Journal of Machine Learning Research 11 (2010) 1517-1561 Submitted 8/09; Revised 3/10; Published 4/10

Hilbert Space Embeddings and Metrics on Probability Measures

Bharath K. Sriperumbudur BHARATHSV@UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407, USA

Arthur Gretton* ARTHUR@TUEBINGEN.MPG.DE
MPI for Biological Cybernetics
Spemannstraße 38
72076, Tübingen, Germany

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa
Tokyo 190-8562, Japan

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
MPI for Biological Cybernetics
Spemannstraße 38
72076, Tübingen, Germany

Gert R. G. Lanckriet GERT@ECE.UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407, USA

Editor: Ingo Steinwart

Abstract

A Hilbert space embedding for probability measures has recently been proposed, with applications including dimensionality reduction, homogeneity testing, and independence testing. This embedding represents any probability measure as a mean element in a reproducing kernel Hilbert space (RKHS). A pseudometric on the space of probability measures can be defined as the distance between distribution embeddings: we denote this as γ_k, indexed by the kernel function k that defines the inner product in the RKHS.

We present three theoretical properties of γ_k. First, we consider the question of determining the conditions on the kernel k for which γ_k is a metric: such k are denoted *characteristic kernels*. Unlike pseudometrics, a metric is zero only when two distributions coincide, thus ensuring the RKHS embedding maps all distributions uniquely (i.e., the embedding is injective). While previously published conditions may apply only in restricted circumstances (e.g., on compact domains), and are difficult to check, our conditions are straightforward and intuitive: *integrally strictly positive definite kernels* are characteristic. Alternatively, if a bounded continuous kernel is translation-invariant on $\mathbb{R}^d$, then it is characteristic if and only if the support of its Fourier transform is the entire $\mathbb{R}^d$. Second, we show that the distance between distributions under γ_k results from an interplay between the properties of the kernel and the distributions, by demonstrating that distributions are close in the embedding space when their differences occur at higher frequencies. Third, to understand the

*. Also at Carnegie Mellon University, Pittsburgh, PA 15213, USA.

nature of the topology induced by γ_k, we relate γ_k to other popular metrics on probability measures, and present conditions on the kernel k under which γ_k metrizes the weak topology.

Keywords: probability metrics, homogeneity tests, independence tests, kernel methods, universal kernels, characteristic kernels, Hilbertian metric, weak topology

1. Introduction

The concept of distance between probability measures is a fundamental one and has found many applications in probability theory, information theory and statistics (Rachev, 1991; Rachev and Rüschendorf, 1998; Liese and Vajda, 2006). In statistics, distances between probability measures are used in a variety of applications, including hypothesis tests (homogeneity tests, independence tests, and goodness-of-fit tests), density estimation, Markov chain monte carlo, etc. As an example, homogeneity testing, also called the two-sample problem, involves choosing whether to accept or reject a null hypothesis $H_0 : \mathbb{P} = \mathbb{Q}$ versus the alternative $H_1 : \mathbb{P} \neq \mathbb{Q}$, using random samples $\{X_j\}_{j=1}^m$ and $\{Y_j\}_{j=1}^n$ drawn i.i.d. from probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on a topological space $(M, \mathcal{A})$. It is easy to see that solving this problem is equivalent to testing $H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0$ versus $H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0$, where γ is a metric (or, more generally, a semi-metric[1]) on the space of all probability measures defined on M. The problems of testing independence and goodness-of-fit can be posed in an analogous form. In non-parametric density estimation, $\gamma(p_n, p_0)$ can be used to study the quality of the density estimate, p_n, that is based on the samples $\{X_j\}_{j=1}^n$ drawn i.i.d. from p_0. Popular examples for γ in these statistical applications include the *Kullback-Leibler divergence*, the *total variation distance*, the *Hellinger distance* (Vajda, 1989)—these three are specific instances of the generalized ϕ-divergence (Ali and Silvey, 1966; Csiszár, 1967)—the *Kolmogorov distance* (Lehmann and Romano, 2005, Section 14.2), the *Wasserstein distance* (del Barrio et al., 1999), etc.

In probability theory, the distance between probability measures is used in studying limit theorems, the popular example being the central limit theorem. Another application is in metrizing the weak convergence of probability measures on a separable metric space, where the *Lévy-Prohorov distance* (Dudley, 2002, Chapter 11) and *dual-bounded Lipschitz distance* (also called the *Dudley metric*) (Dudley, 2002, Chapter 11) are commonly used.

In the present work, we will consider a particular pseudometric[1] on probability distributions which is an instance of an *integral probability metric* (IPM) (Müller, 1997). Denoting $\mathscr{P}$ the set of all Borel probability measures on $(M, \mathcal{A})$, the IPM between $\mathbb{P} \in \mathscr{P}$ and $\mathbb{Q} \in \mathscr{P}$ is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int_M f \, d\mathbb{P} - \int_M f \, d\mathbb{Q} \right|, \tag{1}$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on M. In addition to the general application domains discussed earlier for metrics on probabilities, IPMs have been used in proving central limit theorems using Stein's method (Stein, 1972; Barbour and Chen, 2005), and are popular in empirical process theory (van der Vaart and Wellner, 1996). Since most of the applications listed

1. Given a set M, a *metric* for M is a function $\rho : M \times M \to \mathbb{R}_+$ such that *(i)* $\forall x, \rho(x,x) = 0$, *(ii)* $\forall x, y, \rho(x,y) = \rho(y,x)$, *(iii)* $\forall x, y, z, \rho(x,z) \leq \rho(x,y) + \rho(y,z)$, and *(iv)* $\rho(x,y) = 0 \Rightarrow x = y$. A semi-metric only satisfies *(i)*, *(ii)* and *(iv)*. A pseudometric only satisfies *(i)-(iii)* of the properties of a metric. Unlike a metric space (M, ρ), points in a pseudometric space need not be distinguishable: one may have $\rho(x,y) = 0$ for $x \neq y$.

Now, in the two-sample test, though we mentioned that γ is a metric/semi-metric, it is sufficient that γ satisfies *(i)* and *(iv)*.

above require $\gamma_{\mathcal{F}}$ to be a metric on $\mathscr{P}$, the choice of $\mathcal{F}$ is critical (note that irrespective of $\mathcal{F}$, $\gamma_{\mathcal{F}}$ is a pseudometric on $\mathscr{P}$). The following are some examples of $\mathcal{F}$ for which $\gamma_{\mathcal{F}}$ is a metric.

(a) $\mathcal{F} = C_b(M)$, the space of bounded continuous functions on (M,ρ), where ρ is a metric (Shorack, 2000, Chapter 19, Definition 1.1).

(b) $\mathcal{F} = C_{bu}(M)$, the space of bounded ρ-uniformly continuous functions on (M,ρ)—Portmonteau theorem (Shorack, 2000, Chapter 19, Theorem 1.1).

(c) $\mathcal{F} = \{f : \|f\|_\infty \leq 1\} =: \mathcal{F}_{TV}$, where $\|f\|_\infty = \sup_{x\in M} |f(x)|$. $\gamma_{\mathcal{F}}$ is called the *total variation distance* (Shorack, 2000, Chapter 19, Proposition 2.2), which we denote as TV, that is, $\gamma_{\mathcal{F}_{TV}} =: TV$.

(d) $\mathcal{F} = \{f : \|f\|_L \leq 1\} =: \mathcal{F}_W$, where $\|f\|_L := \sup\{|f(x) - f(y)|/\rho(x,y) : x \neq y \text{ in } M\}$. $\|f\|_L$ is the Lipschitz semi-norm of a real-valued function f on M and $\gamma_{\mathcal{F}}$ is called the *Kantorovich metric*. If (M,ρ) is separable, then $\gamma_{\mathcal{F}}$ equals the *Wasserstein distance* (Dudley, 2002, Theorem 11.8.2), denoted as $W := \gamma_{\mathcal{F}_W}$.

(e) $\mathcal{F} = \{f : \|f\|_{BL} \leq 1\} =: \mathcal{F}_\beta$, where $\|f\|_{BL} := \|f\|_L + \|f\|_\infty$. $\gamma_{\mathcal{F}}$ is called the *Dudley metric* (Shorack, 2000, Chapter 19, Definition 2.2), denoted as $\beta := \gamma_{\mathcal{F}_\beta}$.

(f) $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\} =: \mathcal{F}_{KS}$. $\gamma_{\mathcal{F}}$ is called the *Kolmogorov distance* (Shorack, 2000, Theorem 2.4).

(g) $\mathcal{F} = \{e^{\sqrt{-1}\langle \omega, \cdot \rangle} : \omega \in \mathbb{R}^d\} =: \mathcal{F}_c$. This choice of $\mathcal{F}$ results in the maximal difference between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$. That $\gamma_{\mathcal{F}_c}$ is a metric on $\mathscr{P}$ follows from the *uniqueness theorem* for characteristic functions (Dudley, 2002, Theorem 9.5.1).

Recently, Gretton et al. (2007b) and Smola et al. (2007) considered $\mathcal{F}$ to be the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ (Aronszajn, 1950), with k as its reproducing kernel (r.k.), that is, $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\} =: \mathcal{F}_k$ (also see Chapter 4 of Berlinet and Thomas-Agnan, 2004, and references therein for related work): we denote $\gamma_{\mathcal{F}_k} =: \gamma_k$. While we have seen many possible $\mathcal{F}$ for which $\gamma_{\mathcal{F}}$ is a metric, $\mathcal{F}_k$ has a number of important advantages:

- **Estimation of $\gamma_{\mathcal{F}}$:** In applications such as hypothesis testing, $\mathbb{P}$ and $\mathbb{Q}$ are known only through the respective random samples $\{X_j\}_{j=1}^m$ and $\{Y_j\}_{j=1}^n$ drawn i.i.d. from each, and $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ is estimated based on these samples. One approach is to compute $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ using the empirical measures $\mathbb{P}_m = \frac{1}{m}\sum_{j=1}^m \delta_{X_j}$ and $\mathbb{Q}_n = \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}$, where δ_x represents a Dirac measure at x. It can be shown that choosing $\mathcal{F}$ as $C_b(M)$, $C_{bu}(M)$, $\mathcal{F}_{TV}$ or $\mathcal{F}_c$ results in this approach not yielding consistent estimates of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ for all $\mathbb{P}$ and $\mathbb{Q}$ (Devroye and Györfi, 1990). Although choosing $\mathcal{F} = \mathcal{F}_W$ or $\mathcal{F}_\beta$ yields consistent estimates of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ for all $\mathbb{P}$ and $\mathbb{Q}$ when $M = \mathbb{R}^d$, the rates of convergence are dependent on d and become slow for large d (Sriperumbudur et al., 2009b). On the other hand, $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma_k(\mathbb{P},\mathbb{Q})$ if k is measurable and bounded, for all $\mathbb{P}$ and $\mathbb{Q}$. If k is translation invariant on $M = \mathbb{R}^d$, the rate is independent of d (Gretton et al., 2007b; Sriperumbudur et al., 2009b), an important property when dealing with high dimensions. Moreover, $\gamma_{\mathcal{F}}$ is not straightforward to compute when $\mathcal{F}$ is $C_b(M), C_{bu}(M), \mathcal{F}_W$ or $\mathcal{F}_\beta$ (Weaver, 1999, Section 2.3): by contrast, $\gamma_k^2(\mathbb{P},\mathbb{Q})$ is simply a sum of expectations of the kernel k (see (9) and Theorem 1).

- **Comparison to ϕ-divergences:** Instead of using $\gamma_{\mathcal{F}}$ in statistical applications, one can also use ϕ-divergences. However, the estimators of ϕ-divergences (especially the Kullback-Leibler divergence) exhibit arbitrarily slow rates of convergence depending on the distributions (see Wang et al., 2005; Nguyen et al., 2008, and references therein for details), while, as noted above, $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ exhibits good convergence behavior.

- **Structured domains:** Since γ_k is dependent only on the kernel (see Theorem 1) and kernels can be defined on arbitrary domains M (Aronszajn, 1950), choosing $\mathcal{F} = \mathcal{F}_k$ provides the flexibility of measuring the distance between probability measures defined on structured domains (Borgwardt et al., 2006) like graphs, strings, etc., unlike $\mathcal{F} = \mathcal{F}_{KS}$ or $\mathcal{F}_c$, which can handle only $M = \mathbb{R}^d$.

The distance measure γ_k has appeared in a wide variety of applications. These include statistical hypothesis testing, of homogeneity (Gretton et al., 2007b), independence (Gretton et al., 2008), and conditional independence (Fukumizu et al., 2008); as well as in machine learning applications including kernel independent component analysis (Bach and Jordan, 2002; Gretton et al., 2005a; Shen et al., 2009) and kernel based dimensionality reduction for supervised learning (Fukumizu et al., 2004). In these applications, kernels offer a linear approach to deal with higher order statistics: given the problem of homogeneity testing, for example, differences in higher order moments are encoded as differences in the means of nonlinear features of the variables. To capture all nonlinearities that are relevant to the problem at hand, the embedding RKHS therefore has to be "sufficiently large" that differences in the embeddings correspond to differences of interest in the distributions. Thus, a natural question is how to guarantee k provides a sufficiently rich RKHS so as to detect *any* difference in distributions. A second problem is to determine what properties of distributions result in their being proximate or distant in the embedding space. Finally, we would like to compare γ_k to the classical integral probability metrics listed earlier, when used to measure convergence of distributions. In the following section, we describe the contributions of the present paper, addressing each of these three questions in turn.

1.1 Contributions

The contributions in this paper are three-fold and explained in detail below.

1.1.1 WHEN IS $\mathcal{H}$ CHARACTERISTIC?

Recently, Fukumizu et al. (2008) introduced the concept of a *characteristic kernel*, that is, a reproducing kernel for which $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, that is, γ_k is a metric on $\mathscr{P}$. The corresponding RKHS, $\mathcal{H}$ is referred to as a *characteristic RKHS*. The following are two characterizations for characteristic RKHSs that have already been studied in literature:

1. When M is compact, Gretton et al. (2007b) showed that $\mathcal{H}$ is characteristic if k is *universal* in the sense of Steinwart (2001, Definition 4), that is, $\mathcal{H}$ is dense in the Banach space of bounded continuous functions with respect to the supremum norm. Examples of such $\mathcal{H}$ include those induced by the Gaussian and Laplacian kernels on every compact subset of $\mathbb{R}^d$.

2. Fukumizu et al. (2008, 2009a) extended this characterization to non-compact M and showed that $\mathcal{H}$ is characteristic if and only if the direct sum of $\mathcal{H}$ and $\mathbb{R}$ is dense in the Banach space of r-integrable (for some $r \geq 1$) functions. Using this characterization, they showed

that the RKHSs induced by the Gaussian and Laplacian kernels (supported on the entire $\mathbb{R}^d$) are characteristic.

In the present study, we provide alternative conditions for characteristic RKHSs which address several limitations of the foregoing. First, it can be difficult to verify the conditions of denseness in both of the above characterizations. Second, universality is in any case an overly restrictive condition because universal kernels assume M to be compact, that is, they induce a metric only on the space of probability measures that are supported on compact M.

In Section 3.1, we present the simple characterization that *integrally strictly positive definite* (pd) kernels (see Section 1.2 for the definition) are characteristic, that is, the induced RKHS is characteristic (also see Sriperumbudur et al., 2009a, Theorem 4). This condition is more natural—strict pd is a natural property of interest for kernels, unlike the denseness condition—and much easier to understand than the characterizations mentioned above. Examples of integrally strictly pd kernels on $\mathbb{R}^d$ include the Gaussian, Laplacian, inverse multiquadratics, Matérn kernel family, B_{2n+1}-splines, etc.

Although the above characterization of integrally strictly pd kernels being characteristic is simple to understand, it is only a sufficient condition and does not provide an answer for kernels that are not integrally strictly pd,[2] for example, a Dirichlet kernel. Therefore, in Section 3.2, we provide an easily checkable condition, after making some assumptions on the kernel. We present a complete characterization of characteristic kernels when the kernel is translation invariant on $\mathbb{R}^d$. We show that a bounded continuous translation invariant kernel on $\mathbb{R}^d$ is characteristic if and only if the support of the Fourier transform of the kernel is the entire $\mathbb{R}^d$. This condition is easy to check compared to the characterizations described above. An earlier version of this result was provided by Sriperumbudur et al. (2008): by comparison, we now present a simpler and more elegant proof. We also show that all compactly supported translation invariant kernels on $\mathbb{R}^d$ are characteristic. Note, however, that the characterization of integral strict positive definiteness in Section 3.1 does not assume M to be $\mathbb{R}^d$ nor k to be translation invariant.

We extend the result of Section 3.2 to M being a d-Torus, that is, $\mathbb{T}^d = S^1 \times \overset{d}{\ldots} \times S^1 \equiv [0, 2\pi)^d$, where S^1 is a circle. In Section 3.3, we show that a translation invariant kernel on $\mathbb{T}^d$ is characteristic if and only if the Fourier series coefficients of the kernel are positive, that is, the support of the Fourier spectrum is the entire $\mathbb{Z}^d$. The proof of this result is similar in flavor to the one in Section 3.2. As examples, the Poisson kernel can be shown to be characteristic, while the Dirichlet kernel is not.

Based on the discussion so far, it is clear that the characteristic property of k can be determined in many ways. In Section 3.4, we summarize the relations between various kernel families (e.g., the universal kernels and the strictly pd kernels), and show how they relate in turn to characteristic kernels. A summary is depicted in Figure 1.

1.1.2 DISSIMILAR DISTRIBUTIONS WITH SMALL γ_k

As we have seen, the characteristic property of a kernel is critical in distinguishing between distinct probability measures. Suppose, however, that for a given characteristic kernel k and for any $\varepsilon > 0$, there exist $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{P} \neq \mathbb{Q}$, such that $\gamma_k(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Though k distinguishes between such $\mathbb{P}$ and $\mathbb{Q}$, it can be difficult to tell the distributions apart in applications (even with characteristic kernels), since $\mathbb{P}$ and $\mathbb{Q}$ are then replaced with finite samples, and the distance between them may not be

2. It can be shown that integrally strictly pd kernels are strictly pd (see Footnote 4). Therefore, examples of kernels that are not integrally strictly pd include those kernels that are not strictly pd.

statistically significant (Gretton et al., 2007b). Therefore, given a characteristic kernel, it is of interest to determine the properties of distributions $\mathbb{P}$ and $\mathbb{Q}$ that will cause their embeddings to be close. To this end, in Section 4, we show that given a kernel k (see Theorem 19 for conditions on the kernel), for any $\varepsilon > 0$, there exists $\mathbb{P} \neq \mathbb{Q}$ (with non-trivial differences between them) such that $\gamma_k(\mathbb{P},\mathbb{Q}) < \varepsilon$. These distributions are constructed so as to differ at a sufficiently high frequency, which is then penalized by the RKHS norm when computing γ_k.

1.1.3 WHEN DOES γ_k METRIZE THE WEAK TOPOLOGY ON $\mathscr{P}$?

Given γ_k, which is a metric on $\mathscr{P}$, a natural question of theoretical and practical importance is "how is γ_k related to other probability metrics, such as the Dudley metric (β), Wasserstein distance (W), total variation metric (TV), etc?" For example, in applications like density estimation, where the unknown density is estimated based on finite samples drawn i.i.d. from it, the quality of the estimate is measured by computing the distance between the true density and the estimated density. In such a setting, given two probability metrics, ρ_1 and ρ_2, one might want to use the *stronger*[3] of the two to determine this distance, as the convergence of the estimated density to the true density in the stronger metric implies the convergence in the weaker metric, while the converse is not true. On the other hand, one might need to use a metric of weaker topology (i.e., coarser topology) to show convergence of some estimators, as the convergence might not occur w.r.t. a metric of strong topology. Clarifying and comparing the topology of a metric on the probabilities is, thus, important in the analysis of density estimation. Based on this motivation, in Section 5, we analyze the relation between γ_k and other probability metrics, and show that γ_k is weaker than all these other metrics.

It is well known in probability theory that β is weaker than W and TV, and it metrizes the weak topology (we will provide formal definitions in Section 5) on $\mathscr{P}$ (Shorack, 2000; Gibbs and Su, 2002). Since γ_k is weaker than all these other probability metrics, that is, the topology induced by γ_k is coarser than the one induced by these metrics, the next interesting question to answer would be, "When does γ_k metrize the weak topology on $\mathscr{P}$?" In other words, for what k, does the topology induced by γ_k coincide with the weak topology? Answering this question would show that γ_k is equivalent to β, while it is weaker than W and TV. In probability theory, the metrization of weak topology is of prime importance in proving results related to the weak convergence of probability measures. Therefore, knowing the answer to the above question will help in using γ_k as a theoretical tool in probability theory. To this end, in Section 5, we show that universal kernels on compact (M,ρ) metrize the weak topology on $\mathscr{P}$. For the non-compact setting, we assume $M = \mathbb{R}^d$ and provide sufficient conditions on the kernel such that γ_k metrizes the weak topology on $\mathscr{P}$.

In the following section, we introduce the notation and some definitions that are used throughout the paper. Supplementary results used in proofs are collected in Appendix A.

1.2 Definitions and Notation

For a measurable space, M and μ a Borel measure on M, $L^r(M,\mu)$ denotes the Banach space of r-power ($r \geq 1$) μ-integrable functions. We will also use $L^r(M)$ for $L^r(M,\mu)$ and dx for $d\mu(x)$ if μ is

3. Two metrics $\rho_1 : Y \times Y \to \mathbb{R}_+$ and $\rho_2 : Y \times Y \to \mathbb{R}_+$ are said to be equivalent if $\rho_1(x,y) = 0 \Leftrightarrow \rho_2(x,y) = 0, \forall x,y \in Y$. On the other hand, ρ_1 is said to be stronger than ρ_2 if $\rho_1(x,y) = 0 \Rightarrow \rho_2(x,y) = 0, \forall x,y \in Y$ but not vice-versa. If ρ_1 is stronger than ρ_2, then we say ρ_2 is weaker than ρ_1. Note that if ρ_1 is stronger (*resp.* weaker) than ρ_2, then the topology induced by ρ_1 is finer (*resp.* coarser) than the one induced by ρ_2.

the Lebesgue measure on $M \subset \mathbb{R}^d$. $C_b(M)$ denotes the space of all bounded, continuous functions on M. The space of all r-continuously differentiable functions on M is denoted by $C^r(M), 0 \leq r \leq \infty$. For $x \in \mathbb{C}$, $\bar{x}$ represents the complex conjugate of x. We denote as i the imaginary unit $\sqrt{-1}$.

