-of-fit can be posed as variations of the two-sample problem. If $\gamma$ is a metric (or more generally a semi-metric) on $\mathfrak{S}$, then $\gamma(P, Q)$ can be used as a test statistic to address the two-sample problem. This is because the test statistic, in a population setting, takes a unique and distinctive value of zero only when $P = Q$ and therefore, the test can be reduced to testing $H_0 : \gamma(P, Q) = 0$ versus the alternative $H_1 : \gamma(P, Q) > 0$. Examples for $\gamma$ include the *variational distance* (*total variation*), the *Hellinger distance*, etc. which are specific instances of the generalized $f$-divergence introduced by Csiszár [3]. When $M = \mathbb{R}$, the popular choice for $\gamma$ is the *Kolmogorov distance*, which measures the maximal difference between the cumulative distribution functions (cdf) of $P$ and $Q$.

Recently, Smola *et al.* [20] proposed a Hilbert space embedding for probability measures, which involves embedding any probability measure as a mean element in some reproducing kernel Hilbert space (RKHS). Formally, let $(\mathcal{H}, k)$ be an RKHS with $k$ as its reproducing kernel (see [1] for details on positive definite kernels and RKHS). Then the embedding of $P \in \mathfrak{S}$ into $\mathcal{H}$ is defined by the mapping, $\Pi : \mathfrak{S} \to \mathcal{H}$ with $\Pi[P] = \int_M k(\cdot, x)\, dP(x)$, which is a mean element in $\mathcal{H}$. $\Pi$ is shown to be injective when $k$ is a universal kernel.[1] Therefore, $\Pi$ can be used to define a metric on $\mathfrak{S}$ by letting $\gamma(P, Q) = \|\Pi[P] - \Pi[Q]\|_{\mathcal{H}}$, referred to as the *maximum mean discrepancy* (MMD) [11], where $\|.\|_{\mathcal{H}}$ represents the RKHS norm. This approach has been successfully applied in testing for homogeneity [11], independence [12], conditional independence [8], kernel dimensionality reduction [7] etc. One drawback of this approach is that if $k$ is non-universal or if the domain of $k$ is non-compact, $\Pi$ is not guaranteed to be injective and so $\gamma(.,.)$ may not be a metric but a pseudometric,[2] which means $\gamma(P, Q) = 0$ does not imply $P = Q$. In such a case, this approach is not particularly useful in aforementioned applications like homogeneity testing, independence testing etc., as they rely on the fact that $\gamma$ is a metric (or atleast a semi-metric). This means, when $\Pi$ is not injective, the two-sample test can decide $H_0 : P = Q$ even when $P \neq Q$, i.e., the test fails to distinguish between two different probability distributions.

Given a positive definite kernel, it is not always easy to verify its universality and therefore it is not clear whether the given kernel yields a injective map, $\Pi$ or not. So, one would like to have a simple verifiable rule that characterizes the injective nature of $\Pi$. In this paper, we address this issue by determining explicitly the properties of the feature space for which the embedding function is injective (or non-injective), considering in particular non-universal kernels. Broadly, we derive an equivalence between the set of kernels and the set of probability measures defined on a metric space (not necessarily compact) for which $\gamma(P, Q) = 0$ implies $P = Q$. To this end, in §4, we derive a new formulation for the maximum mean discrepancy by restricting ourselves to convolution kernels and relate it to the Fourier spectrum of the kernel and characteristic functions of probability measures. Using this formulation, in §5, we deduce conditions on the kernel and the set of probability measures for which the embedding function, $\Pi$ is injective (or non-injective). We show that the support of the kernel spectrum is an important quantity that determines the behavior of $\Pi$ and $\Pi$ is injective if and only if the kernel spectrum has the entire domain as its support. We present a fundamental result in §6 that captures the limitation of MMD. In §**??**, we study some applications of MMD, wherein the results derived in §4 and §5 are applied to problems of homogeneity and independence testing. §7 concludes the main body of the paper with some pointers to future work. The results presented in this paper use tools from the *distribution theory* and the Fourier transform theory and so, a brief introduction to these theories is presented in Appendix A, while few technical lemmas are collected in Appendix B.

## 2 Notation

For $M \subset \mathbb{R}^d$ and $\mu$ a Borel measure on $M$, $L^p(M, \mu) := \{f : M \to \mathbb{C} : \int_M |f(x)|^p\, d\mu(x) < \infty\}$, $1 \leq p < \infty$. We will also use $L^p(M)$ for $L^p(M, \mu)$ and $dx$ for $d\mu(x)$ if $\mu$ is the Lebesgue measure on $M$. $C_b(M)$

[1] A continuous positive definite kernel $k$ on a compact metric space $(M, \rho)$ is called *universal* if the corresponding RKHS is dense in the Banach space of continuous functions with sup norm [21, Definition 4]. Examples include the Gaussian and Laplacian kernels which are universal on every compact subset of $\mathbb{R}^d$. The RKHS induced by a universal kernel is called the *universal RKHS*.

[2] A pseudometric space $(M, \rho)$ is a set $M$ together with a non-negative real-valued function, $\rho : M \times M \to \mathbb{R}_+$ (called a pseudometric) such that, for every $x, y, z \in M$, $(i)\, \rho(x, x) = 0$, $(ii)\, \rho(x, y) = \rho(y, x)$ and $(iii)\, \rho(x, z) \leq \rho(x, y) + \rho(y, z)$. Unlike a metric space, points in a pseudometric space need not be distinguishable, that is one may have $\rho(x, y) = 0$ for distinct values $x \neq y$.

denotes the space of all bounded, continuous, real-valued functions on $M$. The space of all $p$-continuously differentiable real-valued functions on $M$ is denoted by $C^p(M)$, $0 \leq p \leq \infty$. For $x \in \mathbb{C}$, $\overline{x}$ represents the complex conjugate of $x$. $j$ denotes the complex number, $\sqrt{-1}$.

The functional space of infinitely differentiable functions in $\mathbb{R}^d$ whose support is compact 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 $U \in \mathbb{R}^d$, $\mathscr{D}(U)$ 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). Refer to Appendix A for details on $\mathscr{D}_d$, $\mathscr{D}'_d$ $\mathscr{S}_d$, $\mathscr{S}'_d$. $m_d$ 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 $P$, $Pf := \int f\, dP = \int_M f(x)\, dP(x)$. $\delta_x$ represents the Dirac measure at $x$. The symbol $\delta$ 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

In this section, we briefly review the method of Hilbert space embedding of probability distributions proposed by Smola *et al.* [20] and motivate our study to understand its dependence on the properties of the kernel. The maximum mean discrepancy (MMD) defined in §1 can be derived starting from the following result [4, Lemma 9.3.2] related to the weak convergence of probability measures on metric spaces.

**Lemma 1** ([4])**.** *Let $(M, \rho)$ be a metric space, and $P$, $Q$ be two Borel probability measures defined on $M$. Then $P = Q$ if and only if $Pf = Qf$, $\forall\, f \in C_b(M)$.*

Originally, Gretton *et al.* [11] defined the *maximum mean discrepancy* as follows.

**Definition 2.** *Let $\mathcal{F}$ be a class of functions $f : M \to \mathbb{R}$ and $P, Q$ be probability measures defined on $(M, \rho, \mathcal{M}_\rho)$. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples drawn from $P$ and $Q$. Then the maximum mean discrepancy (MMD) and its empirical estimate are defined as*

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

$$\widehat{\gamma_{\mathcal{F}}}(m, n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(X_i) - \frac{1}{n} \sum_{i=1}^{n} f(Y_i) \right|. \tag{2}$$

With this definition, one can derive various metrics (mentioned in §1) that are used to define the 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 on $\mathfrak{S}$. Therefore, the choice of $\mathcal{F}$ determines whether $\gamma_{\mathcal{F}}(P, Q) = 0$ implies $P = Q$ or not. 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, $\rho$-uniformly continuous functions on $M$, by the Portmanteau theorem [19, Chapter 19, Theorem 1.1], $\gamma_{\mathcal{F}}$ is a metric on $\mathfrak{S}$. $\gamma_{\mathcal{F}}$ is a *Dudley metric* [19, Chapter 19, Definition 2.2] if $\mathcal{F}$ is the set of all real-valued, bounded Lipschitz functions defined on $(M, \rho)$ that lie in a unit ball defined by the bounded Lipschitz norm. $\gamma_{\mathcal{F}}$ is the *total variation metric* when $\mathcal{F} = \{\mathbb{1}_A : A \in \mathcal{M}_\rho\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$. If $\mathcal{F} = \{\exp(j\langle t, .\rangle) : t \in \mathbb{R}^d\}$, then $\gamma_{\mathcal{F}}(P, Q)$ reduces to finding the maximal difference between the characteristic functions of $P$ and $Q$. By the uniqueness theorem[3] [4, Theorem 9.5.1], we have $\gamma_{\mathcal{F}}(P, Q) = 0 \Leftrightarrow \phi_P = \phi_Q \Leftrightarrow P = Q$, where $\phi_P$ and $\phi_Q$ represent the characteristic functions of $P$ and $Q$ respectively.[4] Therefore, the function class, $\mathcal{F} = \{\exp(j\langle t, .\rangle) : t \in \mathbb{R}^d\}$ induces a metric on $\mathfrak{S}$. Gretton *et al.* [11, 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 (see footnote 1 for the definition of universal RKHS) defined on a compact metric space. This choice of $\mathcal{F}$ yields an injective embedding function, $\Pi$ as proposed by Smola *et al.* [20].

[3] The uniqueness theorem ensures that the characteristic function uniquely determines the probability distribution.

[4] The characteristic function of a random variable $X$ with distribution $F$ is the complex valued function of a real variable $\omega$ defined by $\phi(\omega) := \int_{\mathbb{R}} \exp(j\omega x)\, dF(x)$, $\forall\, \omega \in \mathbb{R}$.

As aforementioned in §1, $\gamma_{\mathcal{F}}$ can be used as a test statistic in many statistical applications like testing for homogeneity or independence. In a practical scenario, one has access only to the random samples drawn from $P, Q$. In such a case, choosing $\mathcal{F}$ such that $\gamma_{\mathcal{F}}$ in Eq. (1) is a metric on $\mathfrak{S}$ is not sufficient. In addition, $\mathcal{F}$ should be chosen such that $\widehat{\gamma_{\mathcal{F}}}(m,n)$ is not only a consistent estimate of $\gamma_{\mathcal{F}}(P,Q)$ but also has a fast rate of convergence. The fast rate of convergence is important so that $\widehat{\gamma_{\mathcal{F}}}(m,n)$ is a meaningful test statistic in practice where access to data samples is limited. It can be shown that if $\mathcal{F}$ is a universal Glivenko-Cantelli class [5], then $\widehat{\gamma_{\mathcal{F}}}(m,n) \overset{a.s.}{\to} \gamma_{\mathcal{F}}(P,Q)$. In addition, if the Rademacher complexity of $\mathcal{F}$ is well behaved, fast rates of convergence can be achieved. Many of the function classes discussed before are not practical to work in the finite sample setting. The Kolmogorov distance, popularly used as the *Kolmogorov-Smirnov* test statistic is widely used in practice when $M = \mathbb{R}$. The function class, $\mathcal{F} = \{\exp(j\langle t, . \rangle) : t \in \mathbb{R}^d\}$ is not useful in practice as the empirical characteristic function is an extremely poor characteristic function estimator. Though it is an unbiased estimator of the characteristic function, it has unacceptably high variance [6]. [11, Theorem 4] proved that if $\mathcal{F}$ is a unit ball in a universal RKHS with a bounded kernel defined on a compact metric space, then $\widehat{\gamma_{\mathcal{F}}}(m,n)$ converges to $\gamma_{\mathcal{F}}(P,Q)$ at the rate $O(m^{-\frac{1}{2}} + n^{-\frac{1}{2}})$. Therefore, a universal RKHS defined on a compact metric space not only provides a metric structure for $\gamma_{\mathcal{F}}$ but also provides a good finite sample behavior for $\widehat{\gamma_{\mathcal{F}}}(m,n)$. The following result provides a different representation for $\gamma_{\mathcal{F}}$ in Eq. (1) by exploiting the reproducing property of $\mathcal{F}$.

**Theorem 3.** *Let $(\mathcal{H}, k)$ be a RKHS (not necessarily universal) defined on $(M, \rho)$ where $M$ is not necessarily compact with $|k(x,y)| \leq K < \infty, \forall\, x, y \in M$. Suppose $\mathcal{F} = \{\|f\|_{\mathcal{H}} \leq 1 : f \in \mathcal{H}\}$. Then*

$$\gamma_{\mathcal{F}}(P,Q) = \|Pk - Qk\|_{\mathcal{H}}, \tag{3}$$

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

*Proof.* Let $T_P : \mathcal{H} \to \mathbb{R}$ be a functional defined as $T_P[f] = Pf = \int_M f(x)\, dP(x)$ with $\|T_P\| := \sup_{f \in \mathcal{H}} \frac{|T_P[f]|}{\|f\|_{\mathcal{H}}}$. Consider $|T_P[f]| = \left|\int_M f(x)\, dP(x)\right| \leq \int_M |f(x)|\, dP(x) = \int_M |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}|\, dP(x) \leq \int_M \sqrt{k(x,x)}\|f\|_{\mathcal{H}}\, dP(x) \leq \sqrt{K}\|f\|_{\mathcal{H}} < \infty$. We have exploited the reproducing property and boundedness of the kernel to show that $T_P$ is a bounded linear functional on $\mathcal{H}$. Therefore, by the Riesz representation theorem [16, 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 $u \in M$. So, $T_P[k(\cdot,u)] = \langle k(\cdot,u), \lambda_P\rangle_{\mathcal{H}} = \lambda_P(u)$. Therefore, $\lambda_P = T_P[k] = Pk = \int_M k(\cdot, x)\, dP(x)$. Now, consider $\gamma_{\mathcal{F}}(P,Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |Pf - Qf| = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\langle f, \lambda_P - \lambda_Q\rangle_{\mathcal{H}}| = \|\lambda_P - \lambda_Q\|_{\mathcal{H}} = \|Pk - Qk\|_{\mathcal{H}}$. □

**Remark 4.** *If $M$ is compact, the result is straight forward as $Pf = \int_M f(x)\, dP(x) = \int_M \langle f, k(\cdot,x)\rangle_{\mathcal{H}}\, dP(x) = \langle f, \int_M k(\cdot,x)\, dP(x)\rangle_{\mathcal{H}} = \langle f, Pk\rangle_{\mathcal{H}}$, wherein the expectation and the RKHS inner product are interchanged. The result therefore follows by applying the same simplification to $Qf$ and invoking the Cauchy-Schwartz inequality.*

The representation of $\gamma_{\mathcal{F}}$ in Eq. (3) yields the embedding function, $\Pi[P] = \int_M k(\cdot, x)\, dP(x)$ as proposed by Smola *et al.* [20]. When $k$ is universal, $\gamma_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow P = Q, \forall\, P, Q \in \mathfrak{S}$ which means $\Pi[P] = \Pi[Q] \Leftrightarrow P = Q, \forall\, P, Q \in \mathfrak{S}$ and therefore, $\Pi$ is an injective mapping from $\mathfrak{S}$ to $\mathcal{H}$. Now, what if $k$ is non-universal or $M$ is non-compact. The representation of $\gamma_{\mathcal{F}}$ in Eq. (3) still holds but it need not have to be a metric on $\mathfrak{S}$, in other words, $\Pi$ is not guaranteed to be injective. As we argued in §1 that such a choice of $k$ would make the homogeneity or independence test to fail. One obvious question to ask is "For what class of kernels is $\Pi$ injective?". To understand this in detail, we are interested in the following questions which we address in this paper.

