Journal of Machine Learning Research 19 (2018) 1-29 Submitted 3/16; Revised 5/18; Published 9/18

Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions

Carl-Johann Simon-Gabriel CJSIMON@TUEBINGEN.MPG.DE
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
MPI for Intelligent Systems
Spemannstrasse 41,
72076 Tübingen, Germany

Editor: Ingo Steinwart

Abstract

Kernel mean embeddings have become a popular tool in machine learning. They map probability measures to functions in a reproducing kernel Hilbert space. The distance between two mapped measures defines a semi-distance over the probability measures known as the maximum mean discrepancy (MMD). Its properties depend on the underlying kernel and have been linked to three fundamental concepts of the kernel literature: universal, characteristic and strictly positive definite kernels.

The contributions of this paper are three-fold. First, by slightly extending the usual definitions of universal, characteristic and strictly positive definite kernels, we show that these three concepts are essentially equivalent. Second, we give the first complete characterization of those kernels whose associated MMD-distance metrizes the weak convergence of probability measures. Third, we show that kernel mean embeddings can be extended from probability measures to generalized measures called Schwartz-distributions and analyze a few properties of these distribution embeddings.

Keywords: kernel mean embedding, universal kernel, characteristic kernel, Schwartz-distributions, kernel metrics on distributions, metrisation of the weak topology

1. Introduction

During the past decades, kernel methods have risen to a major tool across various areas of machine learning. They were originally introduced via the "kernel trick" to generalize linear regression and classification tasks by effectively transforming the optimization over a set of linear functions into an optimization over a so-called reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$, which is entirely defined by the kernel k. This lead to kernel (ridge) regression, kernel SVM and many other now standard algorithms. Besides these regression-type algorithms, another major family of kernel methods rely on kernel mean embeddings (KMEs). A KME is a mapping Φ_k that maps probability measures to functions in an RKHS via $\Phi_k : \; P \;\longmapsto\; \int_{\mathcal{X}} k(.,x)\,\mathrm{d}P(x)$. The RKHS-distance between two mapped measures defines a semi-distance over the set of probability measures, known as the Maximum Mean Discrepancy (MMD). It has numerous applications, ranging from homogeneity (Gretton et al., 2007), distribution comparison (Gretton et al., 2007, 2012) and (conditional) independence tests (Gretton et al., 2005, 2008; Fukumizu et al., 2008; Gretton and Györfi, 2010; Lopez-Paz et al., 2013) to generative adversarial networks (Dziugaite et al., 2015; Li et al., 2015). While KMEs have already been extended to embed not only probability measures, but also signed finite measures, a first contribution of this paper is to show that they can be extended even further to embed generalized measures called *Schwartz-distributions*. For an introduction to Schwartz-distributions—which we will now simply call a distribution, as opposed to a (signed) measure—see Appendix B. Furthermore, we show that for smooth and translation-invariant kernels, if the KME is injective over the set of probability measures, then it remains injective when extended to some Schwartz-distribution sets.

Our second contribution concerns the notions of universal, characteristic and strictly positive definite (s.p.d.) kernels. They are of prime importance to guarantee the consistency of many regression-type or MMD-based algorithms (Steinwart, 2001; Steinwart and Christmann, 2008). While these notions were originally introduced in very different contexts, they were shown to be connected in many ways which were eventually summarized in Figure 1 of Sriperumbudur et al. (2011). But by handling separately all the many variants of universal, characteristic and s.p.d. kernels that had been introduced, this figure—and the general machine learning literature—somehow missed the underlying very general duality principle that connects these notions. By giving a unified definition of these three concepts, we will make their link explicit, easy to remember, and immediate to generalize to Schwartz-distributions and other spaces.

Our third contribution concerns the MMD semi-metric. Through a series of articles, Sriperumbudur et al. (2010b; 2016) gave various sufficient conditions for a kernel to *metrize the weak-convergence of probability measures*, which means that a sequence of probability measures converges in MMD distance if and only if (iff) it converges weakly. Here, we generalize these results and give the first complete characterization of the kernels that metrize weak convergence when the underlying space $\mathcal{X}$ is locally compact.

Finally, we develop a few calculus rules to work with KMEs of Schwartz distributions. In particular, we prove the following formulae:

$$\left\langle f\,,\, \int k(.,x)\,\mathrm{d}D(x) \right\rangle_k = \int \langle f\,,\, k(.,x)\rangle_k \,\mathrm{d}D(x) \tag{Definition of KME}$$

$$\left\langle \int k(.,y)\,\mathrm{d}D(y)\,,\, \int k(.,x)\,\mathrm{d}T(x) \right\rangle_k = \int k(x,y)\,\mathrm{d}D(y)\,\mathrm{d}\bar{T}(x) \tag{Fubini}$$

$$\int k(.,x)\,\mathrm{d}(\partial^p S)(x) = (-1)^{|p|} \int \partial^{(0,p)} k(.,x)\,\mathrm{d}S(x). \tag{Differentiation}$$

The first and second lines are standard calculus rules for KMEs when applied with two probability measures D and T. We extend them to distributions. The third line however is specific to distributions. It uses the distributional derivative ('∂') which extends the usual derivative of functions to signed measures and distributions. For a quick introduction to Schwartz distributions and their derivatives see Appendix B.

The structure of this paper roughly follows this exposition. After fixing our notations, Section 2 introduces KMEs of measures and distributions. In Section 3 we define the concepts of universal, characteristic and s.p.d. kernels and prove their equivalence. Section 4 compares convergence in MMD with other modes of convergence for measures and distributions. Section 5 focuses specifically on KMEs of Schwartz-distributions, and Section 6 gives a brief overview of the related work and concludes.

1.1. Definitions and Notations

Let $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ be the sets of non-negative integers, of reals and of complex numbers. The input set $\mathscr{X}$ of all considered kernels and functions will be locally compact and Hausdorff. This includes any Euclidian spaces or smooth manifolds, but no infinite-dimensional Banach-space. Whenever referring to differentiable functions or to distributions of order ≥ 1, we will *implicitly* assume that $\mathscr{X}$ is an open subset of $\mathbb{R}^d$ for some $d > 0$.

A *kernel* $k : \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{C}$ is a positive definite function, meaning that for all $n \in \mathbb{N}\backslash\{0\}$, all $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, and all $x_1, x_2, \ldots x_n \in \mathscr{X}$, $\sum_{i,j=1}^n \lambda_i k(x_i, x_j)\overline{\lambda_j} \geq 0$. For $p = (p_1, p_2, \ldots, p_d) \in \mathbb{N}^d$ and $f : \mathscr{X} \longrightarrow \mathbb{C}$, we define $|p| := \sum_{i=1}^d p_i$ and $\partial^p f := \frac{\partial^{|p|} f}{\partial^{p_1} x_1 \partial^{p_2} x_2 \cdots \partial^{p_d} x_d}$. For $m \in \mathbb{N} \cup \{\infty\}$, we say that f (resp. k) is m-times (resp. (m, m)-times) continuously differentiable and write $f \in \mathscr{C}^m$ (resp. $k \in \mathscr{C}^{(m,m)}$), if for any p with $|p| = m$, $\partial^p f$ (resp. $\partial^{(p,p)} k$) exists and is continuous. $\mathscr{C}^m_b$ (resp. $\mathscr{C}^m_0$, $\mathscr{C}^m_c$) is the subsets of $\mathscr{C}^m$ for which $\partial^p f$ is bounded (resp. converges to 0 at infinity, resp. has compact support) whenever $|p| \leq m$. Whenever $m = 0$, we may drop the superscript m. By default, we equip $\mathscr{C}^m_*$ ($* \in \{\emptyset, b, 0, c\}$) with their natural topologies (see Introduction of Simon-Gabriel and Schölkopf 2016 or Treves 1967). We write $k \in \mathscr{C}_0^{(m,m)}$ whenever k is bounded, (m, m)-times continuously differentiable and for all $|p| \leq m$ and $x \in \mathscr{X}$, $\partial^{(p,p)} k(., x) \in \mathscr{C}_0$.

We call *space of functions* and denote by $\mathscr{F}$ any locally convex (loc. cv.) topological vector space (TVS) of functions (see Appendix C and Treves 1967). Loc. cv. TVSs include all Banach- or Fréchet-spaces and all function spaces defined in this paper.

The dual $\mathscr{F}'$ of a space of functions $\mathscr{F}$ is the space of *continuous* linear forms over $\mathscr{F}$. We denote $\mathscr{M}_\delta$, $\mathscr{E}^m$, $\mathscr{D}^m_{L^1}$ and $\mathscr{D}^m$ the duals of $\mathbb{C}^{\mathscr{X}}$, $\mathscr{C}^m$, $\mathscr{C}^m_0$ and $\mathscr{C}^m_c$ respectively. By identifying each signed measure μ with a linear functional of the form $f \longmapsto \int f \, \mathrm{d}\mu$, the Riesz-Markov-Kakutani representation theorem (see Appendix C) identifies $\mathscr{D}^0$ (resp. $\mathscr{D}^0_{L^1}$, $\mathscr{E}^0$ and $\mathscr{M}_\delta$) with the set $\mathscr{M}_r$ (resp. $\mathscr{M}_f$, $\mathscr{M}_c$, $\mathscr{M}_\delta$) of signed regular Borel measures (resp. with finite total variation, with compact support, with finite support). By definition, $\mathscr{D}^\infty$ is the set of all Schwartz-distributions, but all duals defined above can be seen as subsets of $\mathscr{D}^\infty$ and are therefore sets of Schwartz-distributions. Any element μ of $\mathscr{M}_r$ will be called a measure, any element of $\mathscr{D}^\infty$ a distribution. See Appendix B for a brief introduction to distributions and their connection to measures. We extend the usual notation $\mu(f) := \int f(x) \, \mathrm{d}\mu(x)$ for measures μ to distributions D: $D(f) =: \int f(x) \, \mathrm{d}D(x)$. Given a KME Φ_k and two embeddable distributions D, T (see Definition 1), we define

$$\langle D \, , \, T \rangle_k := \langle \Phi_k(D) \, , \, \Phi_k(T) \rangle_k \quad \text{and} \quad \|D\|_k := \|\Phi_k(D)\|_k \ .$$

where $\langle . \, , . \rangle_k$ is the inner product of the RKHS $\mathscr{H}_k$ of k. To avoid introducing a new name, we call $\|D\|_k$ the maximum mean discrepancy (MMD) of D, even though the term "discrepancy" usually specifically designates a distance between two distributions rather than the norm of a single one. Given two topological sets $\mathscr{S}_1, \mathscr{S}_2$, we write

$$\mathscr{S}_1 \hookrightarrow \mathscr{S}_2$$

and say that $\mathscr{S}_1$ *is continuously contained in* $\mathscr{S}_2$ if $\mathscr{S}_1 \subset \mathscr{S}_2$ and if the topology of $\mathscr{S}_1$ is stronger than the topology induced by $\mathscr{S}_2$. For a general introduction to topology, TVSs and distributions, we recommend Treves (1967).