For a measurable function f and a signed measure $\mathbb{P}$, $\mathbb{P}f := \int f \, d\mathbb{P} = \int_M f(x) \, d\mathbb{P}(x)$. δ_x represents the Dirac measure at x. The symbol δ is overloaded to represent the Dirac measure, the Dirac-delta distribution, and the Kronecker-delta, which should be distinguishable from the context. For $M = \mathbb{R}^d$, the characteristic function, $\phi_{\mathbb{P}}$ of $\mathbb{P} \in \mathscr{P}$ is defined as $\phi_{\mathbb{P}}(\omega) := \int_{\mathbb{R}^d} e^{i\omega^T x} \, d\mathbb{P}(x), \omega \in \mathbb{R}^d$.

Support of a Borel measure: The support of a finite regular Borel measure, μ on a Hausdorff space, M is defined to be the closed set,

$$\text{supp}(\mu) := M \backslash \bigcup \{U \subset M : U \text{ is open}, \ \mu(U) = 0\}. \tag{2}$$

Vanishing at infinity and $C_0(M)$: A complex function f on a locally compact Hausdorff space M is said to *vanish at infinity* if for every $\varepsilon > 0$ there exists a compact set $K \subset M$ such that $|f(x)| < \varepsilon$ for all $x \notin K$. The class of all continuous f on M which vanish at infinity is denoted as $C_0(M)$.

Holomorphic and entire functions: Let $D \subset \mathbb{C}^d$ be an open subset and $f : D \to \mathbb{C}$ be a function. f is said to be *holomorphic* (or *analytic*) at the point $z_0 \in D$ if

$$f'(z_0) := \lim_{z \to z_0} \frac{f(z_0) - f(z)}{z_0 - z}$$

exists. Moreover, f is called holomorphic if it is holomorphic at every $z_0 \in D$. f is called an *entire function* if f is holomorphic and $D = \mathbb{C}^d$.

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

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

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

Integrally strictly positive definite: Let M be a topological space. A measurable and bounded kernel, k is said to be integrally strictly positive definite if

$$\int\int_M k(x,y) \, d\mu(x) \, d\mu(y) > 0,$$

for all finite non-zero signed Borel measures μ defined on M.

The above definition is a generalization of *integrally strictly positive definite functions* on $\mathbb{R}^d$ (Stewart, 1976, Section 6): $\iint_{\mathbb{R}^d} k(x,y) f(x) f(y) \, dx \, dy > 0$ for all $f \in L^2(\mathbb{R}^d)$, which is the strictly positive definiteness of the integral operator given by the kernel. Note that the above definition is *not* equivalent to the definition of strictly pd kernels: if k is integrally strictly pd, then it is strictly pd, while the converse is not true.[4]

4. Suppose k is not strictly pd. This means for some $n \in \mathbb{N}$ and for mutually distinct $x_1, \ldots, x_n \in M$, there exists $\mathbb{R} \ni \alpha_j \neq 0$ for some $j \in \{1, \ldots, n\}$ such that $\sum_{j,l=1}^{n} \alpha_j \alpha_l k(x_j, x_l) = 0$. By defining $\mu = \sum_{j=1}^{n} \alpha_j \delta_{x_j}$, it is easy to see that there exists $\mu \neq 0$ such that $\iint_M k(x,y) \, d\mu(x) \, d\mu(y) = 0$, which means k is not integrally strictly pd. Therefore, if k is integrally strictly pd, then it is strictly pd. However, the converse is not true. See Steinwart and Christmann (2008, Proposition 4.60, Theorem 4.62) for an example.

Fourier transform in $\mathbb{R}^d$: For $f \in L^1(\mathbb{R}^d)$, $\widehat{f}$ and $f^\vee$ represent the Fourier transform and inverse Fourier transform of f respectively, defined as

$$\widehat{f}(y) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-iy^T x} f(x)\, dx,\ y \in \mathbb{R}^d, \tag{4}$$

$$f^\vee(x) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{ix^T y} f(y)\, dy,\ x \in \mathbb{R}^d. \tag{5}$$

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

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

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

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

Rapidly decaying functions, $\mathscr{D}_d$ and $\mathscr{S}_d$: Let $\mathscr{D}_d$ be the space of compactly supported infinitely differentiable functions on $\mathbb{R}^d$, that is, $\mathscr{D}_d = \{f \in C^\infty(\mathbb{R}^d) \,|\, \mathrm{supp}(f) \text{ is bounded}\}$, where $\mathrm{supp}(f) = \mathrm{cl}\left(\{x \in \mathbb{R}^d \,|\, f(x) \neq 0\}\right)$. A function $f : \mathbb{R}^d \to \mathbb{C}$ is said to decay rapidly, or be rapidly decreasing, if for all $N \in \mathbb{N}$,

$$\sup_{\|\alpha\|_1 \leq N} \sup_{x \in \mathbb{R}^d} (1 + \|x\|_2^2)^N |(T_\alpha f)(x)| < \infty,$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is an ordered d-tuple of non-negative α_j, $\|\alpha\|_1 = \sum_{j=1}^d \alpha_j$ and $T_\alpha = \left(\frac{1}{i}\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{1}{i}\frac{\partial}{\partial x_d}\right)^{\alpha_d}$. $\mathscr{S}_d$, called the Schwartz class, denotes the vector space of rapidly decreasing functions. Note that $\mathscr{D}_d \subset \mathscr{S}_d$. Also, for any $p \in [1, \infty]$, $\mathscr{S}_d \subset L^p(\mathbb{R}^d)$. It can be shown that for any $f \in \mathscr{S}_d$, $\widehat{f} \in \mathscr{S}_d$ and $f^\vee \in \mathscr{S}_d$ (see Folland, 1999, Chapter 9 and Rudin, 1991, Chapter 6 for details).

Distributions, $\mathscr{D}'_d$: A linear functional on $\mathscr{D}_d$ which is continuous with respect to the Fréchet topology (see Rudin, 1991, Definition 6.3) is called a *distribution* in $\mathbb{R}^d$. The space of all distributions in $\mathbb{R}^d$ is denoted by $\mathscr{D}'_d$.

As examples, 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 $K \subset \mathbb{R}^d$), then the functional D_f defined by

$$D_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)\, dx,\ \varphi \in \mathscr{D}_d, \tag{6}$$

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

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

defines a distribution D_μ in $\mathbb{R}^d$, which is identified with μ.

Support of a distribution: For an open set $U \subset \mathbb{R}^d$, $\mathscr{D}_d(U)$ denotes the subspace of $\mathscr{D}_d$ consisting of the functions with support contained in U. Suppose $D \in \mathscr{D}'_d$. If U is an open set of $\mathbb{R}^d$ and if

$D(\varphi) = 0$ for every $\varphi \in \mathscr{D}_d(U)$, then D is said to *vanish* or be *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.

Tempered distributions, $\mathscr{S}'_d$ and Fourier transform on $\mathscr{S}'_d$: A linear continuous functional (with respect to the Fréchet topology) over the space $\mathscr{S}_d$ is called a *tempered distribution* and the space of all tempered distributions in $\mathbb{R}^d$ is denoted by $\mathscr{S}'_d$. For example, every compactly supported distribution is tempered.

For any $f \in \mathscr{S}'_d$, the Fourier and inverse Fourier transforms are defined as

$$\begin{aligned} \widehat{f}(\varphi) &:= f(\hat{\varphi}), \varphi \in \mathscr{S}_d, \\ f^{\vee}(\varphi) &:= f(\varphi^{\vee}), \varphi \in \mathscr{S}_d, \end{aligned}$$

respectively. The Fourier transform is a linear, one-to-one, bicontinuous mapping from $\mathscr{S}'_d$ to $\mathscr{S}'_d$.

For complete details on distribution theory and Fourier transforms of distributions, we refer the reader to Folland (1999, Chapter 9) and Rudin (1991, Chapter 6).

2. Hilbert Space Embedding of Probability Measures

Embeddings of probability distributions into reproducing kernel Hilbert spaces were introduced in the late 70's and early 80's, generalizing the notion of mappings of individual points: see Berlinet and Thomas-Agnan (2004, Chapter 4) for a survey. Following Gretton et al. (2007b) and Smola et al. (2007), γ_k can be alternatively expressed as a pseudometric between such distribution embeddings. The following theorem describes this relation.

Theorem 1 *Let $\mathscr{P}_k := \{\mathbb{P} \in \mathscr{P} : \int_M \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty\}$, where k is measurable on M. Then for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_k$,*

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \left\| \int_M k(\cdot, x)\, d\mathbb{P}(x) - \int_M k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}} =: \|\mathbb{P}k - \mathbb{Q}k\|_{\mathcal{H}}, \tag{7}$$

where $\mathcal{H}$ is the RKHS generated by k.

Proof Let $T_{\mathbb{P}} : \mathcal{H} \to \mathbb{R}$ be the linear functional defined as $T_{\mathbb{P}}[f] := \int_M f(x)\, d\mathbb{P}(x)$ with $\|T_{\mathbb{P}}\| := \sup_{f \in \mathcal{H}, f \neq 0} \frac{|T_{\mathbb{P}}[f]|}{\|f\|_{\mathcal{H}}}$. Consider

$$|T_{\mathbb{P}}[f]| = \left| \int_M f\, d\mathbb{P} \right| \leq \int_M |f(x)|\, d\mathbb{P}(x) = \int_M |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}|\, d\mathbb{P}(x) \leq \int_M \sqrt{k(x,x)} \|f\|_{\mathcal{H}}\, d\mathbb{P}(x),$$

which implies $\|T_{\mathbb{P}}\| < \infty, \forall \mathbb{P} \in \mathscr{P}_k$, that is, $T_{\mathbb{P}}$ is a bounded linear functional on $\mathcal{H}$. Therefore, by the Riesz representation theorem (Reed and Simon, 1980, Theorem II.4), for each $\mathbb{P} \in \mathscr{P}_k$, there exists a unique $\lambda_{\mathbb{P}} \in \mathcal{H}$ such that $T_{\mathbb{P}}[f] = \langle f, \lambda_{\mathbb{P}}\rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$. Let $f = k(\cdot, u)$ for some $u \in M$. Then, $T_{\mathbb{P}}[k(\cdot, u)] = \langle k(\cdot, u), \lambda_{\mathbb{P}}\rangle_{\mathcal{H}} = \lambda_{\mathbb{P}}(u)$, which implies $\lambda_{\mathbb{P}} = \int_M k(\cdot, x)\, d\mathbb{P}(x) =: \mathbb{P}k$. Therefore, with

$$|\mathbb{P}f - \mathbb{Q}f| = |T_{\mathbb{P}}[f] - T_{\mathbb{Q}}[f]| = |\langle f, \lambda_{\mathbb{P}}\rangle_{\mathcal{H}} - \langle f, \lambda_{\mathbb{Q}}\rangle_{\mathcal{H}}| = |\langle f, \lambda_{\mathbb{P}} - \lambda_{\mathbb{Q}}\rangle_{\mathcal{H}}|,$$

we have

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathbb{P}f - \mathbb{Q}f| = \|\lambda_{\mathbb{P}} - \lambda_{\mathbb{Q}}\|_{\mathcal{H}} = \|\mathbb{P}k - \mathbb{Q}k\|_{\mathcal{H}}.$$

Note that this holds for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_k$. ■

Given a kernel, k, (7) holds for all $\mathbb{P} \in \mathscr{P}_k$. However, in practice, especially in statistical inference applications, it is not possible to check whether $\mathbb{P} \in \mathscr{P}_k$ as $\mathbb{P}$ is not known. Therefore, one would prefer to have a kernel such that

$$\int_M \sqrt{k(x,x)}\, d\mathbb{P}(x) < \infty, \forall \mathbb{P} \in \mathscr{P}. \tag{8}$$

The following proposition shows that (8) is equivalent to the kernel being bounded. Therefore, combining Theorem 1 and Proposition 2 shows that if k is measurable and bounded, then $\gamma_k(\mathbb{P},\mathbb{Q}) = \|\mathbb{P}k - \mathbb{Q}k\|_{\mathcal{H}}$ for any $\mathbb{P},\mathbb{Q} \in \mathscr{P}$.

Proposition 2 *Let f be a measurable function on M. Then $\int_M f(x)\, d\mathbb{P}(x) < \infty$ for all $\mathbb{P} \in \mathscr{P}$ if and only if f is bounded.*

Proof One direction is straightforward because if f is bounded, then $\int_M f(x)\, d\mathbb{P}(x) < \infty$ for all $\mathbb{P} \in \mathscr{P}$. Let us consider the other direction. Suppose f is not bounded. Then there exists a sequence $\{x_n\} \subset M$ such that $f(x_n) \overset{n\to\infty}{\longrightarrow} \infty$. By taking a subsequence, if necessary, we can assume $f(x_n) > n^2$ for all n. Then, $A := \sum_{n=1}^{\infty} \frac{1}{f(x_n)} < \infty$. Define a probability measure $\mathbb{P}$ on M by $\mathbb{P} = \sum_{n=1}^{\infty} \frac{1}{Af(x_n)} \delta_{x_n}$, where δ_{x_n} is a Dirac measure at x_n. Then, $\int_M f(x)\, d\mathbb{P}(x) = \frac{1}{A}\sum_{n=1}^{\infty} \frac{f(x_n)}{f(x_n)} = \infty$, which means if f is not bounded, then there exists a $\mathbb{P} \in \mathscr{P}$ such that $\int_M f(x)\, d\mathbb{P}(x) = \infty$. ■

The representation of γ_k in (7) yields the embedding,

$$\Pi : \mathscr{P} \to \mathcal{H} \qquad \mathbb{P} \mapsto \int_M k(\cdot,x)\, d\mathbb{P}(x),$$

as proposed by Berlinet and Thomas-Agnan (2004, Chapter 4, Section 1.1) and Gretton et al. (2007b); Smola et al. (2007). Berlinet and Thomas-Agnan derived this embedding as a generalization of $\delta_x \mapsto k(\cdot,x)$, while Gretton et al. arrived at the embedding by choosing $\mathcal{F} = \mathcal{F}_k$ in (1). Since $\gamma_k(\mathbb{P},\mathbb{Q}) = \|\Pi[\mathbb{P}] - \Pi[\mathbb{Q}]\|_{\mathcal{H}}$, the question "When is γ_k a metric on $\mathscr{P}$?" is equivalent to the question "When is Π injective?" Addressing these questions is the central focus of the paper and is discussed in Section 3.

Before proceeding further, we present a number of equivalent representations of γ_k, which will improve our understanding of γ_k and be helpful in its computation. First, Gretton et al. have shown the reproducing property of k leads to

$$\begin{aligned}
\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \left\| \int_M k(\cdot,x)\, d\mathbb{P}(x) - \int_M k(\cdot,x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}^2 \\
&= \left\langle \int_M k(\cdot,x)\, d\mathbb{P}(x) - \int_M k(\cdot,x)\, d\mathbb{Q}(x), \int_M k(\cdot,y)\, d\mathbb{P}(y) - \int_M k(\cdot,y)\, d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&= \left\langle \int_M k(\cdot,x)\, d\mathbb{P}(x), \int_M k(\cdot,y)\, d\mathbb{P}(y) \right\rangle_{\mathcal{H}} + \left\langle \int_M k(\cdot,x)\, d\mathbb{Q}(x), \int_M k(\cdot,y)\, d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&\qquad -2\left\langle \int_M k(\cdot,x)\, d\mathbb{P}(x), \int_M k(\cdot,y)\, d\mathbb{Q}(y) \right\rangle_{\mathcal{H}} \\
&\overset{(a)}{=} \iint_M k(x,y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \iint_M k(x,y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\
&\qquad -2\iint_M k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \qquad (9) \\
&= \iint_M k(x,y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y), \qquad (10)
\end{aligned}$$

where (a) follows from the fact that $\int_M f(x)\,d\mathbb{P}(x) = \langle f, \int_M k(\cdot,x)\,d\mathbb{P}(x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $\mathbb{P} \in \mathscr{P}$ (see proof of Theorem 1), applied with $f = \int_M k(\cdot,y)\,d\mathbb{P}(y)$. As motivated in Section 1, γ_k^2 is a straightforward sum of expectations of k, and can be computed easily, for example, using (9) either in closed form or using numerical integration techniques, depending on the choice of k, $\mathbb{P}$ and $\mathbb{Q}$. It is easy to show that, if k is a Gaussian kernel with $\mathbb{P}$ and $\mathbb{Q}$ being normal distributions on $\mathbb{R}^d$, then γ_k can be computed in a closed form (see Song et al., 2008 and Sriperumbudur et. al., 2009b, Section III-C for examples). In the following corollary to Theorem 1, we prove three results which provide a nice interpretation for γ_k when $M = \mathbb{R}^d$ and k is translation invariant, that is, $k(x,y) = \psi(x-y)$, where ψ is a positive definite function. We provide a detailed explanation for Corollary 4 in Remark 5. Before stating the results, we need a famous result due to Bochner, that characterizes ψ. We quote this result from Wendland (2005, Theorem 6.6).

Theorem 3 (Bochner) *A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$, that is,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\,d\Lambda(\omega),\ x \in \mathbb{R}^d. \tag{11}$$

Corollary 4 (Different interpretations of γ_k) *(i) Let $M = \mathbb{R}^d$ and $k(x,y) = \psi(x-y)$, where $\psi : M \to \mathbb{R}$ is a bounded, continuous positive definite function. Then for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$,*

$$\gamma_k(\mathbb{P},\mathbb{Q}) = \sqrt{\int_{\mathbb{R}^d} |\phi_{\mathbb{P}}(\omega) - \phi_{\mathbb{Q}}(\omega)|^2\,d\Lambda(\omega)} =: \|\phi_{\mathbb{P}} - \phi_{\mathbb{Q}}\|_{L^2(\mathbb{R}^d,\Lambda)}, \tag{12}$$

where $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ represent the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ respectively.

(ii) Suppose $\theta \in L^1(\mathbb{R}^d)$ is a continuous bounded positive definite function and $\int_{\mathbb{R}^d} \theta(x)\,dx = 1$. Let $\psi(x) := \psi_t(x) = t^{-d}\theta(t^{-1}x)$, $t > 0$. Assume that p and q are bounded uniformly continuous Radon-Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the Lebesgue measure, that is, $d\mathbb{P} = p\,dx$ and $d\mathbb{Q} = q\,dx$. Then,

$$\lim_{t\to 0} \gamma_k(\mathbb{P},\mathbb{Q}) = \|p - q\|_{L^2(\mathbb{R}^d)}. \tag{13}$$

In particular, if $|\theta(x)| \le C(1+\|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$, then (13) holds for all bounded p and q (not necessarily uniformly continuous).

(iii) Suppose $\psi \in L^1(\mathbb{R}^d)$ and $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$. Then,

$$\gamma_k(\mathbb{P},\mathbb{Q}) = (2\pi)^{-d/4}\|\Phi * \mathbb{P} - \Phi * \mathbb{Q}\|_{L^2(\mathbb{R}^d)}, \tag{14}$$

where $\Phi := \left(\sqrt{\widehat{\psi}}\right)^{\vee}$ and $d\Lambda = (2\pi)^{-d/2}\widehat{\psi}\,d\omega$. Here, $\Phi * \mathbb{P}$ represents the convolution of Φ and $\mathbb{P}$.

Proof *(i)* Let us consider (10) with $k(x,y) = \psi(x-y)$. Then, we have

$$\begin{aligned}
\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \iint_{\mathbb{R}^d} \psi(x-y)\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y) \\
&\overset{(a)}{=} \iiint_{\mathbb{R}^d} e^{-i(x-y)^T\omega}\,d\Lambda(\omega)\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y) \\
&\overset{(b)}{=} \iint_{\mathbb{R}^d} e^{-ix^T\omega}\,d(\mathbb{P}-\mathbb{Q})(x) \int_{\mathbb{R}^d} e^{iy^T\omega}\,d(\mathbb{P}-\mathbb{Q})(y)\,d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} (\phi_{\mathbb{P}}(\omega) - \phi_{\mathbb{Q}}(\omega))\left(\overline{\phi_{\mathbb{P}}(\omega)} - \overline{\phi_{\mathbb{Q}}(\omega)}\right) d\Lambda(\omega) = \int_{\mathbb{R}^d} |\phi_{\mathbb{P}}(\omega) - \phi_{\mathbb{Q}}(\omega)|^2\,d\Lambda(\omega),
\end{aligned}$$

where Bochner's theorem (Theorem 3) is invoked in (a), while Fubini's theorem (Folland, 1999, Theorem 2.37) is invoked in (b).

(ii) Consider (9) with $k(x,y) = \psi_t(x-y)$,

$$\begin{aligned}\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \iint_{\mathbb{R}^d} \psi_t(x-y)p(x)p(y)\,dx\,dy + \iint_{\mathbb{R}^d} \psi_t(x-y)q(x)q(y)\,dx\,dy \\ &\qquad -2\iint_{\mathbb{R}^d} \psi_t(x-y)p(x)q(y)\,dx\,dy \\ &= \int_{\mathbb{R}^d} (\psi_t * p)(x)p(x)\,dx + \int_{\mathbb{R}^d} (\psi_t * q)(x)q(x)\,dx - 2\int_{\mathbb{R}^d} (\psi_t * q)(x)p(x)\,dx. \qquad (15)\end{aligned}$$

Note that $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d}\lim_{t\to 0}(\psi_t * p)(x)p(x)\,dx$, by invoking the dominated convergence theorem. Since p is bounded and uniformly continuous, by Theorem 25 (see Appendix A), we have $p * \psi_t \to p$ uniformly as $t \to 0$, which means $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d} p^2(x)\,dx$. Using this in (15), we have

$$\lim_{t\to 0}\gamma_k^2(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d}(p^2(x)+q^2(x)-2p(x)q(x))\,dx = \|p-q\|^2_{L^2(\mathbb{R}^d)}.$$

Suppose $|\theta(x)| \le (1+\|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$. Since $p \in L^1(\mathbb{R}^d)$, by Theorem 26 (see Appendix A), we have $(p * \psi_t)(x) \to p(x)$ as $t \to 0$ for almost every x. Therefore $\lim_{t\to 0}\int_{\mathbb{R}^d}(\psi_t * p)(x)p(x)\,dx = \int_{\mathbb{R}^d} p^2(x)\,dx$ and the result follows.