1. Are there kernels apart from universal kernels for which $\Pi$ is injective on $\mathfrak{S}$? If so, what is their characterization?

2. Are there kernels for which $\Pi$ is not injective on $\mathfrak{S}$? In other words, for what class of kernels can we construct $P \neq Q$, $P, Q \in \mathfrak{S}$ such that $\gamma_{\mathcal{F}}(P,Q) = 0$?

3. Let $\mathfrak{D} \subset \mathfrak{S}$ be a set of probability distributions defined on $(M, \rho, \mathcal{M}_\rho)$. What are the conditions on $\mathfrak{D}$ for which $\Pi$ is injective even for non-universal kernels, i.e., $\gamma_{\mathcal{F}}(P,Q) = 0 \Leftrightarrow P = Q$ for $P, Q \in \mathfrak{D}$?

Note that question 3 is a restriction of question 1 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, $\mathfrak{D} \subset \mathfrak{S}$. So, there is a kind of equivalence between the class of kernels and the class of probability measures which make $\Pi$ to be injective. We would like to characterize this relationship between these classes of kernels and probability distributions, both defined on $(M, \rho)$. In the following section, we derive a new representation for $\gamma_{\mathcal{F}}$ in Eq. (3) to answer the above questions. From now on, we refer to $\gamma_{\mathcal{F}}$ in Eq. (3) as the maximum mean discrepancy (MMD).

# 4 A New Formulation of Maximum Mean Discrepancy

Based on the discussion so far, it should be clear that the embedding function, $\Pi$ is injective if the kernel defined on a compact metric space is universal. So, if we know that a given kernel is universal, we can use it applications like testing for homogeneity or independence. Suppose, if we do not know about the universality of the kernel, then checking for its universality is not easy. Therefore, one would like to have a simple verifiable rule that characterizes the injective or non-injective behavior of $\Pi$. To answer these questions, we propose a different formulation of MMD by making the following assumption.

*Assumption A-1. $k$ is a measurable positive definite convolution kernel on $\mathbb{R}^d$, i.e., $k(x, y) = \psi(x - y)$ where $\psi$ is a bounded continuous positive definite function.*

The above assumption of $k$ being a convolution kernel on $\mathbb{R}^d$ is not very restrictive as 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. We quote this result from [24, Theorem 6.6]. Since assumption (*A-1*) completely defines $k$ in terms of $\psi$, from now onwards, we interchangeably use $k$ and $\psi$ to mean a positive definite kernel.

**Theorem 5** (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 $\Lambda$ on $\mathbb{R}^d$, i.e.*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-jx^T\omega} \, d\Lambda(\omega), \quad x \in \mathbb{R}^d. \tag{4}$$

Using the assumption (*A-1*) and the Fourier characterization of $\psi$ given by Eq. (4), 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.[5] For the paper to be self-contained, a brief introduction to distribution theory is provided in Appendix A, which closely follows [17, Chapters 6,7]. Another good and basic reference on distribution theory is [22] (see also [9, Chapters 6–9]).

**Theorem 6** (Fourier representation of MMD)**.** *Let $(\mathcal{H}, k)$ be a RKHS defined on $\mathbb{R}^d$ with $k$ satisfying assumption (A-1). Let $\phi_P$ and $\phi_Q$ be the characteristic functions corresponding to probability distributions $P$ and $Q$ defined on $\mathbb{R}^d$. Suppose $\mathcal{F} = \{\|f\|_{\mathcal{H}} \leq 1 : f \in \mathcal{H}\}$.[6] Then*

$$\gamma_{\mathcal{F}}(P, Q) = \left\| \left[ (\overline{\phi}_P - \overline{\phi}_Q)\Lambda \right]^{\vee} \right\|_{\mathcal{H}}, \tag{5}$$

*where $-$ represents complex conjugation, $\vee$ represents the Fourier inversion and $\Lambda$ represents the finite non-negative Borel measure on $\mathbb{R}^d$ as defined in Eq. (4). $(\overline{\phi}_P - \overline{\phi}_Q)\Lambda$ represents a finite Borel measure defined by*

$$\left( (\overline{\phi}_P - \overline{\phi}_Q)\Lambda \right)(E) = \int_{\mathbb{R}^d} I_E(x)(\overline{\phi}_P(x) - \overline{\phi}_Q(x)) \, d\Lambda(x), \tag{6}$$

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

[5] Here, the term *distribution* should not be confused with probability distributions. In short, distribution refers 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. See Appendix A for details.

[6] Equivalently, we refer to $\mathcal{F}$ as a unit ball in a RKHS $(\mathcal{H}, k)$.

*Proof.* From Theorem 3, we have $\gamma_{\mathcal{F}}(P,Q) = \|Pk - Qk\|_{\mathcal{H}} = \left\|\int_{\mathbb{R}^d} k(\cdot,x)\,dP(x) - \int_{\mathbb{R}^d} k(\cdot,x)\,dQ(x)\right\|_{\mathcal{H}} = \left\|\int_{\mathbb{R}^d} \psi(.-x)\,dP(x) - \int_{\mathbb{R}^d} \psi(.-x)\,dQ(x)\right\|_{\mathcal{H}}$. By Eq. (34), $\int_{\mathbb{R}^d} \psi(.-x)\,dP(x)$ represents the convolution of $\psi$ and $P$ denoted as $\psi * P$. By appealing to the convolution theorem (Lemma 28 in Appendix B), we have $(\psi * P)^\wedge = \hat{P}\Lambda$, where $\hat{P}(\omega) = \int_{\mathbb{R}^d} e^{-j\omega^T x}\,dP(x)$, $\forall\,\omega \in \mathbb{R}^d$ (by Lemma 26). Complex conjugating $\hat{P}$ on both sides, we get $\overline{\hat{P}}(\omega) = \int_{\mathbb{R}^d} e^{j\omega^T x}\,dP(x) = \phi_P(\omega)$, $\forall\,\omega \in \mathbb{R}^d$, which implies $\hat{P} = \overline{\phi}_P$. Therefore, $\gamma_{\mathcal{F}}(P,Q) = \|\psi * P - \psi * Q\|_{\mathcal{H}} = \left\|(\overline{\phi}_P\Lambda)^\vee - (\overline{\phi}_Q\Lambda)^\vee\right\|_{\mathcal{H}} = \left\|\left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee\right\|_{\mathcal{H}}$ (by the linearity of Fourier inverse). □

**Remark 7.** *In Eq. (5), the term inside the RKHS norm is the Fourier inverse of a finite Borel measure, which is defined according to Eq. (6). If $\Psi$ is the distributional derivative[7] of $\Lambda$, then Eq. (5) can also be written as $\gamma_{\mathcal{F}}(P,Q) = \left\|\left[(\overline{\phi}_P - \overline{\phi}_Q)\Psi\right]^\vee\right\|_{\mathcal{H}}$, where the term inside the RKHS norm is the Fourier inverse of a tempered distribution.*

The representation of MMD in terms of the kernel spectrum as in Eq. (5) will be the central equation in our work which helps to understand its dependence on the properties of the kernel. By making integrability assumptions on $\psi$, in the following, we provide a preliminary result to show the dependence of $\gamma_{\mathcal{F}}$ on the support of the kernel spectrum.

**Proposition 8.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H},k)$ defined on $\mathbb{R}^d$ and let $P$, $Q$ be probability distributions defined on $\mathbb{R}^d$. Suppose $k$ satisfies assumption (A-1) with $\psi \in L^1(\mathbb{R}^d)$ and $\Psi(\omega) > 0$ a.e. Then $\gamma_{\mathcal{F}}(P,Q) = 0$ if and only if $P = Q$.*

*Proof.* The "if" part is trivial. The "only if" part is proved as follows. By Theorem 6, we have $\gamma_{\mathcal{F}}(P,Q) = \left\|\left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee\right\|_{\mathcal{H}}$. Since, $\psi \in L^1(\mathbb{R}^d)$, the Fourier inversion theorem [4, Theorem 9.5.4] ensures that $\Lambda$ is absolutely continuous w.r.t. the Lebesgue measure, i.e., $d\Lambda = \Psi\,dm_d$ and $\Psi(\omega) = \int_{\mathbb{R}^d} e^{j\omega^T x}\psi(x)\,dm_d(x)$ a.e. To show $\gamma_{\mathcal{F}}(P,Q) = 0 \Rightarrow P = Q$, let us consider $\gamma_{\mathcal{F}}(P,Q) = 0 \Rightarrow \left[(\overline{\phi}_P - \overline{\phi}_Q)\Psi\right]^\vee = 0$ a.e., which is equivalent to $(\overline{\phi}_P - \overline{\phi}_Q)\Psi = 0$ a.e. Since $\Psi$ is strictly positive almost everywhere on $\mathbb{R}^d$, we have $\phi_P = \phi_Q$ a.e. and therefore, the result follows from the uniqueness theorem for characteristic functions [4, Theorem 9.5.1]. □

The above proposition provides a simple verifiable rule on $\psi$ for which $\gamma_{\mathcal{F}}$ is a metric on $\mathfrak{S}$. Given an integrable positive definite kernel, $\psi$ defined on $\mathbb{R}^d$, if the support of its Fourier spectrum is the entire domain, i.e., $\text{supp}(\Psi) = \mathbb{R}^d$, then $\gamma_{\mathcal{F}}$ is guaranteed to be a metric on $\mathfrak{S}$. In other words, the map $\Pi : \mathfrak{S} \to \mathcal{H}$ is injective. However, Proposition 8 is restrictive because of the integrability assumption on $\psi$. What if $\psi$ is not integrable in which case $\Psi$ is not defined as the inverse Fourier transform of $\psi$ in the usual sense. We mention that Proposition 8 is only a preliminary result to motivate the following discussion and a more general result will be provided in §5. It is clear from Proposition 8 that the behavior of MMD is dependent on the kernel through the support of its Fourier spectrum. Table 1 shows some popular convolution kernels (Gaussian, Laplacian, $B_{2n+1}$-spline and Sinc are aperiodic while Poisson, Dirichlet, Féjer and Cosine are periodic) along with their spectra, $\Psi$ and its support. A plot of these kernels along with their spectra is shown in Figure 1. It has to be noted that, of all the kernels shown in Table 1, only Gaussian, Laplacian and $B_{2n+1}$-spline kernels are integrable and their corresponding $\Psi$ are computed using the Fourier transform in $L^1$ sense. The other kernels shown in Table 1 are not integrable and their corresponding $\Psi$ is computed in distribution sense using the tools from distribution theory, except for the Sinc kernel whose Fourier transform can be computed in $L^2$ sense. It can be seen from Table 1 that two scenarios can occur: (a) $\text{supp}(\Psi) = \mathbb{R}^d$ and (b) $\text{supp}(\Psi) \subsetneq \mathbb{R}^d$. We address these scenarios separately in §5.1 and §5.2. To gain an insight, if we restrict ourselves to $\mathbb{R}$, informally we can reduce these two scenarios to the following three: *(i)* $\{\omega : \Psi(\omega) = 0\}$ is empty, *(ii)* $\{\omega : \Psi(\omega) = 0\}$ is countable (e.g. $B_{2n+1}$-splines though $\text{supp}(\Psi) = \mathbb{R}$)

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

| Kernel | $\psi(x)$ | $\Psi(\omega)$ | supp($\Psi$) |
|---|---|---|---|
| Gaussian | $\exp\left(-\frac{x^2}{2\sigma^2}\right)$ | $\sigma\exp\left(-\frac{\sigma^2\omega^2}{2}\right)$ | $\mathbb{R}$ |
| Laplacian | $\exp(-\sigma\lvert x\rvert)$ | $\sqrt{\frac{2}{\pi}}\frac{\sigma}{\sigma^2+\omega^2}$ | $\mathbb{R}$ |
| $B_{2n+1}$-spline[8] [18] | $*_1^{n+1}\mathbb{1}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$ | $\left(\frac{2}{\pi}\right)^{\frac{n+1}{2}}\frac{\sin^{n+1}\left(\frac{\omega}{2}\right)}{\omega^{n+1}}$ | $\mathbb{R}$ |
| Sinc | $\frac{\sin(\sigma x)}{x}$ | $\sqrt{\frac{\pi}{2}}\mathbb{1}_{[-\sigma,\sigma]}(\omega)$ | $[-\sigma,\sigma]$ |
| Poisson [2, 21, 23] | $\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$, $0<\sigma<1$ | $\sum_{n=-\infty}^{\infty}\sigma^{\lvert n\rvert}\,\delta(\omega-n)$ | $\mathbb{Z}$ |
| Dirichlet [2, 18] | $\frac{\sin\frac{(2n+1)x}{2}}{\sin\frac{x}{2}}$ | $\sum_{i=-n}^{n}\delta(\omega-i)$ | $\{0,\pm1,\ldots,\pm n\}$ |
| Féjer [2] | $\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\frac{x}{2}}$ | $\sum_{i=-n}^{n}\left(1-\frac{\lvert i\rvert}{n+1}\right)\delta(\omega-i)$ | $\{0,\pm1,\ldots,\pm n\}$ |
| Cosine | $\cos(\sigma x)$ | $\frac{1}{2}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$ | $\{-\sigma,\sigma\}$ |

Table 1: Convolution kernels defined by $\psi$, their spectra, $\Psi$ and its support, supp($\Psi$) (refer to Appendix A for the definition of support of a distribution). The first four are aperiodic kernels while the last four are periodic. Here, the domain is considered to be $\mathbb{R}$ for simplicity. For $x\in\mathbb{R}^d$, the above formulae can be extended by computing $\psi(x)=\prod_{i=1}^{d}\psi(x_i)$ where $x=(x_1,\ldots,x_d)$ and $\Psi(\omega)=\prod_{i=1}^{d}\Psi(\omega_i)$ where $\omega=(\omega_1,\ldots,\omega_d)$. $\delta$ represents the Dirac-delta function.[9]

and *(iii)* $\{\omega:\Psi(\omega)=0\}$ is uncountable (e.g. sinc kernel and all periodic kernels) which is equivalent to $\text{supp}(\Psi)\subsetneq\mathbb{R}$.