2. Kernel Mean Embeddings of Distributions

In this section, we show how to embed general distribution spaces into an RKHS. To do so, we redefine the integral $\int k(.,x)\,\mathrm{d}\mu(x)$ so as to be well-defined even if μ is a distribution. It is often defined as a Bochner-integral; here we instead use the *weak*- (or *Pettis*-) integral:

Definition 1 (Weak Integral and KME) *Let D be a linear form over a space of functions $\mathcal{F}$. Let $\vec{\varphi}: \mathcal{X} \longrightarrow \mathcal{H}_k$ be an RKHS-valued function such that for any $f \in \mathcal{H}_k$, $x \longmapsto \langle f\,,\,\vec{\varphi}(x)\rangle_k \in \mathcal{F}$. Then $\vec{\varphi}: \mathcal{X} \longrightarrow \mathcal{H}_k$ is weakly integrable with respect to (w.r.t.) D if there exists a function in $\mathcal{H}_k$, written $\int \vec{\varphi}(x)\,\mathrm{d}D(x)$, such that*

$$\forall f \in \mathcal{H}_k, \quad \left\langle f\,,\, \int \vec{\varphi}(x)\,\mathrm{d}D(x) \right\rangle_k = \int \langle f\,,\,\vec{\varphi}(x)\rangle_k \,\mathrm{d}\bar{D}(x)\,, \tag{1}$$

where the right-hand-side stands for $\bar{D}\,(x \mapsto \langle f\,,\,\vec{\varphi}(x)\rangle_k)$ and $\bar{D}$ denotes the complex-conjugate of D. If $\vec{\varphi}(x) = k(.,x)$, we call $\int k(.,x)\,\mathrm{d}D(x)$ the kernel mean embedding (KME) of D and say that D embeds into $\mathcal{H}_k$. We denote $\Phi_{\vec{\varphi}}$ the map $\Phi_{\vec{\varphi}}: \ D \longmapsto \int \vec{\varphi}(x)\,\mathrm{d}D(x)$.

This definition extends the usual Bochner-integral: if $\vec{\varphi}$ is Bochner-integrable w.r.t. a measure $\mu \in \mathcal{M}_r$, then $\vec{\varphi}$ is weakly integrable w.r.t. μ and the integrals coincide (Schwabik, 2005, Prop. 2.3.1). In particular, if $x \longmapsto \|\vec{\varphi}(x)\|_k$ is Lebesgue-integrable, then $\vec{\varphi}$ is Bochner integrable, thus weakly integrable.

The general definition with $\vec{\varphi}$ instead of $k(.,x)$ will be useful in Section 5. But for now, let us concentrate on KMEs where $\vec{\varphi}(x) = k(.,x)$. Kernels satisfy the so-called *reproducing property*: for any $f \in \mathcal{H}_k$, $f(x) = \langle f\,,\,k(.,x)\rangle_k$. Therefore, the condition for all $f \in \mathcal{H}_k$ $x \longmapsto \langle f\,,\,\vec{\varphi}(x)\rangle_k \in \mathcal{F}$ reduces to $\mathcal{H}_k \subset \mathcal{F}$, and Equation (1) reads:

$$\forall f \in \mathcal{H}_k, \qquad \left\langle f\,,\, \int k(.,x)\,\mathrm{d}D(x) \right\rangle_k = \bar{D}(f)\,. \tag{2}$$

Thus, by the Riesz representation theorem (see Appendix C), D embeds into $\mathcal{H}_k$ iff it defines a continuous linear form over $\mathcal{H}_k$. And in that case, its KME $\int k(.,x)\,\mathrm{d}D(x)$ is the Riesz-representer of $\bar{D}$ restricted to $\mathcal{H}_k$. Thus, for an embeddable space of distributions $\mathcal{D}$, the embedding Φ_k can be decomposed as follows:

$$\Phi_k : \left\{ \begin{array}{ccccc} \mathcal{D} & \longrightarrow & \mathcal{H}_k' & \longrightarrow & \mathcal{H}_k \\ & \text{Conjugate restriction} & & \text{Riesz representer} & \\ D & \longmapsto & \bar{D}\big|_{\mathcal{H}_k} & \longmapsto & \int k(.,x)\,\mathrm{d}D(x) \end{array} \right. . \tag{3}$$

To know if D is continuous over $\mathcal{H}_k$, we use the following lemma, and its applications.

Lemma 2 *If $\mathcal{H}_k \hookrightarrow \mathcal{F}$, then $\mathcal{F}'$ embeds into $\mathcal{H}_k$.*

Proof Suppose that $\mathcal{H}_k \hookrightarrow \mathcal{F}$. Let $D \in \mathcal{F}'$ and let $f, f_1, f_2, \ldots \in \mathcal{H}_k$. If $f_n \to f$ in $\mathcal{H}_k$ then $f_n \to f$ in $\mathcal{F}$, thus $D(f_n) \to D(f)$. Thus D is a continuous linear form over $\mathcal{H}_k$. ■

In practice we typically use one of the following two corollaries (proofs in Appendices A.1 and A.2). The space $(\mathscr{C}_b)_c$ that they mention will be introduced in the discussions following Theorem 6. It has the same elements as $\mathscr{C}_b$, but carries a weaker topology.

Corollary 3 (Embedding of Measures) *$\mathscr{H}_k \subset \mathscr{C}_0$ (resp. $\mathscr{H}_k \subset \mathscr{C}_b$, resp. $\mathscr{H}_k \subset \mathscr{C}$) iff the two following conditions hold.*

(i) For all $x \in \mathcal{X}$, $k(.,x) \in \mathscr{C}_0$ (resp. $k(.,x) \in \mathscr{C}_b$, resp. $k(.,x) \in \mathscr{C}$).

(ii) $x \longmapsto k(x,x)$ is bounded (resp. bounded, resp. locally *bounded, meaning that, for each $y \in \mathcal{X}$, there exists a (compact) neighborhood of y on which $x \longmapsto k(x,x)$ is bounded.).*

If so, then $\mathscr{H}_k \hookrightarrow \mathscr{C}_0$ (resp. $\mathscr{H}_k \hookrightarrow \mathscr{C}_b$, thus $\mathscr{H}_k \hookrightarrow (\mathscr{C}_b)_c$, resp. $\mathscr{H}_k \hookrightarrow \mathscr{C}$) and $\mathcal{M}_f$ (resp. $\mathcal{M}_f$, resp. $\mathcal{M}_c$) embeds into $\mathscr{H}_k$.

Corollary 4 (Embedding of Distributions)

If $k \in \mathscr{C}^{(m,m)}$, then $\mathscr{H}_k \hookrightarrow \mathscr{C}^m$, thus $\mathscr{E}^m$ embeds into $\mathscr{H}_k$.
If $k \in \mathscr{C}_0^{(m,m)}$, then $\mathscr{H}_k \hookrightarrow \mathscr{C}_0^m$, thus $\mathscr{D}_{L^1}^m$ embeds into $\mathscr{H}_k$.
If $k \in \mathscr{C}_b^{(m,m)}$, then $\mathscr{H}_k \hookrightarrow \mathscr{C}_b^m$, thus $\mathscr{H}_k \hookrightarrow (\mathscr{C}_b^m)_c$, thus $\mathscr{D}_{L^1}^m$ embeds into $\mathscr{H}_k$.

Corollary 3 applied to $\mathscr{C}_b$ shows that $\mathscr{H}_k$ is (continuously) contained in $\mathscr{C}_b$ iff k is bounded and separately continuous. As discovered by Lehtö (1952), there also exist kernels which are not continuous but whose RKHS $\mathscr{H}_k$ is contained in $\mathscr{C}_b$. So the conditions in Corollary 4 are sufficient, but in general not necessary. Concerning Lemma 2, note that it not only requires $\mathscr{H}_k \subset \mathscr{F}$, but also that $\mathscr{H}_k$ carries a stronger topology than $\mathscr{F}$. Otherwise there might exist a continuous form over $\mathscr{F}$ that is defined but non-continuous over $\mathscr{H}_k$. However, Corollary 3 shows that this cannot happen for $\mathscr{C}_*$, because if $\mathscr{H}_k \subset \mathscr{C}_*$ then $\mathscr{H}_k \hookrightarrow \mathscr{C}_*$. Although this also holds for $m = \infty$ (Simon-Gabriel and Schölkopf, 2016, Prop.4 & Comments), we do not know whether it extends to any $m > 0$.

3. Universal, Characteristic and S.P.D. Kernels

The literature distinguishes various variants of universal, characteristic and s.p.d. kernels, such as c-, $cc-$ or c_0-universal kernels, s.p.d. and integrally strictly positive definite ($\int$s.p.d.) kernels. They are all special cases of the following unifying definitions.

Definition 5 *Let k be a kernel, $\mathscr{F}$ be a space of functions such that $\mathscr{H}_k \subset \mathscr{F}$, and $\mathscr{D}$ be an embeddable subset of $\mathscr{F}'$ (e.g. an embeddable set of distributions). We say that k is*

- ▷ universal over $\mathscr{F}$ *if $\mathscr{H}_k$ is dense in $\mathscr{F}$.*
- ▷ characteristic to $\mathscr{D}$ *if the KME Φ_k is injective over $\mathscr{D}$.*
- ▷ strictly positive definite (s.p.d.) over $\mathscr{D}$ *if: $\forall D \in \mathscr{D}$, $\|\Phi_k(D)\|_k^2 = 0 \Rightarrow D = 0$.*

A universal kernel over $\mathscr{C}^m$ (resp. $\mathscr{C}_0^m$) will be said c^m- (resp. c_0^m-) universal (without the superscript when $m = 0$). A characteristic kernel to the set $\mathscr{P}$ of probability measures will simply be called characteristic.