(iii) Since ψ is positive definite, $\widehat{\psi}$ is nonnegative and therefore $\sqrt{\widehat{\psi}}$ is valid. Since $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$, Φ exists. Define $\phi_{\mathbb{P},\mathbb{Q}} := \phi_{\mathbb{P}} - \phi_{\mathbb{Q}}$. Now, consider

$$\begin{aligned}\|\Phi * \mathbb{P} - \Phi * \mathbb{Q}\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} |(\Phi * (\mathbb{P}-\mathbb{Q}))(x)|^2\,dx \\ &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\Phi(x-y)\,d(\mathbb{P}-\mathbb{Q})(y)\right|^2 dx \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\,e^{i(x-y)^T\omega}\,d\omega\,d(\mathbb{P}-\mathbb{Q})(y)\right|^2 dx \\ &\overset{(c)}{=} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}(\overline{\phi_{\mathbb{P}}(\omega)} - \overline{\phi_{\mathbb{Q}}(\omega)})\,e^{ix^T\omega}\,d\omega\right|^2 dx \\ &= \frac{1}{(2\pi)^d}\iiint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\sqrt{\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\,e^{i(\omega-\xi)^Tx}\,d\omega\,d\xi\,dx \\ &\overset{(d)}{=} \iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\sqrt{\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\left[\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i(\omega-\xi)^Tx}\,dx\right] d\omega\,d\xi \\ &= \iint_{\mathbb{R}^d}\sqrt{\widehat{\psi}(\omega)}\sqrt{\widehat{\psi}(\xi)}\,\overline{\phi_{\mathbb{P},\mathbb{Q}}(\omega)}\,\phi_{\mathbb{P},\mathbb{Q}}(\xi)\,\delta(\omega-\xi)\,d\omega\,d\xi \\ &= \int_{\mathbb{R}^d}\widehat{\psi}(\omega)\,|\phi_{\mathbb{P}}(\omega)-\phi_{\mathbb{Q}}(\omega)|^2\,d\omega \\ &= (2\pi)^{d/2}\gamma_k^2(\mathbb{P},\mathbb{Q}),\end{aligned}$$

where (c) and (d) are obtained by invoking Fubini's theorem. ■

Remark 5 *(a) (12) shows that γ_k is the L^2-distance between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$ computed w.r.t. the non-negative finite Borel measure, Λ, which is the Fourier transform of ψ. If $\psi \in L^1(\mathbb{R}^d)$, then (12) rephrases the well known fact (Wendland, 2005, Theorem 10.12) that for any $f \in \mathcal{H}$,*

$$\|f\|_{\mathcal{H}}^2 = \int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{\widehat{\psi}(\omega)}\, d\omega. \tag{16}$$

Choosing $f = (\mathbb{P} - \mathbb{Q}) * \psi$ in (16) yields $\widehat{f} = (\phi_{\mathbb{P}} - \phi_{\mathbb{Q}})\widehat{\psi}$ and therefore the result in (12).

(b) Suppose $d\Lambda(\omega) = (2\pi)^{-d}\, d\omega$. Assume $\mathbb{P}$ and $\mathbb{Q}$ have p and q as Radon-Nikodym derivatives w.r.t. the Lebesgue measure, that is, $d\mathbb{P} = p\, dx$ and $d\mathbb{Q} = q\, dx$. Using these in (12), it can be shown that $\gamma_k(\mathbb{P}, \mathbb{Q}) = \|p - q\|_{L^2(\mathbb{R}^d)}$. However, this result should be interpreted in a limiting sense as mentioned in Corollary 4(ii) because the choice of $d\Lambda(\omega) = (2\pi)^{-d}\, d\omega$ implies $\psi(x) = \delta(x)$, which does not satisfy the conditions of Corollary 4(i). It can be shown that $\psi(x) = \delta(x)$ is obtained in a limiting sense (Folland, 1999, Proposition 9.1): $\psi_t \to \delta$ in $\mathscr{D}_d'$ as $t \to 0$.

(c) Choosing $\theta(x) = (2\pi)^{-d/2} e^{-\|x\|_2^2/2}$ in Corollary 4(ii) corresponds to ψ_t being a Gaussian kernel (with appropriate normalization such that $\int_{\mathbb{R}^d} \psi_t(x)\, dx = 1$). Therefore, (13) shows that as the bandwidth, t of the Gaussian kernel approaches zero, γ_k approaches the L^2-distance between the densities p and q. The same result also holds for choosing ψ_t as the Laplacian kernel, B_{2n+1}-spline, inverse multiquadratic, etc. Therefore, $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be seen as a generalization of the L^2-distance between probability measures, $\mathbb{P}$ and $\mathbb{Q}$.

(d) The result in (13) holds if p and q are bounded and uniformly continuous. Since any condition on $\mathbb{P}$ and $\mathbb{Q}$ is usually difficult to check in statistical applications, it is better to impose conditions on ψ rather than on $\mathbb{P}$ and $\mathbb{Q}$. In Corollary 4(ii), by imposing additional conditions on ψ_t, the result in (13) is shown to hold for all $\mathbb{P}$ and $\mathbb{Q}$ with bounded densities p and q. The condition, $|\theta(x)| \le C(1 + \|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$, is, for example, satisfied by the inverse multiquadratic kernel, $\theta(x) = \widetilde{C}(1 + \|x\|_2^2)^{-\tau}$, $x \in \mathbb{R}^d$, $\tau > d/2$, where $\widetilde{C} = \left(\int_{\mathbb{R}^d} (1 + \|x\|_2^2)^{-\tau}\, dx\right)^{-1}$.

(e) The result in Corollary 4(ii) has connections to the kernel density estimation in L^2-sense using Parzen windows (Rosenblatt, 1975), where ψ can be chosen as the Parzen window: see Gretton et al. (2007a, Section 7.1) for further discussion. Note in particular that when γ_k is used in a homogeneity test, a constant kernel bandwidth results in a faster decrease of the Type II error with increasing sample size (Anderson et al., 1994, p. 43). A decreasing bandwidth is required for a consistent estimate of $\|p - q\|_{L^2(\mathbb{R}^d)}$, however.

(f) (14) shows that γ_k is proportional to the L^2-distance between $\Phi * \mathbb{P}$ and $\Phi * \mathbb{Q}$. Let Φ be such that Φ is nonnegative and $\Phi \in L^1(\mathbb{R}^d)$. Then, defining $\widetilde{\Phi} := \left(\int_{\mathbb{R}^d} \Phi(x)\, dx\right)^{-1} \Phi = \Phi / \sqrt{\widehat{\psi}(0)} = \left(\int_{\mathbb{R}^d} \psi(x)\, dx\right)^{-1/2} \Phi$ and using this in (14), we have

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = (2\pi)^{-d/4} \sqrt{\widehat{\psi}(0)} \left\| \widetilde{\Phi} * \mathbb{P} - \widetilde{\Phi} * \mathbb{Q} \right\|_{L^2(\mathbb{R}^d)}. \tag{17}$$

The r.h.s. of (17) can be interpreted as follows. Let X, Y and N be independent random variables such that $X \sim \mathbb{P}$, $Y \sim \mathbb{Q}$ and $N \sim \widetilde{\Phi}$. This means γ_k is proportional to the L^2-distance computed between the densities associated with the perturbed random variables, $X + N$ and $Y + N$. Note that $\|p - q\|_{L^2(\mathbb{R}^d)}$ is the L^2-distance between the densities of X and Y. Examples of ψ that satisfy the conditions in Corollary 4(iii) in addition to the conditions on Φ as mentioned here include the

Gaussian and Laplacian kernels on $\mathbb{R}^d$. The result in (14) holds even if $\sqrt{\widehat{\psi}} \notin L^1(\mathbb{R}^d)$ as the proof of (iii) can be handled using distribution theory. However, we assumed $\sqrt{\widehat{\psi}} \in L^1(\mathbb{R}^d)$ to keep the proof simple, without delving into distribution theory.

Although we will not be using all the results of Corollary 4 in deriving our main results in the following sections, Corollary 4 was presented to provide a better intuitive understanding of γ_k. To summarize, the core results of this section are Theorem 1 (combined with Proposition 2), which provides a closed form expression for γ_k in terms of the measurable and bounded k, and Corollary 4*(i)*, which provides an alternative representation for γ_k when k is bounded, continuous and translation invariant on $\mathbb{R}^d$.

3. Conditions for Characteristic Kernels

In this section, we address the question, "When is γ_k a metric on $\mathscr{P}$?". In other words, "When is Π injective?" or "Under what conditions is k characteristic?". To this end, we start with the definition of characteristic kernels and provide some examples where k is such that γ_k is not a metric on $\mathscr{P}$. As discussed in Section 1.1.1, although some characterizations are available for k so that γ_k is a metric on $\mathscr{P}$, they are difficult to check in practice. In Section 3.1, we provide the characterization that if k is integrally strictly pd, then γ_k is a metric on $\mathscr{P}$. In Section 3.2, we present more easily checkable conditions wherein we show that if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ (see (2) for the definition of the support of a Borel measure), then γ_k is a metric on $\mathscr{P}$. This result is extended in a straightforward way to $\mathbb{T}^d$ (d-Torus) in Section 3.3. The main results of this section are summarized in Table 1.

We start by defining characteristic kernels.

Definition 6 (Characteristic kernel) *A bounded measurable positive definite kernel k is characteristic to a set $\mathscr{Q} \subset \mathscr{P}$ of probability measures defined on $(M, \mathcal{A})$ if for $\mathbb{P}, \mathbb{Q} \in \mathscr{Q}$, $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$. k is simply said to be characteristic if it is characteristic to $\mathscr{P}$. The RKHS $\mathcal{H}$ induced by such a k is called a characteristic RKHS.*

As mentioned before, the injectivity of Π is related to the characteristic property of k. If k is characteristic, then $\gamma_k(\mathbb{P}, \mathbb{Q}) = \|\Pi[\mathbb{P}] - \Pi[\mathbb{Q}]\|_{\mathcal{H}} = 0 \Rightarrow \mathbb{P} = \mathbb{Q}$, which means $\mathbb{P} \mapsto \int_M k(\cdot, x)\, d\mathbb{P}(x)$, that is, Π is injective. Therefore, when $M = \mathbb{R}^d$, the embedding of a distribution to a characteristic RKHS can be seen as a generalization of the characteristic function, $\phi_{\mathbb{P}} = \int_{\mathbb{R}^d} e^{i\langle \cdot, x\rangle}\, d\mathbb{P}(x)$. This is because, by the uniqueness theorem for characteristic functions (Dudley, 2002, Theorem 9.5.1), $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$, which means $\mathbb{P} \mapsto \int_{\mathbb{R}^d} e^{i\langle \cdot, x\rangle}\, d\mathbb{P}(x)$ is injective. So, in this context, intuitively $e^{i\langle y, x\rangle}$ can be treated as the characteristic kernel, k, although, formally, this is not true as $e^{i\langle y, x\rangle}$ is not a pd kernel.

Before we get to the characterization of characteristic kernels, the following examples show that there exist bounded measurable kernels that are not characteristic.

Example 1 (Trivial kernel) *Let $k(x, y) = \psi(x - y) = C$, $\forall x, y \in \mathbb{R}^d$ with $C > 0$. Using this in (9), we have $\gamma_k^2(\mathbb{P}, \mathbb{Q}) = C + C - 2C = 0$ for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, which means k is not characteristic.*

Example 2 (Dot product kernel) *Let $k(x, y) = x^T y$, $x, y \in \mathbb{R}^d$. Using this in (9), we have*

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \mu_{\mathbb{P}}^T \mu_{\mathbb{P}} + \mu_{\mathbb{Q}}^T \mu_{\mathbb{Q}} - 2\mu_{\mathbb{P}}^T \mu_{\mathbb{Q}} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_2^2,$$

where $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ represent the means associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, that is, $\mu_{\mathbb{P}} := \int_{\mathbb{R}^d} x\, d\mathbb{P}(x)$. It is clear that k is not characteristic as $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0 \Rightarrow \mu_{\mathbb{P}} = \mu_{\mathbb{Q}} \not\Rightarrow \mathbb{P} = \mathbb{Q}$ for all $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$.

Summary of Main Results				
Domain	Property	$\mathscr{Q}$	Characteristic	Reference
M	k is integrally strictly pd	$\mathscr{P}$	Yes	Theorem 7
$\mathbb{R}^d$	$\Omega = \mathbb{R}^d$	$\mathscr{P}$	Yes	Theorem 9
$\mathbb{R}^d$	$\text{supp}(\psi)$ is compact	$\mathscr{P}$	Yes	Corollary 10
$\mathbb{R}^d$	$\Omega \subsetneq \mathbb{R}^d$, $\text{int}(\Omega) \neq \emptyset$	$\mathscr{P}_1$	Yes	Theorem 12
$\mathbb{R}^d$	$\Omega \subsetneq \mathbb{R}^d$	$\mathscr{P}$	No	Theorem 9
$\mathbb{T}^d$	$A_\psi(0) \geq 0, A_\psi(n) > 0, \forall n \neq 0$	$\mathscr{P}$	Yes	Theorem 14
$\mathbb{T}^d$	$\exists n \neq 0 \,\vert\, A_\psi(n) = 0$	$\mathscr{P}$	No	Theorem 14

Table 1: The table should be read as: If "Property" is satisfied on "Domain", then k is characteristic (or not) to $\mathscr{Q}$. $\mathscr{P}$ is the set of all Borel probability measures on a topological space, M. See Section 1.2 for the definition of integrally strictly pd kernels. When $M = \mathbb{R}^d$, $k(x,y) = \psi(x-y)$, where ψ is a bounded, continuous positive definite function on $\mathbb{R}^d$. ψ is the Fourier transform of a finite nonnegative Borel measure, Λ, and $\Omega := \text{supp}(\Lambda)$ (see Theorem 3 and (2) for details). $\mathscr{P}_1 := \{\mathbb{P} \in \mathscr{P} : \phi_\mathbb{P} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d), \mathbb{P} \ll \lambda \text{ and } \text{supp}(\mathbb{P}) \text{ is compact}\}$, where $\phi_\mathbb{P}$ is the characteristic function of $\mathbb{P}$ and λ is the Lebesgue measure. $\mathbb{P} \ll \lambda$ denotes that $\mathbb{P}$ is absolutely continuous w.r.t. λ. When $M = \mathbb{T}^d$, $k(x,y) = \psi(x-y)$, where ψ is a bounded, continuous positive definite function on $\mathbb{T}^d$. $\{A_\psi(n)\}_{n\in\mathbb{Z}^d}$ are the Fourier series coefficients of ψ which are nonnegative and summable (see Theorem 13 for details).

Example 3 (Polynomial kernel of order 2) *Let $k(x,y) = (1+x^T y)^2$, $x,y \in \mathbb{R}^d$. Using this in (10), we have*

$$\begin{aligned}\gamma_k^2(\mathbb{P},\mathbb{Q}) &= \int\!\!\int_{\mathbb{R}^d} (1 + 2x^T y + x^T y y^T x)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) \\ &= 2\|\mu_\mathbb{P} - \mu_\mathbb{Q}\|_2^2 + \|\Sigma_\mathbb{P} - \Sigma_\mathbb{Q} + \mu_\mathbb{P}\mu_\mathbb{P}^T - \mu_\mathbb{Q}\mu_\mathbb{Q}^T\|_F^2,\end{aligned}$$

where $\Sigma_\mathbb{P}$ and $\Sigma_\mathbb{Q}$ represent the covariance matrices associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, that is, $\Sigma_\mathbb{P} := \int_{\mathbb{R}^d} xx^T\, d\mathbb{P}(x) - \mu_\mathbb{P}\mu_\mathbb{P}^T$. $\|\cdot\|_F$ represents the Frobenius norm. Since $\gamma_k(\mathbb{P},\mathbb{Q}) = 0 \Rightarrow (\mu_\mathbb{P} = \mu_\mathbb{Q}$ and $\Sigma_\mathbb{P} = \Sigma_\mathbb{Q}) \not\Rightarrow \mathbb{P} = \mathbb{Q}$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}$, k is not characteristic.

In the following sections, we address the question of when k is characteristic, that is, for what k is γ_k a metric on $\mathscr{P}$?

3.1 Integrally Strictly Positive Definite Kernels are Characteristic

Compared to the existing characterizations in literature (Gretton et al., 2007b; Fukumizu et al., 2008, 2009a), the following result provides a more natural and easily understandable characterization for characteristic kernels, namely that integrally strictly pd kernels are characteristic to $\mathscr{P}$.

Theorem 7 (Integrally strictly pd kernels are characteristic) *Let k be an integrally strictly positive definite kernel on a topological space M. Then k is characteristic to $\mathscr{P}$.*

Before proving Theorem 7, we provide a supplementary result in Lemma 8 that provides necessary and sufficient conditions for a kernel *not* to be characteristic. We show that choosing k to be integrally strictly pd violates the conditions in Lemma 8, and k is therefore characteristic to $\mathscr{P}$.

Lemma 8 *Let k be measurable and bounded on a topological space, M. Then $\exists \mathbb{P} \neq \mathbb{Q}$ where $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if there exists a finite non-zero signed Borel measure μ that satisfies:*

(i) $\int\int_M k(x,y)\, d\mu(x)\, d\mu(y) = 0$,

(ii) $\mu(M) = 0$.

Proof ($\Leftarrow$) Suppose there exists a finite non-zero signed Borel measure, μ that satisfies *(i)* and *(ii)* in Lemma 8. By the Jordan decomposition theorem (Dudley, 2002, Theorem 5.6.1), there exist unique positive measures μ^+ and μ^- such that $\mu = \mu^+ - \mu^-$ and $\mu^+ \perp \mu^-$ (μ^+ and μ^- are singular). By *(ii)*, we have $\mu^+(M) = \mu^-(M) =: \alpha$. Define $\mathbb{P} = \alpha^{-1}\mu^+$ and $\mathbb{Q} = \alpha^{-1}\mu^-$. Clearly, $\mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}$. Then, by (10), we have

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \int\int_M k(x,y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) = \alpha^{-2} \int\int_M k(x,y)\, d\mu(x)\, d\mu(y) \overset{(a)}{=} 0,$$

where *(a)* is obtained by invoking *(i)*. So, we have constructed $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$.
($\Rightarrow$) Suppose $\exists \mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Let $\mu = \mathbb{P} - \mathbb{Q}$. Clearly μ is a finite non-zero signed Borel measure that satisfies $\mu(M) = 0$. Note that by (10),

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \int\int_M k(x,y)\, d(\mathbb{P}-\mathbb{Q})(x)\, d(\mathbb{P}-\mathbb{Q})(y) = \int\int_M k(x,y)\, d\mu(x)\, d\mu(y),$$

and therefore *(i)* follows. ■

Proof (of Theorem 7) Since k is integrally strictly pd on M, we have

$$\int\int_M k(x,y)\, d\eta(x) d\eta(y) > 0,$$

for any finite non-zero signed Borel measure η. This means there does not exist a finite non-zero signed Borel measure that satisfies *(i)* in Lemma 8. Therefore, by Lemma 8, there does not exist $\mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$, which implies k is characteristic. ■

Examples of integrally strictly pd kernels on $\mathbb{R}^d$ include the Gaussian, $\exp(-\sigma\|x-y\|_2^2), \sigma > 0$; the Laplacian, $\exp(-\sigma\|x-y\|_1), \sigma > 0$; inverse multiquadratics, $(\sigma^2 + \|x-y\|_2^2)^{-c}, c > 0, \sigma > 0$,

etc, which are translation invariant kernels on $\mathbb{R}^d$. A *translation variant* integrally strictly pd kernel, $\widetilde{k}$, can be obtained from a translation invariant integrally strictly pd kernel, k, as $\widetilde{k}(x,y) = f(x)k(x,y)f(y)$, where $f : M \to \mathbb{R}$ is a bounded continuous function. A simple example of a translation variant integrally strictly pd kernel on $\mathbb{R}^d$ is $\widetilde{k}(x,y) = \exp(\sigma x^T y)$, $\sigma > 0$, where we have chosen $f(\cdot) = \exp(\sigma\|\cdot\|_2^2/2)$ and $k(x,y) = \exp(-\sigma\|x-y\|_2^2/2)$, $\sigma > 0$. Clearly, this kernel is characteristic on compact subsets of $\mathbb{R}^d$. The same result can also be obtained from the fact that $\widetilde{k}$ is universal on compact subsets of $\mathbb{R}^d$ (Steinwart, 2001, Section 3, Example 1), recalling that universal kernels are characteristic (Gretton et al., 2007b, Theorem 3).

Although the condition for characteristic k in Theorem 7 is easy to understand compared to other characterizations in literature, it is not always easy to check for integral strict positive definiteness of k. In the following section, we assume $M = \mathbb{R}^d$ and k to be translation invariant and present a complete characterization for characteristic k which is simple to check.

3.2 Characterization for Translation Invariant k on $\mathbb{R}^d$

The complete, detailed proofs of the main results in this section are provided in Section 3.5. Compared to Sriperumbudur et al. (2008), we now present simple proofs for these results without resorting to distribution theory. Let us start with the following assumption.