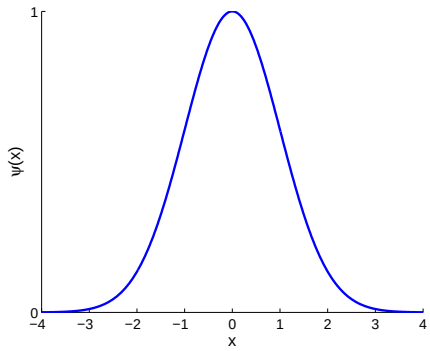
As aforementioned, universal kernels guarantee the map $\Pi$ to be injective, whose examples include the Gaussian and Laplacian kernels. Proposition 8 also provides a characterization for the map $\Pi$ to be injective, which is satisfied by the Gaussian and Laplacian kernels. This makes one to suspect some kind of a relation between the spectrum of the kernel and its universality, which however is not clear to us at this point of time. Steinwart [21] did not use the kernel spectrum in establishing the universality of Gaussian and Laplacian kernels (see [21, Corollary 10]). However, for rotational invariant periodic convolution kernels on $[0,2\pi)^d$, Steinwart considered their Fourier series coefficients in establishing the universality (see [21, Corollary 11]).

# 5 Characteristic Kernels & Main Theorems

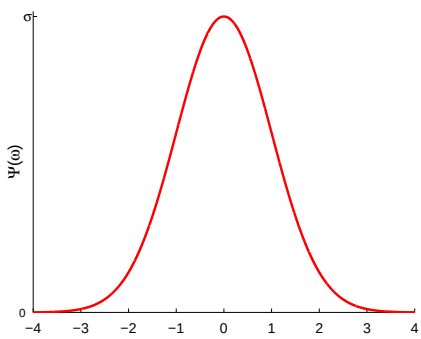
In this section, we present main results related to the behavior of MMD. First, we start with the following definition of characteristic kernels, which was recently introduced by Fukumizu *et al.* [8] in the context of measuring conditional (in)dependence using positive definite kernels.

[8] $B_{2n+1}$-spline is a $B_n$-spline of odd order. Only $B_{2n+1}$-splines are admissible, i.e. $B_n$ splines of odd order are positive definite kernels whereas the ones of even order have negative components in their Fourier spectrum, $\Psi$ and so are not admissible kernels. In Table 1, the symbol $*_1^{n+1}$ represents the $(n+1)$-fold convolution. An important point to be noted with $B_{2n+1}$-spline kernel is that its spectrum, $\Psi$ has vanishing points at $\omega=2\pi\alpha$, $\alpha\in\mathbb{Z}\backslash\{0\}$ unlike Gaussian and Laplacian kernels which do not have any vanishing points in their spectrum. However, all these kernels have spectrum whose support is $\mathbb{R}$.

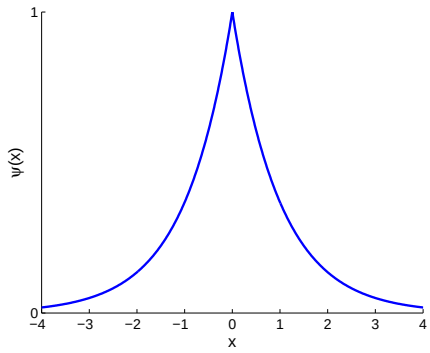
[9] The Dirac-delta function is defined $\delta(x)=0$, for $x\neq0$ and $\int_{-\infty}^{\infty}\delta(x)\,dx=1$. As aforementioned, it should not be treated as any other function but as a distribution. $\delta$ is just a symbolic representation of a function with the above characteristics and makes sense only under the integral sign. In the distribution sense, it is defined as $D_\delta(\varphi)=\varphi(0)$ for $\varphi\in\mathscr{D}_d$ (see Appendix A).



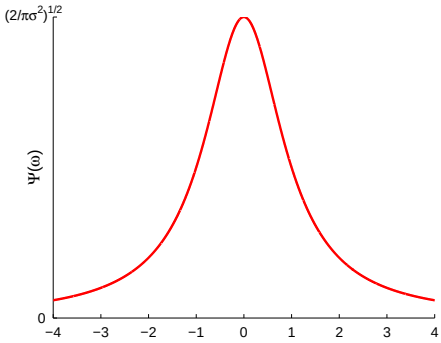

(a)



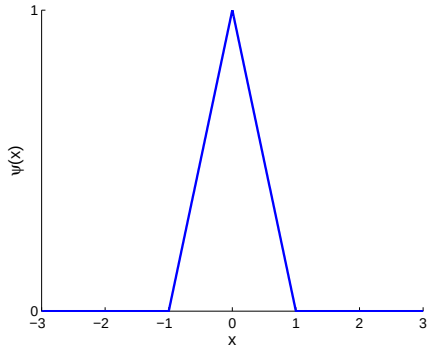

(a′)



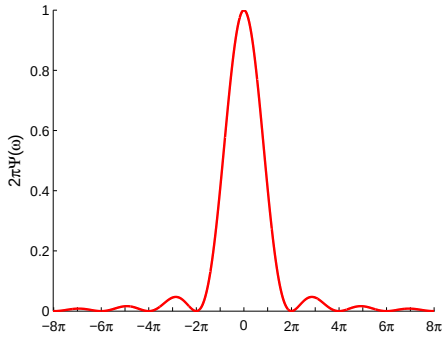

(b)



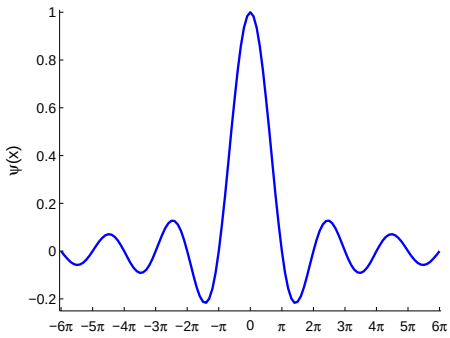

(b′)



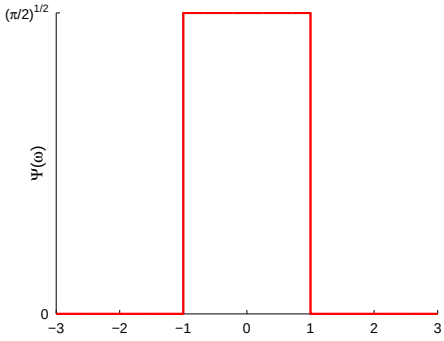

(c)




(c′)




(d)




(d′)

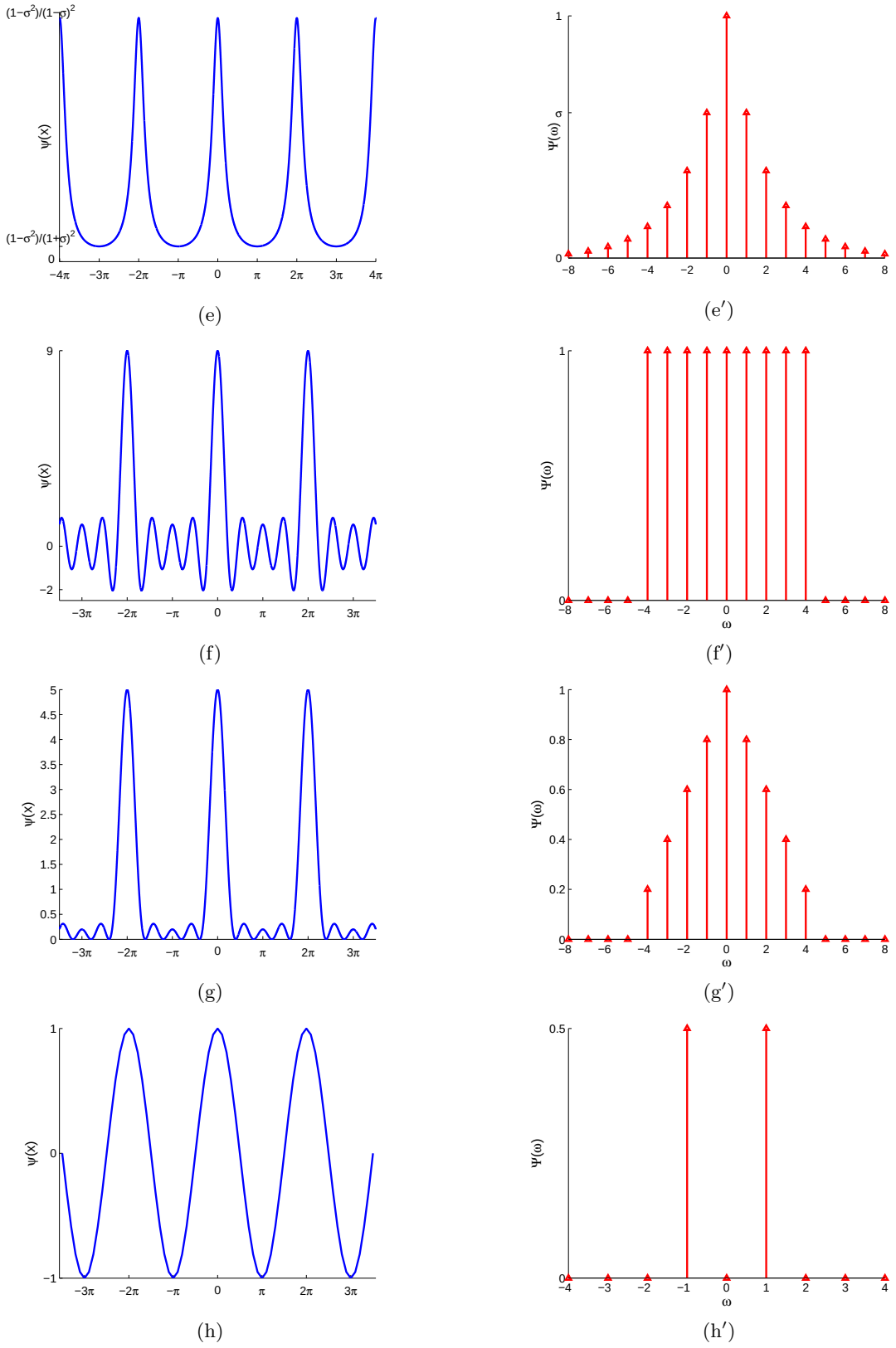



Figure 1: (a) Gaussian kernel (b) Laplacian kernel (c) $B_1$-spline kernel (d) Sinc kernel (e) Poisson kernel (f) Dirichlet kernel with $n = 4$ (g) Féjer kernel with $n = 4$ and (h) Cosine kernel. (a′)−(h′) represent their corresponding spectra, $\Psi$.

**Definition 9** (Characteristic kernel)**.** *A positive definite kernel, $k$ is a characteristic kernel to a set $\mathfrak{D}$ of probability measures defined on $(M, \rho, \mathcal{M}_\rho)$ if $\gamma_{\mathcal{F}}(P, Q) = 0 \Leftrightarrow P = Q$ for $P, Q \in \mathfrak{D}$.*

**Remark 10.** *Equivalently, $k$ is said to be characteristic to $\mathfrak{D}$ if the map, $\Pi : \mathfrak{D} \to \mathcal{H}$, $\Pi[P] = \int_M k(\cdot, x)\, dP(x)$, $P \in \mathfrak{D}$ is injective. When $M = \mathbb{R}^d$, the notion of characteristic kernel is a generalization of the characteristic function, $\phi_P(x) = \int_{\mathbb{R}^d} e^{jx^T\omega}\, dP(\omega)$, $\forall\, x \in \mathbb{R}^d$, which is the expectation of the complex-valued positive definite kernel, $k(x, \omega) = e^{jx^T\omega}$. Thus, the definition of a characteristic kernel generalizes the well-known property of the characteristic function that $\phi_P$ uniquely determines a Borel probability measure $P$ on $\mathbb{R}^d$. See [8] for more details.*

It is obvious from Definition 9 that universal kernels defined on a compact set $M$ 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 all positive definite convolution kernels on $\mathbb{R}^d$ that are characteristic to $\mathfrak{S}$. Here, $\mathfrak{S}$ represents the family of all probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, where $\mathcal{B}(\mathbb{R}^d)$ is the Borel $\sigma$-field on $\mathbb{R}^d$. Among the kernels that are not characteristic to $\mathfrak{S}$, we would like to determine those kernels that are characteristic to some subset, $\mathfrak{D}$ of $\mathfrak{S}$, for appropriately chosen $\mathfrak{D}$. Intuitively, one can get convinced that smaller the set $\mathfrak{D}$ is, larger is the family of kernels that are characteristic to $\mathfrak{D}$. Before we provide a characterization for $\psi$ that are characteristic to $\mathfrak{S}$, in the following example, we show that there exists kernels that are not characteristic to $\mathfrak{S}$.

**Example 11** (Trivial kernel)**.** *Let $k(x, y) = \psi(x - y) = 1$, $\forall\, x, y \in \mathbb{R}^d$. From Eq. (3),*

$$\gamma_{\mathcal{F}}(P, Q) = \|Pk - Qk\|_{\mathcal{H}} = \left\| \int_{\mathbb{R}^d} \psi(. - x)\, dP(x) - \int_{\mathbb{R}^d} \psi(. - x) dQ(x) \right\|_{\mathcal{H}} = \left\| \int_{\mathbb{R}^d} dP(x) - \int_{\mathbb{R}^d} dQ(x) \right\|_{\mathcal{H}} = 0,$$

*for any $P, Q \in \mathfrak{S}$. The same result can be obtained by the Fourier representation of $\gamma_{\mathcal{F}}$. By assumption (A-1), $\psi \in C_b(\mathbb{R}^d)$. Therefore, $\psi$ is a tempered distribution whose inverse Fourier transform is a tempered distribution given by $\Psi = \delta$, the Dirac-delta function. Consider*

$$\begin{aligned} \left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee(x) &= \int_{\mathbb{R}^d} e^{jx^T\omega}(\overline{\phi}_P(\omega) - \overline{\phi}_Q(\omega))\, d\Lambda(\omega) = \int_{\mathbb{R}^d} e^{jx^T\omega}(\overline{\phi}_P(\omega) - \overline{\phi}_Q(\omega))\Psi(\omega)\, dm_d(\omega) \\ &= \int_{\mathbb{R}^d} e^{jx^T\omega}(\overline{\phi}_P(\omega) - \overline{\phi}_Q(\omega))\delta(\omega)\, dm_d(\omega) = \overline{\phi}_P(0) - \overline{\phi}_Q(0) = 0,\ \forall\, x \in \mathbb{R}^d, \end{aligned} \tag{7}$$

*since $\phi_P(0) = \int_{\mathbb{R}^d} dP(x) = 1 = \phi_Q(0)$. Therefore, $\gamma_{\mathcal{F}}(P, Q) = \| \left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee \|_{\mathcal{H}} = 0$ for any $P, Q \in \mathfrak{S}$.*

The kernel in the above example is an un-interesting one which cannot detect the difference between any two probability distributions defined on $\mathbb{R}^d$. This means, for such a kernel, one can construct probability distributions $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Are there any non-trivial kernels for which $\exists\, P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$? The following result answers this question by providing conditions for $\gamma_{\mathcal{F}}(P, Q) = 0$ when $P \neq Q$.