In general, instead of writing $\|\Phi_k(D)\|_k$ and $\langle \Phi_k(D) \,, \Phi_k(T) \rangle_k$, we will write $\|D\|_k$ and $\langle D \,, T \rangle_k$. These definitions encompass the usual s.p.d. definitions. Denoting δ_x the Dirac measure concentrated on x, what is usually called

- ▷ s.p.d. corresponds to $\mathscr{D} = \mathcal{M}_\delta$, i.e.:

$$\forall \mu = \sum_{i=1}^{n} \lambda_i \delta_{x_i} \in \mathcal{M}_\delta : \quad \|\mu\|_k^2 = \sum_{i,j=1}^{n} \lambda_i k(x_i, x_j) \bar{\lambda}_j = 0 \quad \Rightarrow \quad \lambda_1 = \ldots = \lambda_n = 0 \ .$$

- ▷ conditionally s.p.d. corresponds to $\mathscr{D} = \mathcal{M}_\delta^0$ where $\mathcal{M}_\delta^0 := \{\mu \in \mathcal{M}_\delta : \mu(\mathcal{X}) = 0\}$, i.e.:

$$\left.\begin{array}{l} \forall \mu = \sum_{i=1}^{n} \lambda_i \delta_{x_i} \in \mathcal{M}_\delta \\ \text{s.t. } \sum_{i=1}^{n} \lambda_i = 0 \end{array}\right\} : \ \|\mu\|_k^2 = \sum_{i,j=1}^{n} \lambda_i k(x_i, x_j) \bar{\lambda}_j = 0 \quad \Rightarrow \quad \lambda_1 = \ldots = \lambda_n = 0 \ .$$

- ▷ $\int$s.p.d. corresponds to $\mathscr{D} = \mathcal{M}_f$, i.e.:

$$\forall \mu \in \mathcal{M}_f : \quad \|\mu\|_k^2 = \iint k(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\bar{\mu}(y) = 0 \quad \Rightarrow \quad \mu = 0 \ .$$

Let us now state the general link between universal, characteristic and s.p.d. kernels, which is the key that underlies Figure 1 of Sriperumbudur et al. (2011).

Theorem 6 *If $\mathcal{H}_k \hookrightarrow \mathcal{F}$, then the following statements are equivalent.*

(i) k is universal over $\mathcal{F}$.
(ii) k is characteristic to $\mathcal{F}'$.
(iii) k is strictly positive definite over $\mathcal{F}'$.

Proof Equivalence of (ii) & (iii): Saying that $\|\Phi_k(D)\|_k = 0$ is equivalent to saying $\Phi_k(D) = 0$. Thus Φ_k is s.p.d. over $\mathcal{F}'$ iff the $\mathrm{Ker}(\Phi_k)$ (meaning the vector space that is mapped to 0 via Φ_k) is reduced to $\{0\}$, which happens iff Φ_k is injective over $\mathcal{F}'$.

Equivalence of (i) & (ii): Φ_k is the conjugate restriction operator $|_{\mathcal{H}_k} : \ D \longmapsto \bar{D}|_{\mathcal{H}_k}$ composed with the Riesz representer mapping (Diagram Eq.3). The Riesz representer map is injective, so Φ_k is injective iff $|_{\mathcal{H}_k}$ is injective. Now, if $\mathcal{H}_k$ is dense in $\mathcal{F}$, then, by continuity, any $D \in \mathcal{F}'$ is uniquely defined by its values taken on $\mathcal{H}_k$. Thus $|_{\mathcal{H}_k}$ is injective. Reciprocally, if $\mathcal{H}_k$ is not dense in $\mathcal{F}$, then, by the Hahn-Banach theorem (Treves, 1967, Thm.18.1, Cor.3), there exists two different elements in $\mathcal{F}'$ that coincide on $\mathcal{H}_k$ but not on the entire space $\mathcal{F}$. So $|_{\mathcal{H}_k}$ is not injective. Thus $|_{\mathcal{H}_k}$ is injective iff $\mathcal{H}_k$ is dense in $\mathcal{F}$. ■

To apply this theorem it suffices to find so-called *duality pairs* $(\mathcal{F}, \mathcal{F}')$ such that $\mathcal{H}_k \hookrightarrow \mathcal{F}$. Table 1 lists several such pairs. It shows in particular the well-known equivalence between c- (resp. c_0-) universal kernels and characteristic kernels to $\mathcal{M}_c$ (resp. $\mathcal{M}_f$) (Sriperumbudur et al., 2008). But we now discover that s.p.d. kernels over $\mathcal{M}_\delta$ can also be characterized in terms of universality over $\mathbb{C}^{\mathcal{X}}$, because $(\mathbb{C}^{\mathcal{X}})' = \mathcal{M}_\delta$ (Duc-Jacquet, 1973, p.II.35). And we directly get the generalization to distributions and c_*^m-universality.

However, Theorem 6 leaves open the important case where k is characteristic (to $\mathcal{P}$). Of course, as $\mathcal{P}$ is contained in $\mathcal{M}_f$, it shows that a c_0-universal kernel must be characteristic. But to really characterize characteristic kernels in terms of universality, we would need to find a predual of $\mathcal{P}$, meaning a space $\mathcal{F}$ such that $\mathcal{F}' = \mathcal{P}$. This is hardly possible, as $\mathcal{P}$ is not even a vector space. However, we will see in Theorem 8 that k is characteristic iff k is

Universal	Characteristic	S.P.D.	Name	Proof
$\mathscr{F}$	$\mathscr{F}'$	$\mathscr{F}'$	/	Thm. 6
$\mathbb{C}^{\mathcal{X}}$	$\mathcal{M}_\delta$	$\mathcal{M}_\delta$	s.p.d.	Thm. 6
$\mathbb{C}^{\mathcal{X}}/\mathbb{1}$	$\mathcal{M}_\delta^0$	$\mathcal{M}_\delta^0$	conditionally s.p.d.	Prop. 7
$\mathscr{C}$	$\mathcal{M}_c$	$\mathcal{M}_c$	c-universal (or cc-universal)	Thm. 6
$\mathscr{C}_0$	$\mathcal{M}_f$	$\mathcal{M}_f$	c_0-universal	Thm. 6
$(\mathscr{C}_b)_c$	$\mathcal{M}_f$	$\mathcal{M}_f$	$\int$spd	Thm. 6
$((\mathscr{C}_b)_c)/\mathbb{1}$	$\mathscr{P}$ (or $\mathcal{M}_f^0$)	$\mathcal{M}_f^0$	characteristic	Prop. 7
$\mathscr{C}^m$	$\mathscr{E}^m$	$\mathscr{E}^m$	c^m-universal	Thm. 6
$\mathscr{C}_0^m$	$\mathscr{D}_{L^1}^m$	$\mathscr{D}_{L^1}^m$	c_0^m-universal	Thm. 6
$(\mathscr{C}_b^m)_c$	$\mathscr{D}_{L^1}^m$	$\mathscr{D}_{L^1}^m$	/	Thm. 6

Table 1: Equivalence between the notions of universal, characteristic and s.p.d. kernels.

characteristic to the vector space $\mathcal{M}_f^0 := \{\mu \in \mathcal{M}_f : \mu(\mathcal{X}) = 0\}$. So if we find a predual of $\mathcal{M}_f^0$, then we get an analog of Theorem 6 applied to $\mathscr{P}$. Let us do so now.

As $\mathcal{M}_f^0$ is the hyperplane of $\mathcal{M}_f$ that is given by the equation $\int 1 \, \mathrm{d}\mu = 0$, our idea is to take a predual $\mathscr{F}$ of $\mathcal{M}_f$ and consider the quotient $\mathscr{F}/\mathbb{1}$ of $\mathscr{F}$ divided by the constant function $\mathbb{1}$. Proposition 35.5 of Treves (1967) would then show that $(\mathscr{F}/\mathbb{1})' = \mathcal{M}_f^0$. But if we take the usual predual of $\mathcal{M}_f$, $\mathscr{F} = \mathscr{C}_0$, then $\mathbb{1} \notin \mathscr{F}$, so the quotient $\mathscr{F}/\mathbb{1}$ is undefined. However, preduals are not unique, so let us try with another space $\mathscr{F}$ that contains $\mathbb{1}$, for example $\mathscr{F} = \mathscr{C}_b$. This time $\mathbb{1} \in \mathscr{F}$, but now the problem is that $\mathscr{F}'$ is in general strictly bigger than $\mathcal{M}_f$ (Fremlin et al., 1972, Sec. 2, §2) whereas we want $\mathscr{F}' = \mathcal{M}_f$. The trick now is to keep $\mathscr{C}_b$, but equip it with a weaker topology than the usual one, so that $\mathscr{F}'$ becomes smaller. Intuitively, the reason for this decrease of $\mathscr{F}'$ is that, by weakening the topology of $\mathscr{F}$, we let more sequences converge in $\mathscr{F}$. This makes it more difficult for a functional over $\mathscr{F}$ to be continuous, because for any converging sequence in $\mathscr{F}$, its images need to converge. Thus some of the linear functionals that were continuous for the original topology of $\mathscr{F}$ get "kicked out" of $\mathscr{F}'$ when $\mathscr{F}$ carries a weaker topology. Now the only remaining step is to find a topology such that $\mathscr{F}'$ shrinks exactly to $\mathcal{M}_f$. There are at least two such topologies: one defined by Schwartz (1954, p.100-101) and another, called the strict topology, whose definition can be found in Fremlin et al. (1972). Denoting τ_c either of these topologies, and $(\mathscr{C}_b)_c$ the space $\mathscr{C}_b$ equipped with τ_c, we finally get $((\mathscr{C}_b)_c)' = \mathcal{M}_f$, and thus:

Proposition 7 $((\mathscr{C}_b)_c/\mathbb{1})' = \mathcal{M}_f^0$. *Thus, if $\mathcal{H}_k \hookrightarrow (\mathscr{C}_b)_c$, then k is characteristic to $\mathscr{P}$ iff k is universal over the quotient space $((\mathscr{C}_b)_c/\mathbb{1})$.*

Proof That $((\mathscr{C}_b)_c)' = \mathcal{M}_f$ is proven in Fremlin et al. (1972, Thm. 1) or Schwartz (1954, p.100-101). Proposition 35.5 of Treves (1967) then implies $((\mathscr{C}_b)_c/\mathbb{1})' = \mathcal{M}_f^0$ (because $\mathcal{M}_f^0$

is the so-called *polar set* of $\mathbb{1}$; see Treves 1967). Theorem 6 implies the rest. ■

For our purposes, the exact definition of τ_c does not matter. What matters more is that τ_c is weaker than the usual topology of $\mathscr{C}_b$, so that if $\mathcal{H}_k \hookrightarrow \mathscr{C}_b$, then $\mathcal{H}_k \hookrightarrow (\mathscr{C}_b)_c$. Proposition 7 thus applies every time that $\mathcal{H}_k \subset \mathscr{C}_b$ (see Corollaries 3 and 4). However, we do not know of any practical application of Proposition 7, except that it completes our overall picture of the equivalences between universal, characteristic and s.p.d. kernels. Let us also mention that, similarly to Proposition 7, as $(\mathbb{C}^{\mathcal{X}})' = \mathcal{M}_\delta$, we also have $(\mathbb{C}^{\mathcal{X}}/\mathbb{1})' = \mathcal{M}_\delta^0$. So conditionnally s.p.d. kernels (meaning s.p.d. over $\mathcal{M}_\delta^0$) are universal to $\mathbb{C}^{\mathcal{X}}/\mathbb{1}$.