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

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

Theorem 9 *Suppose k satisfies Assumption 1. Then k is characteristic if and only if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$, *where Λ is defined as in (11).*

First, note that the condition $\operatorname{supp}(\Lambda) = \mathbb{R}^d$ is easy to check compared to all other, aforementioned characterizations for characteristic k. Table 2 shows some popular translation invariant kernels on $\mathbb{R}$ along with their Fourier spectra, $\widehat{\psi}$ and its support: Gaussian, Laplacian, B_{2n+1}-spline[5] (Schölkopf and Smola, 2002) and Sinc kernels are aperiodic while Poisson (Brémaud, 2001; Steinwart, 2001; Vapnik, 1998), Dirichlet (Brémaud, 2001; Schölkopf and Smola, 2002), Féjer (Brémaud, 2001) and cosine kernels are periodic. Although the Gaussian and Laplacian kernels are shown to be characteristic by all the characterizations we have mentioned so far, the case of B_{2n+1}-splines is addressed only by Theorem 9, which shows them to be characteristic (note that B_{2n+1}-splines being integrally strictly pd also follows from Theorem 9). In fact, one can provide a more general result on compactly supported translation invariant kernels, which we do later in Corollary 10. The Matérn class of kernels (Rasmussen and Williams, 2006, Section 4.2.1), given by

$$k(x,y) = \psi(x-y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x-y\|_2}{\sigma}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x-y\|_2}{\sigma}\right), \nu > 0, \sigma > 0, \qquad (18)$$

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

Kernel	$\psi(x)$	$\widehat{\psi}(\omega)$	$\mathrm{supp}(\widehat{\psi})$
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	$*_1^{(2n+2)}\mathbb{1}_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$	$\frac{4^{n+1}}{\sqrt{2\pi}}\frac{\sin^{2n+2}\left(\frac{\omega}{2}\right)}{\omega^{2n+2}}$	$\mathbb{R}$
Sinc	$\frac{\sin(\sigma x)}{x}$	$\sqrt{\frac{\pi}{2}}\mathbb{1}_{[-\sigma,\sigma]}(\omega)$	$[-\sigma,\sigma]$
Poisson	$\frac{1-\sigma^2}{\sigma^2-2\sigma\cos(x)+1}$, $0<\sigma<1$	$\sqrt{2\pi}\sum_{j=-\infty}^{\infty}\sigma^{\lvert j\rvert}\delta(\omega-j)$	$\mathbb{Z}$
Dirichlet	$\frac{\sin\frac{(2n+1)x}{2}}{\sin\frac{x}{2}}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\delta(\omega-j)$	$\{0,\pm1,\ldots,\pm n\}$
Féjer	$\frac{1}{n+1}\frac{\sin^2\frac{(n+1)x}{2}}{\sin^2\frac{x}{2}}$	$\sqrt{2\pi}\sum_{j=-n}^{n}\left(1-\frac{\lvert j\rvert}{n+1}\right)\delta(\omega-j)$	$\{0,\pm1,\ldots,\pm n\}$
Cosine	$\cos(\sigma x)$	$\sqrt{\frac{\pi}{2}}\left[\delta(\omega-\sigma)+\delta(\omega+\sigma)\right]$	$\{-\sigma,\sigma\}$

Table 2: Translation invariant kernels on $\mathbb{R}$ defined by ψ, their spectra, $\widehat{\psi}$ and its support, $\mathrm{supp}(\widehat{\psi})$. The first four are aperiodic kernels while the last four are periodic. The domain is considered to be $\mathbb{R}$ for simplicity. For $x\in\mathbb{R}^d$, the above formulae can be extended by computing $\psi(x)=\prod_{j=1}^d\psi(x_j)$ where $x=(x_1,\ldots,x_d)$ and $\widehat{\psi}(\omega)=\prod_{j=1}^d\widehat{\psi}(\omega_j)$ where $\omega=(\omega_1,\ldots,\omega_d)$. δ represents the Dirac-delta distribution.

is characteristic as the Fourier spectrum of ψ, given by

$$\widehat{\psi}(\omega)=\frac{2^{d+\nu}\pi^{d/2}\Gamma(\nu+d/2)\nu^\nu}{\Gamma(\nu)\sigma^{2\nu}}\left(\frac{2\nu}{\sigma^2}+4\pi^2\|\omega\|_2^2\right)^{-(\nu+d/2)},\ \omega\in\mathbb{R}^d, \tag{19}$$

is positive for any $\omega\in\mathbb{R}^d$. Here, Γ is the Gamma function, K_ν is the modified Bessel function of the second kind of order ν, where ν controls the smoothness of k. The case of $\nu=\frac{1}{2}$ in the Matérn class gives the exponential kernel, $k(x,y)=\exp(-\|x-y\|_2/\sigma)$, while $\nu\to\infty$ gives the Gaussian kernel. Note that $\widehat{\psi}(x-y)$ in (19) is actually the inverse multiquadratic kernel, which is characteristic both by Theorem 7 and Theorem 9.

By Theorem 9, the Sinc kernel in Table 2 is not characteristic, which is not easy to show using other characterizations. By combining Theorem 7 with Theorem 9, it can be shown that the Sinc, Poisson, Dirichlet, Féjer and cosine kernels are not integrally strictly pd. Therefore, for translation invariant kernels on $\mathbb{R}^d$, the integral strict positive definiteness of the kernel (or the lack of it) can be tested using Theorems 7 and 9.

Of all the kernels shown in Table 2, only the Gaussian, Laplacian and B_{2n+1}-spline kernels are integrable and their corresponding $\widehat{\psi}$ are computed using (4). The other kernels shown in Table 2

are not integrable and their corresponding $\widehat{\psi}$ have to be treated as distributions (see Folland, 1999, Chapter 9 and Rudin, 1991, Chapter 6 for details), except for the Sinc kernel whose Fourier transform can be computed in the L^2 sense.[6]

Proof (Theorem 9) We provide an outline of the complete proof, which is presented in Section 3.5. The sufficient condition in Theorem 9 is simple to prove and follows from Corollary 4*(i)*, whereas we need a supplementary result to prove its necessity, which is presented in Lemma 16 (see Section 3.5). Proving the necessity of Theorem 9 is equivalent to showing that if $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, then $\exists \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. In Lemma 16, we present equivalent conditions for the existence of $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, using which we prove the necessity of Theorem 9. ■

The whole family of compactly supported translation invariant continuous bounded kernels on $\mathbb{R}^d$ is characteristic, as shown by the following corollary to Theorem 9.

Corollary 10 *Suppose $k \neq 0$ satisfies Assumption 1 and $\text{supp}(\psi)$ is compact. Then k is characteristic.*

Proof Since $\psi \in C_b(\mathbb{R}^d)$ is compactly supported on $\mathbb{R}^d$, by (6), $\psi \in \mathscr{D}'_d$. Therefore, by the Paley-Wiener theorem (Theorem 29 in Appendix A), $\widehat{\psi}$ is the restriction to $\mathbb{R}^d$ of an entire function on $\mathbb{C}^d$, which means $\widehat{\psi}$ is an analytic function on $\mathbb{R}^d$. Suppose $\text{supp}(\widehat{\psi})$ is compact, which means there exists an open set, $U \subset \mathbb{R}^d$ such that $\widehat{\psi}(x) = 0, \forall x \in U$. But being analytic, this implies that $\widehat{\psi}(x) = 0, \forall x \in \mathbb{R}^d$, that is, $\psi = 0$, which leads to a contradiction. Therefore, $\widehat{\psi}$ cannot be compactly supported, that is, $\text{supp}(\widehat{\psi}) = \mathbb{R}^d$, and the result follows from Theorem 9. ■

The above result is interesting in practice because of the computational advantage in dealing with compactly supported kernels. Note that proving such a general result for compactly supported kernels on $\mathbb{R}^d$ is not straightforward (maybe not even possible) with the other characterizations.

As a corollary to Theorem 9, the following result provides a method to construct new characteristic kernels from a given one.

Corollary 11 *Let k, k_1 and k_2 satisfy Assumption 1. Suppose k is characteristic and $k_2 \neq 0$. Then $k + k_1$ and $k \cdot k_2$ are characteristic.*

Proof Since k, k_1 and k_2 satisfy Assumption 1, $k + k_1$ and $k_2 \cdot k$ also satisfy Assumption 1. In addition,

$$\begin{aligned}
(k+k_1)(x,y) &:= k(x,y) + k_1(x,y) = \psi(x-y) + \psi_1(x-y) = \int_{\mathbb{R}^d} e^{-i(x-y)^T\omega}\, d(\Lambda + \Lambda_1)(\omega), \\
(k \cdot k_2)(x,y) &:= k(x,y)k_2(x,y) = \psi(x-y)\psi_2(x-y) = \iint_{\mathbb{R}^d} e^{-i(x-y)^T(\omega+\xi)}\, d\Lambda(\omega)\, d\Lambda_2(\xi) \\
&\overset{(a)}{=}: \int_{\mathbb{R}^d} e^{-i(x-y)^T\omega}\, d(\Lambda * \Lambda_2)(\omega),
\end{aligned}$$

6. If $f \in L^2(\mathbb{R}^d)$, the Fourier transform $\mathcal{F}[f] := \widehat{f}$ of f is defined to be the limit, in the L^2-norm, of the sequence $\{\widehat{f}_n\}$ 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 $\widehat{f}$ is defined almost everywhere on $\mathbb{R}^d$ and belongs to $L^2(\mathbb{R}^d)$. Thus, $\mathcal{F}$ is a linear operator, mapping $L^2(\mathbb{R}^d)$ into $L^2(\mathbb{R}^d)$. See Gasquet and Witomski (1999, Chapter IV, Lesson 22) for details.

where (a) follows from the definition of convolution of measures (see Rudin 1991, Section 9.14 for details). Since k is characteristic, that is, $\text{supp}(\Lambda) = \mathbb{R}^d$, and $\text{supp}(\Lambda) \subset \text{supp}(\Lambda + \Lambda_1)$, we have $\text{supp}(\Lambda + \Lambda_1) = \mathbb{R}^d$ and therefore $k + k_1$ is characteristic. Similarly, since $\text{supp}(\Lambda) \subset \text{supp}(\Lambda * \Lambda_2)$, we have $\text{supp}(\Lambda * \Lambda_2) = \mathbb{R}^d$ and therefore, $k \cdot k_2$ is characteristic. ■

Note that in the above result, we do not need k_1 or k_2 to be characteristic. Therefore, one can generate all sorts of kernels that are characteristic by starting with a characteristic kernel, k.

So far, we have considered characterizations for k such that it is characteristic to $\mathscr{P}$. We showed in Theorem 9 that kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $\mathscr{P}$. Now, we can question whether such kernels can be characteristic to some proper subset $\mathscr{Q}$ of $\mathscr{P}$. The following result addresses this. Note that these kernels, that is, the kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are usually not useful in practice, especially in statistical inference applications, because the conditions on $\mathscr{Q}$ are usually not easy to check. On the other hand, the following result is of theoretical interest: along with Theorem 9, it completes the characterization of characteristic kernels that are translation invariant on $\mathbb{R}^d$. Before we state the result, we denote $\mathbb{P} \ll \mathbb{Q}$ to mean that $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$.

Theorem 12 *Let $\mathscr{P}_1 := \{\mathbb{P} \in \mathscr{P} : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d),\ \mathbb{P} \ll \lambda$ and $\text{supp}(\mathbb{P})$ is compact$\}$, where λ is the Lebesgue measure. Suppose k satisfies Assumption 1 and $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior, where Λ is defined as in (11). Then k is characteristic to $\mathscr{P}_1$.*

Proof See Section 3.5. ■

Although, by Theorem 9, the kernels with $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ are not characteristic to $\mathscr{P}$, Theorem 12 shows that there exists a subset of $\mathscr{P}$ to which a subset of these kernels are characteristic. This type of result is not available for the previously mentioned characterizations. An example of a kernel that satisfies the conditions in Theorem 12 is the Sinc kernel, $\psi(x) = \frac{\sin(\sigma x)}{x}$ which has $\text{supp}(\Lambda) = [-\sigma, \sigma]$. The condition that $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ has a non-empty interior is important for Theorem 12 to hold. If $\text{supp}(\Lambda)$ has an empty interior (examples include periodic kernels), then one can construct $\mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. This is illustrated in Example 5 of Section 3.5.

So far, we have characterized the characteristic property of kernels that satisfy (a) $\text{supp}(\Lambda) = \mathbb{R}^d$ or (b) $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ with $\text{int}(\text{supp}(\Lambda)) \neq \emptyset$. In the following section, we investigate kernels that have $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$ with $\text{int}(\text{supp}(\Lambda)) = \emptyset$, examples of which include periodic kernels on $\mathbb{R}^d$. This discussion uses the fact that a periodic function on $\mathbb{R}^d$ can be treated as a function on $\mathbb{T}^d$, the d-Torus.

3.3 Characterization for Translation Invariant k on $\mathbb{T}^d$

Let $M = \times_{j=1}^d [0, \tau_j)$ and $\tau := (\tau_1, \ldots, \tau_d)$. A function defined on M with periodic boundary conditions is equivalent to considering a periodic function on $\mathbb{R}^d$ with period τ. With no loss of generality, we can choose $\tau_j = 2\pi, \forall\, j$ which yields $M = [0, 2\pi)^d =: \mathbb{T}^d$, called the d-Torus. The results presented here hold for any $0 < \tau_j < \infty, \forall\, j$ but we choose $\tau_j = 2\pi$ for simplicity. Similar to Assumption 1, we now make the following assumption.

Assumption 2 *$k(x, y) = \psi((x - y)_{mod\ 2\pi})$, where ψ is a continuous real-valued positive definite function on $M = \mathbb{T}^d$.*

Similar to Theorem 3, we now state Bochner's theorem on $M = \mathbb{T}^d$.

Theorem 13 (Bochner) *A continuous function $\psi : \mathbb{T}^d \to \mathbb{R}$ is positive definite if and only if*

$$\psi(x) = \sum_{n \in \mathbb{Z}^d} A_\psi(n)\, e^{ix^T n},\ x \in \mathbb{T}^d, \tag{20}$$

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

Examples for ψ include the Poisson, Dirichlet, Féjer and cosine kernels, which are shown in Table 2. We now state the result that defines characteristic kernels on $\mathbb{T}^d$.

Theorem 14 *Suppose k satisfies Assumption 2. Then k is characteristic (to the set of all Borel probability measures on $\mathbb{T}^d$) if and only if $A_\psi(0) \geq 0$, $A_\psi(n) > 0, \forall\, n \neq 0$.*

The proof is provided in Section 3.5 and the idea is similar to that of Theorem 9. Based on the above result, one can generate characteristic kernels by constructing an infinite sequence of positive numbers that are summable and then using them in (20). It can be seen from Table 2 that the Poisson kernel on $\mathbb{T}$ is characteristic while the Dirichlet, Féjer and cosine kernels are not. Some examples of characteristic kernels on $\mathbb{T}$ are:

(1) $k(x,y) = e^{\alpha\cos(x-y)}\cos(\alpha\sin(x-y)),\ 0 < \alpha \leq 1 \leftrightarrow A_\psi(0) = 1, A_\psi(n) = \frac{\alpha^{|n|}}{2|n|!},\ \forall\, n \neq 0.$

(2) $k(x,y) = -\log(1 - 2\alpha\cos(x-y) + \alpha^2),\ |\alpha| < 1 \leftrightarrow A_\psi(0) = 0, A_\psi(n) = \frac{\alpha^n}{n},\ \forall\, n \neq 0.$

(3) $k(x,y) = (\pi - (x-y)_{mod\ 2\pi})^2 \leftrightarrow A_\psi(0) = \frac{\pi^2}{3}, A_\psi(n) = \frac{2}{n^2},\ \forall\, n \neq 0.$

(4) $k(x,y) = \frac{\sinh\alpha}{\cosh\alpha - \cos(x-y)},\ \alpha > 0 \leftrightarrow A_\psi(0) = 1, A_\psi(n) = e^{-\alpha|n|},\ \forall\, n \neq 0.$

(5) $k(x,y) = \frac{\pi\cosh(\alpha(\pi - (x-y)_{mod\,2\pi}))}{\alpha\sinh(\pi\alpha)} \leftrightarrow A_\psi(0) = \frac{1}{\alpha^2}, A_\psi(n) = \frac{1}{n^2+\alpha^2},\ \forall\, n \neq 0.$

The following result relates characteristic kernels and universal kernels defined on $\mathbb{T}^d$.

Corollary 15 *Let k be a characteristic kernel satisfying Assumption 2 with $A_\psi(0) > 0$. Then k is also universal.*

Proof Since k is characteristic with $A_\psi(0) > 0$, we have $A_\psi(n) > 0, \forall\, n$. Therefore, by Corollary 11 of Steinwart (2001), k is universal. ■

Since k being universal implies that it is characteristic, the above result shows that the converse is not true (though almost true except that $A_\psi(0)$ can be zero for characteristic kernels). The condition on A_ψ in Theorem 14, that is, $A_\psi(0) \geq 0, A_\psi(n) > 0, \forall\, n \neq 0$ can be equivalently written as $\text{supp}(A_\psi) = \mathbb{Z}^d$ or $\text{supp}(A_\psi) = \mathbb{Z}^d \backslash \{0\}$. Therefore, Theorems 9 and 14 are of similar flavor. In fact, these results can be generalized to locally compact Abelian groups. Fukumizu et al. (2009b) shows that a bounded continuous translation invariant kernel on a locally compact Abelian group G is characteristic to the set of all probability measures on G if and only if the support of the Fourier transform of the translation invariant kernel is the dual group of G. In our case, $(\mathbb{R}^d, +)$ and $(\mathbb{T}^d, +)$ are locally compact Abelian groups with $(\mathbb{R}^d, +)$ and $(\mathbb{Z}^d, +)$ as their respective dual groups. In Fukumizu et al. (2009b), these results are also extended to translation invariant kernels on non-Abelian compact groups and the semigroup $\mathbb{R}^d_+$.

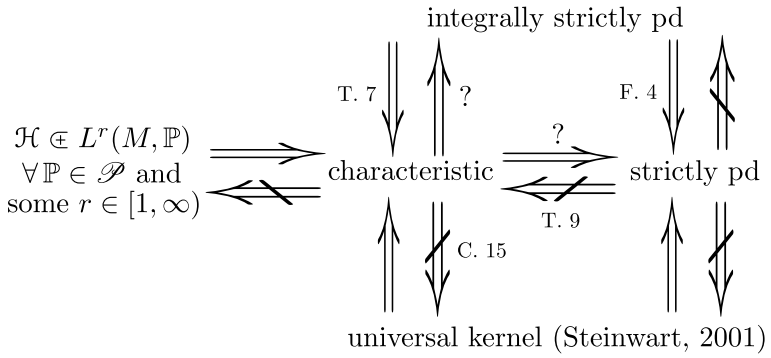

Figure 1: Summary of the relations between various families of kernels is shown along with the reference. The letters "C", "F", and "T" refer to Corollary, Footnote and Theorem respectively. For example, T. 7 refers to Theorem 7. The implications which are open problems are shown with "?". $A \Subset B$ indicates that A is a dense subset of B. Refer to Section 3.4 for details.

3.4 Overview of Relations Between Families of Kernels

So far, we have presented various characterizations of characteristic kernels, which are easily checkable compared with characterizations proposed in the earlier literature (Gretton et al., 2007b; Fukumizu et al., 2008, 2009b). We now provide an overview of various useful conditions one can impose on kernels (to be universal, strictly pd, integrally strictly pd, or characteristic), and the implications that relate some of these conditions. A summary is provided in Figure 1.

Characteristic kernels vs. Integrally strictly pd kernels: It is clear from Theorem 7 that integrally strictly pd kernels on a topological space M are characteristic, whereas the converse remains undetermined. When k is translation invariant on $\mathbb{R}^d$, however, then the converse holds. This is because if k is characteristic, then by Theorem 9, $\text{supp}(\Lambda) = \mathbb{R}^d$, where Λ is defined as in (11). It is easy to check that if $\text{supp}(\Lambda) = \mathbb{R}^d$, then k is integrally strictly pd.

Integrally strictly pd kernels vs. Strictly pd kernels: The relation between integrally strictly pd and strictly pd kernels shown in Figure 1 is straightforward, as one direction follows from Footnote 4, while the other direction is not true, which follows from Steinwart and Christmann (2008, Proposition 4.60, Theorem 4.62). However, if M is a finite set, then k being strictly pd also implies it is integrally strictly pd.

Characteristic kernels vs. Strictly pd kernels: Since integrally strictly pd kernels are characteristic and are also strictly pd, a natural question to ask is, "What is the relation between characteristic and strictly pd kernels?" It can be seen that strictly pd kernels need not be characteristic because the sinc-squared kernel, $k(x,y) = \frac{\sin^2(\sigma(x-y))}{(x-y)^2}$ on $\mathbb{R}$, which has $\text{supp}(\Lambda) = [-\sigma,\sigma] \subsetneq \mathbb{R}$ is strictly pd (Wendland, 2005, Theorem 6.11), while it is not characteristic by Theorem 9. However, for any general M, it is not clear whether k being characteristic implies that it is strictly pd. As a special case, if $M = \mathbb{R}^d$ or $M = \mathbb{T}^d$, then by Theorems 9 and 12, it follows that a translation invariant k being characteristic also implies that it is strictly pd.

Universal kernels vs. Characteristic kernels: Gretton et al. (2007b) have shown that if k is universal in the sense of Steinwart (2001), then it is characteristic. As mentioned in Section 3.3, the converse is not true, that is, if a kernel is characteristic, then it need not be universal, which

follows from Corollary 15. Note that in this case, M is assumed to be a compact metric space. The notion of universality of kernels was extended to non-compact domains by Micchelli et al. (2006): k is said to be universal on a non-compact Hausdorff space, M, if for any compact $Z \subset M$, the set $K(Z) := \overline{\text{span}}\{k(\cdot, y) : y \in Z\}$ is dense in $C_b(Z)$ w.r.t. the supremum norm. It is to be noted that when M is compact, this notion of universality is same as that of Steinwart (2001). Micchelli et al. (2006, Proposition 15) have provided a characterization of universality for translation invariant kernels on $\mathbb{R}^d$: k is universal if $\lambda(\text{supp}(\Lambda)) > 0$, where λ is the Lebesgue measure and Λ is defined as in (11). This means if a translation invariant kernel on $\mathbb{R}^d$ is characteristic, that is, $\text{supp}(\Lambda) = \mathbb{R}^d$, then it is also universal in the sense of Micchelli et al. (2006), while the converse is not true (e.g., sinc-squared kernel is not characteristic as $\text{supp}(\Lambda) = [-\sigma, \sigma] \subsetneq \mathbb{R}$ but universal in the sense of Micchelli as $\lambda(\text{supp}(\Lambda)) = 2\sigma > 0$). The relation between these notions for a general non-compact Hausdorff space M (other than $\mathbb{R}^d$) remains to be determined (whether or not the kernel is translation invariant).