**Theorem 12.** *Let $\mathcal{F}$ be a unit ball in a 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 (A-1) and supp$(\Lambda) \subset \mathbb{R}^d$. Then $\gamma_{\mathcal{F}}(P, Q) = 0$ if and only if there exists a tempered distribution $\theta$ 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.*

*Proof.* The proof directly follows from the formulation of $\gamma_{\mathcal{F}}$ in Eq. (5).

($\Rightarrow$) Let $\theta$ be a tempered distribution satisfying $(i)$ and $(ii)$. $(i)$ implies that $(p - q)^\wedge = \hat{\hat{\theta}} = \theta$ since $\theta$ is a tempered distribution. Therefore, $\theta = \hat{p} - \hat{q} = \overline{\phi}_P - \overline{\phi}_Q$. So, $\gamma_{\mathcal{F}}(P, Q) = \left\| \left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee \right\|_{\mathcal{H}} = \left\| [\theta\Lambda]^\vee \right\|_{\mathcal{H}} = 0$ (by $(ii)$).

($\Leftarrow$) Let $\gamma_{\mathcal{F}}(P,Q) = \left\| \left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee \right\|_{\mathcal{H}} = 0 \Rightarrow \left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee = 0$. Since $(\overline{\phi}_P - \overline{\phi}_Q)\Lambda$ is a finite Borel measure as defined by Eq. (6), it is therefore a tempered distribution and so $\left[\left[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\right]^\vee\right]^\wedge = (\overline{\phi}_P - \overline{\phi}_Q)\Lambda = 0$. Let $\theta = \overline{\phi}_P - \overline{\phi}_Q$. Clearly $\theta$ is a tempered distribution as by Lemma 26, $\overline{\phi}_P$ and $\overline{\phi}_Q$ are tempered distributions. So, $\check{\theta} = (\overline{\phi}_P - \overline{\phi}_Q)^\vee = p - q$. □

$\theta = 0$ trivially satisfies $(ii)$ in Theorem 12. 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 $\theta$, which we present in the following proposition.

**Proposition 13** (Properties of $\theta$)**.** *$\theta$ in Theorem 12 satisfies the following properties:*

*(a) $\theta$ 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 $\Omega := supp(\Lambda)$. In addition, if $\Omega = \{a_1, a_2, \ldots\}$, then $\theta(a_i) = 0, \forall\, a_i \in \Omega$.*

*Proof.* $(a)$ By Lemma 26, $\phi_P$ is a conjugate symmetric, bounded and uniformly continuous function on $\mathbb{R}^d$. Since $\theta = \overline{\phi}_P - \overline{\phi}_Q$, the result in $(a)$ follows.

$(b)$ By Lemma 26, $\phi_P(0) = \phi_Q(0) = 1$. 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 $\varphi \in \mathscr{D}(V)$ with $\Lambda(\varphi) \neq 0$. Take $h \in \mathscr{D}(U)$ 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 $\mathrm{supp}(h) \subset U$ and $\theta \neq 0$ on $U$. By $(ii)$ of Theorem 12, $\theta\Lambda = 0$, where $\theta\Lambda$ is a finite Borel measure on $\mathbb{R}^d$ as defined by Eq. (38). Therefore,

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

But the left hand side is

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

which causes contradiction. Therefore, $\mathrm{supp}(\theta) \subset \overline{\mathbb{R}^d \backslash \Omega}$.

If $\Omega = \{a_1, a_2, \ldots\}$, then $\Lambda = \sum_{a_i \in \Omega} \beta_i \delta_{a_i}$, $\beta_i > 0$ and $\sum_i \beta_i < \infty$. Since $\theta\Lambda = 0$, $\int_{\mathbb{R}^d} \alpha(x)\theta(x)\, d\Lambda(x) = \sum_i \beta_i \alpha(a_i)\theta(a_i) = 0$ for any continuous function $\alpha$ in $\mathbb{R}^d$. This implies $\theta(a_i) = 0, \forall\, a_i \in \Omega$. □

Theorem 12 provides conditions under which $\gamma_{\mathcal{F}}(P,Q) = 0$ when $P \neq Q$. However, it would be nice to have a constructive procedure, i.e., a way to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P,Q) = 0$. Condition $(ii)$ in Theorem 12 says that $\theta$ has to be chosen such that its support is disjoint with that of the kernel spectrum. From condition $(i)$, we have $p = q + \check{\theta}$. So, for a given $Q$, we can choose $\theta$ that satisfies the properties in Proposition 13 so that $\gamma_{\mathcal{F}}(q + \check{\theta}, q) = 0$. However, $p$ should be a positive distribution so that it corresponds to a positive measure.[10] Therefore, $\theta$ should also be such that $q + \check{\theta}$ is a positive distribution. Imposing such a constraint on $\theta$ is not straightforward and therefore Theorem 12 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 $\theta$ can be explicitly specified, thereby yielding a procedure to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P,Q) = 0$.

[10] A positive distribution is defined to be as the one that takes nonnegative values on nonnegative test functions. So, $D \in \mathscr{D}'(M)$ is a positive distribution if $D(\varphi) \geq 0$ for $0 \leq \varphi \in \mathscr{D}(M)$. If $\mu$ 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 [22, §6.4].

**Theorem 14.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of probability measures on $\mathbb{R}^d$ whose characteristic functions are 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 (A-1) and supp$(\Lambda) \subsetneq \mathbb{R}^d$. Then for any $Q \in \mathfrak{D}$, there exists $P \neq Q$, $P \in \mathfrak{D}$ given by*

$$p = q + \check{\theta} \tag{8}$$

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

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

*(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 $\theta$ 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 some $Q \in \mathfrak{D}$, $\phi_Q \in \big(L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\big) \cap C_b(\mathbb{R}^d)$. When $\phi_Q \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$, Riemann-Lebesgue lemma (Lemma 29) 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 $L^2$ sense[11] implies that $q = [\overline{\phi}_Q]^\vee \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Therefore, $q \in L^1(\mathbb{R}^d) \cap \big(L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d)\big)$. Define $p := q + \check{\theta}$. Clearly $p \in L^1(\mathbb{R}^d) \cap \big(L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d)\big)$. In addition, $\overline{\phi}_P = \hat{p} = \hat{q} + \hat{\check{\theta}} = \overline{\phi}_Q + \theta \in \big(L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\big) \cap C_b(\mathbb{R}^d)$. Consider, $\int_{\mathbb{R}^d} p(x)\, dx = \int_{\mathbb{R}^d} q(x)\, dx + \int_{\mathbb{R}^d} \check{\theta}(x)\, dx = 1 + \theta(0) = 1$ (by $(iv)$). $(v)$ implies that $p(x) \geq 0, \forall\, x$. Since $\theta$ is conjugate symmetric, $\check{\theta}$ is real valued and so is $p$. 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 $\gamma_{\mathcal{F}}(P, Q) = \Big\| \big[(\overline{\phi}_P - \overline{\phi}_Q)\Lambda\big]^\vee \Big\|_{\mathcal{H}} = \|[\theta\Lambda]^\vee\|_{\mathcal{H}} = 0$ (by $(iii)$).

$(\Leftarrow)$ Suppose that $P, Q \in \mathfrak{D}$ and $p = q + \check{\theta}$ gives $\gamma_{\mathcal{F}}(P, Q) = 0$. We need to show that $\theta$ satisfies $(i) - (v)$. $P, Q \in \mathfrak{D}$ implies $\phi_P, \phi_Q \in \big(L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\big) \cap C_b(\mathbb{R}^d)$ and $p, q \in L^1(\mathbb{R}^d) \cap \big(L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d)\big)$. Therefore, $\theta = \overline{\phi}_P - \overline{\phi}_Q \in \big(L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\big) \cap C_b(\mathbb{R}^d)$ and $\check{\theta} = p - q \in L^1(\mathbb{R}^d) \cap \big(L^2(\mathbb{R}^d) \cup C_b(\mathbb{R}^d)\big)$. By Lemma 26, $\phi_P$ and $\phi_Q$ are conjugate symmetric and so $\theta = \overline{\phi}_P - \overline{\phi}_Q$ is conjugate symmetric. Therefore $\theta$ satisfies $(i)$ and $\check{\theta}$ satisfies $(ii)$. $\theta$ 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 $(iii)$ whose proof is similar to that of Theorem 12. □

**Remark 15.** *Conditions $(iii)$ and $(iv)$ are same as that of Proposition 13. Conditions $(i)$ and $(ii)$ are required to satisfy our assumption $P, Q \in \mathfrak{D}$ and Eq. (8). Condition $(v)$ ensures that $P$ is a positive measure, which was the condition difficult to be imposed in Theorem 12.*

In the above result, we restricted ourselves to probability measures, $P$ whose characteristic functions, $\phi_P$ are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. This ensures that the inverse Fourier transform of $\phi_P$ exists in $L^1$ or $L^2$ sense. Without this assumption, $\phi_P \in C_b(\mathbb{R}^d)$ is not guaranteed to have Fourier transform in $L^1$ or $L^2$ sense and therefore has to be treated as a tempered distribution while computing its Fourier transform. So, $\theta = \overline{\phi}_P - \overline{\phi}_Q$ has to be treated as a tempered distribution, which is the setting in Theorem 12. Since we wanted to avoid dealing with distributions wherein the required positivity constraint is difficult to impose,

[11] 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 $L^2(\mathbb{R}^d)$. So, $\mathcal{F}$ is a linear operator, mapping $L^2(\mathbb{R}^d)$ into $L^2(\mathbb{R}^d)$.

we restricted ourselves to $\mathfrak{D}$.[12] Though this result explicitly captures the conditions on $\theta$, it is a very restricted result as it only deals with continuous (a.e.) probability measures. However, we use this result later to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

Theorem 12 and Theorem 14 are main results that provide conditions for the existence of $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. In the following, we derive specific results relating to a class of kernels that are characteristic (or non-characteristic) to a class of probability measures defined on $\mathbb{R}^d$. It is clear from Theorem 12 and Theorem 14 that the dependence of $\gamma_{\mathcal{F}}$ on the kernel appears in the form of the support of the kernel spectrum. So, as aforementioned in §4, two scenarios exist: $(i)\,\text{supp}(\Lambda) = \mathbb{R}^d$ and $(ii)\,\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$.

## 5.1 $\textbf{supp}(\Lambda) = \mathbb{R}^d$

**Theorem 16.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Suppose $k$ satisfies assumption (A-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$(\Lambda) = \mathbb{R}^d$.*

*Proof.* $(\Rightarrow)$ 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 Theorem 12, $\exists\, \theta$ satisfying $(i)$ and $(ii)$ given in Theorem 12. By Proposition 13, $\theta\Lambda = 0$ implies $\text{supp}(\theta) \subset \overline{\mathbb{R}^d \backslash \text{supp}(\Lambda)}$. Since $\text{supp}(\Lambda) = \mathbb{R}^d$ and $\theta$ is uniformly continuous function in $\mathbb{R}^d$, we have $\text{supp}(\theta) = \emptyset$ which means $\theta = 0$ a.e. Therefore, by $(i)$ of Theorem 12, we have $P = Q$, leading to a contradiction. So, $\nexists\, P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.

$(\Leftarrow)$ Suppose $k$ is characteristic to $\mathfrak{S}$. Then we need to show that $\text{supp}(\Lambda) = \mathbb{R}^d$. This is equivalent to proving that $k$ is not characteristic to $\mathfrak{S}$ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. So, 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 $\mathbb{R}^d$ whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. The claim is that $\exists\, P \neq Q$, $P, Q \in \mathfrak{D} \subsetneq \mathfrak{S}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. This claim is later proved in Theorem 21. Therefore, $k$ is not characteristic to $\mathfrak{S}$ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. □

The above result shows that the embedding function, $\Pi$ associated with a positive definite convolution 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 $\Pi$ to be injective. Theorem 16 is a stronger version of Proposition 8 as it does not make any integrability assumptions on the kernel. However, both these results show that the support of the Fourier spectrum of a positive definite convolution kernel in $\mathbb{R}^d$ characterizes the injective or non-injective behavior of $\Pi$. Examples of kernels that are characteristic to $\mathfrak{S}$ include the Gaussian, Laplacian and $B_{2n+1}$-splines (see Table 1). A more useful result in practice is provided by the following corollary to Theorem 16.

**Corollary 17.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Suppose $k$ satisfies assumption (A-1) and supp$(\psi) \subset \mathbb{R}^d$ is compact. Then $k$ is a characteristic kernel to $\mathfrak{S}$.*

*Proof.* Since $\text{supp}(\psi)$ is compact in $\mathbb{R}^d$, by Lemma 31 which is a corollary of the Paley-Wiener theorem (see also [9, Theorem 31.5.2, Proposition 31.5.4]), we have $\text{supp}(\Lambda) = \mathbb{R}^d$. Therefore, the result directly follows from Theorem 16. □

This is a useful result in practice as any compactly supported convolution kernel in $\mathbb{R}^d$ is not only characteristic to $\mathfrak{S}$ but also is computationally advantageous compared with non-compact kernels like Gaussian and Laplacian. So, in practice, one can simply use $B_{2n+1}$-spline kernels instead of Gaussian or Laplacian kernel.

[12] Choosing $\mathfrak{D}$ to be the set of all probability measures whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$ is the best possible restriction that avoids treating $\theta$ 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.

## 5.2 $\mathbf{supp}(\Lambda) \subsetneq \mathbb{R}^d$

Based on the results shown in §5.1, it is clear that kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $\mathfrak{S}$. One example is the trivial kernel (see Example 11) which has $\text{supp}(\Lambda) = \{0\}$. The following result shows that periodic kernels in $\mathbb{R}^d$ are not characteristic to the set of discrete probability measures whose support is related to the period of the kernel.