We now prove what we announced and used earlier: a kernel is characteristic to $\mathcal{P}$ iff it is characteristic to $\mathcal{M}_f^0$. We add a few other characterisations which are probably more useful in practice. They rely on the following observation: as $\mathcal{M}_f^0$ is a hyperplane of $\mathcal{M}_f$, saying that k is characteristic to $\mathcal{P}$ is almost the same than saying that it is characteristic to $\mathcal{M}_f$, i.e. $\int$s.p.d. (Thm. 6): after all, there is only one dimension needed to go from $\mathcal{M}_f^0$ to $\mathcal{M}_f$. Thus there should be a way to construct an $\int$s.p.d. kernel out of any characteristic kernel. This is what is described here and proven in Appendix A.3.

Theorem 8 (Characteristic Kernels) *Let k_0 be a kernel. The following is equivalent.*

(i) *k_0 is characteristic to $\mathcal{P}$.*
(ii) *k_0 is characteristic to $\mathcal{M}_f^0$.*
(iii) *There exists $\epsilon \in \mathbb{R}$ such that the kernel $k(x,y) := k_0(x,y) + \epsilon^2$ is $\int$s.p.d..*
(iv) *For all $\epsilon \in \mathbb{R}\backslash\{0\}$, the kernel $k(x,y) := k_0(x,y) + \epsilon^2$ is $\int$s.p.d..*
(v) *There exists an RKHS $\mathcal{H}_k$ with kernel k and a measure $\nu_0 \in \mathcal{M}_f \backslash \mathcal{M}_f^0$ such that k is characteristic to $\mathcal{M}_f$ and $k_0(x,y) = \langle \delta_x - \nu_0 \,,\, \delta_y - \nu_0 \rangle_k$.*

Under these conditions, k_0 and k induce the same MMD semi-metric in $\mathcal{M}_f^0$ and in $\mathcal{P}$.

We will use this theorem to prove Theorem 12. Intuitively, a characteristic kernel guarantees that any two different signed measures μ_1, μ_2 with same total mass get mapped to two different functions in the RKHS. This is captured by (ii) which arbitrarily focuses on the special case where the total mass is 0. When they have different total masses however, they may still get mapped to a same function f, except if, like in (iii) and (iv), we add a positive constant to the kernel. In that case, μ_1 and μ_2 get mapped to the functions $f + \mu_1(\mathcal{X})\mathbb{1}$ and $f + \mu_2(\mathcal{X})\mathbb{1}$ which are now different, because $\mu_1(\mathcal{X}) \neq \mu_2(\mathcal{X})$. Intuively, by adding a positive constant to our kernel, we added one dimension to the RKHS (carried by the function $\mathbb{1}$) that explicitly 'checks' if two measures have the same mass. Finally, (v) tells us that, out of any $\int$s.p.d. kernel k, we can construct a characteristic kernel k_0 that is not $\int$s.p.d. anymore and vice-versa.

4. Topology Induced by k

Remember that for any distribution D of a set of embeddable distributions $\mathcal{D}$ we defined $\|D\|_k := \|\Phi_k(D)\|_k$ and called $\|D\|_k$ the Maximum Mean Discrepancy (MMD) of D. Doing this defines a new topology on $\mathcal{D}$, in which a net D_α converges to D iff $\|D_\alpha - D\|_k$ converges to 0. (A reader unfamiliar with nets may think of them as sequences where the index α

can be continuous; see Berg et al. 1984.) In this section, we investigate how convergence in MMD compares with other types of convergences defined on $\mathscr{D}$ that we now shortly present.

We defined $\mathscr{D}$ as a subset of a dual space $\mathscr{F}'$, so $\mathscr{D}$ will carry the topology induced by $\mathscr{F}'$. Many topologies can be defined on dual spaces, but the two most prominent ones, which we will consider here, are the *weak-$*$* and the *strong* topology, denoted $w(\mathscr{F}', \mathscr{F})$ and $b(\mathscr{F}', \mathscr{F})$ respectively, or simply $w*$ and b. The weak-$*$ topology is the topology of pointwise convergence (where by 'point', we mean a function in $\mathscr{F}$), while the strong topology corresponds to the uniform convergence over the bounded subsets of $\mathscr{F}$ (see Eq. 4). Bounded sets of a TVS are defined in Appendix C (Definition 24). By default, we equip $\mathscr{F}'$ with the strong topology and sometimes write $\mathscr{F}'_b$ to emphasize it. When $\mathscr{F}$ is a Banach space, the strong topology of $\mathscr{F}'$ is the topology of the operator norm $\|D\|_{\mathscr{F}'} := \sup_{\|f\|_{\mathscr{F}} \leq 1} |D(f)|$. In particular, strong convergence in $\mathcal{M}_f = (\mathscr{C}_0)'$ means convergence in total variation (TV) norm and weak-$*$ convergence in $\mathcal{M}_f$ means convergence for any function $f \in \mathscr{C}_0$. On $\mathcal{M}_f$, we will also consider the topology of pointwise convergence over $\mathscr{C}_b$ (instead of $\mathscr{C}_0$). It is widely used in probability theory where it is known as the *weak* (or narrow) convergence topology. We will denote it by σ. Importantly, the weak and weak-$*$ topologies of $\mathcal{M}_f$ coincide on $\mathscr{P}$ (but not on $\mathcal{M}_f$) (Berg et al., 1984, Chap. 2, Cor. 4.3). Finally, we define the weak RKHS convergence of embeddable distributions, denoted by $w{-}k$, as the pointwise convergence over $\mathscr{H}_k$. Note that D_α converges in $w{-}k$ to D iff their embeddings converge weakly (or equivalently weakly-$*$) in $\mathscr{H}_k$, in the sense that, for any $f \in \mathscr{H}_k$, $\langle f\,,\,\Phi_k(D_\alpha)\rangle_k$ converges to $\langle f\,,\,\Phi_k(D)\rangle_k$. The following summarizes the different convergence types.

$$\begin{array}{rrlll}
D_\alpha \xrightarrow{b} D := & \sup_{f \in \mathscr{B}} |D_\alpha(f) - D(f)| \longrightarrow 0 & \forall \text{ bounded } \mathscr{B} \subset \mathscr{F} & D_\alpha \in \mathscr{F}' & \\
D_\alpha \xrightarrow{w*} D := & |D_\alpha(f) - D(f)| \longrightarrow 0 & \forall f \in \mathscr{F} & D_\alpha \in \mathscr{F}' & \\
\mu_\alpha \xrightarrow{\sigma} \mu := & |\mu_\alpha(f) - \mu(f)| \longrightarrow 0 & \forall f \in \mathscr{C}_b & \mu_\alpha \in \mathcal{M}_f & (4) \\
D_\alpha \xrightarrow{w-k} D := & |D_\alpha(f) - D(f)| \longrightarrow 0 & \forall f \in \mathscr{H}_k & D_\alpha \text{ embeddable} & \\
D_\alpha \xrightarrow{\|\cdot\|_k} D := & \|D_\alpha - D\|_k \longrightarrow 0 & & D_\alpha \text{ embeddable} &
\end{array}$$

4.1. Embeddings of Dual Spaces are Continuous

In this section, we show that the MMD topology is often weaker than other topologies τ defined on $\mathscr{D}$, meaning that if D_α converges to D in τ, then it also converges to D in MMD. Note that this is equivalent to saying that the KME of $\mathscr{D}_\tau$ (read '$\mathscr{D}$ equipped with τ') is continuous. We start with the following pretty coarse, yet very general result.

Proposition 9 *If $\mathscr{H}_k \hookrightarrow \mathscr{F}$, then $D_\alpha \xrightarrow{b} D \;\Rightarrow\; D_\alpha \xrightarrow{\|\cdot\|_k} D$ and $D_\alpha \xrightarrow{w*} D \;\Rightarrow\; D_\alpha \xrightarrow{w-k} D$.*

Proof Proposition 9 states that the KME is continuous when both $\mathscr{F}'$ and $\mathscr{H}_k$ carry their strong or their weak-$*$ topology, which we now show. From Diagram Eq.(3), we know that the KME is the composition of the conjugate restriction operator with the Riesz representer map. The Riesz representer map is a topological (anti-)isomorphism between $\mathscr{H}'_k$ and $\mathscr{H}_k$, thus continuous (see Appendix C). And the restriction map is the adjoint (or transpose) of the canonical embedding map $\imath : \begin{array}{ccc} \mathscr{H}_k & \longrightarrow & \mathscr{F} \\ f & \longmapsto & f \end{array}$, thus continuous when both $\mathscr{F}'$ and $\mathscr{H}'_k$ carry their weak-$*$ or strong topologies (Treves, 1967, Prop.19.5 & Corollary). ■

Let us briefly comment on this result. The statement $D_\alpha \xrightarrow{w*} D \;\Rightarrow\; D_\alpha \xrightarrow{w-k} D$ is actually obvious, because $\mathscr{H}_k \subset \mathscr{F}$. Concerning strong convergence, Proposition 9 implies that, if $\mathscr{F}$ is a Banach space, then any net that converges for the dual norm $\|\cdot\|_{\mathscr{F}'}$ converges in MMD. Applying this with $\mathscr{F} = \mathscr{C}_0$ and $\mathscr{F}' = \mathcal{M}_f$ shows that convergence in TV norm implies convergence in MMD, or equivalently, that the TV norm is stronger than the MMD. Similar reasoning can be used to show that the MMD is weaker than the so-called Kantorovich(-Wasserstein) and the Dudley norms (see Example 1 in Simon-Gabriel and Schölkopf 2016). These results can also be found in Sriperumbudur et al. (2010b). However, the authors there directly bounded the MMD semi-norm by the target norm. This has the advantage of giving concrete bounds, but is more difficult to generalize if $\mathscr{F}$ is not a Banach space.

Though very general, Proposition 9 is pretty weak, as it only compares a strong with a strong and a weak-$*$ with a weak(-$*$) topology. But how does the weak-$*$ topology on $\mathscr{F}'$ compare with the strong topology of $\mathscr{H}_k$: does weak-$*$ convergence imply convergence in MMD? This question is discussed in details in Simon-Gabriel and Schölkopf (2016, Sec.7). The short answer is: not always, but sometimes; it depends on the space $\mathscr{F}'$. For example, if $k \in \mathscr{C}^{(m,m)}$, then weak-$*$ convergence in $\mathscr{E}^m$ implies convergence in MMD; but weak-$*$ convergence in $\mathscr{D}^m_{L^1}$ usually does not imply MMD convergence when $\mathcal{X}$ is non-compact. For us, the only thing we will need later is to know what happens on $\mathcal{M}_+$, the set of finite positive measures. The following lemma shows that weak convergence in $\mathcal{M}_+$ usually implies MMD convergence.