Fukumizu et al. (2008, 2009b) have shown that k is characteristic if and only if $\mathcal{H} + \mathbb{R}$ is dense in $L^r(M, \mathbb{P})$ for all $\mathbb{P} \in \mathscr{P}$ and for some $r \in [1, \infty)$. Using this, it is easy to see that if $\mathcal{H}$ is dense in $L^r(M, \mathbb{P})$ for all $\mathbb{P} \in \mathscr{P}$ and for some $r \in [1, \infty)$, then k is characteristic. Clearly, the converse is not true. However, if constant functions are included in $\mathcal{H}$, then it is easy to see that the converse is also true.

Universal kernels vs. Strictly pd kernels: If a kernel is universal, then it is strictly pd, which follows from Steinwart and Christmann (2008, Definition 4.53, Proposition 4.54, Exercise 4.11). On the other hand, if a kernel is strictly pd, then it need not be universal, which follows from the results due to Dahmen and Micchelli (1987) and Pinkus (2004) for Taylor kernels (Steinwart and Christmann, 2008, Lemma 4.8, Corollary 4.57). Refer to Steinwart and Christmann (2008, Section 4.7, p. 161) for more details.

Recently, Sriperumbudur et al. (2010a,b) carried out a thorough study relating characteristic kernels to various notions of universality, addressing some open questions mentioned in the above discussion and Figure 1. This is done by relating universality to the injective embedding of regular Borel measures into an RKHS, which can therefore be seen as a generalization of the notion of characteristic kernels, as the latter deal with the injective RKHS embedding of probability measures.

3.5 Proofs

First, we present a supplementary result in Lemma 16 that will be used to prove Theorem 9. The idea of Lemma 16 is to characterize the equivalent conditions for the existence of $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$. Its proof relies on the properties of characteristic functions, which we have collected in Theorem 27 in Appendix A.

Lemma 16 *Let $\mathscr{P}_0 := \{\mathbb{P} \in \mathscr{P} : \phi_{\mathbb{P}} \in L^1(\mathbb{R}^d) \cup L^2(\mathbb{R}^d)$ and $\mathbb{P} \ll \lambda\}$, where λ is the Lebesgue measure. Suppose k satisfies Assumption 1 and $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, where Λ is defined as in (11). Then, for any $\mathbb{Q} \in \mathscr{P}_0$, $\exists\, \mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathscr{P}_0$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$ if and only if there exists a non-zero function $\theta : \mathbb{R}^d \to \mathbb{C}$ that satisfies the following conditions:*

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

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

(iii) $\int_{\mathbb{R}^d} |\theta(x)|^2 \, d\Lambda(x) = 0$,

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

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

Proof Define $L^1 := L^1(\mathbb{R}^d)$, $L^2 := L^2(\mathbb{R}^d)$ and $C_b := C_b(\mathbb{R}^d)$.

($\Leftarrow$) Suppose there exists a non-zero function θ satisfying *(i) – (v)*. For any $\mathbb{Q} \in \mathscr{P}_0$, we have $\phi_\mathbb{Q} \in L^1 \cup L^2$ and $\phi_\mathbb{Q} \in C_b$ (by Theorem 27), that is, $\phi_\mathbb{Q} \in (L^1 \cup L^2) \cap C_b$. Now, consider the case of $\phi_\mathbb{Q} \in L^1 \cap C_b$. Since $\phi_\mathbb{Q} \in L^1$, by the inversion theorem for characteristic functions (Dudley, 2002, Theorem 9.5.4), $\mathbb{Q}$ is absolutely continuous w.r.t. λ. If q is the Radon-Nikodym derivative of $\mathbb{Q}$ w.r.t. λ, then $q = [\overline{\phi_\mathbb{Q}}]^\vee \in L^1$. In addition, by the Riemann-Lebesgue lemma (Lemma 28 in Appendix A), we have $q \in C_0(\mathbb{R}^d) \subset C_b$, which therefore implies $q \in L^1 \cap C_b$. When $\phi_\mathbb{Q} \in L^2 \cap C_b$, the Fourier transform in the L^2 sense (see Footnote 6) implies that $q = [\overline{\phi_\mathbb{Q}}]^\vee \in L^1 \cap L^2$. Therefore, $q \in L^1 \cap (L^2 \cup C_b)$. Define $p := q + \theta^\vee$. Clearly $p \in L^1 \cap (L^2 \cup C_b)$. In addition, $\overline{\phi_\mathbb{P}} = \widehat{p} = \widehat{q} + \widehat{\theta^\vee} = \overline{\phi_\mathbb{Q}} + \theta \in (L^1 \cup L^2) \cap C_b$. Since θ is conjugate symmetric, $\theta^\vee$ is real valued and so is p. Consider

$$\int_{\mathbb{R}^d} p(x)\,dx = \int_{\mathbb{R}^d} q(x)\,dx + \int_{\mathbb{R}^d} \theta^\vee(x)\,dx = 1 + \theta(0) = 1.$$

(v) implies that p is non-negative. Therefore, p is the Radon-Nikodym derivative of a probability measure $\mathbb{P}$ w.r.t. λ, where $\mathbb{P}$ is such that $\mathbb{P} \neq \mathbb{Q}$ and $\mathbb{P} \in \mathscr{P}_0$. By (12), we have

$$\gamma_k^2(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} |\phi_\mathbb{P}(x) - \phi_\mathbb{Q}(x)|^2\, d\Lambda(x) = \int_{\mathbb{R}^d} |\theta(x)|^2\, d\Lambda(x) = 0.$$

($\Rightarrow$) Suppose that there exists $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P},\mathbb{Q} \in \mathscr{P}_0$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. Define $\theta := \phi_\mathbb{P} - \phi_\mathbb{Q}$. We need to show that θ satisfies *(i) – (v)*. Recalling Theorem 27 in the appendix, $\mathbb{P},\mathbb{Q} \in \mathscr{P}_0$ implies $\phi_\mathbb{P}, \phi_\mathbb{Q} \in (L^1 \cup L^2) \cap C_b$ and $p, q \in L^1 \cap (L^2 \cup C_b)$. Therefore, $\theta = \overline{\phi}_\mathbb{P} - \overline{\phi}_\mathbb{Q} \in (L^1 \cup L^2) \cap C_b$ and $\theta^\vee = p - q \in L^1 \cap (L^2 \cup C_b)$. By Theorem 27 (see Appendix A), $\phi_\mathbb{P}$ and $\phi_\mathbb{Q}$ are conjugate symmetric and so is θ. Therefore θ satisfies *(i)* and $\theta^\vee$ satisfies *(ii)*. θ satisfies *(iv)* as

$$\theta(0) = \int_{\mathbb{R}^d} \theta^\vee(x)\,dx = \int_{\mathbb{R}^d} (p(x) - q(x))\,dx = 0.$$

Non-negativity of p yields *(v)*. By (12), $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$ implies *(iii)*. ■

Remark 17 *Note that the dependence of* θ *on the kernel appears in the form of (iii) in Lemma 16. This condition shows that* $\lambda(\mathrm{supp}(\theta) \cap \mathrm{supp}(\Lambda)) = 0$*, that is, the supports of* θ *and* Λ *are disjoint w.r.t. the Lebesgue measure,* λ*. In other words,* $\mathrm{supp}(\theta) \subset \mathrm{cl}(\mathbb{R}^d \backslash \mathrm{supp}(\Lambda))$*. So, the idea is to introduce the perturbation,* θ *over an open set,* U *where* $\Lambda(U) = 0$*. The remaining conditions characterize the nature of this perturbation so that the constructed measure,* $p = q + \theta^\vee$*, is a valid probability measure. Conditions (i), (ii) and (iv) simply follow from* $\theta = \phi_\mathbb{P} - \phi_\mathbb{Q}$*, while (v) ensures that* $p(x) \geq 0, \forall x$*.*

Using Lemma 16, we now present the proof of Theorem 9.

Proof(Theorem 9) The sufficiency follows from (12): if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, then $\gamma_k^2(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}^d} |\phi_\mathbb{P}(x) - \phi_\mathbb{Q}(x)|^2\, d\Lambda(x) = 0 \Rightarrow \phi_\mathbb{P} = \phi_\mathbb{Q}$, a.e. Recalling from Theorem 27 that $\phi_\mathbb{P}$ and $\phi_\mathbb{Q}$ are uniformly continuous on $\mathbb{R}^d$, we have that $\mathbb{P} = \mathbb{Q}$, and therefore k is characteristic. To prove necessity, we need to show that if $\mathrm{supp}(\Lambda) \subsetneq \mathbb{R}^d$, then there exists $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P},\mathbb{Q} \in \mathscr{P}$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. By Lemma 16, this is equivalent to showing that there exists a non-zero θ satisfying the conditions in

Lemma 16. Below, we provide a constructive procedure for such a θ when $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, thereby proving the result.

Consider the following function, $f_{\beta,\omega_0} \in C^\infty(\mathbb{R}^d)$ supported in $[\omega_0 - \beta, \omega_0 + \beta]$,

$$f_{\beta,\omega_0}(\omega) = \prod_{j=1}^{d} h_{\beta_j,\omega_{0,j}}(\omega_j) \text{ with } h_{a,b}(y) := \mathbb{1}_{[-a,a]}(y-b)\, e^{-\frac{a^2}{a^2-(y-b)^2}},$$

where $\omega = (\omega_1, \ldots, \omega_d)$, $\omega_0 = (\omega_{0,1}, \ldots, \omega_{0,d})$, $\beta = (\beta_1, \ldots, \beta_d)$, $a \in \mathbb{R}_{++}$, $b \in \mathbb{R}$ and $y \in \mathbb{R}$. Since $\text{supp}(\Lambda) \subsetneq \mathbb{R}^d$, there exists an open set $U \subset \mathbb{R}^d$ such that $\Lambda(U) = 0$. So, there exists $\beta \in \mathbb{R}^d_{++}$ and $\omega_0 > \beta$ (element-wise inequality) such that $[\omega_0 - \beta, \omega_0 + \beta] \subset U$. Let

$$\theta = \alpha(f_{\beta,\omega_0} + f_{\beta,-\omega_0}), \alpha \in \mathbb{R}\backslash\{0\},$$

which implies $\text{supp}(\theta) = [-\omega_0 - \beta, -\omega_0 + \beta] \cup [\omega_0 - \beta, \omega_0 + \beta]$ is compact. Clearly $\theta \in \mathscr{D}_d \subset \mathscr{S}_d$ which implies $\theta^\vee \in \mathscr{S}_d \subset L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Therefore, by construction, θ satisfies *(i) – (iv)* in Lemma 16. Since $\int_{\mathbb{R}^d} \theta^\vee(x)\, dx = \theta(0) = 0$ (by construction), $\theta^\vee$ will take negative values, so we need to show that there exists $\mathbb{Q} \in \mathscr{P}_0$ such that *(v)* in Lemma 16 holds. Let $\mathbb{Q}$ be such that it has a density given by

$$q(x) = C_l \prod_{j=1}^{d} \frac{1}{(1+|x_j|^2)^l}, l \in \mathbb{N} \text{ where } C_l = \prod_{j=1}^{d} \left(\int_{\mathbb{R}} (1+|x_j|^2)^{-l}\, dx_j \right)^{-1},$$

and $x = (x_1, \ldots, x_d)$. It can be verified that choosing α such that

$$0 < |\alpha| \leq \frac{C_l}{2 \sup_x \left| \prod_{j=1}^{d} h^\vee_{\beta_j,0}(x_j)(1+|x_j|^2)^l \cos(\omega_0^T x) \right|} < \infty,$$

ensures that θ satisfies *(v)* in Lemma 16. The existence of finite α is guaranteed as $h_{a,0} \in \mathscr{D}_1 \subset \mathscr{S}_1$ which implies $h^\vee_{a,0} \in \mathscr{S}_1, \forall a$. We conclude there exists a non-zero θ as claimed earlier, which completes the proof. ■

To elucidate the necessity part in the above proof, in the following, we present a simple example that provides an intuitive understanding about the construction of θ such that for a given $\mathbb{Q}$, $\mathbb{P} \neq \mathbb{Q}$ can be constructed with $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$.

Example 4 *Let $\mathbb{Q}$ be a Cauchy distribution in $\mathbb{R}$, that is, $q(x) = \frac{1}{\pi(1+x^2)}$ with characteristic function, $\phi_{\mathbb{Q}}(\omega) = \frac{1}{\sqrt{2\pi}} e^{-|\omega|}$ in $L^1(\mathbb{R})$. Let ψ be a Sinc kernel, that is, $\psi(x) = \sqrt{\frac{2}{\pi}} \frac{\sin(\beta x)}{x}$ with Fourier transform given by $\widehat{\psi}(\omega) = \mathbb{1}_{[-\beta,\beta]}(\omega)$ and $\text{supp}(\widehat{\psi}) = [-\beta, \beta] \subsetneq \mathbb{R}$. Let θ be*

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

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. Note that θ is such that $\text{supp}(\theta) \cap \text{supp}(\widehat{\psi})$ is a null set w.r.t. the Lebesgue measure, which satisfies (iii) in Lemma 16. It is easy to verify that $\theta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ also satisfies conditions (i) and (iv) in Lemma 16. $\theta^\vee$ can be computed as

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

and $\theta^\vee \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ *satisfies* (ii) *in Lemma 16. Choose*

$$0 < |\alpha| \leq \frac{\sqrt{2}}{\sqrt{\pi}\beta^N \sup_x \left| (1+x^2)\sin(\omega_0 x)\mathrm{sinc}^N\left(\frac{\beta x}{2\pi}\right)\right|},$$

where $\mathrm{sinc}(x) := \frac{\sin(\pi x)}{\pi x}$. *Define* $g(x) := \sin(\omega_0 x)\mathrm{sinc}^N\left(\frac{\beta x}{2\pi}\right)$. *Since* $g \in \mathscr{S}_1$, $0 < \sup_x |(1+x^2)g(x)| < \infty$ *and, therefore,* α *is a finite non-zero number. It is easy to see that* θ *satisfies* (v) *of Lemma 16. Then, by Lemma 16, there exists* $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P} \in \mathscr{P}_0$, *given by*

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

with $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} + \theta = \phi_{\mathbb{Q}} + i\theta_I$ *where* $\theta_I = \mathrm{Im}[\theta]$ *and* $\phi_{\mathbb{P}} \in L^1(\mathbb{R})$. *So, we have constructed* $\mathbb{P} \neq \mathbb{Q}$, *such that* $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. *Figure 2 shows the plots of* ψ, $\widehat{\psi}$, θ, $\theta^\vee$, q, $\phi_{\mathbb{Q}}$, p *and* $|\phi_{\mathbb{P}}|$ *for* $\beta = 2\pi$, $N = 2$, $\omega_0 = 4\pi$ *and* $\alpha = \frac{1}{50}$.

We now prove Theorem 12.

Proof(Theorem 12) Suppose $\exists\, \mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$. Since any positive Borel measure on $\mathbb{R}^d$ is a distribution (Rudin, 1991, p. 157), $\mathbb{P}$ and $\mathbb{Q}$ can be treated as distributions with compact support. By the Paley-Wiener theorem (Theorem 29 in Appendix A), $\phi_{\mathbb{P}}$ and $\phi_{\mathbb{Q}}$ are restrictions to $\mathbb{R}^d$ of entire functions on $\mathbb{C}^d$. Let $\theta := \phi_{\mathbb{P}} - \phi_{\mathbb{Q}}$. Since $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$, we have from (12) that $\int_{\mathbb{R}^d} |\theta(\omega)|^2 d\Lambda(\omega) = 0$. From Remark 17, it follows that $\mathrm{supp}(\theta) \subset \mathrm{cl}(\mathbb{R}^d \backslash \mathrm{supp}(\Lambda))$. Since $\mathrm{supp}(\Lambda)$ has a non-empty interior, we have $\mathrm{supp}(\theta) \subsetneq \mathbb{R}^d$. Thus, there exists an open set, $U \subset \mathbb{R}^d$ such that $\theta(x) = 0, \forall x \in U$. Since θ is analytic on $\mathbb{R}^d$, we have $\theta = 0$, which means $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} \Rightarrow \mathbb{P} = \mathbb{Q}$, leading to a contradiction. So, there does not exist $\mathbb{P} \neq \mathbb{Q}, \mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$, and k is therefore characteristic to $\mathscr{P}_1$. ■

The condition that $\mathrm{supp}(\Lambda)$ has a non-empty interior is important for Theorem 12 to hold. In the following, we provide a simple example to show that $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$ can be constructed such that $\gamma_k(\mathbb{P},\mathbb{Q}) = 0$, if k is a periodic translation invariant kernel for which $\mathrm{int}(\mathrm{supp}(\Lambda)) = \emptyset$.

Example 5 *Let* $\mathbb{Q}$ *be a uniform distribution on* $[-\beta,\beta] \subset \mathbb{R}$, *that is,* $q(x) = \frac{1}{2\beta}\mathbb{1}_{[-\beta,\beta]}(x)$ *with its characteristic function,* $\phi_{\mathbb{Q}}(\omega) = \frac{1}{\beta\sqrt{2\pi}}\frac{\sin(\beta\omega)}{\omega} \in L^2(\mathbb{R})$. *Let* ψ *be the Dirichlet kernel with period* τ, *where* $\tau \leq \beta$, *that is,* $\psi(x) = \frac{\sin\frac{(2l+1)\pi x}{\tau}}{\sin\frac{\pi x}{\tau}}$ *and* $\widehat{\psi}(\omega) = \sqrt{2\pi}\sum_{j=-l}^{l}\delta\left(\omega - \frac{2\pi j}{\tau}\right)$ *with* $\mathrm{supp}(\widehat{\psi}) = \left\{\frac{2\pi j}{\tau} : j \in \{0, \pm 1, \ldots, \pm l\}\right\}$. *Clearly,* $\mathrm{supp}(\widehat{\psi})$ *has an empty interior. Let* θ *be*

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

with $\alpha \leq \frac{1}{2\beta}$. *It is easy to verify that* $\theta \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$, *so* θ *satisfies* (i) *in Lemma 16. Since* $\theta(\omega) = 0$ *at* $\omega = \frac{2\pi l}{\tau}$, $l \in \mathbb{Z}$, $\mathrm{supp}(\theta) \cap \mathrm{supp}(\widehat{\psi}) \subset \mathrm{supp}(\widehat{\psi})$ *is a set of Lebesgue measure zero, so* (iii) *and* (iv) *in Lemma 16 are satisfied.* $\theta^\vee$ *is given by*

$$\theta^\vee(x) = \begin{cases} \frac{2\alpha\left|x+\frac{\tau}{2}\right|}{\tau} - \alpha, & -\tau \leq x \leq 0 \\ \alpha - \frac{2\alpha\left|x-\frac{\tau}{2}\right|}{\tau}, & 0 \leq x \leq \tau \\ 0, & \text{otherwise,} \end{cases}$$

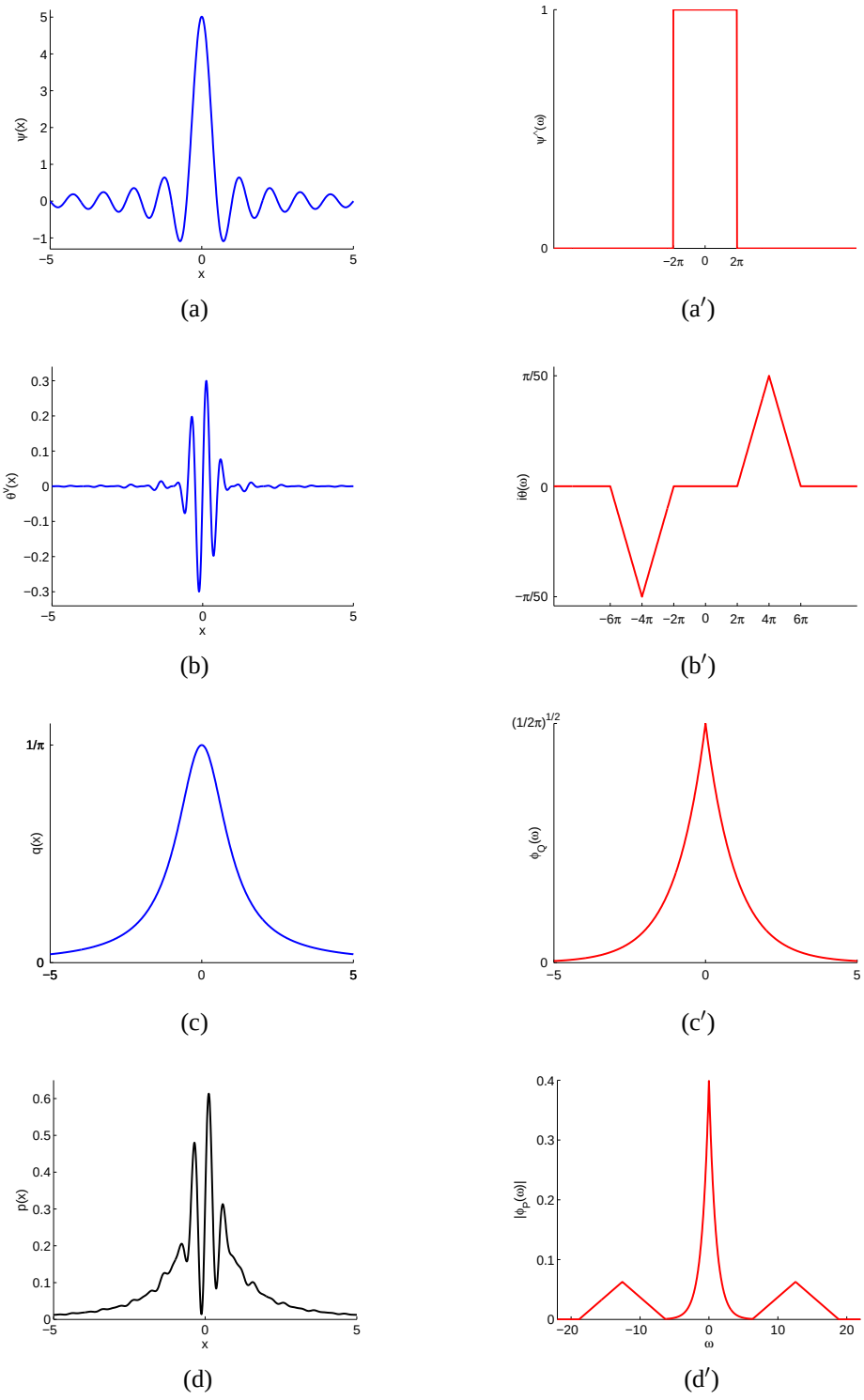

Figure 2: (a-a′) ψ and its Fourier spectrum $\widehat{\psi}$, (b-b′) $\theta^{\vee}$ and $i\theta$, (c-c′) the Cauchy distribution, q and its characteristic function $\phi_{\mathbb{Q}}$, and (d-d′) $p = q + \theta^{\vee}$ and $|\phi_{\mathbb{P}}|$. See Example 4 for details.

where $\theta^\vee \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap C_b(\mathbb{R})$ *satisfies* (ii) *in Lemma 16. Now, consider* $p = q + \theta^\vee$, *which is given as*

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

Clearly, $p(x) \geq 0, \forall x$ *and* $\int_{\mathbb{R}} p(x)\,dx = 1$. $\phi_{\mathbb{P}} = \phi_{\mathbb{Q}} + \theta = \phi_{\mathbb{Q}} + i\theta_I$ *where* $\theta_I = \mathrm{Im}[\theta]$ *and* $\phi_{\mathbb{P}} \in L^2(\mathbb{R})$. *We have therefore constructed* $\mathbb{P} \neq \mathbb{Q}$, *such that* $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$, *where* $\mathbb{P}$ *and* $\mathbb{Q}$ *are compactly supported in* $\mathbb{R}$ *with characteristic functions in* $L^2(\mathbb{R})$, *that is,* $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$. *Figure 3 shows the plots of* ψ, $\widehat{\psi}$, θ, $\theta^\vee$, q, $\phi_{\mathbb{Q}}$, p *and* $|\phi_{\mathbb{P}}|$ *for* $\tau = 2$, $l = 2$, $\beta = 3$ *and* $\alpha = \frac{1}{8}$.