**Proposition 18.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$ where $k$ satisfies assumption (A-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) $\psi$ is $\tau$-periodic in $\mathbb{R}^d$, i.e., $\psi(x) = \psi(x + \eta \bullet \tau)$, $\eta \in \mathbb{Z}^d$, $\tau \in \mathbb{R}^d_+$.*[13]

*(ii) $x_i - x_j = l_{ij} \bullet \tau$, $l_{ij} \in \mathbb{Z}^d$, $\forall\, i, j$.*

*where $\bullet$ represents the Hadamard multiplication.*

*Proof.* Let $\psi$ be $\tau$-periodic and $x_i - x_j = l_{ij} \bullet \tau$, $l_{ij} \in \mathbb{Z}^d$, $\forall\, i, j$ and $l \in \mathbb{Z}^d$. Consider $P$, $Q \in \mathcal{D}$ 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

$$\begin{aligned}
\gamma_{\mathcal{F}}(P, Q) &= \|Pk - Qk\|_{\mathcal{H}} = \left\| \int_{\mathbb{R}^d} \psi(. - x)\, dP(x) - \int_{\mathbb{R}^d} \psi(. - x)\, dQ(x) \right\|_{\mathcal{H}} = \left\| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_n) \right\|_{\mathcal{H}} \\
&= \left\| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_1 - l_{n1} \bullet \tau) \right\|_{\mathcal{H}} = \left\| \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \psi(. - x_1) \right\|_{\mathcal{H}} \quad (\because \psi \text{ is } \tau\text{-periodic}) \\
&= \left\| \psi(. - x_1) \sum_{n=1}^{\infty} (\tilde{p}_n - \tilde{q}_n) \right\|_{\mathcal{H}} = 0.
\end{aligned}$$

This holds for any $P, Q \in \mathfrak{D}$. □

In order to prove the converse, we need to show that $(i)$ and $(ii)$ in Proposition 18 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 $\{r_s\}_{s=1}^{\infty} \in [-1, 1]$. So, to prove the converse, we need to characterize all $\psi$, $\{\tilde{r}_n\}_{n=1}^{\infty}$ and $\{x_n\}_{n=1}^{\infty}$ that satisfy $\mathcal{R} = \{\sum_{s,t=1}^{\infty} \tilde{r}_t \tilde{r}_s \psi(x_s - x_t) = 0 : \sum_{s=1}^{\infty} \tilde{r}_s = 0, \{r_s\}_{s=1}^{\infty} \in [-1, 1]\}$, which is not easy. However, choosing some $\psi$, $\{\tilde{r}_n\}_{n=1}^{\infty}$ and $\{x_n\}_{n=1}^{\infty}$ is easy, as is shown in Proposition 18. Suppose there exists a class, $\mathcal{K}$ of positive definite convolution 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 $\mathfrak{E}$.

The following two results address the behavior of $\gamma_{\mathcal{F}}$ on the set of probability measures whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$ for the case of $k$ with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. The first one deals with compactly supported probability measures while the second one deals with probability measures whose support is non-compact.

**Theorem 19.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of all compactly supported probability measures on $\mathbb{R}^d$ whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Suppose $k$ satisfies assumption (A-1) and supp$(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior. Then $k$ is a characteristic kernel to $\mathfrak{D}$.*

*Proof.* Suppose $\exists P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Then by Theorem 12, there exists a tempered distribution, $\theta$ 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 Theorem 14 and so $\theta$ is a non-zero function that satisfies the conditions $(i) - (v)$ mentioned

[13] $\tau$-periodic $\psi$ 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 $\sum_{n=-\infty}^{\infty} \alpha_n < \infty$. So, $\text{supp}(\Lambda) = \{\frac{2\pi n}{\tau} : n \in \mathbb{Z}\} \subsetneq \mathbb{R}$. $\{\alpha_n\}_{n=-\infty}^{\infty}$ are called the Fourier series coefficients of $\psi$.

in Theorem 14. The condition $\theta\Lambda = 0$ implies that $\text{supp}(\theta) \subset \overline{\mathbb{R}^d \backslash \text{supp}(\Lambda)}$. Since $\text{supp}(\Lambda)$ has a non-empty interior, we have $\text{supp}(\theta) \subsetneq \mathbb{R}^d$. So, there exists an open set, $U \subset \mathbb{R}^d$ such that $\theta(x) = 0, \forall\, x \in U$. By Lemma 31, this means that $\check{\theta}$ is not compactly supported in $\mathbb{R}^d$. Condition $(iv)$ implies that $\int_{\mathbb{R}^d} \check{\theta}(x)\, dx = 0$ 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 Theorem 14. So, there does not exist a non-zero $\theta$ that satisfies the conditions $(i) - (v)$ in Theorem 14, thereby leading to a contradiction. □

The assumption that $\text{supp}(\Lambda)$ has a non-empty interior is important for the above result to hold. If $\text{supp}(\Lambda)$ has an empty interior (examples include periodic kernels), then $\text{supp}(\theta) = \mathbb{R}^d$. In principle, one can choose such a $\theta$ by selecting $\theta \in \mathscr{S}_d$ so that it satisfies the conditions $(i) - (iv)$ of Theorem 14 while satisfying the decay conditions (Eq. (40) and Eq. (41)) given in the Paley-Wiener theorem (see Lemma 30 in Appendix B). Therefore, by Paley-Wiener theorem, $\check{\theta}$ is a $C^\infty$ function with a compact support. If $\theta$ is chosen such that $\text{supp}(\check{\theta}) \subset \text{supp}(Q)$, then condition $(v)$ of Theorem 14 will be satisfied. Thus, one can construct $P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. Note that the conditions $(i)$ and $(ii)$ of Theorem 14 are automatically satisfied (except for conjugate symmetry) by choosing $\theta \in \mathscr{S}_d$. However, choosing $\theta$ such that it is also an entire function is not straightforward. In the following, we provide a simple example to show that $P \neq Q$, $P, Q \in \mathcal{D}$ can be constructed such that $\gamma_{\mathcal{F}}(P, Q) = 0$, wherein $\mathcal{F}$ corresponds to a unit ball in a RKHS $(\mathcal{H}, k)$ induced by a periodic convolution kernel whose $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has an empty interior.

**Example 20.** *Let $Q$ be a uniform distribution on $[-\beta, \beta] \subset \mathbb{R}$, i.e., $q(x) = \frac{1}{2M} \mathbb{1}_{[-\beta,\beta]}(x)$ whose characteristic function, $\phi_Q(\omega) = \frac{1}{\beta\sqrt{2\pi}} \frac{\sin(\beta\omega)}{\omega}$ is in $L^2(\mathbb{R})$. Let $\psi$ be the Dirichlet kernel with period $\tau$, where $\tau \leq \beta$, i.e., $\psi(x) = \frac{\sin \frac{(2l+1)\pi x}{\tau}}{\sin \frac{\pi x}{\tau}}$ and $\Psi(\omega) = \sum_{i=-l}^{l} \delta\left(\omega - \frac{2\pi i}{\tau}\right)$ with $\text{supp}(\Psi) = \{\frac{2\pi i}{\tau},\, i \in \{0, \pm 1, \ldots, \pm l\}\}$. Clearly, $\text{supp}(\Psi)$ has an empty interior. Let $\theta$ be*

$$\theta(\omega) = \frac{8j\alpha}{\pi} \sin\left(\frac{\omega\tau}{2}\right) \frac{\sin^2\left(\frac{\omega\tau}{4}\right)}{\tau\omega^2}, \tag{9}$$

*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 $\theta$ satisfies $(i)$ in Theorem 14. Since $\theta(\omega) = 0$ at $\omega = \frac{2\pi l}{\tau}$, $l \in \mathbb{Z}$, $\theta$ also satisfies $(iii)$ and $(iv)$ in Theorem 14. $\check{\theta}$ is given by*

$$\check{\theta}(x) = \begin{cases} \frac{2\alpha\left|x + \frac{\tau}{2}\right|}{\tau} - \alpha, & -\tau \leq x \leq 0 \\ A - \frac{2\alpha\left|x - \frac{\tau}{2}\right|}{\tau}, & 0 \leq x \leq \tau \\ 0, & \textit{otherwise} \end{cases}, \tag{10}$$

*where $\check{\theta} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ satisfies $(ii)$ in Theorem 14. Now, consider $p = q + \check{\theta}$ which is given as*

$$p(x) = \begin{cases} \frac{1}{2\beta}, & x \in [-\beta, -\tau] \cup [\tau, \beta] \\ \frac{2\alpha\left|x + \frac{\tau}{2}\right|}{\tau} + \frac{1}{2\beta} - \alpha, & x \in [-\tau, 0] \\ \alpha + \frac{1}{2\beta} - \frac{2\alpha\left|x - \frac{\tau}{2}\right|}{\tau}, & x \in [0, \tau] \\ 0, & \textit{otherwise} \end{cases}. \tag{11}$$

*Clearly, $p(x) \geq 0$, $\forall\, x$ and $\int_{\mathbb{R}} p(x)\, dx = 1$. $\phi_P = \phi_Q + \theta = \phi_Q + j\theta_I$ where $\theta_I = Im[\theta]$ and $\phi_P \in L^2(\mathbb{R})$. So, we have constructed $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, where $P$ and $Q$ are compactly supported in $\mathbb{R}$ with their characteristic functions in $L^2(\mathbb{R})$. Figure 2 shows the plots of $\psi$, $\Psi$, $\theta$, $\check{\theta}$, $q$, $\phi_Q$, $p$ and $|\phi_P|$ for $\tau = 2$, $l = 2$, $\beta = 3$ and $\alpha = \frac{1}{8}$.*

**Theorem 21.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Let $\mathfrak{D}$ be the set of all probability measures with non-compact support on $\mathbb{R}^d$ whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Suppose $k$ satisfies assumption (A-1) and $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Then $\exists\, P \neq Q$, $P, Q \in \mathfrak{D}$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$.*

*Proof.* We claim that there exists a non-zero function, $\theta$ satisfying $(i) - (v)$ in Theorem 14 which therefore proves the result. Consider the following function, $g_{\beta,\omega_0} \in C^\infty(\mathbb{R}^d)$ supported in $[\omega_0 - \beta, \omega_0 + \beta]$,

$$g_{\beta,\omega_0}(\omega) = \prod_{i=1}^{d} \mathbb{1}_{[-\beta_i,\beta_i]}(\omega_i - \omega_{0,i}) \exp\left(-\frac{\beta_i^2}{\beta_i^2 - (\omega_i - \omega_{0,i})^2}\right), \tag{12}$$

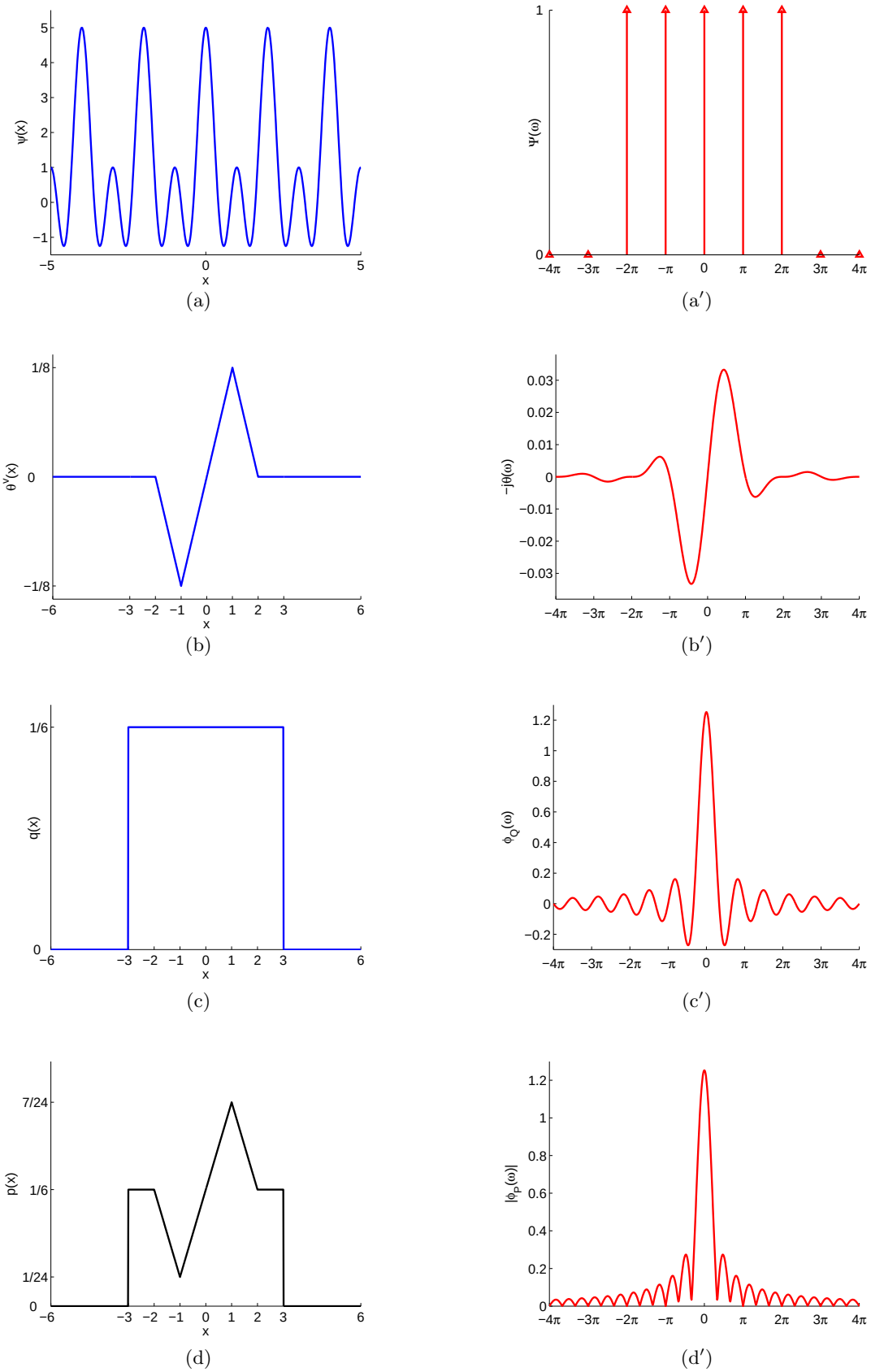



Figure 2: (a-a′) $\psi$ and its spectrum $\Psi$, (b-b′) $\breve{\theta}$ and $-j\theta$, (c-c′) uniform distribution, $q$ and its characteristic function $\phi_Q$ and (d-d′) $p = q + \breve{\theta}$ and $|\phi_P|$. See Example 20 for details.

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 $\Lambda$ is null. So, $\exists\, \beta, \omega_0 \neq 0$ such that $[\omega_0 - \beta, \omega_0 + \beta] \subset U$. Choose $\theta = \alpha(g_{\beta,\omega_0} + g_{\beta,-\omega_0})$, $\alpha \neq 0$ which implies $\text{supp}(\theta) = [-\omega_0 - \beta, -\omega_0 + \beta] \cup [\omega_0 - \beta, \omega_0 + \beta]$ is compact. Therefore, by Paley-Wiener theorem (Lemma 30), $\check{\theta}$ 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 $q(x) \propto \prod_{i=1}^{d} \frac{1}{1+|x_i|^{k+\epsilon}}$, $\forall\, k \in \mathbb{N}$, $\epsilon > 0$ where $x = (x_1, \ldots, x_d)$. It can be verified that $\theta$ satisfies the conditions $(i) - (v)$ in Theorem 14. So, there exists a non-zero $\theta$ as we claimed earlier and hence the result. □

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 $\mathbb{R}^d$ whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. Though, we provided a constructive procedure for $\theta$ in the proof of Theorem 21, the method is not straightforward as $\check{g}_{\beta,\omega_0}$ is not known in closed form and therefore, $\theta$ and $p$. So, instead of using this procedure, in the following, we provide a simple example which establishes that $k$ with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ is not characteristic to the class $\mathfrak{D}$ considered in Theorem 21.