Lemma 10 *A bounded kernel k is continuous iff: $\forall\, \mu_\alpha, \mu \in \mathcal{M}_+$, $\mu_\alpha \xrightarrow{\sigma} \mu \implies \mu_\alpha \xrightarrow{\|\cdot\|_k} \mu$.*

Proof We assume k bounded to ensure that any probability measure is embeddable. Now, suppose that weak convergence implies MMD convergence and take $x, y, x_0, y_0 \in \mathcal{X}$ such that $x \to x_0$ and $y \to y_0$. Then $\delta_x \overset{\sigma}{\to} \delta_{x_0}$ and $\delta_y \overset{\sigma}{\to} \delta_{y_0}$, so $\Phi_k(\delta_x) \to \Phi_k(\delta_{x_0})$ and $\Phi_k(\delta_y) \to \Phi_k(\delta_{y_0})$ in $\mathscr{H}_k$. And by continuity of the inner product:

$$k(x,y) = \left\langle \Phi_k(\delta_y)\,,\, \Phi_k(\delta_x) \right\rangle_k \to \left\langle \Phi_k(\delta_{y_0})\,,\, \Phi_k(\delta_{x_0}) \right\rangle_k = k(x_0, y_0)\ ,$$

so k is continuous. Conversely, suppose that k is continuous, and let $\mu_\alpha \overset{\sigma}{\to} \mu$ in $\mathcal{M}_+$. The tensor-product mapping $\begin{array}{ccc} \mathcal{M}_+(\mathcal{X}) & \longrightarrow & \mathcal{M}_+(\mathcal{X} \times \mathcal{X}) \\ \mu & \longmapsto & \mu \otimes \mu \end{array}$ is weakly continuous (Berg et al., 1984, Chap.2, Thm.3.3). So by applying $\bar{\mu}_\alpha \otimes \mu_\alpha$ to a bounded continuous kernel k, we get

$$\begin{aligned} \|\Phi_k(\mu_\alpha) - \Phi_k(\mu)\|_k^2 &= \iint k(x,y)\,\mathrm{d}(\mu_\alpha - \mu)(y)\,\mathrm{d}(\bar{\mu}_\alpha - \bar{\mu})(x) \\ &= [\bar{\mu}_\alpha \otimes \mu_\alpha](k) - [\bar{\mu} \otimes \mu_\alpha](k) - [\bar{\mu}_\alpha \otimes \mu](k) + [\bar{\mu} \otimes \mu](k) \longrightarrow 0\ . \quad \blacksquare \end{aligned}$$

4.2. When Does k Metrize the Topology of $\mathscr{F}'$?

So far we focused on the question: when does convergence in $\mathscr{D}$ imply convergence in MMD. We now seek the opposite: when does MMD-convergence imply convergence in $\mathscr{D}$?

First, the kernel *must* be characteristic to $\mathscr{D}$. Otherwise, the MMD does not define a distance but only a semi-distance, so that the induced topology would not be Hausdorff. Second, we will suppose that $\mathscr{F}$ is barreled. This is a technical, yet very general assumption that we use in the next theorem. The definition of a barreled space is given in Appendix C for completeness, but all that the reader should remember is that all Banach, Fréchet, Limit-Fréchet and *all function spaces defined in this paper are barreled,*[1] except $(\mathscr{C}_b^m)_c$.

Lemma 11 *Suppose that $\mathscr{F}$ is barreled, k is universal over $\mathscr{F}$, $\mathscr{H}_k \hookrightarrow \mathscr{F}$ and let $(D_\alpha)_\alpha$ be a bounded net in $\mathscr{F}'_b$. Then $D_\alpha \xrightarrow{w-k} D$ iff $D_\alpha \xrightarrow{w*} D$. Hence $D_\alpha \xrightarrow{\|.\|_k} D \;\Rightarrow\; D_\alpha \xrightarrow{w*} D$.*

Proof Proposition 32.5 of Treves (1967) shows that the weak topologies of $\mathscr{F}'$ and of $\mathscr{H}'_k$ coincide on so-called *equicontinuous* sets of $\mathscr{F}'$, and the Banach-Steinhaus theorem (see Appendix C) states that if $\mathscr{F}$ is barreled, then the equicontinuous sets of $\mathscr{F}'$ are exactly its bounded sets. This precisely means that if the net D_α is bounded in $\mathscr{F}'$, then $D_\alpha(f) \to D(f)$ for all $f \in \mathscr{F}$ iff it converges for all $f \in \mathscr{H}_k$. Now, if $\|D_\alpha - D\|_k \to 0$, then, by continuity of the inner product, $D_\alpha(f) - D(f) = \langle f \,,\, D_\alpha - D \rangle_k \to 0$ for any $f \in \mathscr{H}_k$. ■

Lemma 11 says that the weak-$*$ topologies of $\mathscr{F}'$ and of $\mathscr{H}_k$ coincide on subsets of $\mathscr{F}'$ that are bounded in the strong topology. But from the Banach-Steinhaus theorem (see App. C) we know that on barreled spaces it is equivalent to be bounded in strong or in weak topology. Hence the net D_α of Lemma 11 is bounded iff $\sup_\alpha |D_\alpha(f)| < \infty$ for all $f \in \mathscr{F}$. Nevertheless, it is not enough in general to show that $\sup_\alpha \|D_\alpha\|_k < \infty$. A bounded set in $\mathcal{M}_f$ is also a set whose measures have uniformly bounded total variation. The total variation of any probability measure being 1, $\mathscr{P}$ is bounded. So Lemma 11 shows that for continuous c_0-universal kernels, convergence of probability measures in MMD distance implies weak-$*$ convergence, which on $\mathscr{P}$ is the same as weak-convergence. But by Lemma 10 the reverse is true as well. Thus, *for a continuous c_0-universal kernel k, probability measures converge weakly iff they converge in MMD distance.* Such kernels are said to *metrize* the weak convergence on $\mathscr{P}$.

However, the condition that k be c_0-universal seems slightly too restrictive. Indeed, it is needed in Lemma 11 to ensure that the KME be characteristic to $\mathcal{M}_f$ (by Thm. 6 applied to $\mathscr{F} = \mathscr{C}_0$) so that the MMD be a metric over $\mathcal{M}_f$ (not only a semi-metric). But, to be a metric over $\mathscr{P}$, it would suffice that k be characteristic to $\mathscr{P}$, which is a slightly coarser assumption than c_0-universality. Is this condition enough to guarantee the metrization of weak-convergence in $\mathscr{P}$? The following theorem shows that it is.

Theorem 12 *A bounded kernel over a locally compact Hausdorff space $\mathcal{X}$ metrizes the weak convergence of probability measures iff it is continuous and characteristic (to $\mathscr{P}$).*

Proof [Theorem 12] If k metrizes the weak convergence over $\mathscr{P}$, then, by Lemma 10, k is continuous, and, for $\|.\|_k$ to be a norm, k needs to be characteristic. Conversely, if k is continuous, then by Lemma 10 weak convergence implies convergence in MMD. So it remains to show that MMD convergence implies weak convergence. To do so, we use Lemma 20 of the appendix, which states that for an $\int$s.p.d. kernel, MMD convergence of probability measures implies their weak convergence. Now k might not be $\int$s.p.d., but using Theorem 8(iv), we can transform it to a kernel $k_1 := k + 1$ which induces the same MMD metric

[1] $\mathbb{C}^{\mathcal{X}}$ is barreled, because it is a topological product $\prod_{\mathcal{X}} \mathbb{C}$ of barreled spaces. All other mentioned spaces are either Banach, Fréchet or Limit-Fréchet spaces, thus barreled (Treves, 1967, Prop. 33.2 & Cor.1-3).

over probability measures than k, but which is $\int$s.p.d. This concludes. ■

To the best of our knowledge, this is the first characterization of the class of kernels that metrize the weak-convergence of probability measures. For example Gaussian, Laplace, inverse-multiquadratic or Matérn kernels are continuous and characteristic, so they all metrize the weak convergence over $\mathscr{P}$. In general however, even if a kernel metrizes the weak convergence over $\mathscr{P}$, it usually does not metrize weak convergence over $\mathcal{M}_+$ or $\mathcal{M}_f$ (see Simon-Gabriel and Schölkopf 2016).

5. Kernel Mean Embeddings of Schwartz-Distributions

We extended KMEs of measures to Schwartz-distributions and showed that they are continuous, but we hardly said anything about what to do and how to work with distributions. We will now catch up by focusing on distributions only. In Section 5.1, we discuss and prove the Fubini and the Differentiation formulae featured in the introduction. In Section 5.2 we provide sufficient conditions for a translation-invariant kernel to be c_*^m-universal.

5.1. Distributional Calculus

Proposition 13 (Fubini) *Let D, T be two embeddable distributions into $\mathscr{H}_k$. Then:*

$$\begin{aligned}\langle D\,,\,T\rangle_k &= \iint k(x,y)\,\mathrm{d}D(y)\,\mathrm{d}\bar{T}(x) = \iint k(x,y)\,\mathrm{d}\bar{T}(x)\,\mathrm{d}D(y) \\ \|D\|_k^2 &= \iint k(x,y)\,\mathrm{d}D(y)\,\mathrm{d}\bar{D}(x) = \iint k(x,y)\,\mathrm{d}\bar{D}(x)\,\mathrm{d}D(y)\,,\end{aligned} \tag{5}$$

where $\iint k(x,y)\,\mathrm{d}D(y)\,\mathrm{d}\bar{T}(x)$ is to be understood as $\bar{T}(\mathscr{I})$ with $\mathscr{I}(x) = \int k(x,y)\,\mathrm{d}D(y)$.

Proof Definition 1 of a KME, together with the property that $k(y,x) = \overline{k(x,y)}$ leads to:

$$\begin{aligned}\langle D\,,\,T\rangle_k &= \int_x \left\langle \int_y k(.,y)\,\mathrm{d}D(y)\,,\,k(.,x)\right\rangle_k \mathrm{d}\bar{T}(x) \\ &= \int_x \overline{\left\langle k(.,x)\,,\,\int_y k(.,y)\,\mathrm{d}D(y)\right\rangle_k}\,\mathrm{d}\bar{T}(x) \\ &= \int_x \overline{\int_y \langle k(.,x)\,,\,k(.,y)\rangle_k\,\mathrm{d}\bar{D}(y)}\,\mathrm{d}\bar{T}(x) \\ &= \iint k(x,y)\,\mathrm{d}D(y)\,\mathrm{d}\bar{T}(x).\end{aligned}$$

To prove the right-most part of (5), use $\langle D\,,\,T\rangle_k = \overline{\langle T\,,\,D\rangle}_k$. ■

These formulae are well-known when D and T are probability measures. They show that if you know how to integrate a function (the kernel) w.r.t. a measure or a distribution, then you can compute its MMD norm. However, integrating w.r.t. a distribution that is not a measure can be tedious. But the following proposition gives us a way to convert an integration w.r.t. a distribution into an integration w.r.t. a measure.