We now present the proof of Theorem 14, which is similar to that of Theorem 9.

Proof (Theorem 14) ($\Leftarrow$) From (10), we have

$$\begin{aligned} \gamma_k^2(\mathbb{P}, \mathbb{Q}) &= \iint_{\mathbb{T}^d} \psi(x-y)\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(a)}{=} \iint_{\mathbb{T}^d} \sum_{n\in\mathbb{Z}^d} A_\psi(n)\, e^{i(x-y)^T n}\,d(\mathbb{P}-\mathbb{Q})(x)\,d(\mathbb{P}-\mathbb{Q})(y) \\ &\overset{(b)}{=} \sum_{n\in\mathbb{Z}^d} A_\psi(n) \left| \int_{\mathbb{T}^d} e^{-ix^T n}\,d(\mathbb{P}-\mathbb{Q})(x) \right|^2 \\ &\overset{(c)}{=} (2\pi)^{2d} \sum_{n\in\mathbb{Z}^d} A_\psi(n)\,|A_{\mathbb{P}}(n) - A_{\mathbb{Q}}(n)|^2, \end{aligned} \tag{21}$$

where we have invoked Bochner's theorem (Theorem 13) in (a), Fubini's theorem in (b) and

$$A_{\mathbb{P}}(n) := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} e^{-in^T x}\,d\mathbb{P}(x),\ n \in \mathbb{Z}^d,$$

in (c). $A_{\mathbb{P}}$ is the Fourier transform of $\mathbb{P}$ in $\mathbb{T}^d$. Since $A_\psi(0) \geq 0$ and $A_\psi(n) > 0, \forall n \neq 0$, we have $A_{\mathbb{P}}(n) = A_{\mathbb{Q}}(n), \forall n$. Therefore, by the uniqueness theorem of Fourier transform, we have $\mathbb{P} = \mathbb{Q}$.

($\Rightarrow$) Proving the necessity is equivalent to proving that if $A_\psi(0) \geq 0, A_\psi(n) > 0, \forall n \neq 0$ is violated, then k is not characteristic, which is equivalent to showing that $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Let $\mathbb{Q}$ be a uniform probability measure with $q(x) = \frac{1}{(2\pi)^d}, \forall x \in \mathbb{T}^d$. Let k be such that $A_\psi(n) = 0$ for some $n = n_0 \neq 0$. Define

$$A_{\mathbb{P}}(n) := \begin{cases} A_{\mathbb{Q}}(n), & n \neq \pm n_0 \\ A_{\mathbb{Q}}(n) + \theta(n), & n = \pm n_0 \end{cases},$$

where $A_{\mathbb{Q}}(n) = \frac{1}{(2\pi)^d}\delta_{0n}$ and $\theta(-n_0) = \overline{\theta(n_0)}$. So,

$$p(x) = \sum_{n\in\mathbb{Z}^d} A_{\mathbb{P}}(n) e^{ix^T n} = \frac{1}{(2\pi)^d} + \theta(n_0) e^{ix^T n_0} + \theta(-n_0) e^{-ix^T n_0}.$$

Choose $\theta(n_0) = i\alpha$, $\alpha \in \mathbb{R}$. Then, $p(x) = \frac{1}{(2\pi)^d} - 2\alpha \sin(x^T n_0)$. It is easy to check that p integrates to one. Choosing $|\alpha| \leq \frac{1}{2(2\pi)^d}$ ensures that $p(x) \geq 0, \forall x \in \mathbb{T}^d$. By using $A_{\mathbb{P}}(n)$ in (21), it is clear that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$. Therefore, $\exists\, \mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0$, which means k is not characteristic. ■

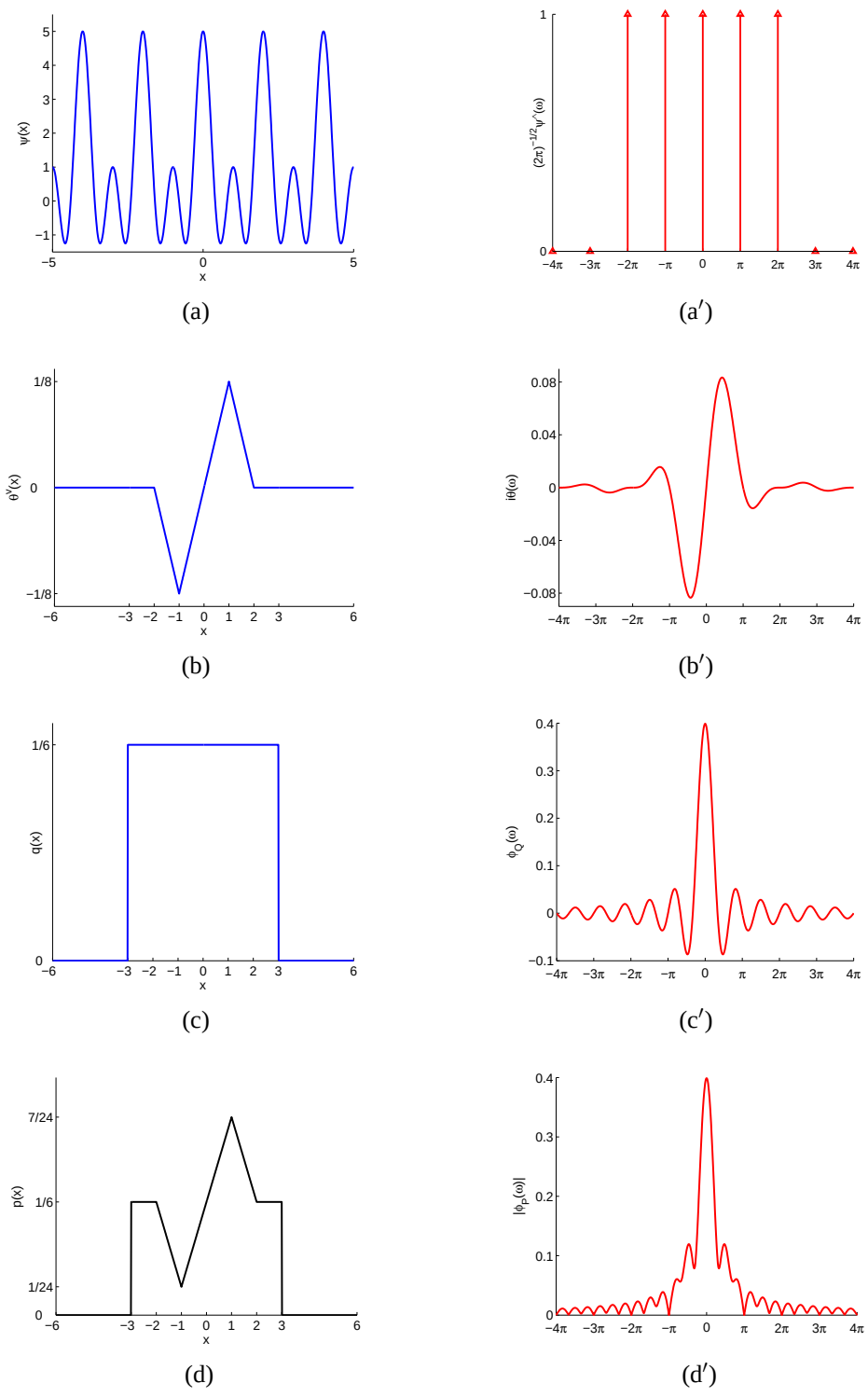

Figure 3: (a-a′) ψ and its Fourier spectrum $\widehat{\psi}$, (b-b′) $\theta^\vee$ and $i\theta$, (c-c′) the uniform distribution, q and its characteristic function $\phi_{\mathbb{Q}}$, and (d-d′) $p = q + \theta^\vee$ and $|\phi_{\mathbb{P}}|$. See Example 5 for details.

4. Dissimilar Distributions with Small γ_k

So far, we have studied different characterizations for the kernel k such that γ_k is a metric on $\mathscr{P}$. As mentioned in Section 1, the metric property of γ_k is crucial in many statistical inference applications like hypothesis testing. Therefore, in practice, it is important to use characteristic kernels. However, in this section, we show that characteristic kernels, while guaranteeing γ_k to be a metric on $\mathscr{P}$, may nonetheless have difficulty in distinguishing certain distributions on the basis of finite samples. More specifically, in Theorem 19 we show that for a given kernel k and for any $\varepsilon > 0$, there exist $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P},\mathbb{Q}) < \varepsilon$. Before proving the result, we motivate it through the following example.

Example 6 *Let $\mathbb{P}$ be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}$ with the Radon-Nikodym derivative defined as*

$$p(x) = q(x) + \alpha q(x)\sin(\nu\pi x), \tag{22}$$

where q is the Radon-Nikodym derivative of $\mathbb{Q}$ w.r.t. the Lebesgue measure satisfying $q(x) = q(-x), \forall x$ and $\alpha \in [-1,1]\backslash\{0\}$, $\nu \in \mathbb{R}\backslash\{0\}$. It is obvious that $\mathbb{P} \neq \mathbb{Q}$. The characteristic function of $\mathbb{P}$ is given as

$$\phi_{\mathbb{P}}(\omega) = \phi_{\mathbb{Q}}(\omega) - \frac{i\alpha}{2}\left[\phi_{\mathbb{Q}}(\omega-\nu\pi) - \phi_{\mathbb{Q}}(\omega+\nu\pi)\right], \omega \in \mathbb{R},$$

where $\phi_{\mathbb{Q}}$ is the characteristic function associated with $\mathbb{Q}$. Note that with increasing $|\nu|$, p has higher frequency components in its Fourier spectrum, as shown in Figure 4. In Figure 4, (a-c) show the plots of p when $q = \mathcal{U}[-1,1]$ (uniform distribution) and (a′-c′) show the plots of p when $q = \mathcal{N}(0,2)$ (zero mean normal distribution with variance 2) for $\nu = 0, 2$ and 7.5 with $\alpha = \frac{1}{2}$.

Consider the B_1-spline kernel on $\mathbb{R}$ given by $k(x,y) = \psi(x-y)$ where

$$\psi(x) = \begin{cases} 1-|x|, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{23}$$

with its Fourier transform given by

$$\widehat{\psi}(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{\sin^2\frac{\omega}{2}}{\omega^2}.$$

Since ψ is characteristic to $\mathscr{P}$, $\gamma_k(\mathbb{P},\mathbb{Q}) > 0$ (see Theorem 9). However, it would be of interest to study the behavior of $\gamma_k(\mathbb{P},\mathbb{Q})$ as a function of ν. We study the behavior of $\gamma_k^2(\mathbb{P},\mathbb{Q})$ through its unbiased, consistent estimator,[7] $\gamma_{k,u}^2(m,m)$ as considered by Gretton et al. (2007b, Lemma 7).

Figure 5(a) shows the behavior of $\gamma_{k,u}^2(m,m)$ as a function of ν for $q = \mathcal{U}[-1,1]$ and $q = \mathcal{N}(0,2)$ using the B_1-spline kernel in (23). Since the Gaussian kernel, $k(x,y) = e^{-(x-y)^2}$ is also a characteristic kernel, its effect on the behavior of $\gamma_{k,u}^2(m,m)$ is shown in Figure 5(b) in comparison to that of the B_1-spline kernel.

In Figure 5, we observe two circumstances under which γ_k^2 may be small. First, $\gamma_{k,u}^2(m,m)$ decays with increasing $|\nu|$, and can be made as small as desired by choosing a sufficiently large $|\nu|$. Second,

7. Let $\{X_j\}_{j=1}^m$ and $\{Y_j\}_{j=1}^m$ be random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ respectively. An unbiased *empirical estimate* of $\gamma_k^2(\mathbb{P},\mathbb{Q})$, denoted as $\gamma_{k,u}^2(m,m)$ is given by $\gamma_{k,u}^2(m,m) = \frac{1}{m(m-1)}\sum_{l\neq j}^m h(Z_l,Z_j)$, which is a one-sample U-statistic with $h(Z_l,Z_j) := k(X_l,X_j) + k(Y_l,Y_j) - k(X_l,Y_j) - k(X_j,Y_l)$, where $Z_1,\ldots,Z_m$ are m i.i.d. random variables with $Z_j := (X_j,Y_j)$. See Gretton et al. (2007b, Lemma 7) for details.

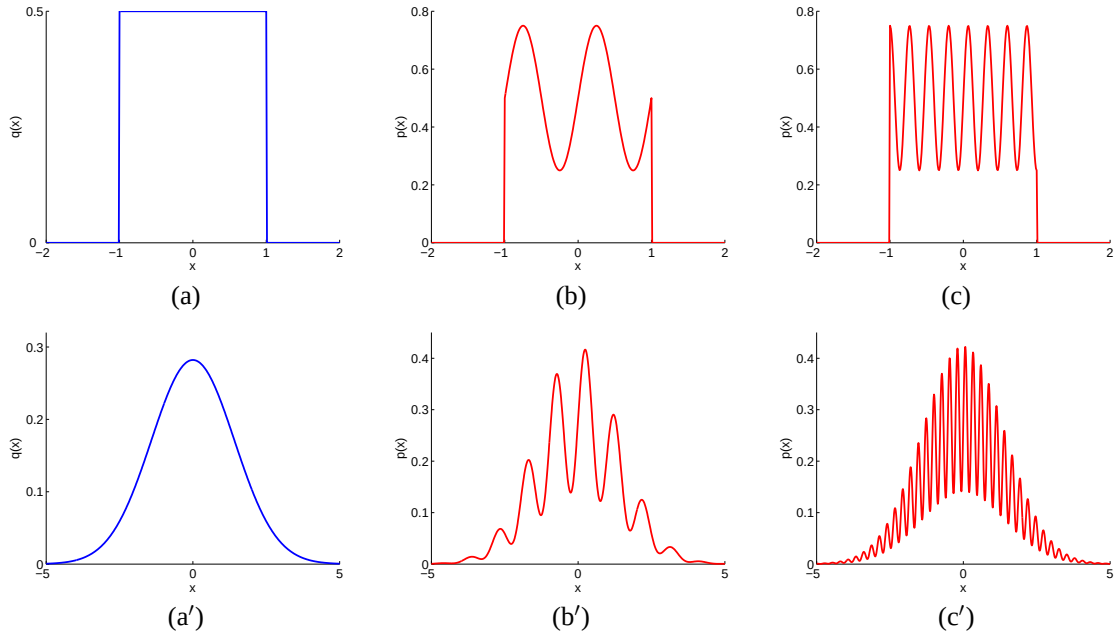

Figure 4: (a) $q = \mathcal{U}[-1,1]$, (a′) $q = \mathcal{N}(0,2)$. (b-c) and (b′-c′) denote $p(x)$ computed as $p(x) = q(x) + \frac{1}{2}q(x)\sin(\nu\pi x)$ with $q = \mathcal{U}[-1,1]$ and $q = \mathcal{N}(0,2)$ respectively. ν is chosen to be 2 in (b,b′) and 7.5 in (c,c′). See Example 6 for details.

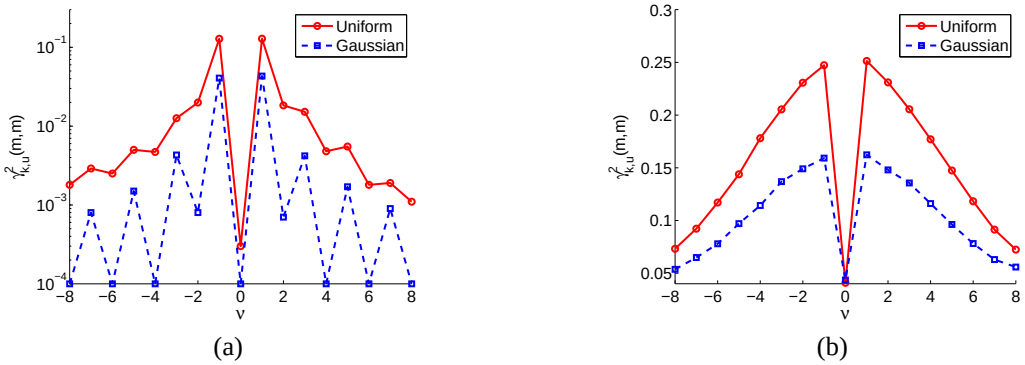

Figure 5: Behavior of the empirical estimate of $\gamma_k^2(\mathbb{P},\mathbb{Q})$ w.r.t. ν for (a) the B_1-spline kernel and (b) the Gaussian kernel. $\mathbb{P}$ is constructed from $\mathbb{Q}$ as defined in (22). "Uniform" corresponds to $\mathbb{Q} = \mathcal{U}[-1,1]$ and "Gaussian" corresponds to $\mathbb{Q} = \mathcal{N}(0,2)$. $m = 1000$ samples are generated from $\mathbb{P}$ and $\mathbb{Q}$ to estimate $\gamma_k^2(\mathbb{P},\mathbb{Q})$ through $\gamma_{k,u}^2(m,m)$. This is repeated 100 times and the average $\gamma_{k,u}^2(m,m)$ is plotted in both figures. Since the quantity of interest is the average behavior of $\gamma_{k,u}^2(m,m)$, we omit the error bars. See Example 6 for details.

in Figure 5(a), $\gamma_{k,u}^2(m,m)$ has troughs at $\nu = \frac{\omega_0}{\pi}$ where $\omega_0 = \{\omega : \widehat{\psi}(\omega) = 0\}$. Since $\gamma_{k,u}^2(m,m)$ is a consistent estimate of $\gamma_k^2(\mathbb{P},\mathbb{Q})$, one would expect similar behavior from $\gamma_k^2(\mathbb{P},\mathbb{Q})$. This means that, although the B_1-spline kernel is characteristic to $\mathscr{P}$, in practice, it becomes harder to distinguish

between $\mathbb{P}$ and $\mathbb{Q}$ with finite samples, when $\mathbb{P}$ is constructed as in (22) with $\nu = \frac{\omega_0}{\pi}$. In fact, one can observe from a straightforward spectral argument that the troughs in $\gamma_k^2(\mathbb{P},\mathbb{Q})$ can be made arbitrarily deep by widening q, when q is Gaussian.

For characteristic kernels, although $\gamma_k(\mathbb{P},\mathbb{Q}) > 0$ when $\mathbb{P} \neq \mathbb{Q}$, Example 6 demonstrates that one can construct distributions such that $\gamma_{k,u}^2(m,m)$ is indistinguishable from zero with high probability, for a given sample size m. Below, in Theorem 19, we explicitly construct $\mathbb{P} \neq \mathbb{Q}$ such that $|\mathbb{P}\varphi_l - \mathbb{Q}\varphi_l|$ is large for some large l, but $\gamma_k(\mathbb{P},\mathbb{Q})$ is arbitrarily small, making it hard to detect a non-zero value of $\gamma_k(\mathbb{P},\mathbb{Q})$ based on finite samples. Here, $\varphi_l \in L^2(M)$ represents the bounded orthonormal eigenfunctions of a positive definite integral operator associated with k. Based on this theorem, for example, in Example 6, the decay mode of γ_k for large $|\nu|$ can be investigated.

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

$$\gamma_k(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{H}} \frac{|\mathbb{P}f - \mathbb{Q}f|}{\|f\|_{\mathcal{H}}}.$$

When $\mathbb{P} \neq \mathbb{Q}$, $|\mathbb{P}f - \mathbb{Q}f|$ can be large for some $f \in \mathcal{H}$. However, $\gamma_k(\mathbb{P},\mathbb{Q})$ can be made small by selecting $\mathbb{P}$ such that the maximization of $\frac{|\mathbb{P}f-\mathbb{Q}f|}{\|f\|_{\mathcal{H}}}$ over $\mathcal{H}$ requires an f with large $\|f\|_{\mathcal{H}}$. More specifically, higher order eigenfunctions of the kernel (φ_l for large l) have large RKHS norms, so, if they are prominent in $\mathbb{P}$ and $\mathbb{Q}$ (i.e., highly non-smooth distributions), one can expect $\gamma_k(\mathbb{P},\mathbb{Q})$ to be small even when there exists an l for which $|\mathbb{P}\varphi_l - \mathbb{Q}\varphi_l|$ is large. To this end, we need the following lemma, which we quote from Gretton et al. (2005b, Lemma 4).