**Example 22.** *Let $Q$ be a Cauchy distribution in $\mathbb{R}$, i.e., $q(x) = \frac{1}{\pi(1+x^2)}$ whose characteristic function, $\phi_Q(\omega) = \frac{1}{\sqrt{2\pi}} \exp(-|\omega|)$ is in $L^1(\mathbb{R})$. Let $\psi$ be a sinc kernel, i.e., $\psi(x) = \sqrt{\frac{2}{\pi}} \frac{\sin(\beta x)}{x}$ whose Fourier transform is given by $\Psi(\omega) = \mathbb{1}_{[-\beta,\beta]}(\omega)$ with supp$(\Psi) = [-\beta, \beta] \subsetneq \mathbb{R}$. Let $\theta$ be*

$$\theta(\omega) = \frac{\alpha}{2j} \left[ *_1^N \mathbb{1}_{\left[-\frac{\beta}{2}, \frac{\beta}{2}\right]}(\omega) \right] * \left[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)\right], \tag{13}$$

*where $|\omega_0| \geq \left(\frac{N+2}{2}\right) \beta$, $N \geq 2$ and $\alpha \neq 0$. $*_1^N$ represents the $N$-fold convolution. It is easy to verify that $\theta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ satisfies the conditions $(i)$, $(iii)$ and $(iv)$ in Theorem 14. Using the convolution theorem (Lemma 28), $\check{\theta}$ is given by*

$$\check{\theta}(x) = \alpha \left(\frac{2}{\pi}\right)^{\frac{N}{2}} \sin(\omega_0 x) \frac{\sin^N\left(\frac{\beta x}{2}\right)}{x^N}, \tag{14}$$

*and $\check{\theta} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ satisfies $(ii)$ in Theorem 14. $N$ determines the rate of decay of $\check{\theta}$ as $|\check{\theta}(x)| \leq \alpha \left(\frac{2}{\pi}\right)^{\frac{N}{2}} \frac{1}{1+|x|^N}$ and it has to be chosen according to the rate of decay of $q$. When $q$ is Cauchy, $N \geq 2$ and $\alpha \leq \frac{1}{\pi} \left(\frac{\pi}{2}\right)^{\frac{N}{2}} \inf_{|x| \leq 1} \frac{1+|x|^N}{1+x^2}$. It is easy to see that $\theta$ satisfies $(v)$ of Theorem 14, wherein $p$ given by*

$$p(x) = \frac{1}{\pi(1+x^2)} + \alpha \left(\frac{2}{\pi}\right)^{\frac{N}{2}} \sin(\omega_0 x) \frac{\sin^N\left(\frac{\beta x}{2}\right)}{x^N} \tag{15}$$

*is a probability density with $\phi_P = \phi_Q + \theta = \phi_Q - j\theta_I$ where $\theta_I = Im[\theta]$ and $\phi_P \in L^1(\mathbb{R})$. So, we have constructed $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$, where $P$ and $Q$ have non-compact support in $\mathbb{R}$ with their characteristic functions in $L^1(\mathbb{R})$. Figure 3 shows the plots of $\psi$, $\Psi$, $\theta$, $\check{\theta}$, $q$, $\phi_Q$, $p$ and $|\phi_P|$ for $\beta = 2\pi$, $N = 2$, $\omega_0 = 4\pi$ and $\alpha = \frac{1}{2}$.*

To summarize the results we discussed so far in this section, we return to the questions posed in §3. Theorem 16 answers the first and second question by showing that the embedding function, $\Pi$ is injective on $\mathfrak{S}$ if and only if the spectrum of the convolution kernel on $\mathbb{R}^d$ has the entire domain as its support. While answering the second question, Theorem 21 provides a procedure to construct $P \neq Q$ such that $\gamma_{\mathcal{F}}(P, Q) = 0$. The third question is answered by Theorem 19 wherein the kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ which are not characteristic to $\mathfrak{S}$ are characteristic to the set of compactly supported probability measures on $\mathbb{R}^d$ whose characteristic functions are in $L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$. A summary of these results is given in Table 2.

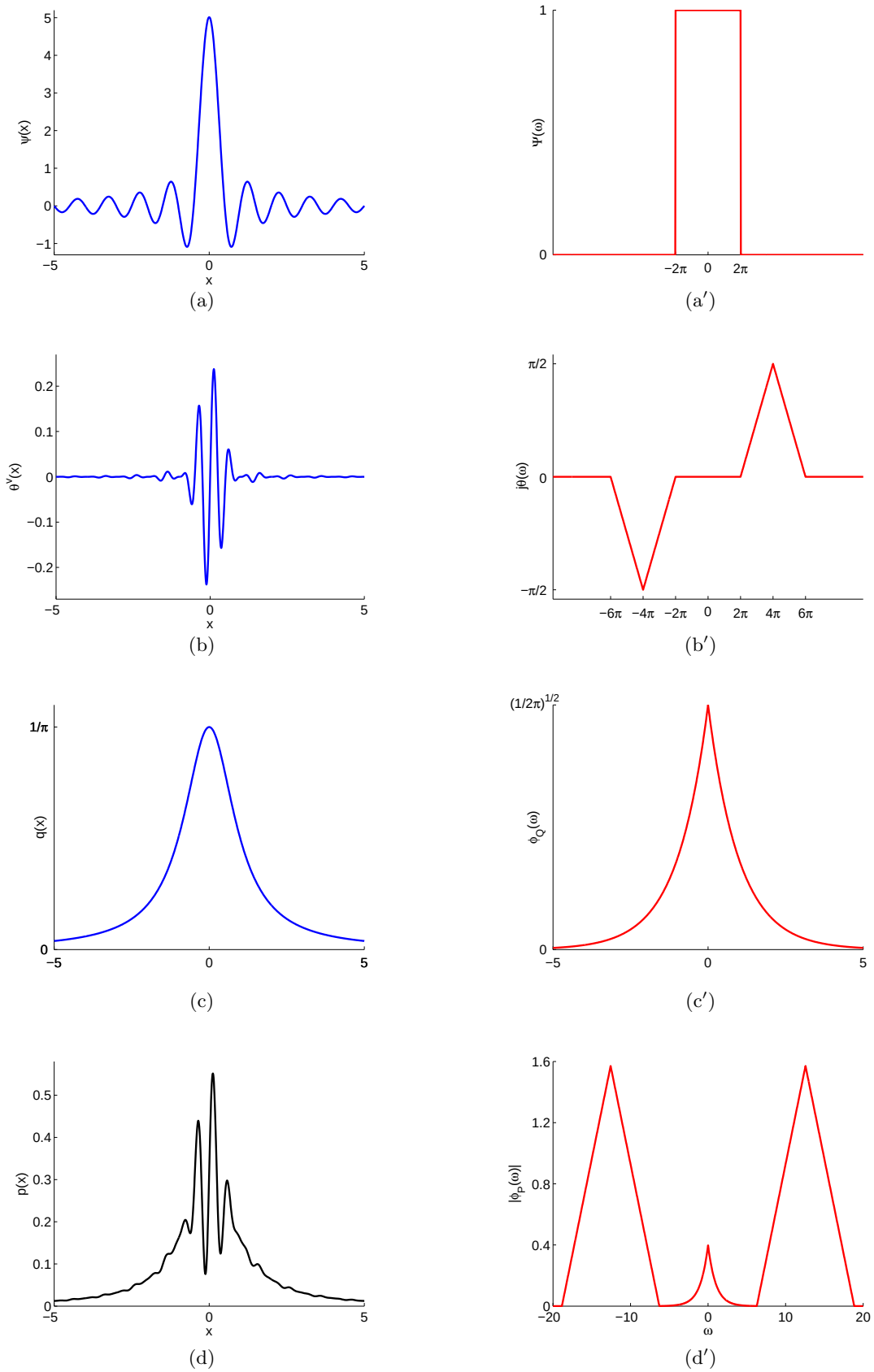



Figure 3: (a-a′) $\psi$ and its spectrum $\Psi$, (b-b′) $\check{\theta}$ and $\theta$, (c-c′) the Cauchy distribution, $q$ and its characteristic function $\phi_Q$ and (d-d′) $p = q + \check{\theta}$ and $|\phi_P|$. See Example 22 for details.

| $\psi(x)$, $\Omega = \text{supp}(\Lambda)$ | $\mathfrak{D}$ | Characteristic | $\gamma_{\mathcal{F}}$ | Reference |
|---|---|---|---|---|
| $\Omega = \mathbb{R}^d$ | $\mathfrak{S}$ | Yes | Metric | Theorem 16 |
| $\text{supp}(\psi)$ is compact | $\mathfrak{S}$ | Yes | Metric | Corollary 17 |
| $\Omega \subsetneq \mathbb{R}^d$ has a non-empty interior | $\{P : \text{supp}(P) \text{ is compact}, \phi_P \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\}$ | Yes | Metric | Theorem 19 |
| $\Omega \subsetneq \mathbb{R}^d$ | $\{P : \text{supp}(P) \text{ is not compact}, \phi_P \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)\}$ | No | Pseudometric | Theorem 21 |

Table 2: $k$ satisfies assumption *(A-1)* and is the Fourier transform of a finite nonnegative Borel measure $\Lambda$ 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 $\phi_P$ 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 a RKHS $(\mathcal{H}, k)$.

# 6 A Limitation of Maximum Mean Discrepancy

Till now, we have studied the behavior of $\gamma_{\mathcal{F}}$ and showed that it depends on the support of the spectrum of the kernel. Basically, we showed that the mapping, $\Pi$ induced by the kernel with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ is not injective. Therefore, if these kernels are used in applications like homogeneity testing or independence testing, these tests fail. So, in practice, one should use kernels that guarantee $\Pi$ to be injective, which in our case is provided by positive definite convolution kernels on $\mathbb{R}^d$ with $\text{supp}(\Lambda) = \mathbb{R}^d$. However, in the following, we present a general result about the behavior of $\gamma_{\mathcal{F}}$ that holds for all positive definite kernels. To this end, we motivate this result through the following example.

**Example 23.** *Let $P$ be defined as $p(x) = q(x) + \alpha q(x)\sin(\nu\pi x)$, $\alpha \in \mathbb{R}$, $\nu \in \mathbb{R}\backslash\{0\}$ where $q$ is a symmetric probability density function. Consider a $B_1$-spline kernel on $\mathbb{R}$ given by $k(x, y) = \psi(x - y)$ where*

$$\psi(x) = \begin{cases} 1 - |x|, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{16}$$

*whose Fourier transform is given by $\Psi(\omega) = \frac{2}{\pi}\frac{\sin^2\frac{\omega}{2}}{\omega^2}$. Since, $\psi$ is characteristic to $\mathfrak{S}$, $\gamma_{\mathcal{F}}(P, Q) > 0$ (see Theorem 16). However, it would be of interest to study the behavior of $\gamma_{\mathcal{F}}(P, Q)$ as a function of $\nu$. Note that with increasing $|\nu|$, $p$ has higher frequency components in its Fourier spectrum and therefore appears more noisy as shown in Figure 4. In Figure 4, (a-c) show the plots of $p$ when $q = \mathcal{U}[-1, 1]$ and (a′-c′) show the plots of $p$ when $q = \mathcal{N}(0, 2)$ for $\nu = 0, 2$ and $7.5$ with $\alpha = \frac{1}{2}$.*

*Since the empirical estimate of MMD, $\widehat{\gamma_{\mathcal{F}}}$ is a consistent estimate of $\gamma_{\mathcal{F}}$, we study the behavior of $\gamma_{\mathcal{F}}$ as a function of $\nu$ through $\widehat{\gamma_{\mathcal{F}}}$.[14] Figure 5(a) shows the behavior of the empirical estimate of $\gamma^2_{\mathcal{F}}$ as a function of $\nu$ for $q = \mathcal{U}[-1, 1]$ and $q = \mathcal{N}(0, 2)$ using the $B_1$-spline kernel in Eq. (16). Since the Gaussian kernel, $\exp\left(-(x - y)^2\right)$ is also a characteristic kernel, its effect on the behavior of $\gamma^2_{\mathcal{F},u}(m, m)$ is shown in Figure 5(b) in comparison to that of $B_1$-spline kernel.*

*Two observations are in place from Figure 5. The first observation is that $\gamma^2_{\mathcal{F},u}(m, m)$ decays with increasing $|\nu|$ and can be made as small as possible by choosing sufficiently large $|\nu|$. The second observation is from Figure 5(a) wherein $\gamma^2_{\mathcal{F},u}(m, m)$ has troughs at $\nu = \frac{\omega_0}{\pi}$ where $\omega_0 = \{\omega : \Psi(\omega) = 0\}$. This means*

[14] Starting from the expression for $\gamma_{\mathcal{F}}$ in Eq. (3), 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 $\gamma^2_{\mathcal{F},u}(m, m) = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} h(Z_i, Z_j)$, which is a one-sample $U$-statistic with $h(Z_i, Z_j) := k(X_i, X_j) + k(Y_i, Y_j) - k(X_i, Y_j) - k(X_j, Y_i)$, where $Z_1, \ldots, Z_m$ are $m$ i.i.d. random variables with $Z_i := (X_i, Y_i)$ (see [10, Lemma 8]).