Proposition 14 (Differentiation) *Let $k \in \mathscr{C}^{(m,m)}$ and $p \in \mathbb{N}^d$ such that $|p| \leq m$. A distribution D embeds into $\mathscr{H}_k$ via $\partial^{(0,p)}k$ iff $\partial^p D$ embeds into $\mathscr{H}_k$ via k. In that case,*

$$\Phi_k(\partial^p D) = (-1)^{|p|} \int [\partial^{(0,p)}k](.,x)\, \mathrm{d}D(x) = (-1)^{|p|}\, \Phi_{\partial^{(0,p)}k}(D)\,. \tag{6}$$

If moreover k is translation-invariant, then

$$\Phi_k(\partial^p D) = \partial^p[\Phi_k(D)]. \tag{7}$$

Proof The proof holds in the following equalities. For any $f \in \mathscr{H}_k$,

$$\begin{aligned}\left\langle f\,,\, \int k(.,x)\, \mathrm{d}[\partial^p D](x) \right\rangle_k &= \int \langle f\,,\, k(.,x)\rangle_k\, \mathrm{d}[\partial^p \bar{D}](x) = [\partial^p \bar{D}](f)\\ &= (-1)^{|p|}\bar{D}(\partial^p f)\\ &= (-1)^{|p|}\bar{D}(\left\langle f\,,\, \partial^{(0,p)}k(.,x)\right\rangle_k)\\ &= (-1)^{|p|}\int \left\langle f\,,\, \partial^{(0,p)}k(.,x)\right\rangle_k \mathrm{d}\bar{D}(x)\\ &= \left\langle f\,,\, (-1)^{|p|}\int \partial^{(0,p)}k(.,x)\, \mathrm{d}D(x)\right\rangle_k\,.\end{aligned}$$

The first line uses the definition of KMEs (1), the second the definition of distributional derivatives (see App. B), the third Lemma 19, the fourth line rewrites the previous line with our notation convention, and the fifth one uses again the definition of a weak integral (1). ■

Equation (7) describes a commutative diagram pictured in Figure 1: it states that with translation-invariant kernels, it is equivalent to take the (distributional) derivative of a distribution and embed it, or to embed it and take the (usual) derivative of the embedding. See Appendix B for an introduction to distributional derivatives. Note that for a signed measure μ with a $|p|$-times differentiable density q, the distributional derivative $\partial^p\mu$ is the signed measure with density $\partial^p_u q$, where ∂^p_u is the usual partial derivative operator. However, Proposition 14 becomes most useful when μ has no differentiable density, for example when μ is an empirical measure. Then there is no analytical formula for the derivative of μ, but we can still compute its KME analytically by using (6) or (7).

Example 1 *Let us illustrate Proposition 14 on KMEs of Gaussian probability measures μ_σ with density $q_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/\sigma^2}$ using a Gaussian kernel $k(x,y) = e^{-(x-y)^2}$. When σ goes to zero, μ_σ gets more and more peaked around 0 and converges weakly to the Dirac measure $\mu_0 := \delta_0$. The KME of μ_σ is easy to compute and using (7) we get*

$$\begin{aligned}\Phi_k(\mu_\sigma)(x) &= \frac{1}{\sqrt{1+2\sigma^2}}\, e^{-\frac{x^2}{1+2\sigma^2}}\\ \Phi_k(\partial\mu_\sigma)(x) &= \partial[\Phi_k(\mu_\sigma)] = -\frac{2x}{(1+2\sigma^2)^{3/2}}\, e^{-\frac{x^2}{1+2\sigma^2}}\quad,\end{aligned}$$

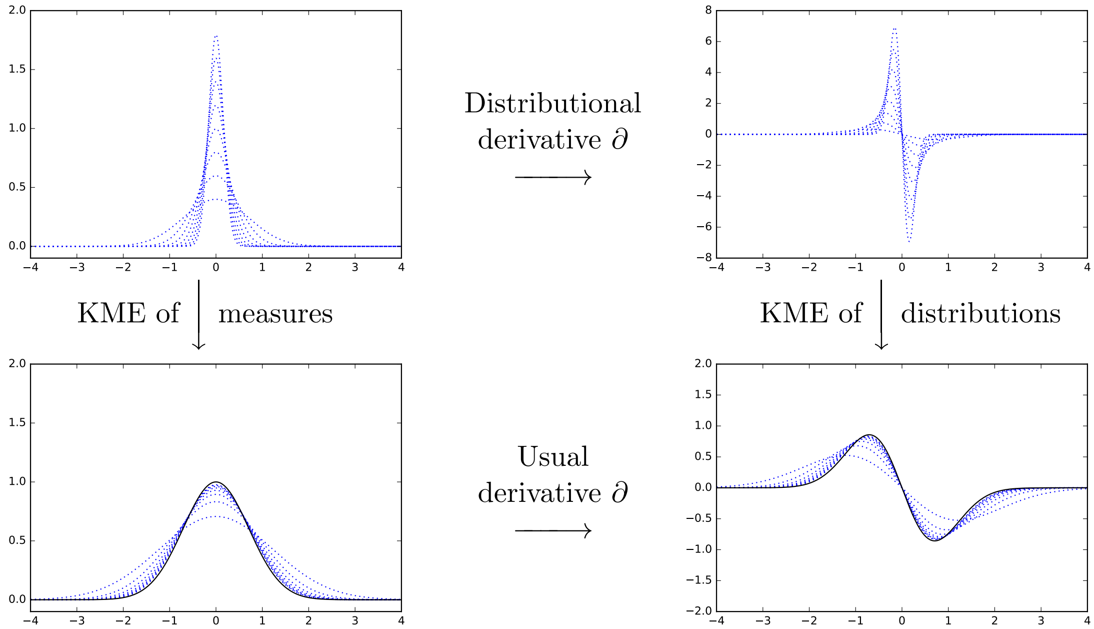

Figure 1: Densities of more and more peaked Gaussian probability measures μ_σ (top left) with their derivatives (top right) and their embeddings (below) using a Gaussian kernel (see Example 1). Equation (7) states that the diagram is commutative. When σ goes to 0, the Gaussians converge (weakly) to a Dirac mass δ_0, which has no density, but who's embedding is the solid black line (bottom left). The derivatives converge (weakly) to the Schwartz-distribution $\partial\delta_0$, which is not even a signed measure, but whose embedding (bottom right, black solid line) can easily be computed using (6) or (7). Moreover, the embeddings of μ_σ and $\partial\mu_\sigma$ converge (weakly) to the embeddings of μ_0 and $\partial\mu_0$, which illustrates Proposition 9.

where the formulae still hold when $\sigma = 0$. Figure 1 plots these embeddings for different σ's. Note that contrary to $\partial\mu_\sigma$ with $\sigma > 0$, $\partial\mu_0$ is not a signed measure (but a Schwartz-distribution) but it has a KME which, moreover, can easily be computed using (7)*. Notice also that on Figure 1 both the embeddings of μ_σ and $\partial\mu_\sigma$ converge (weakly) to the embeddings of μ_0 and $\partial\mu_0$. This illustrates Proposition 9.*

Theoretically, (6) can be used to convert the KME of *any* distribution into a sum of KMEs of measures. In other words, the integral w.r.t. a distribution appearing in (1) can be converted into a sum of integrals w.r.t. signed measures. Here is how. Given a measure $\mu \in \mathcal{M}_f = \mathcal{D}^0_{L^1}(\mathbb{R})$, we may differentiate μ and get a new distribution $\partial\mu$ which may or may not be itself a measure.[2] But in any case, what will follow shows that $\partial\mu$ is in $\mathcal{D}^1_{L^1}(\mathbb{R})$. Thus the space of distributions that can be written as a sum $\mu_0 + \partial\mu_1$ of two finite measures μ_1, μ_2 is a subspace of $\mathcal{D}^1_{L^1}(\mathbb{R})$ and we may wonder how big exactly it is. Schwartz (1954, around p.100) showed that it is exactly the space $\mathcal{D}^1_{L^1}(\mathbb{R})$. More generally, he showed:

Lemma 15 (Schwartz) *For any $m \leq \infty$ and any distribution in $D \in \mathcal{D}^m_{L^1}$ (resp. $D \in \mathcal{E}^m$) there exists a* finite *family of measures $\mu_p \in \mathcal{M}_f$ (resp. $\mu_p \in \mathcal{M}_c$) such that $D = \sum_{|p| \leq m} \partial^p \mu_p$.*

[2] Think for example of the Dirac measure: it is a measure, but not its derivative. See App. B.

Using (6), this means that the KME can be computed as $\sum_{|p|\leq m} \int \partial^{(0,p)} k(.,x)\, \mathrm{d}\mu_p(x)$, which gives a way to numerically compute the KME of distributions. As most distributions encountered in practice happen to be defined as measures or derivatives of some measures, this method is highly relevant in practice.

By combining Propositions 13 and 14, we get the following corollary.

Corollary 16 *Let $k \in \mathscr{C}^{(m,m)}$, $p \in \mathbb{N}^d$ with $|p| \leq m$, and let D, T be two distributions such that $\partial^p D$ and $\partial^p T$ embed into $\mathscr{H}_k$. Then*

$$\langle \partial^p D \,,\, \partial^p T \rangle_k = \langle D \,,\, T \rangle_{\partial^{(p,p)} k} \quad \text{and} \quad \|\partial^p D\|_k = \|D\|_{\partial^{(p,p)} k} \ .$$

Proof The proof reduces to the following equations.

$$\begin{aligned}
\langle \partial^p D \,,\, \partial^p T \rangle_k &\overset{(a)}{=} \left\langle \int \partial^{(0,p)} k(.,x)\, \mathrm{d}D(x) \,,\, \int \partial^{(0,p)} k(.,y)\, \mathrm{d}T(y) \right\rangle_k \\
&\overset{(b)}{=} \int \left\langle \partial^{(0,p)} k(.,y) \,,\, \partial^{(0,p)} k(.,x) \right\rangle_k \mathrm{d}D(y)\, \mathrm{d}\bar{T}(x) \\
&\overset{(c)}{=} \int \partial^{(p,p)} k(x,y)\, \mathrm{d}D(y)\, \mathrm{d}\bar{T}(x) \\
&\overset{(d)}{=} \langle D \,,\, T \rangle_{\partial^{(p,p)} k} \ ,
\end{aligned}$$

Equality (a) uses Proposition 14, (b) uses twice (on the left and on the right of the inner product) the definition of the weak integral (1), (c) uses Equation (9) proven in Appendix A which states that $\left\langle \partial^{(0,p)} k(.,y) \,,\, \partial^{(0,p)} k(.,x) \right\rangle_k = \partial^{(p,p)} k(x,y)$, and (d) uses (5) applied to the kernel $\partial^{(p,p)} k$. ■

Corollary 16 tells us that if we use $\partial^{(p,p)} k$—which is a kernel—to compute the MMD between two probability distributions D, T, then we are actually computing the MMD distance between their derivatives $\partial^p D$ and $\partial^p T$ with the kernel k. One could extend this corollary from (p,p) to (p,q) with $|q| \leq m$, yielding $\langle \partial^p D \,,\, \partial^q T \rangle_k = \int \partial^{(q,p)} k(x,y)\, \mathrm{d}D(y)\, \mathrm{d}\bar{T}(x)$. But in that case, $\partial^{(q,p)} k$ might not be a kernel anymore.