Lemma 18 (Gretton et al., 2005b) *Let $\mathcal{F}$ be the unit ball in an RKHS $(\mathcal{H},k)$ defined on a compact topological space, M, with k being measurable. Let $\varphi_l \in L^2(M,\mu)$ be absolutely bounded orthonormal eigenfunctions and λ_l be the corresponding eigenvalues (arranged in decreasing order for increasing l) of a positive definite integral operator associated with k and a σ-finite measure, μ. Assume λ_l^{-1} increases super-linearly with l. Then, for $f \in \mathcal{F}$ where $f(x) = \sum_{j=1}^{\infty} \widetilde{f}_j \varphi_j(x)$, $\widetilde{f}_j := \langle f, \varphi_j\rangle_{L^2(M,\mu)}$, we have $\sum_{j=1}^{\infty} |\widetilde{f}_j| < \infty$ and for every $\varepsilon > 0$, $\exists\, l_0 \in \mathbb{N}$ such that $|\widetilde{f}_l| < \varepsilon$ if $l > l_0$.*

Theorem 19 ($\mathbb{P} \neq \mathbb{Q}$ can have arbitrarily small γ_k) *Suppose the conditions in Lemma 18 hold. Then there exist probability measures $\mathbb{P} \neq \mathbb{Q}$ defined on M such that $\gamma_k(\mathbb{P},\mathbb{Q}) < \varepsilon$ for any arbitrarily small $\varepsilon > 0$.*

Proof Suppose q be the Radon-Nikodym derivative associated with $\mathbb{Q}$ w.r.t. the σ-finite measure, μ (see Lemma 18). Let us construct $p(x) = q(x) + \alpha_l e(x) + \tau\varphi_l(x)$ where $e(x) = \mathbb{1}_M(x)$. For $\mathbb{P}$ to be a probability measure, the following conditions need to be satisfied:

$$\int_M [\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = 0, \qquad \min_{x\in M} [q(x) + \alpha_l e(x) + \tau\varphi_l(x)] \geq 0. \tag{24}$$

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

$$\begin{aligned} \mathbb{P}f - \mathbb{Q}f &= \int_M f(x)\,[\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) \\ &= \int_M \left[\alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j \varphi_j(x) + \tau\varphi_l(x)\right] \left[\sum_{t=1}^{\infty} \widetilde{f}_t \varphi_t(x)\right] d\mu(x) = \alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j \widetilde{f}_j + \tau \widetilde{f}_l, \end{aligned} \tag{25}$$

where we used the fact that $\langle \varphi_j, \varphi_t \rangle_{L^2(M,\mu)} = \delta_{jt}$ (here, δ is used in the Kronecker sense). Rewriting (24) and substituting for $e(x)$ gives

$$\int_M [\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = \int_M e(x)[\alpha_l e(x) + \tau\varphi_l(x)]\, d\mu(x) = \alpha_l \sum_{j=1}^{\infty} \widetilde{e}_j^2 + \tau\widetilde{e}_l = 0,$$

which implies

$$\alpha_l = -\frac{\tau\widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2}. \tag{26}$$

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

$$\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t = \tau\delta_{tl} - \tau\frac{\widetilde{e}_t\widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2} = \tau\delta_{tl} - \tau\rho_{tl},$$

where $\rho_{tl} := \frac{\widetilde{e}_t\widetilde{e}_l}{\sum_{j=1}^{\infty} \widetilde{e}_j^2}$. By Lemma 18, $\sum_{l=1}^{\infty} |\widetilde{e}_l| < \infty \Rightarrow \sum_{j=1}^{\infty} \widetilde{e}_j^2 < \infty$, and choosing large enough l gives $|\rho_{tl}| < \eta, \forall t$, for any arbitrary $\eta > 0$. Therefore, $|\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t| > \tau - \eta$ for $t = l$ and $|\mathbb{P}\varphi_t - \mathbb{Q}\varphi_t| < \eta$ for $t \neq l$, which means $\mathbb{P} \neq \mathbb{Q}$. In the following, we prove that $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be arbitrarily small, though non-zero.

Recall that $\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathbb{P}f - \mathbb{Q}f|$. Substituting (26) in (25) and replacing $|\mathbb{P}f - \mathbb{Q}f|$ by (25) in $\gamma_k(\mathbb{P}, \mathbb{Q})$, we have

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \sup_{\{\widetilde{f}_j\}_{j=1}^{\infty}} \left\{ \tau \sum_{j=1}^{\infty} \nu_{jl} \widetilde{f}_j \;:\; \sum_{j=1}^{\infty} \frac{\widetilde{f}_j^2}{\lambda_j} \leq 1 \right\}, \tag{27}$$

where we used the definition of RKHS norm as $\|f\|_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\widetilde{f}_j^2}{\lambda_j}$ and $\nu_{jl} := \delta_{jl} - \rho_{jl}$. (27) is a convex quadratically constrained quadratic program in $\{\widetilde{f}_j\}_{j=1}^{\infty}$. Solving the Lagrangian yields $\widetilde{f}_j = \frac{\nu_{jl}\lambda_j}{\sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j}}$. Therefore,

$$\gamma_k(\mathbb{P}, \mathbb{Q}) = \tau \sqrt{\sum_{j=1}^{\infty} \nu_{jl}^2 \lambda_j} = \tau \sqrt{\lambda_l - 2\rho_{ll}\lambda_l + \sum_{j=1}^{\infty} \rho_{jl}^2 \lambda_j} \xrightarrow{l \to \infty} 0,$$

because *(i)* by choosing sufficiently large l, $|\rho_{jl}| < \varepsilon, \forall\, j$, for any arbitrary $\varepsilon > 0$, and *(ii)* $\lambda_l \to 0$ as $l \to \infty$ (Schölkopf and Smola, 2002, Theorem 2.10). Therefore, we have constructed $\mathbb{P} \neq \mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q}) < \varepsilon$ for any arbitrarily small $\varepsilon > 0$. ■

5. Metrization of the Weak Topology

So far, we have shown that a characteristic kernel k induces a metric γ_k on $\mathscr{P}$. As motivated in Section 1.1.3, an important question to consider that is useful both in theory and practice would be: "How strong or weak is γ_k related to other metrics on $\mathscr{P}$?" This question is addressed in Theorem 21, where we compare γ_k to other metrics on $\mathscr{P}$ like the Dudley metric (β), Wasserstein distance (W), total variation distance (TV), and show that γ_k is weaker than all these metrics (see Footnote 3 for the definition of "strong" and "weak" metrics). Since γ_k is weaker than the Dudley metric, which

is known to induce a topology on $\mathscr{P}$ that coincides with the standard topology on $\mathscr{P}$, called the weak-* (weak-star) topology (usually called the weak topology in probability theory), the next question we are interested in is to understand the topology that is being induced by γ_k. In particular, we are interested in determining the conditions on k for which the topology induced by γ_k coincides with the weak topology on $\mathscr{P}$. This is answered in Theorems 23 and 24, where Theorem 23 deals with compact M and Theorem 24 provides a sufficient condition on k when $M = \mathbb{R}^d$. The proofs of all these results are provided in Section 5.1. Before we motivate the need for this study and its implications, we present some preliminaries.

The *weak topology* on $\mathscr{P}$ is the weakest topology such that the map $\mathbb{P} \mapsto \int_M f\, d\mathbb{P}$ is continuous for all $f \in C_b(M)$. For a metric space (M,ρ), a sequence $\mathbb{P}_n$ of probability measures is said to *converge weakly* to $\mathbb{P}$, written as $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$, if and only if $\int_M f\, d\mathbb{P}_n \to \int_M f\, d\mathbb{P}$ for every $f \in C_b(M)$. A metric γ on $\mathscr{P}$ is said to *metrize* the weak topology if the topology induced by γ coincides with the weak topology, which is defined as follows: if, for $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathscr{P}$, $(\mathbb{P}_n \xrightarrow{w} \mathbb{P} \Leftrightarrow \gamma(\mathbb{P}_n, \mathbb{P}) \xrightarrow{n\to\infty} 0)$ holds, then the topology induced by γ coincides with the weak topology.

In the following, we collect well-known results on the relation between various metrics on $\mathscr{P}$, which will be helpful in understanding the behavior of these metrics, both with respect to each other and to ours. Let (M,ρ) be a separable metric space. The *Prohorov metric* on (M,ρ), defined as

$$\varsigma(\mathbb{P},\mathbb{Q}) := \inf\{\varepsilon > 0 : \mathbb{P}(A) \le \mathbb{Q}(A^\varepsilon) + \varepsilon, \forall \text{Borel sets } A\},$$

metrizes the weak topology on $\mathscr{P}$ (Dudley, 2002, Theorem 11.3.3), where $\mathbb{P},\mathbb{Q} \in \mathscr{P}$ and $A^\varepsilon := \{y \in M : \rho(x,y) < \varepsilon \text{ for some } x \in A\}$. Since the Dudley metric is related to the Prohorov metric as

$$\frac{1}{2}\beta(\mathbb{P},\mathbb{Q}) \le \varsigma(\mathbb{P},\mathbb{Q}) \le 2\sqrt{\beta(\mathbb{P},\mathbb{Q})}, \tag{28}$$

it also metrizes the weak topology on $\mathscr{P}$ (Dudley, 2002, Theorem 11.3.3). The Wasserstein distance and total variation distance are related to the Prohorov metric as

$$\varsigma^2(\mathbb{P},\mathbb{Q}) \le W(\mathbb{P},\mathbb{Q}) \le (\operatorname{diam}(M)+1)\varsigma(\mathbb{P},\mathbb{Q}), \tag{29}$$

and

$$\varsigma(\mathbb{P},\mathbb{Q}) \le TV(\mathbb{P},\mathbb{Q}),$$

where $\operatorname{diam}(M) := \sup\{\rho(x,y) : x,y \in M\}$ (Gibbs and Su, 2002, Theorem 2). This means W and TV are stronger than ς, while W and ς are equivalent (i.e., induce the same topology) when M is bounded. By Theorem 4 in Gibbs and Su (2002), TV and W are related as

$$W(\mathbb{P},\mathbb{Q}) \le \operatorname{diam}(M) TV(\mathbb{P},\mathbb{Q}),$$

which means W and TV are comparable if M is bounded. See Shorack (2000, Chapter 19, Theorem 2.4) and Gibbs and Su (2002) for further detail on the relationship between various metrics on $\mathscr{P}$.

Let us now consider a sequence of of probability measures on $\mathbb{R}$, $\mathbb{P}_n := \left(1-\frac{1}{n}\right)\delta_0 + \frac{1}{n}\delta_n$ and let $\mathbb{P} := \delta_0$. It can be shown that $\beta(\mathbb{P}_n,\mathbb{P}) \to 0$ as $n \to \infty$ which means $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$, while $W(\mathbb{P}_n,\mathbb{P}) = 1$ and $TV(\mathbb{P}_n,\mathbb{P}) = 1$ for all n. $\gamma_k(\mathbb{P}_n,\mathbb{P})$ can be computed as

$$\gamma_k^2(\mathbb{P}_n,\mathbb{P}) = \frac{1}{n^2}\iint_{\mathbb{R}} k(x,y)\, d(\delta_0-\delta_n)(x)\, d(\delta_0-\delta_n)(y) = \frac{k(0,0)+k(n,n)-2k(0,n)}{n^2}.$$

If k is, for example, a Gaussian, Laplacian or inverse multiquadratic kernel, then $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$. This example shows that γ_k is weaker than W and TV. It also shows that, for certain choices of k, γ_k behaves similarly to β, which leads to several questions: Does γ_k metrize the weak topology on $\mathscr{P}$? What is the general behavior of γ_k compared to other metrics? In other words, depending on k, how weak or strong is γ_k compared to other metrics on $\mathscr{P}$? Understanding the answer to these questions is important both in theory and practice. If k is such that γ_k metrizes the weak topology on $\mathscr{P}$, then it can be used as a theoretical tool in probability theory, similar to the Prohorov and Dudley metrics. On the other hand, the answer to these questions is critical in applications as it will have a bearing on the choice of kernels to be used. In applications like density estimation, one would need a strong metric to ascertain that the density estimate is a good representation of the true underlying density. For this reason, the total variation distance, Hellinger distance or Kullback-Leibler distance are generally used. However, it is not always possible to show the convergence of a density estimate to the true underlying density using a stronger metric and so, in such cases, one would need a weak metric to analyze the quality of estimate. Therefore, studying the relation between γ_k and these other metrics will provide an understanding of the choice of kernels to be used, depending on the application.

With the above motivation, we first compare γ_k to β, W and TV. Since β is equivalent to ς, we do not compare γ_k to ς. Before we provide the main result in Theorem 21 that compares γ_k to other metrics, we present an upper bound on γ_k in terms of the coupling formulation (Dudley, 2002, Section 11.8), which is not only useful in deriving the main result but also interesting in its own right.

Proposition 20 (Coupling bound) *Let k be measurable and bounded on M. Then, for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$,*

$$\gamma_k(\mathbb{P}, \mathbb{Q}) \leq \inf_{\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})} \iint_M \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \, d\mu(x, y), \tag{30}$$

where $\mathcal{L}(\mathbb{P}, \mathbb{Q})$ represents the set of all laws on $M \times M$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.

Proof For any $\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})$, we have

$$\begin{aligned} \left| \int_M f \, d(\mathbb{P} - \mathbb{Q}) \right| &= \left| \iint_M (f(x) - f(y)) \, d\mu(x, y) \right| \leq \iint_M |f(x) - f(y)| \, d\mu(x, y) \\ &= \iint_M |\langle f, k(\cdot, x) - k(\cdot, y) \rangle_{\mathcal{H}}| \, d\mu(x, y) \leq \|f\|_{\mathcal{H}} \iint_M \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \, d\mu(x, y). \end{aligned} \tag{31}$$

Taking the supremum over $f \in \mathcal{F}_k$ and the infimum over $\mu \in \mathcal{L}(\mathbb{P}, \mathbb{Q})$ in (31), where $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, gives the result in (30). ■

We now present the main result that compares γ_k to β, W and TV.

Theorem 21 (Comparison of γ_k to β, W and TV) *Assume $\sup_{x \in M} k(x, x) \leq C < \infty$, where k is measurable on M. Let*

$$\widetilde{\rho}(x, y) = \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}}. \tag{32}$$

Then, for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$,

(i) $\gamma_k(\mathbb{P}, \mathbb{Q}) \leq W(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\gamma_k^2(\mathbb{P}, \mathbb{Q}) + 4C}$ *if $(M, \widetilde{\rho})$ is separable.*

(ii) $\frac{\gamma_k(\mathbb{P},\mathbb{Q})}{(1+\sqrt{C})} \leq \beta(\mathbb{P},\mathbb{Q}) \leq 2(\gamma_k^2(\mathbb{P},\mathbb{Q})+4C)^{\frac{1}{3}}$ *if* $(M,\widetilde{\rho})$ *is separable.*

(iii) $\gamma_k(\mathbb{P},\mathbb{Q}) \leq \sqrt{C}\,TV(\mathbb{P},\mathbb{Q})$.

The proof is provided in Section 5.1. Below are some remarks on Theorem 21.

Remark 22 *(a) First, note that, since k is bounded,* $(M,\widetilde{\rho})$ *is a bounded metric space. In addition, the metric,* $\widetilde{\rho}$*, which depends on the kernel as in (32), is a Hilbertian metric*[8] *(Berg et al., 1984, Chapter 3, Section 3) on M. A popular example of such a metric is* $\widetilde{\rho}(x,y) = \|x-y\|_2$*, which can be obtained by choosing M to be a compact subset of* $\mathbb{R}^d$ *and* $k(x,y) = x^T y$*.*

(b) Theorem 21 shows that γ_k *is weaker than* β*, W and TV for the assumptions being made on k and* $\widetilde{\rho}$*. Note that the result holds irrespective of whether or not the kernel is characteristic, as we have not assumed anything about the kernel except it being measurable and bounded. Also, it is important to remember that the result holds when* $\widetilde{\rho}$ *is Hilbertian, as mentioned in (32) (see Remark 22(d)).*

(c) Apart from showing that γ_k *is weaker than* β*, W and TV, the result in Theorem 21 can be used to bound these metrics in terms of* γ_k*. For* β*, which is primarily of theoretical interest, we do not know a closed form expression, and likewise a closed form expression for W is known only for* $\mathbb{R}$ *(Vallander, 1973).*[9] *Since* γ_k *is easy to compute (see (9) and (10)), bounds on W can be obtained from Theorem 21 in terms of* γ_k*. A closed form expression for TV is available if* $\mathbb{P}$ *and* $\mathbb{Q}$ *have Radon-Nikodym derivatives w.r.t. a* σ*-finite measure. However, from Theorem 21, a simple lower bound can be obtained on TV in terms of* γ_k *for any* $\mathbb{P},\mathbb{Q} \in \mathscr{P}$*.*

(d) In Theorem 21, the kernel is fixed and $\widetilde{\rho}$ *is defined as in (32), which is a Hilbertian metric. On the other hand, suppose a Hilbertian metric* $\widetilde{\rho}$ *is given. Then the associated kernel k can be obtained from* $\widetilde{\rho}$ *(Berg et al., 1984, Chapter 3, Lemma 2.1) as*

$$k(x,y) = \frac{1}{2}[\widetilde{\rho}^2(x,x_0)+\widetilde{\rho}^2(y,x_0)-\widetilde{\rho}^2(x,y)],\ x,y,x_0 \in M, \tag{33}$$

which can then be used to compute γ_k*.*

The discussion so far has been devoted to relating γ_k to β, W and TV to understand the strength or weakness of γ_k w.r.t. these metrics. In a next step, we address the second question of when γ_k metrizes the weak topology on $\mathscr{P}$. This question would have been answered had the result in Theorem 21 shown that under some conditions on k, γ_k is equivalent to β. Since Theorem 21 does not help in this regard, we approach the problem differently. In the following, we provide two results related to the question. The first result states that when (M,ρ) is compact, γ_k induced by universal kernels metrizes the weak topology. In the second result, we relax the assumption of compactness but restrict ourselves to $M = \mathbb{R}^d$ and provide a sufficient condition on k such that γ_k metrizes the weak topology on $\mathscr{P}$. The proofs of both theorems are provided in Section 5.1.

Theorem 23 (Weak convergence-I) *Let* (M,ρ) *be a compact metric space. If k is universal, then* γ_k *metrizes the weak topology on* $\mathscr{P}$*.*

8. A metric ρ on M is said to be *Hilbertian* if there exists a Hilbert space, H and a mapping Φ such that $\rho(x,y) = \|\Phi(x)-\Phi(y)\|_H, \forall x,y \in M$. In our case, $H = \mathcal{H}$ and $\Phi : M \to \mathcal{H}, x \mapsto k(\cdot,x)$.
9. The explicit form for the Wasserstein distance is known for $(M,\rho(x,y)) = (\mathbb{R},|x-y|)$, and is $W(\mathbb{P},\mathbb{Q}) = \int_{\mathbb{R}} |F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)|\,dx$, where $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty,x])$. It is easy to show that this explicit form can be extended to $(\mathbb{R}^d,\|\cdot\|_1)$.

From Theorem 23, it is clear that γ_k is equivalent to ς, β and W (see (28) and (29)) when M is compact and k is universal.

Theorem 24 (Weak convergence-II) *Let $M = \mathbb{R}^d$ and $k(x,y) = \psi(x-y)$, where $\psi \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is a real-valued bounded strictly positive definite function. If there exists an $l \in \mathbb{N}$ such that*

$$\int_{\mathbb{R}^d} \frac{1}{\widehat{\psi}(\omega)(1+\|\omega\|_2)^l}\, d\omega < \infty, \tag{34}$$

then γ_k metrizes the weak topology on $\mathscr{P}$.

The entire Matérn class of kernels in (18) satisfies the conditions of Theorem 24 and, therefore, the corresponding γ_k metrizes the weak topology on $\mathscr{P}$. Note that Gaussian kernels on $\mathbb{R}^d$ do not satisfy the condition in Theorem 24. The characterization of k for general non-compact domains M (not necessarily $\mathbb{R}^d$), such that γ_k metrizes the weak topology on $\mathscr{P}$, still remains an open problem.

5.1 Proofs

We now present the proofs of Theorems 21, 23 and 24.

Proof (Theorem 21) *(i)* When (M,ρ) is separable, $W(\mathbb{P},\mathbb{Q})$ has a coupling formulation (Dudley, 2002, p. 420), given as

$$W(\mathbb{P},\mathbb{Q}) = \inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \iint_M \rho(x,y)\, d\mu(x,y), \tag{35}$$

where $\mathbb{P},\mathbb{Q} \in \{\mathbb{P} \in \mathscr{P} : \int_M \rho(x,y)\, d\mathbb{P}(y) < \infty, \forall x \in M\}$. In our case $\rho(x,y) = \|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}}$. In addition, (M,ρ) is bounded, which means (35) holds for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}$. The lower bound therefore follows from (30). The upper bound can be obtained as follows. Consider $W(\mathbb{P},\mathbb{Q}) = \inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \iint_M \|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} d\mu(x,y)$, which can be bounded as

$$\begin{aligned}
W(\mathbb{P},\mathbb{Q}) &\leq \iint_M \|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} d\mathbb{P}(x)\, d\mathbb{Q}(y) \\
&\overset{(a)}{\leq} \left[\iint_M \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}} d\mathbb{P}(x)\, d\mathbb{Q}(y)\right]^{\frac{1}{2}} \\
&\leq \left[\int_M k(x,x)\, d(\mathbb{P}+\mathbb{Q})(x) - 2\iint_M k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y)\right]^{\frac{1}{2}} \\
&\leq \left[\gamma_k^2(\mathbb{P},\mathbb{Q}) + \iint_M (k(x,x) - k(x,y))\, d(\mathbb{P}\otimes\mathbb{P} + \mathbb{Q}\otimes\mathbb{Q})(x,y)\right]^{\frac{1}{2}} \\
&\leq \sqrt{\gamma_k^2(\mathbb{P},\mathbb{Q}) + 4C},
\end{aligned} \tag{36}$$

where we have used Jensen's inequality (Folland, 1999, p. 109) in (a).