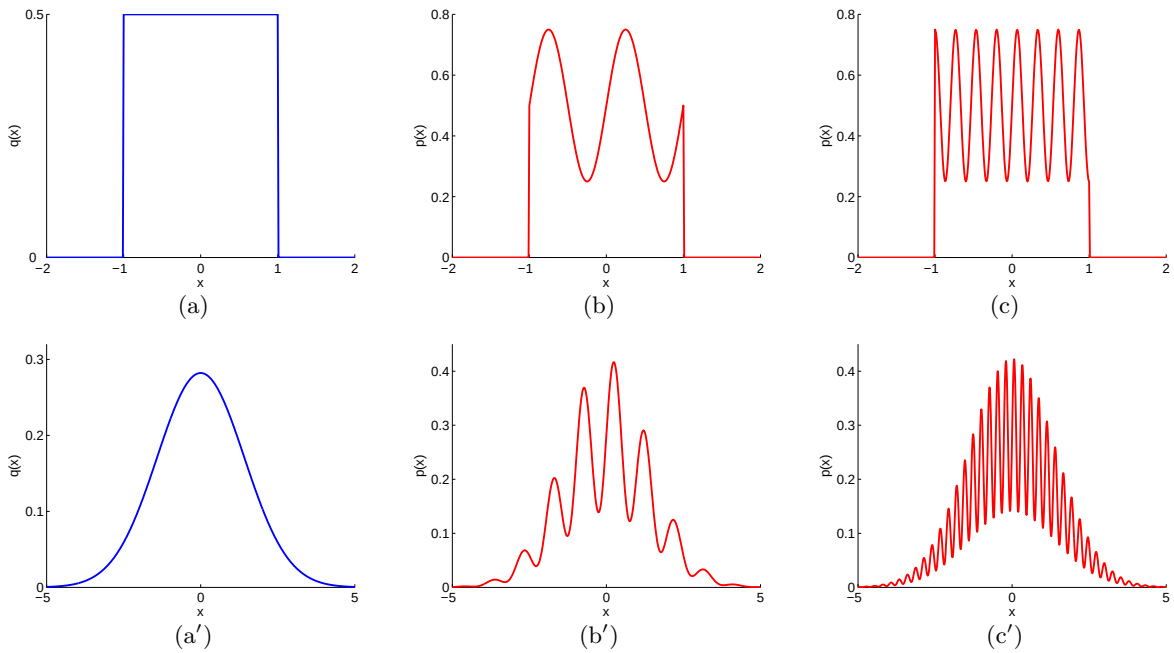



Figure 4: (a) $q = \mathcal{U}[-1, 1]$, (a′) $q = \mathcal{N}(0, 2)$. (b-c) and (b′-c′) denote $p(x)$ computed as $p(x) = q(x) + \frac{1}{2}q(x)\sin(\nu\pi x)$ with $q = \mathcal{U}[-1, 1]$ and $q = \mathcal{N}(0, 2)$ respectively. $\nu$ is chosen to be 2 in (b,b′) and 7.5 in (c,c′). See Example 23 for details.

*that though the $B_1$-spline kernel is characteristic to $\mathfrak{S}$, in practice, it becomes harder to distinguish between $P$ and $Q$ when $P$ is constructed as defined before with $\nu = \frac{\omega_0}{\pi}$.*

For characteristic kernels, although $\gamma_{\mathcal{F}}(P, Q) > 0$ when $P \neq Q$, the previous example demonstrates that one can construct distributions such that $\widehat{\gamma_{\mathcal{F}}}(m, m)$ is indistinguishable from zero for a given sample size $m$. In other words, although $P \neq Q$, they are not distinguishable by the two-sample test based on MMD and one needs a large data sample to distinguish them. Since $\widehat{\gamma_{\mathcal{F}}}$ is a consistent estimate of $\gamma_{\mathcal{F}}$, one would expect similar behavior from $\gamma_{\mathcal{F}}$.

We prove this phenomenon in Theorem 25 by 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 small, therefore making it hard to detect a non-zero value of the *population* MMD on the basis of a *finite sample*, as in Example 23. Here, $\varphi_l \in L^2(M)$ represents bounded orthonormal eigenfunctions of a positive definite integral operator associated with $k$.[15] The result in Theorem 25 is more general in the sense that no assumption about the translation invariance of the kernel is made (see assumption (*A-1*)).

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

$$\gamma_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{H}} \frac{|Pf - Qf|}{\|f\|_{\mathcal{H}}}. \tag{17}$$

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 $\mathcal{H}$ chooses $f$ whose $\|f\|_{\mathcal{H}}$ is large. More specifically, higher order eigenfunctions of the kernel ($\varphi_l$ 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 start, we need the following lemma which we quote from [13, Lemma 6].

[15] See [18, Theorem 2.10] for definition of positive definite integral operator and its corresponding eigenfunctions.

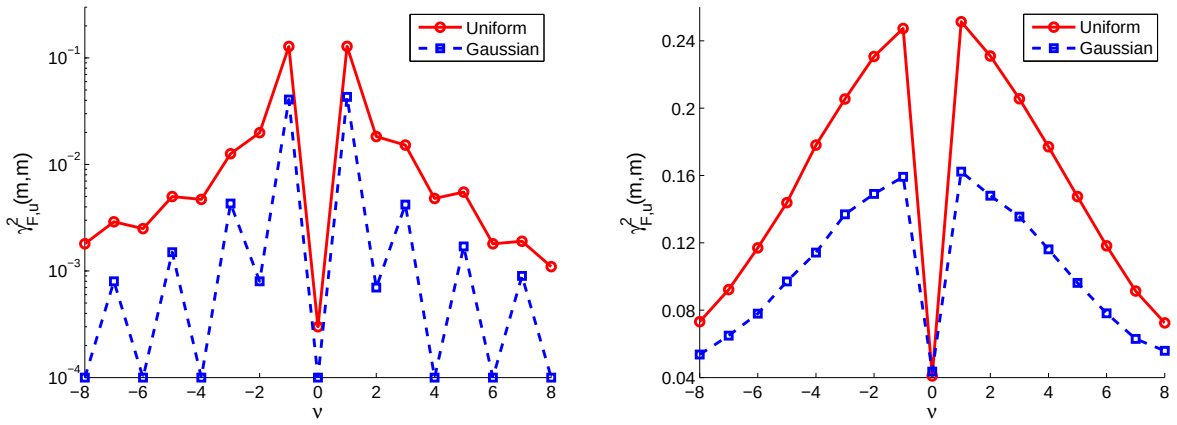



Figure 5: Behavior of $\gamma^2_{\mathcal{F},u}(m,m)$ w.r.t. $\nu$ for (a) $B_1$-spline kernel and (b) Gaussian kernel. "Uniform" corresponds to $q(x) = \mathcal{U}[-1,1]$ and "Gaussian" corresponds to $q(x) = \mathcal{N}(0,2)$. Here $m = 1000$. See Example 23 for details.

**Lemma 24** (Rate of decay of expansion coefficients [13])**.** *Let $\mathcal{F}$ be a unit ball in a RKHS $(\mathcal{H},k)$ defined on $M$. Let $\varphi_l \in L^2(M)$ be absolutely bounded orthonormal eigenfunctions and $\lambda_l$ be the corresponding eigenvalues (arranged in a decreasing order for increasing $l$) of a positive definite integral operator associated with $k$. Assume that $\lambda_l^{-1}$ increases superlinearly with $l$. Then for $f \in \mathcal{F}$ where $f(x) := \sum_{i=1}^{\infty} \tilde{f}_i \varphi_i(x)$, we have $\{|\tilde{f}_i|\}_{i=1}^{\infty} \in \ell_1$ and for every $\epsilon > 0$, $\exists\, l_0 \in \mathbb{N}$ such that $|\tilde{f}_l| < \epsilon$ if $l > l_0$.*

**Theorem 25** ($P \neq Q$ can give small MMD)**.** *Let us assume that the conditions mentioned in Lemma 24 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 $\beta$ 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 $e(x) = I_M(x)$. For $P$ to be a probability distribution, the following conditions need to be satisfied.

$$\int_M [\alpha_l e(x) + \beta\varphi_l(x)]\, dx = 0, \tag{18}$$

$$\min_{x\in M} [q(x) + \alpha_l e(x) + \beta\varphi_l(x)] \geq 0. \tag{19}$$

Expanding $e(x)$ and $f(x)$ in the orthonormal basis set of $\{\varphi_l\}_{l=1}^{\infty}$, we get $e(x) = \sum_{l=1}^{\infty} \varphi_l(x)\langle e, \varphi_l\rangle_{L^2(M)} =: \sum_{l=1}^{\infty} \tilde{e}_l\varphi_l(x)$ and $f(x) = \sum_{l=1}^{\infty} \varphi_l(x)\langle f, \varphi_l\rangle_{L^2(M)} =: \sum_{l=1}^{\infty} \tilde{f}_l\varphi_l(x)$. Therefore,

$$\begin{aligned} Pf - Qf &= \int_M f(x)\,[\alpha_l e(x) + \beta\varphi_l(x)]\, dx = \int_M \left[\alpha_l \sum_{i=1}^{\infty} \tilde{e}_i\varphi_i(x) + \beta\varphi_l(x)\right]\left[\sum_{t=1}^{\infty} \tilde{f}_t\varphi_t(x)\right] dx \\ &= \alpha_l \sum_{i=1}^{\infty} \tilde{e}_i\tilde{f}_i + \beta\tilde{f}_l, \end{aligned} \tag{20}$$

where we used the fact that $\langle\varphi_i,\varphi_t\rangle_{L^2(M)} = \int_M \varphi_i(x)\varphi_t(x)\,dx = \delta_{it}$.[16] Rewriting Eq. (18) and substituting for $e(x)$ gives

$$\int_M [\alpha_l e(x) + \beta\varphi_l(x)]\, dx = \int_M e(x)\,[\alpha_l e(x) + \beta\varphi_l(x)]\, dx = \alpha_l \sum_{i=1}^{\infty} \tilde{e}_i^2 + \beta\tilde{e}_l = 0$$

which implies

$$\alpha_l = -\frac{\beta\tilde{e}_l}{\sum_{i=1}^{\infty} \tilde{e}_i^2}. \tag{21}$$

[16] Here $\delta$ is used in the Kronecker sense.

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

$$P\varphi_t - Q\varphi_t = \beta\delta_{tl} - \beta\frac{\tilde{e}_t\tilde{e}_l}{\sum_{i=1}^{\infty}\tilde{e}_i^2} = \beta\delta_{tl} - \beta\tau_{tl}, \tag{22}$$

where $\tau_{tl} := \frac{\tilde{e}_t\tilde{e}_l}{\sum_{i=1}^{\infty}\tilde{e}_i^2}$. By Lemma 24, $\{|\tilde{e}_l|\}_{l=1}^{\infty} \in \ell_1 \Rightarrow \sum_{i=1}^{\infty}\tilde{e}_i^2 < \infty$ and choosing large enough $l$ gives $|\tau_{tl}| < \epsilon$ 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 $\alpha_l$ in Eq. (20), we have

$$\gamma_{\mathcal{F}}(P, Q) = \sup\left\{\beta\sum_{i=1}^{\infty}\nu_{il}\tilde{f}_i \;:\; \sum_{i=1}^{\infty}\frac{\tilde{f}_i^2}{\lambda_i} \leq 1\right\} \tag{23}$$

where we used the definition of RKHS norm as $\|f\|_{\mathcal{H}} := \sum_{i=1}^{\infty}\frac{\tilde{f}_i^2}{\lambda_i}$ and $\nu_{il} := \delta_{il} - \tau_{il}$. Eq. (23) is a convex quadratic program in $\{\tilde{f}_i\}_{i=1}^{\infty}$. Solving the Langrangian yields $\tilde{f}_i = \frac{\nu_{il}\lambda_i}{\sqrt{\sum_{i=1}^{\infty}\nu_{il}^2\lambda_i}}$. Therefore, $\gamma_{\mathcal{F}}(P, Q) = \beta\sqrt{\sum_{i=1}^{\infty}\nu_{il}^2\lambda_i} = \beta\sqrt{\lambda_l - 2\tau_{ll}\lambda_l + \sum_{i=1}^{\infty}\tau_{il}^2\lambda_i} \to 0$ as $l \to \infty$ because (i) by choosing sufficiently large $l$, $|\tau_{il}| < \epsilon$ for any arbitrary $\epsilon > 0$ and (ii) $\lambda_l \to 0$ as $l \to \infty$ [18, Theorem 2.10]. □

# 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. They 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 (also called convolution kernels) on $\mathbb{R}^d$. We showed that the support of the Fourier spectrum of the kernel determines the injective nature of the embedding. In particular, the necessary and sufficient condition for the embedding to be injective is that the Fourier spectrum of the convolution 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 used Fourier analysis in $\mathbb{R}^d$. Since Fourier theory is available for more general groups other than $\mathbb{R}^d$, one direction for our future work is to extend the analysis for more general case of positive definite kernels defined by group operation.

# Appendix A Basics of Distributions and Fourier Transforms

We briefly give definitions and basic properties of distributions and Fourier transforms of distributions. For complete references, see, for example, [17, Chapters 6,7], [25] and [15].

Distribution theory, invented by Laurent Schwartz, frees differential calculus from certain difficulties that arise because nondifferentiable functions exist. This is done by extending it to a class of objects (called *distributions* or *generalized functions*) which is much larger than the class of differentiable functions to which calculus applies in its original form.

The basic idea for generalizing the notion of function in the context of distributions is to reinterpret a measurable function $f$ as being something that assigns the number $\int f\varphi$ to every suitably chosen "test function" $\varphi$, rather than as being something that assigns the number $f(x)$ to each $x \in \mathbb{R}^d$. So, a well-chosen class of test functions must be specified.

**Spaces $\mathscr{D}(\mathbb{R}^d)$ and $\mathscr{D}'(\mathbb{R}^d)$:** One commonly used test function class, denoted by $\mathscr{D}(\mathbb{R}^d)$ (or simply $\mathscr{D}_d$) is the space of all $\varphi \in C^{\infty}(\mathbb{R}^d)$ whose support is compact. In other words,

$$\mathscr{D}(\mathbb{R}^d) = \{\varphi : \mathbb{R}^d \to \mathbb{C} \,|\, \varphi \in C^{\infty}(\mathbb{R}^d), \text{supp}(\varphi) \text{ is bounded}\}$$

where $\text{supp}(\varphi) = \overline{\{x \in \mathbb{R}^d \,|\, \varphi(x) \neq 0\}}$. For an open set $U \subset \mathbb{R}^d$, $\mathscr{D}(U)$ denotes the subspace of $\mathscr{D}_d$ consisting of the functions whose support is contained in $U$. A linear functional on $\mathscr{D}_d$ which is continuous with respect

to a certain topology (see [17, Definition 6.3]) is called a *distribution* in $\mathbb{R}^d$ and the space of all distributions in $\mathbb{R}^d$ is denoted by $\mathscr{D}'_d$. For example, the Dirac-delta function is a distribution defined as $\delta(\varphi) = \varphi(0)$.

**Functions and measures as distributions:** If $f$ is *locally integrable* on $\mathbb{R}^d$ (this means that $f$ is Lebesgue measurable and $\int_K |f(x)|\,dx < \infty$ for every compact $K \in \mathbb{R}^d$), then the functional $D_f$ defined by

$$D_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)\,dx,\ \forall\,\varphi \in \mathscr{D}_d \tag{24}$$

is a distribution. Similarly, if $\mu$ is a Borel measure on $\mathbb{R}^d$, then

$$D_\mu(\varphi) = \int_{\mathbb{R}^d} \varphi\,d\mu,\ \forall\,\varphi \in \mathscr{D}_d \tag{25}$$

defines a distribution $D_\mu$ in $\mathbb{R}^d$, which is usually identified with $\mu$.