5.2. c^m- and c_0^m-Universal Kernels

Theorem 6 shows the equivalence between c_*^m-universality and characteristicness over $\mathscr{D}_{L^1}^m$ or $\mathscr{E}^m$. But neither the universality, nor the characteristic assumption seems easy to check in general. However, for translation-invariant kernels, meaning kernels that can be written as $k(x,y) = \psi(x-y)$ for some function ψ, we will now show that being characteristic to $\mathscr{P}$ or to $\mathscr{D}_{L^1}^m$ is one and the same thing, provided that $k \in \mathscr{C}_b^{(m,m)}$. Thus, any technique to prove that a kernel is characteristic may also be used to prove that it is characteristic to the much wider space $\mathscr{D}_{L^1}^m$. One of these techniques consists in verifying that the distributional Fourier transform $\mathscr{F}\psi$ has full support. The reader unfamiliar with distributional Fourier transforms may think of them as an extension of the usual Fourier transform—which is usually only defined on L^1, L^2 or $\mathcal{M}_f$—to wider function and distribution spaces. Let us mention that $\mathscr{F}\psi$ is exactly the unique positive, symmetric, finite measure appearing in Bochner's theorem (Wendland, 2004, Thm.6.6), and whose (usual) Fourier transform is ψ. We now successively present the result for $\mathscr{D}_{L^1}^m$, then for $\mathscr{E}^m$.

Theorem 17 *Let $k \in \mathscr{C}^{(m,m)}$ be a translation-invariant kernel $k(x,y) = \psi(x-y)$ with $\mathcal{X} = \mathbb{R}^d$, and $\mathscr{F}\psi$ its distributional Fourier transform. Then $\mathscr{D}^m_{L^1}$ embeds into $\mathscr{H}_k$ and the following are equivalent.*

(i) k is characteristic (to $\mathscr{P}$).
(ii) k is characteristic to $\mathscr{D}^m_{L^1}$.
(iii) $\mathscr{F}\psi$ has full support.

If moreover $\psi \in \mathscr{C}^m_0$, then k is c^m_0-universal iff it is c_0-universal.

Theorem 18 *Let $k \in \mathscr{C}^{(m,m)}$ be a translation-invariant kernel $k(x,y) = \psi(x-y)$ with $\mathcal{X} = \mathbb{R}^d$. If the support of $\mathscr{F}\psi$ has Lebesgue-measure > 0, then k is characteristic to $\mathscr{E}^m$.*

Proof [of Theorem 17] First, note that $\partial^{(p,p)}k(x,y) \leq \partial^{(p,p)}k(x,x)\partial^{(p,p)}k(y,y) = (\partial^{2p}\psi(0))^2$ for any $|p| \leq m$ (see Lemma 19 in Appendix A). Hence $k \in \mathscr{C}^{(m,m)}_b$, which, by Corollary 4, proves that $\mathscr{D}^m_{L^1}$ embeds into $\mathscr{H}_k$. Now suppose that (i) and (ii) are equivalent, then they are also equivalent to k being characteristic to $\mathcal{M}_f$. Using Theorem 6, we thus proved the last sentence. Now, (ii) clearly implies (i) and Theorem 9 of Sriperumbudur et al. (2010b) states that (i) and (iii) are equivalent. So it remains to show that (iii) implies (ii). We now sketch its proof and relegate the details to Appendix A.5. Let Λ be the finite positive measure from Bochner's theorem, such that $\psi = \mathscr{F}\Lambda$ and let $D \in \mathscr{D}^m_{L^1}$. Then

$$\begin{aligned} \|D\|_k^2 &= \iint \left(\int e^{i(x-y)\cdot\xi} \,\mathrm{d}\Lambda(\xi) \right) \mathrm{d}\bar{D}(x)\,\mathrm{d}D(y) \\ &\overset{(a)}{=} \int \left(\iint (e^{i(x-y)\cdot\xi})\,\mathrm{d}\bar{D}(x)\,\mathrm{d}D(y) \right) \mathrm{d}\Lambda(\xi) \\ &\overset{(b)}{=} \int |[\mathscr{F}D](\xi)|^2 \,\mathrm{d}\Lambda(\xi)\,, \end{aligned}$$

where $\cdot$ denotes the Euclidian inner-product on $\mathbb{R}^d$. Λ being positive, if it has full support, then $[\mathscr{F}D](\xi) = 0$ for almost all $\xi \in \mathcal{X}$. Thus $D = 0$. Assuming that (a) and (b) indeed hold, we just showed that if (iii), then $\|D\|_k = 0$ implies $D = 0$, meaning that k is s.p.d. to $\mathscr{D}^m_{L^1}$, which, with Theorem 6, proves (ii). We relegate the proof of (a) and (b) to Appendix A.5. ■

Proof [of Theorem 18] For any $D \in \mathscr{E}^m$, we can write, like before: $\|D\|_k^2 = \int |[\mathscr{F}D](\xi)|^2 \,\mathrm{d}\Lambda(\xi)$. But now, the Paley-Wiener-Schwartz theorem (Treves, 1967, Thm. 29.2) states that $\mathscr{F}D$ is an analytical function, so if its set of zeros has Lebesgue-measure > 0, then $\mathscr{F}D$ is the 0 function, so $D = 0$, showing that Φ_k is injective over $\mathscr{E}^m$. ■

These theorems show for example that Gaussian kernels are c^∞_0-universal and that the sinc kernel, defined on $\mathcal{X} = \mathbb{R}$ by $k(x,y) = \sin(x-y)/(x-y)$ (and 1 on the diagonal), is c^∞- but not c^∞_0-universal. When $\mathcal{X} = \mathbb{R}$, one can refine the conditions on the Fourier transform in Theorem 18 so that they become necessary and sufficient (Simon-Gabriel and Schölkopf, 2016, Theorem 41).

6. Conclusion

We first discuss how this work relates and contributes to the existing machine learning literature and then conclude.

6.1. Related Machine Learning Literature

Universal and characteristic kernels play an essential role in kernel methods and their theory. Universal kernels ensure consistency of many RKHS-based estimators in the context of regression and classification (Steinwart, 2001; Steinwart and Christmann, 2008), whereas characteristic kernels are of prime interest in any MMD-based algorithm, such as kernel two-sample tests (Gretton et al., 2007, 2012), HSIC independence tests (Gretton et al., 2008; Gretton and Györfi, 2010; Fukumizu et al., 2008), kernel density estimators (Sriperumbudur, 2016) and MMD-type GANs (Li et al., 2015; Dziugaite et al., 2015). The machine learning community gradually introduced more and more variants of universal kernels (Steinwart, 2001; Micchelli et al., 2006; Carmeli et al., 2006; Caponnetto et al., 2008), but instead of also introducing variants of characteristic kernels, it stuck to the original definition given by Fukumizu et al. (2004) which considered only characteristicness to $\mathscr{P}$. As a result, the literature started proving various links between the various variants of universal kernels and the only notion of characteristic kernels that it had. Eventually these notions were linked to $\int$s.p.d. and conditionally $\int$s.p.d. kernels (Fukumizu et al., 2004, 2008, 2009b,a; Gretton et al., 2007; Sriperumbudur et al., 2008, 2010a,b) and all known relations got summarized in a superb overview article by Sriperumbudur et al. (2011). However, by not introducing the notion of a characteristic kernel to something else than $\mathscr{P}$, the literature oversaw the fundamental dual link between universal, characteristic and s.p.d. kernels shown in Theorem 6 of this paper, which easily explains all the previously reported links.

Concerning the study of kernels that metrize the weak convergence of probability measures, in mathematics it dates back at least to Guilbart (1978), but it got introduced into the machine learning community only many years later by Sriperumbudur et al. (2010b). They gave new sufficient conditions to metrize the weak convergence, which then got improved by Sriperumbudur (2016)[Thm. 2]. However, by generalizing these sufficient conditions even further, Theorem 12 of this work is the first to provide conditions that are both sufficient *and necessary*, and that holds on any locally compact Hausdorff space $\mathcal{X}$ (which is more general than in the existing literature).

6.2. Future Work and Closing Remarks

This paper grouped various notions of universal, characteristic and s.p.d. kernels into three fundamental definitions—one for each—and showed that they are essentially equivalent: they describe the same family of kernels, but from dual perspectives. Using this duality link, we could systematically recover most of the previously known links, but also discovered new ones, such as the equivalence between characteristicness to $\mathscr{P}$ and universality over $(\mathscr{C}_b)_c/\mathbb{1}$; or between strict positive definiteness (over $\mathcal{M}_\delta$) and universality over $\mathbb{C}^{\mathcal{X}}$. We then compared the convergence in MMD with other convergence types of distributions and measures. Importantly, we showed that a bounded kernel metrizes the weak convergence of probability measures iff it is continuous and characteristic. Incidentally, we also showed that

KMEs over probability measures can be extended to generalized measures called Schwartz-distributions. For translation-invariant kernels, this extension preserves characteristicness, in the sense that a characteristic kernel to $\mathscr{P}$ will also be characteristic to $\mathscr{D}^m_{L^1}$. In all this work, we assumed $\mathscr{X}$ to be locally compact. Although this assumption fits many very general spaces, unfortunately, it does not contain any infinite-dimensional Banach space. So a main open question of this paper is whether our characterization of kernels that metrize the weak convergence of probability measures also applies to more general spaces, such as so-called Polish spaces, which are very standard spaces in probability theory. Finally, we also proved a few results that are specific to KMEs of distributions. Proposition 14 and its Corollary 16 on the embedding of derivatives for example show that these KMEs of distributions naturally appear when considering KMEs w.r.t. derivatives of kernels. We hope that they will in future lead to new insights and applications in machine learning.

Acknowledgments

We are deeply indebted to Alphabet for their support via a Google European Fellowship in Causal Inference accorded to Carl-Johann Simon-Gabriel. We thank Bharath Sriperumbudur for enlightening remarks, Jonas Peters for his kind and helpful advice, Ilya Tolstikhin for useful discussions and encouragements, and the two anonymous reviewers for their extensive, useful and contructive feedback.

Appendix A. Proofs

In this section, we gather all the complements to non fully proved theorems, propositions, corollaries or lemmas appearing in the main text. We start with a lemma that essentially follows from Corollary 4.36 of Steinwart and Christmann (2008), and which we will need a few times for the proofs.