(ii) Let $\mathcal{F} := \{f : \|f\|_{\mathcal{H}} < \infty\}$ and $\mathcal{G} := \{f : \|f\|_{BL} < \infty\}$. For $f \in \mathcal{F}$, we have

$$\begin{aligned}
\|f\|_{BL} &= \sup_{x \neq y} \frac{|f(x) - f(y)|}{\rho(x,y)} + \sup_{x \in M} |f(x)| = \sup_{x \neq y} \frac{|\langle f, k(\cdot,x) - k(\cdot,y)\rangle_{\mathcal{H}}|}{\|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}}} + \sup_{x \in M} |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}| \\
&\leq (1+\sqrt{C})\|f\|_{\mathcal{H}} < \infty,
\end{aligned}$$

which implies $f \in \mathcal{G}$ and, therefore, $\mathcal{F} \subset \mathcal{G}$. For any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$,

$$\begin{aligned}\gamma_k(\mathbb{P},\mathbb{Q}) &= \sup\{|\mathbb{P}f - \mathbb{Q}f| : f \in \mathcal{F}_k\} \\ &\leq \sup\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_{BL} \leq (1+\sqrt{C}),\, f \in \mathcal{F}\} \\ &\leq \sup\{|\mathbb{P}f - \mathbb{Q}f| : \|f\|_{BL} \leq (1+\sqrt{C}),\, f \in \mathcal{G}\} \\ &= (1+\sqrt{C})\beta(\mathbb{P},\mathbb{Q}).\end{aligned}$$

The upper bound is obtained as follows. For any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$, by Markov's inequality (Folland, 1999, Theorem 6.17), for all $\varepsilon > 0$, we have

$$\varepsilon^2 \mu(\|k(\cdot,X) - k(\cdot,Y)\|_{\mathcal{H}} > \varepsilon) \leq \iint_M \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}} d\mu(x,y),$$

where X and Y are distributed as $\mathbb{P}$ and $\mathbb{Q}$ respectively. Choose ε such that $\varepsilon^3 = \iint_M \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}} d\mu(x,y)$, such that $\mu(\|k(\cdot,X) - k(\cdot,Y)\|_{\mathcal{H}} > \varepsilon) \leq \varepsilon$. From the proof of Theorem 11.3.5 in Dudley (2002), when (M,ρ) is separable, we have

$$\mu(\rho(X,Y) \geq \varepsilon) < \varepsilon \Rightarrow \varsigma(\mathbb{P},\mathbb{Q}) \leq \varepsilon,$$

which implies that

$$\begin{aligned}\varsigma(\mathbb{P},\mathbb{Q}) &\leq \left(\inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \iint_M \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}} d\mu(x,y)\right)^{\frac{1}{3}} \\ &\leq \left(\iint_M \|k(\cdot,x) - k(\cdot,y)\|^2_{\mathcal{H}} d\mathbb{P}(x)\, d\mathbb{Q}(y)\right)^{\frac{1}{3}} \overset{(b)}{=} \left(\gamma_k^2(\mathbb{P},\mathbb{Q}) + 4C\right)^{\frac{1}{3}},\end{aligned}$$

where (b) follows from (36). The result follows from (28).

(iii) The proof of this result was presented in Sriperumbudur et al. (2009b) and is provided here for completeness. To prove the result, we use (30) and the coupling formulation for TV (Lindvall, 1992, p. 19), given as

$$\frac{1}{2} TV(\mathbb{P},\mathbb{Q}) = \inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \mu(X \neq Y),$$

where $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $M \times M$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Here, X and Y are distributed as $\mathbb{P}$ and $\mathbb{Q}$ respectively. Consider

$$\|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} \leq \mathbb{1}_{\{x \neq y\}} \|k(\cdot,x) - k(\cdot,y)\|_{\mathcal{H}} \leq 2\sqrt{C}\mathbb{1}_{\{x \neq y\}}. \tag{37}$$

Taking expectations w.r.t. μ and the infimum over $\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})$ on both sides of (37) gives the desired result, which follows from (30). ■

Proof (Theorem 23) We need to show that for measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathscr{P}$, $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ if and only if $\gamma_k(\mathbb{P}_n,\mathbb{P}) \to 0$ as $n \to \infty$. One direction is trivial as $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$ implies $\gamma_k(\mathbb{P}_n,\mathbb{P}) \to 0$ as $n \to \infty$. We prove the other direction as follows. Since k is universal, $\mathcal{H}$ is dense in $C_b(M)$, the space of bounded continuous functions, w.r.t. the uniform norm, that is, for any $f \in C_b(M)$ and every $\varepsilon > 0$, there exists a $g \in \mathcal{H}$ such that $\|f - g\|_\infty \leq \varepsilon$. Therefore,

$$\begin{aligned}|\mathbb{P}_n f - \mathbb{P}f| &= |\mathbb{P}_n(f-g) + \mathbb{P}(g-f) + (\mathbb{P}_n g - \mathbb{P}g)| \\ &\leq \mathbb{P}_n|f-g| + \mathbb{P}|f-g| + |\mathbb{P}_n g - \mathbb{P}g| \\ &\leq 2\varepsilon + |\mathbb{P}_n g - \mathbb{P}g| \leq 2\varepsilon + \|g\|_{\mathcal{H}} \gamma_k(\mathbb{P}_n,\mathbb{P}).\end{aligned}$$

Since $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$ and ε is arbitrary, $|\mathbb{P}_n f - \mathbb{P} f| \to 0$ for any $f \in C_b(M)$. ■

Proof (Theorem 24) As mentioned in the proof of Theorem 23, one direction of the proof is straightforward: $\mathbb{P}_n \xrightarrow{w} \mathbb{P} \Rightarrow \gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$. Let us consider the other direction. Since $\psi \in C_0(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ is a strictly positive definite function, any $f \in \mathcal{H}$ satisfies (Wendland, 2005, Theorem 10.12)

$$\int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{\widehat{\psi}(\omega)} d\omega < \infty.$$

Assume that

$$\sup_{\omega \in \mathbb{R}^d} (1 + \|\omega\|_2)^l |\widehat{f}(\omega)|^2 < \infty,$$

for any $l \in \mathbb{N}$, which means $f \in \mathcal{S}_d$. Let (34) be satisfied for some $l = l_0$. Then,

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{\widehat{\psi}(\omega)} d\omega &= \int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2 (1 + \|\omega\|_2)^{l_0}}{\widehat{\psi}(\omega)(1 + \|\omega\|_2)^{l_0}} d\omega \\ &\leq \sup_{\omega \in \mathbb{R}^d} (1 + \|\omega\|_2)^{l_0} |\widehat{f}(\omega)|^2 \int_{\mathbb{R}^d} \frac{1}{\widehat{\psi}(\omega)(1 + \|\omega\|_2)^{l_0}} d\omega < \infty, \end{aligned}$$

which means $f \in \mathcal{H}$, that is, if $f \in \mathcal{S}_d$, then $f \in \mathcal{H}$, which implies $\mathcal{S}_d \subset \mathcal{H}$. Note that $\mathcal{S}(\mathbb{R}^d)$ is dense in $C_0(\mathbb{R}^d)$. Since $\psi \in C_0(\mathbb{R}^d)$, we have $\mathcal{H} \subset C_0(\mathbb{R}^d)$ (see the proof of Theorem 4.61 in Steinwart and Christmann, 2008) and, therefore, $\mathcal{H}$ is dense in $C_0(\mathbb{R}^d)$ w.r.t. the uniform norm. Suppose $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots \in \mathcal{P}$. Using a similar analysis as in the proof of Theorem 23, it can be shown that for any $f \in C_0(\mathbb{R}^d)$ and every $\varepsilon > 0$, there exists a $g \in \mathcal{H}$ such that $|\mathbb{P}_n f - \mathbb{P} f| \leq 2\varepsilon + |\mathbb{P}_n g - \mathbb{P} g|$. Since ε is arbitrary and $\gamma_k(\mathbb{P}_n, \mathbb{P}) \to 0$ as $n \to \infty$, the result follows. ■

6. Conclusion and Discussion

We have studied various properties associated with a pseudometric γ_k on $\mathcal{P}$, which is based on the Hilbert space embedding of probability measures. First, we studied the conditions on the kernel (called the characteristic kernel) under which γ_k is a metric, and showed that apart from universal kernels, a large family of bounded continuous kernels induces a metric on $\mathcal{P}$: (a) integrally strictly pd kernels and (b) translation invariant kernels on $\mathbb{R}^d$ and $\mathbb{T}^d$ that have the support of their Fourier transform to be $\mathbb{R}^d$ and $\mathbb{Z}^d$ respectively. Next, we showed that there exist distinct distributions which will be considered close according to γ_k (whether or not the kernel is characteristic), and thus may be hard to distinguish based on finite samples. Finally, we compared γ_k to other metrics on $\mathcal{P}$ and explicitly presented the conditions under which it induces a weak topology on $\mathcal{P}$. These results together provide a strong theoretical foundation for using the γ_k metric in both statistics and machine learning applications.

We now discuss two topics related to γ_k, concerning the choice of kernel parameter and kernels defined on $\mathcal{P}$.

An important question not covered in the present paper is how to choose a characteristic kernel. Let us consider the following setting: $M = \mathbb{R}^d$ and $k_\sigma(x, y) = \exp(-\sigma \|x - y\|_2^2)$, $\sigma \in \mathbb{R}_+$, a Gaussian kernel with σ as the bandwidth parameter. $\{k_\sigma : \sigma \in \mathbb{R}_+\}$ is the family of Gaussian kernels and $\{\gamma_{k_\sigma} : \sigma \in \mathbb{R}_+\}$ is the associated family of distance measures indexed by the kernel parameter, σ. Note that k_σ is characteristic for any $\sigma \in \mathbb{R}_{++}$ and, therefore, γ_{k_σ} is a metric on $\mathcal{P}$ for any $\sigma \in \mathbb{R}_{++}$.

In practice, one would prefer a single number that defines the distance between $\mathbb{P}$ and $\mathbb{Q}$. The question therefore to be addressed is how to choose an appropriate σ. Note that as $\sigma \to 0$, $k_\sigma \to 1$ and as $\sigma \to \infty$, $k_\sigma \to 0$ a.e., which means $\gamma_{k_\sigma}(\mathbb{P},\mathbb{Q}) \to 0$ as $\sigma \to 0$ or $\sigma \to \infty$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}$. This behavior is also exhibited by $k_\sigma(x,y) = \exp(-\sigma\|x-y\|_1)$, $\sigma > 0$ and $k_\sigma(x,y) = \sigma^2/(\sigma^2 + \|x-y\|_2^2)$, $\sigma > 0$, which are also characteristic. This means choosing *sufficiently small* or *sufficiently large* σ (depending on $\mathbb{P}$ and $\mathbb{Q}$) makes $\gamma_{k_\sigma}(\mathbb{P},\mathbb{Q})$ arbitrarily small. Therefore, σ must be chosen appropriately in applications to effectively distinguish between $\mathbb{P}$ and $\mathbb{Q}$.

To this end, one can consider the following modification to γ_k, which yields a pseudometric on $\mathscr{P}$,

$$\gamma(\mathbb{P},\mathbb{Q}) = \sup\{\gamma_k(\mathbb{P},\mathbb{Q}) : k \in \mathcal{K}\} = \sup\{\|\mathbb{P}k - \mathbb{Q}k\|_{\mathcal{H}} : k \in \mathcal{K}\}. \tag{38}$$

Note that γ is the maximal RKHS distance between $\mathbb{P}$ and $\mathbb{Q}$ over a family, $\mathcal{K}$ of measurable and bounded positive definite kernels. It is easy to check that, if any $k \in \mathcal{K}$ is characteristic, then γ is a metric on $\mathscr{P}$. Examples for $\mathcal{K}$ include:

1. $\mathcal{K}_g := \left\{ e^{-\sigma\|x-y\|_2^2}, x,y \in \mathbb{R}^d : \sigma \in \mathbb{R}_+ \right\}$.
2. $\mathcal{K}_l := \left\{ e^{-\sigma\|x-y\|_1}, x,y \in \mathbb{R}^d : \sigma \in \mathbb{R}_+ \right\}$.
3. $\mathcal{K}_\psi := \left\{ e^{-\sigma\psi(x,y)}, x,y \in M : \sigma \in \mathbb{R}_+ \right\}$, where $\psi : M \times M \to \mathbb{R}$ is a negative definite kernel (Berg et al., 1984, Chapter 3).
4. $\mathcal{K}_{rbf} := \left\{ \int_0^\infty e^{-\lambda\|x-y\|_2^2} d\mu_\sigma(\lambda), x,y \in \mathbb{R}^d, \mu_\sigma \in \mathscr{M}^+ : \sigma \in \Sigma \subset \mathbb{R}^d \right\}$, where $\mathscr{M}^+$ is the set of all finite nonnegative Borel measures, μ_σ on $\mathbb{R}_+$ that are not concentrated at zero, etc.
5. $\mathcal{K}_{lin} := \left\{ k_\lambda = \sum_{j=1}^l \lambda_j k_j \,|\, k_\lambda \text{ is pd}, \sum_{j=1}^l \lambda_j = 1 \right\}$, which is the linear combination of pd kernels $\{k_j\}_{j=1}^l$.
6. $\mathcal{K}_{con} := \left\{ k_\lambda = \sum_{j=1}^l \lambda_j k_j \,|\, \lambda_j \geq 0, \sum_{j=1}^l \lambda_j = 1 \right\}$, which is the convex combination of pd kernels $\{k_j\}_{j=1}^l$.

The idea and validity behind the proposal of γ in (38) can be understood from a Bayesian perspective, where we define a non-negative finite measure λ over $\mathcal{K}$, and average γ_k over that measure, that is, $\alpha(\mathbb{P},\mathbb{Q}) := \int_{\mathcal{K}} \gamma_k(\mathbb{P},\mathbb{Q})\, d\lambda(k)$. This also yields a pseudometric on $\mathscr{P}$. That said, $\alpha(\mathbb{P},\mathbb{Q}) \leq \lambda(\mathcal{K})\gamma(\mathbb{P},\mathbb{Q}), \forall \mathbb{P},\mathbb{Q}$, which means that, if $\mathbb{P}$ and $\mathbb{Q}$ can be distinguished by α, then they can be distinguished by γ, but not vice-versa. In this sense, γ is stronger than α and therefore studying γ makes sense. One further complication with the Bayesian approach is in defining a sensible λ over $\mathcal{K}$. Note that γ_{k_0} can be obtained by defining $\lambda(k) = \delta(k-k_0)$ in $\alpha(\mathbb{P},\mathbb{Q})$. Future work will include analyzing γ and investigating its utility in applications compared to that of γ_k (with a fixed kernel, k). Sriperumbudur et al. (2009a) describes preliminary work, showing that $\gamma(\mathbb{P}_m,\mathbb{Q}_n)$ is a $\sqrt{mn/(m+n)}$-consistent estimator of $\gamma(\mathbb{P},\mathbb{Q})$, for families of kernels $\mathcal{K}$ including those mentioned above.

We now discuss how kernels on $\mathscr{P}$ can be obtained from γ_k. As noted by Gretton et al. (2007b, Section 4), and following Hein et al. (2004), γ_k is a *Hilbertian metric* on $\mathscr{P}$: the associated kernel can be easily computed using (33),

$$K(\mathbb{P},\mathbb{Q}) = \left\langle \int_M k(\cdot,x)\, d\mathbb{P}(x), \int_M k(\cdot,x)\, d\mathbb{Q}(x) \right\rangle_{\mathcal{H}} = \int\int_M k(x,y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y),$$

where the positive definite kernel $K : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ is a dot-product kernel on $\mathscr{P}$. Using the results in Berg et al. (1984, Chapter 3, Theorems 2.2 and 2.3), Gaussian and inverse multi-quadratic kernels on $\mathscr{P}$ can be defined as

$$K(\mathbb{P},\mathbb{Q}) = \exp\left(-\sigma\gamma_k^2(\mathbb{P},\mathbb{Q})\right), \sigma > 0 \text{ and } K(\mathbb{P},\mathbb{Q}) = \left(\sigma+\gamma_k^2(\mathbb{P},\mathbb{Q})\right)^{-1}, \sigma > 0$$

respectively. Further work on Hilbertian metrics and positive definite kernels on probability measures has been carried out by Hein and Bousquet (2005) and Fuglede and Topsøe (2003).

Acknowledgments

The authors thank the editor and reviewers for their constructive comments. B. K. S. and G. R. G. L. wish to acknowledge support from the Max Planck Institute (MPI) for Biological Cybernetics, the National Science Foundation (grant DMS-MSPA 0625409), the Fair Isaac Corporation and the University of California MICRO program. Part of this work was done while B. K. S. was an intern at the MPI, and part was done while A. G. was a project scientist at CMU, under grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747, and NSF NeTS-NOSS CNS-0625518. This work is also supported by the IST Program of the EC, under the FP7 Network of Excellence, ICT-216886-NOE. B. K. S. wishes to thank Agnes Radl for her comments on an earlier version of the manuscript.

Appendix A. Supplementary Results

For completeness, we present the supplementary results that were used to prove the results in this paper. The following result is quoted from Folland (1999, Theorem 8.14).

Theorem 25 *Suppose $\phi \in L^1(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \phi(x)\,dx = a$ and $\phi_t(x) = t^{-d}\phi(t^{-1}x)$ for $t > 0$. If f is bounded and uniformly continuous on $\mathbb{R}^d$, then $f * \phi_t \to af$ uniformly as $t \to 0$.*

By imposing slightly stronger conditions on ϕ, the following result quoted from Folland (1999, Theorem 8.15) shows that $f * \phi_t \to af$ almost everywhere for $f \in L^r(\mathbb{R}^d)$.

Theorem 26 *Suppose $|\phi(x)| \leq C(1+\|x\|_2)^{-d-\varepsilon}$ for some $C, \varepsilon > 0$, and $\int_{\mathbb{R}^d} \phi(x)\,dx = a$. If $f \in L^r(\mathbb{R}^d)$ $(1 \leq r \leq \infty)$, then $f * \phi_t(x) \to af(x)$ as $t \to 0$ for every x in the Lebesgue set of f—in particular, for almost every x, and for every x at which f is continuous.*

Theorem 27 (Fourier transform of a measure) *Let μ be a finite Borel measure on $\mathbb{R}^d$. The Fourier transform of μ is given by*

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

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

(i) $\overline{\widehat{\mu}(\omega)} = \widehat{\mu}(-\omega), \forall\, \omega \in \mathbb{R}^d$, *that is, $\widehat{\mu}$ is conjugate symmetric,*

(ii) $\widehat{\mu}(0) = 1$.

The following result, called the Riemann-Lebesgue lemma, is quoted from Rudin (1991, Theorem 7.5).

Lemma 28 (Riemann-Lebesgue) *If $f \in L^1(\mathbb{R}^d)$, then $\widehat{f} \in C_0(\mathbb{R}^d)$, and $\|\widehat{f}\|_\infty \leq \|f\|_1$.*

The following theorem is a version of the *Paley-Wiener theorem* for distributions, and is proved in Rudin (1991, Theorem 7.23).

Theorem 29 (Paley-Wiener) *If $f \in \mathscr{D}'_d$ has compact support, then $\widehat{f}$ is the restriction to $\mathbb{R}^d$ of an entire function on $\mathbb{C}^d$.*

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological)*, 28:131–142, 1966.

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

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

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

A. D. Barbour and L. H. Y. Chen. *An Introduction to Stein's Method*. Singapore University Press, Singapore, 2005.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Spring Verlag, New York, 1984.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, London, UK, 2004.

K. M. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

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

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

W. Dahmen and C. A. Micchelli. Some remarks on ridge functions. *Approx. Theory Appl.*, 3: 139–143, 1987.

E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Testing of goodness of fit based on the L_2-Wasserstein distance. *Annals of Statistics*, 27:1230–1239, 1999.

L. Devroye and L. Györfi. No empirical probability measure can converge in the total variation sense for all distributions. *Annals of Statistics*, 18(3):1496–1499, 1990.

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

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley-Interscience, New York, 1999.

B. Fuglede and F. Topsøe. Jensen-Shannon divergence and Hilbert space embedding, 2003. Preprint.

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

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

K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(5):1871–1905, 2009a.

K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 473–480, 2009b.

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

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

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 2005a.

A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In Z. Ghahramani and R. Cowell, editors, *Proc. 10^{th} International Workshop on Artificial Intelligence and Statistics*, pages 1–8, 2005b.

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, 2007a.

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

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proc. 10^{th} International Workshop on Artificial Intelligence and Statistics*, pages 136–143, 2005.

M. Hein, T.N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVMs. In *Proceedings of the 26th DAGM Symposium*, pages 270–277, Berlin, 2004. Springer.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypothesis*. Springer-Verlag, New York, 2005.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Information Theory*, 52(10):4394–4412, 2006.

T. Lindvall. *Lectures on the Coupling Method*. John Wiley & Sons, New York, 1992.

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

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

A. Pinkus. Strictly positive definite functions on a real inner product space. *Adv. Comput. Math.*, 20:263–271, 2004.

S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. John Wiley & Sons, Chichester, 1991.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Vol. I Theory, Vol. II Applications*. Probability and its Applications. Springer-Verlag, Berlin, 1998.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

M. Reed and B. Simon. *Methods of Modern Mathematical Physics: Functional Analysis I*. Academic Press, New York, 1980.

M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Annals of Statistics*, 3(1):1–14, 1975.

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

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

H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57:3498 – 3511, 2009.

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

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.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In R. Servedio and T. Zhang, editors, *Proc. of the 21st Annual Conference on Learning Theory*, pages 111–122, 2008.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. G. Lanckriet, and B. Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press, 2009a.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *http://arxiv.org/abs/0901.2698v4*, October 2009b.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In Y. W. Teh and M. Titterington, editors, *Proc. 13th International Conference on Artificial Intelligence and Statistics*, volume 9 of *Workshop and Conference Proceedings*. JMLR, 2010a.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *http://arxiv.org/abs/1003.0887*, March 2010b.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1972.

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

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

J. Stewart. Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409–433, 1976.

I. Vajda. *Theory of Statistical Inference and Information*. Kluwer Academic Publishers, Boston, 1989.

S. S. Vallander. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.*, 18:784–786, 1973.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

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

N. Weaver. *Lipschitz Algebras*. World Scientific Publishing Company, 1999.

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