**Elementary operations on distributions:**

1. *Derivative of a distribution:* If $\alpha$ is a multi-index and $D \in \mathscr{D}'_d$, then

$$(T^\alpha D)(\varphi) = (-1)^{|\alpha|} D(T^\alpha \varphi),\ \forall\,\varphi \in \mathscr{D}_d \tag{26}$$

   defines a linear functional $T^\alpha D$ on $\mathscr{D}_d$.[17]

2. *Multiplication by functions:* Suppose $D \in \mathscr{D}'_d$ and $f \in C^\infty(\mathbb{R}^d)$. Then

$$(fD)(\varphi) = D(f\varphi),\ \forall\,\varphi \in \mathscr{D}_d. \tag{27}$$

3. *Support of a distribution:* Suppose $D \in \mathscr{D}'_d$. If $U$ is an open set of $\mathbb{R}^d$ and if $D(\varphi) = 0$ for every $\varphi \in \mathscr{D}(U)$, then $D$ is said to *vanish* or *null* in $U$. Let $W$ be the union of all open $U \subset \mathbb{R}^d$ in which $D$ vanishes. The complement of $W$ is the *support* of $D$.

**Fourier transforms on $L^1(\mathbb{R}^d)$ :** The normalized Lebesgue measure on $\mathbb{R}^d$ is the measure $m_d$ defined by

$$dm_d(x) = \frac{1}{(2\pi)^{\frac{d}{2}}}\,dx.$$

The Fourier transform of a function $f \in L^1(\mathbb{R}^d)$ is the function $\hat{f}$ defined by

$$\hat{f}(y) = \int_{\mathbb{R}^d} e^{-jy^T x} f(x)\,dm_d(x),\ \forall\,y \in \mathbb{R}^d \tag{28}$$

and the inverse Fourier transform is the function $\check{f}$ defined by

$$\check{f}(x) = \int_{\mathbb{R}^d} e^{jx^T y} f(y)\,dm_d(y),\ \forall\,x \in \mathbb{R}^d \tag{29}$$

where $j = \sqrt{-1}$. To extend the Fourier transform to distributions, consider the following for $f \in L^1(\mathbb{R}^d)$ with $\varphi \in \mathscr{D}_d$. Since $\hat{f}$ is a continuous function, in the sense of distributions we have

$$\begin{aligned} D_{\hat{f}}(\varphi) &= \int_{\mathbb{R}^d} \hat{f}(x)\varphi(x)\,dx = \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d} e^{-jx^T y} f(y)\,dy\right] \varphi(x)\,dx \\ &= \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d} e^{-jy^T x}\varphi(x)\,dx\right] f(y)\,dy \ \text{(by Fubini's theorem)} \\ &= \int_{\mathbb{R}^d} f(y)\hat{\varphi}(y)\,dy = D_f(\hat{\varphi}). \end{aligned}$$

[17] The term multi-index denotes an ordered $d$-tuple $\alpha = (\alpha_1, \ldots, \alpha_d)$ of non-negative $\alpha_i$. With each multi-index $\alpha$ is associated the differential operator, $T^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

But, $D_f(\hat{\varphi})$ will make sense only if $\hat{\varphi} \in \mathscr{D}_d$. Since, supp$(\varphi)$ is compact, $\hat{\varphi} \in C^\infty(\mathbb{R}^d)$. But, $\hat{\varphi}$ is not guaranteed to have compact support for every $\varphi \in \mathscr{D}_d$. Therefore, to extend the Fourier transform to distributions, a suitable space of test functions must be specified.

**Spaces $\mathscr{S}(\mathbb{R}^d)$ and $\mathscr{S}'(\mathbb{R}^d)$:** A function $f : \mathbb{R}^d \to \mathbb{C}$ is said to decay rapidly, or be rapidly decreasing, if for all $N \in \mathbb{N}$,

$$\sup_{|\alpha| \le N} \sup_{x \in \mathbb{R}^d} (1 + |x|^2)^N |(T_\alpha f)(x)| < \infty, \tag{30}$$

where $T_\alpha = j^{-|\alpha|} T^\alpha = \left(\frac{1}{j}\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{1}{j}\frac{\partial}{\partial x_d}\right)^{\alpha_d}$. Clearly, $f \in C^\infty(\mathbb{R}^d)$. $\mathscr{S}(\mathbb{R}^d)$ (or simply $\mathscr{S}_d$), called the Schwartz class, denotes the vector space of rapidly decreasing functions. It can be shown that $\mathscr{S}_d$ is invariant under Fourier transforms, i.e., $f \in \mathscr{S}_d \Rightarrow \hat{f} \in \mathscr{S}_d$. In addition, $\check{\hat{f}} = f$ and $\hat{\check{f}} = f$. Also, $\mathscr{D}_d \subset \mathscr{S}_d$. A linear continuous functional $D$ over the space $\mathscr{S}_d$ of test functions is called a *tempered distribution* and the space of all tempered distributions in $\mathbb{R}^d$ is denoted by $\mathscr{S}'_d$. Every linear continuous functional on $\mathscr{S}_d$ is also a linear continuous functional on $\mathscr{D}_d$ and therefore $\mathscr{S}'_d \subset \mathscr{D}'_d$. It can be shown that $f \in L_p(\mathbb{R}^d)$, $p \ge 1$ is a tempered distribution. Similarly, it can be shown that any finite Borel measure, $\mu$ on $\mathbb{R}^d$ is a tempered distribution.

**Fourier transform on $\mathscr{S}'_d$:** The Fourier transform and the inverse Fourier transform of $f \in \mathscr{S}'_d$ are defined by

$$\widehat{D_f}(\varphi) = D_f(\hat{\varphi}), \forall\, \varphi \in \mathscr{S}_d \tag{31}$$

$$(D_f)^\vee(\varphi) = D_f(\check{\varphi}), \forall\, \varphi \in \mathscr{S}_d \tag{32}$$

respectively. The Fourier transform is a linear, one-to-one, bicontinuous mapping from $\mathscr{S}'_d$ to $\mathscr{S}'_d$. If $f \in \mathscr{S}'_d$, then $\hat{f}$, $\check{f} \in \mathscr{S}'_d$ and $\check{\hat{f}} = \hat{\check{f}} = f$. If $f \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$, then $\widehat{D_f} = D_{\hat{f}}$.

**Convolution:** If $f$ and $g$ are complex functions in $\mathbb{R}^d$, their convolution $f * g$ is defined by

$$(f * g)(x) = \int_{\mathbb{R}^d} f(y) g(x - y)\, dy, \tag{33}$$

provided that the integral exists for all (or at least for almost all) $x \in \mathbb{R}^d$, in the Lebesgue sense. Let $\mu$ 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 $\mu$, $f * \mu$, which is a bounded measurable function is defined by

$$(f * \mu)(x) = \int_{\mathbb{R}^d} f(x - y)\, d\mu(y). \tag{34}$$

# Appendix B Lemmas Used in the Proofs

We show five lemmas used in the proofs in §4 and §5. The first two may be basic but we provide the complete proofs for convenience. The remaining lemmas include the well-known Riemann-Lebesgue lemma, Paley-Wiener theorem and a corollary of Paley-Wiener theorem, whose proofs we do not provide as they are more involved.

**Lemma 26** (Fourier transform of a measure)**.** *Let $\mu$ be a finite Borel measure on $\mathbb{R}^d$. The Fourier transform of $\mu$ is a tempered distribution given by*

$$\hat{\mu}(\omega) = \int_{\mathbb{R}^d} e^{-j\omega^T x}\, d\mu(x), \tag{35}$$

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

*(i)* $\overline{\hat{\mu}(\omega)} = \hat{\mu}(-\omega)$.

*(ii)* $\hat{\mu}(\omega) = \hat{\mu}(-\omega)$ *if and only if* $\mu(x) + \mu(-x) = C < \infty, \forall\, x$.

*Proof.* Let $D_\mu$ denote a tempered distribution defined by $\mu$. We have for $\varphi \in \mathscr{S}_d$,

$$\widehat{D_\mu}(\varphi) = D_\mu(\hat{\varphi}) = \int_{\mathbb{R}^d} \hat{\varphi}(\omega)\, d\mu(\omega) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-j\omega^T x} \varphi(x)\, dm_d(x)\, d\mu(\omega).$$

From Fubini's theorem,

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

which proves Eq. (35). Clearly $\hat{\mu}$ is bounded as $|\hat{\mu}(\omega)| \le \int_{\mathbb{R}^d} |e^{-j\omega^T x}|\, d\mu(x) = \int_{\mathbb{R}^d} d\mu(x) = \hat{\mu}(0) = 1$. $\hat{\mu}$ is uniformly continuous on $\mathbb{R}^d$ as for any $\omega \in \mathbb{R}^d$, we have $\lim_{h\to 0} |\hat{\mu}(\omega + h) - \hat{\mu}(\omega)| \le \lim_{h\to 0} \int_{\mathbb{R}^d} |e^{-jh^T x} - 1|\, d\mu(x) = 0$ (by Lebesgue's dominated convergence theorem).

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

*(ii)* ($\Rightarrow$) If $\mu(x)+\mu(-x) = C$, then for any continuous function $f$, we have $\int_{\mathbb{R}^d} f(x)\, d\mu(x) = \int_{\mathbb{R}^d} f(-x)\, d\mu(x)$. Consider

$$\begin{aligned} \hat{\mu}(\omega) &= \int_{\mathbb{R}^d} e^{-j\omega^T x}\, d\mu(x) = \int_{\mathbb{R}^d} e^{j\omega^T x}\, d\mu(x) \text{ (follows from the assumption)} \\ &= \int_{\mathbb{R}^d} e^{-j(-\omega)^T x}\, d\mu(x) = \hat{\mu}(-\omega). \end{aligned}$$

($\Leftarrow$) If $\hat{\mu}(\omega) = \hat{\mu}(-\omega)$, then $\int_{\mathbb{R}^d} e^{-j\omega^T x}\, d\mu(x) = \int_{\mathbb{R}^d} e^{j\omega^T x}\, d\mu(x) = -\int_{\mathbb{R}^d} e^{-j\omega^T x}\, d\mu(-x)$. Therefore, $\int_{\mathbb{R}^d} e^{-j\omega^T x}\, [d\mu(x) + d\mu(-x)] = \int_{\mathbb{R}^d} e^{-j\omega^T x}\, d(\mu(x) + \mu(-x)) = 0, \forall\, \omega$ implies $\mu(x) + \mu(-x) = C < \infty, \forall\, x$. □

**Remark 27.** *Property (i) in the above lemma 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. Property (ii) shows that real symmetric tempered distributions have real symmetric Fourier transforms.*

The following result is popularly known as the *Convolution theorem.*

**Lemma 28.** *Let* $\mu$ *be a finite Borel measure and* $f$ *be a bounded function on* $\mathbb{R}^d$. *Suppose* $f$ *is expressed by*

$$f(x) = \int_{\mathbb{R}^d} e^{jx^T\omega}\, d\Lambda(\omega), \tag{36}$$

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

$$(f * \mu)^\wedge = \hat{\mu}\Lambda, \tag{37}$$

*where the right hand side is a finite Borel measure*[18] *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

$$f * \mu = (\hat{\mu}\Lambda)^\vee. \tag{39}$$

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

$$\begin{aligned} (\hat{\mu}\Lambda)^\vee(\varphi) = (\hat{\mu}\Lambda)(\check{\varphi}) &= \int_{\mathbb{R}^d} \check{\varphi}(x)\hat{\mu}(x)\, d\Lambda(x) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{jx^T\omega} \varphi(\omega)\, dm_d(\omega) \int_{\mathbb{R}^d} e^{-jx^T y}\, d\mu(y)\, d\Lambda(x). \end{aligned}$$

[18]The finite Borel measure in Eq. (37) is defined in the following sense. Let $\mu$ be a finite Borel measure and $f$ be bounded measurable function on $\mathbb{R}^d$. Define a finite Borel measure $f\mu$ by

$$(f\mu)(E) = \int_{\mathbb{R}^d} I_E(x) f(x)\, d\mu(x), \tag{38}$$

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

By Fubini's theorem, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} e^{j(\omega-y)^T x}\, d\Lambda(x)\right]\varphi(\omega)\, dm_d(\omega)\, d\mu(y) &= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} f(\omega-y)\, d\mu(y)\right]\varphi(\omega)\, dm_d(\omega)\\ &= \int_{\mathbb{R}^d}(f*\mu)(\omega)\varphi(\omega)\, dm_d(\omega),\end{aligned}$$

which proves Eq. (39). □

The following result called the Riemann-Lebesgue lemma is quoted from [17, Theorem 7.5].

**Lemma 29** (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 lemma is a version of the *Paley-Wiener theorem* for $C^\infty$ functions whose proof is given in [22, Theorem 7.2.2].

**Lemma 30** (Paley-Wiener)**.** *Let $f$ be a $C^\infty$ function supported in $[-\beta, \beta]$. Then $\hat{f}(\omega + j\sigma)$ is a entire function of exponential type $\beta$, i.e., $\exists C$ such that*

$$\left|\hat{f}(\omega + j\sigma)\right| \leq Ce^{\beta|\sigma|}, \tag{40}$$

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

$$\left|\hat{f}(\omega)\right| \leq \frac{c_n}{(1+|\omega|)^n}, \forall n \in \mathbb{N}. \tag{41}$$

*Conversely, if $F(\omega + j\sigma)$ is an entire function of exponential type $\beta$, 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 whose proof is given in [14, Theorem 2.6].

**Lemma 31** ([14])**.** *Let $g \neq 0$ has a compact support then its Fourier transform, $\hat{g}$ cannot be zero on a whole interval. Similarly, if $\hat{g} \neq 0$ has a compact support then $g$ cannot be zero on a whole interval.*

# Acknowledgments

BKS wishes to acknowledge the support from the Max Planck Institute for Biological Cybernetics, Tübingen where a part of the research was carried out while the author was an intern. BKS also wishes to acknowledge the support from NSF grant 0625409, the Fair Isaac Corporation and the University of California MICRO program, which partly funded the research.

# References

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

[2] P. Brémaud. *Mathematical Principles of Signal Processing.* Springer-Verlag, New York, 2001.

[3] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, 2:299–318, 1967.

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

[5] R. M. Dudley, E. Gine, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

[6] A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *The Annals of Statistics*, 5(1):88–97, 1977.

[7] K. Fukumizu, F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

[8] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, 2008. To appear.

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

[10] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. Technical Report 157, MPI for Biological Cybernetics, 2007.

[11] 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, Cambridge, MA, 2007. MIT Press.

[12] 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.

[13] 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.

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

[15] R. S. Pathak. *A Course in Distribution Theory and Applications*. Narosa Publishing House, New Delhi, 2001.

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

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

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

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

[20] 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.

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

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

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

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

[25] A. H. Zemanian. *Distribution Theory and Transform Analysis*. McGraw-Hill, New York, 1965.