Lemma 19 *Let* $k \in \mathscr{C}_b^{(m,m)}$ *and let* $\Phi : \begin{array}{ccc} \mathscr{X} & \longrightarrow & \mathscr{H}_k \\ x & \longmapsto & k(.,x) \end{array}$. *Then for any* $p \in \mathbb{N}^d$ *with* $|p| \leq m$, *the partial derivative* $\partial^p \Phi$ *exists, belongs to* $\mathscr{H}_k$, *is continuous and verifies* $\partial^p \Phi(x) = \partial^{(0,p)} k(.,x)$. *Moreover, for any* $f \in \mathscr{H}_k$, $\partial^p f$ *exists, belongs to* $\mathscr{H}_k$ *and verifies:*

$$\partial^p f(x) = \left\langle f \, , \, \partial^{(0,p)} k(.,x) \right\rangle_k \, . \tag{8}$$

Applied with $f = \partial^{(0,q)} k(.,y)$ *where* $|q| \leq m$ *also proves that*

$$\partial^{(p,q)} k(x,y) = \left\langle \partial^{(0,q)} k(.,y) \, , \, \partial^{(0,p)} k(.,x) \right\rangle_k \, . \tag{9}$$

Proof This Lemma is essentially proven in Corollary 4.36 and in its proof of Steinwart and Christmann (2008). We only added Equation (9), which is a straightforward consequence of (8), and the part stating that $\partial^p \Phi(x) = \partial^{(0,p)} k(.,x)$. This can be shown as follows. Steinwart and Christmann (2008) prove that $\partial^p \Phi$ exists and belongs to $\mathscr{H}_k$. Thus

$$[\partial^p \Phi(x)](y) = \langle \partial^p \Phi(x) \, , \, k(.,y) \rangle_k$$

$$\begin{aligned}
&= \left\langle \lim_{h\to 0} (\Phi(x+he_i) - \Phi(x))/h \,,\, k(.,y) \right\rangle_k \\
&= \lim_{h\to 0} (k(y, x+he_i) - k(y,x))/h \\
&= \partial^{(0,p)} k(y,x) \ ,
\end{aligned}$$

where we used the continuity of the inner product to swap limit and bracket signs. ■

A.1. Proof of Corollary 3

Proof Suppose that $\mathscr{H}_k \subset \mathscr{C}_0$. (i) clearly holds. Suppose (ii) was not met. Then let $x_n \in \mathcal{X}$ such that $k(x_n, x_n) = \|k(.,x_n)\|_k^2 \to \infty$. Thus $k(.,x_n)$ is unbounded. But $\langle f \,,\, k(.,x_n)\rangle_k = f(x_n)$ is bounded for any $f \in \mathscr{H}_k$, thus $k(.,x_n)$ is bounded (Banach-Steinhaus Theorem). Contradiction. Thus (ii) is met.

Conversely, suppose that (i) and (ii) hold. Let $\mathscr{H}_k^{\mathrm{pre}} := \operatorname{span}\{k(.,x) \,|\, x \in \mathcal{X}\}$. Then, $\mathscr{H}_k^{\mathrm{pre}} \subset \mathscr{C}_0$, and for any $f, g \in \mathscr{H}_k$, $\|f-g\|_\infty \leq \|f-g\|_k \, \|k\|_\infty$. Thus $\mathscr{H}_k^{\mathrm{pre}}$ continuously embeds into the *closed* $\mathscr{C}_0$, thus so does its $\|.\|_k$-closure, $\mathscr{H}_k$. The proof of the cases $\mathscr{H}_k \subset \mathscr{C}$ and $\mathscr{H}_k \subset \mathscr{C}_b$ are similar (see also Berlinet and Thomas-Agnan, 2004, Thm. 17). ■

A.2. Proof of Corollary 4

Proof Suppose that $k \in \mathscr{C}_b^{(m,m)}$. Then $\mathscr{H}_k^{\mathrm{pre}} \subset \mathscr{C}_b^m$ (Steinwart and Christmann, 2008, Corollary 4.36) and for any $x \in \mathcal{X}$, $f \in \mathscr{H}_k^{\mathrm{pre}}$, and $|p| \leq m$, we have $\|\partial^p f\|_\infty \leq \|f\|_k \left\| \sqrt{\partial^{(p,p)} k} \right\|_\infty$. Thus $\mathscr{H}_k^{\mathrm{pre}}$ continuously embeds into the *closed* space $\mathscr{C}_b^m$, thus so does its $\|.\|_k$-closure, $\mathscr{H}_k$. But, by definition of $(\mathscr{C}_b^m)_c$ is the space $\mathscr{C}_b$ equipped with a weaker topology (see Section 3), thus $\mathscr{C}_b^m \hookrightarrow (\mathscr{C}_b^m)_c$. Thus $\mathscr{H}_k \hookrightarrow (\mathscr{C}_b^m)_c$, which concludes. The proofs when $k \in \mathscr{C}$ or $k \in \mathscr{C}_0$ are similar. ■

A.3. Proof of Theorem 8

Proof Equivalence between *(i)* & *(ii)*. As KMEs are linear over $\mathcal{M}_f$, a kernel k is characteristic to $\mathscr{P}$ iff it is characteristic to $\mathscr{P} - P := \{\mu - P : \mu \in \mathscr{P}\}$, where P can be any fixed probability measure. This is equivalent to being characteristic to the linear span of $\mathscr{P} - P$. But the linear span of $\mathscr{P} - P$ is precisely $\mathcal{M}_f^0$, which concludes.

Equivalence of (ii) & (v): First of all, notice that, if (v), then k and k_0 define the same MMD on $\mathcal{M}_f^0$, because, for any $\mu \in \mathcal{M}_f^0$, $\mu(\mathbb{1}) = 0$, thus:

$$\begin{aligned}
\|\mu\|_{k_0}^2 &= \iint \langle \delta_x - \nu_0 \,,\, \delta_y - \nu_0 \rangle_k \, \mathrm{d}\bar{\mu}(x) \, \mathrm{d}\mu(y) \\
&= \iint k(x,y) \, \mathrm{d}\bar{\mu}(x) \, \mathrm{d}\mu(y) - \int \langle \delta_x \,,\, \nu_0 \rangle_k \, \mathrm{d}\bar{\mu}(x) \int \mathrm{d}\mu(y) \\
&\quad - \int \mathrm{d}\bar{\mu}(x) \int \langle \nu_0 \,,\, \delta_y \rangle_k \, \mathrm{d}\mu(y) - \|\nu_0\|_k^2 \iint \mathrm{d}\bar{\mu}(x) \, \mathrm{d}\mu(y)
\end{aligned}$$

$$= \|\mu\|_k^2 \ ,$$

Thus k_0 is characteristic to $\mathscr{M}_f^0$ iff k is also. Thus (v) implies (ii). Conversely, if k_0 is characteristic to $\mathscr{M}_f^0$, then k_0 is either characteristic to $\mathscr{M}_f$, in which case choosing $k_0 = k$ and $\nu_0 = 0$ fulfills the requirements of (v); or there exists a non zero measure $\nu_0 \in \mathscr{M}_f$ such that $\Phi_{k_0}(\nu_0) = 0$. As Φ_{k_0} is linear, we can choose $\nu_0(\mathbb{1}) = 1$ without loss of generality. Supposing now that we are in the latter case, the proof proceeds as follows.

(a) Show that the constant function $\mathbb{1} \notin \mathscr{H}_{k_0}$.
(b) Construct a new Hilbert space of functions of the form $\mathscr{H}_k = \operatorname{span} \mathbb{1} \oplus \mathscr{H}_{k_0}$.
(c) Show that it has a reproducing kernel k.
(d) Show that k_0 and k fulfill the requirements of (v).

(a) Suppose that $\mathbb{1} \in \mathscr{H}_{k_0}$. Then $1 = \bar{\nu}_0(\mathbb{1}) = \int \langle \mathbb{1} \, , \, k_0(.,x) \rangle_{k_0} \, \mathrm{d}\bar{\nu}_0(x) \overset{(*)}{=} \left\langle \mathbb{1} \, , \, \int k_0(.,x) \, \mathrm{d}\nu_0(x) \right\rangle_{k_0} = \langle \mathbb{1} \, , \, \Phi_{k_0}(\nu_0) \rangle_{k_0} = 0$, where in $(*)$ we use the definition of KMEs (1) . Contradiction. Thus $\mathbb{1} \notin \mathscr{H}_{k_0}$.

(b) Define $\mathscr{H} := \operatorname{span} \mathbb{1} \oplus \mathscr{H}_{k_0}$ and equip it with the inner product $\langle . \, , . \rangle$ that extends the inner product of $\mathscr{H}_{k_0}$ so, that

$$\mathbb{1} \perp \mathscr{H}_{k_0} \quad \text{and} \quad \|\mathbb{1}\| = 1 \, . \tag{10}$$

In other words, for any $f = c_f \mathbb{1} + f^\perp \in \mathscr{H}$ and any $g = c_g \mathbb{1} + g^\perp \in \mathscr{H}$:

$$\langle f \, , \, g \rangle := \left\langle f^\perp \, , \, g^\perp \right\rangle_{k_0} + c_f \bar{c_g}. \tag{11}$$

Obviously $\mathscr{H}$ is a Hilbert space of functions.

(c) We now construct k by first defining an injective embedding Φ and then showing that $k(x,y) := \langle \Phi(\delta_x) \, , \, \Phi(\delta_y) \rangle$ is a reproducing kernel with KME Φ.
As $\mathscr{M}_f^0$ is a hyperplane in $\mathscr{M}_f$ and $\nu_0 \in \mathscr{M}_f \backslash \mathscr{M}_f^0$, each measure $\mu \in \mathscr{M}_f$ can be decomposed uniquely in a sum: $\mu = \mu^\perp + \mu(\mathbb{1})\nu_0$ where $\mu^\perp = \mu - \mu(\mathbb{1})\nu_0 \in \mathscr{M}_f^0$. We may thus define the following linear embedding $\Phi : \ \mathscr{M}_f \ \longrightarrow \ \mathscr{H}$ by

$$\Phi(\mu) := \begin{cases} \Phi_{k_0}(\mu) & \text{if } \mu \in \mathscr{M}_f^0 \\ \mathbb{1} & \text{if } \mu = \nu_0 \end{cases} \quad \text{i.e.} \quad \begin{array}{rcl} \Phi(\mu) & := & \Phi_{k_0}(\mu^\perp) + \mu(\mathbb{1})\mathbb{1} \\ & = & \Phi_{k_0}(\mu) + \mu(\mathbb{1})\mathbb{1} \end{array} \ . \tag{12}$$

Noting that $\Phi(\mu)^\perp = \Phi(\mu^\perp) = \Phi_{k_0}(\mu^\perp) = \Phi_{k_0}(\mu)$ and using (11), we get

$$\forall f \in \mathscr{H}, \, \forall x \in \mathscr{X}, \ \langle f \, , \, \Phi(\delta_x) \rangle = \left\langle f^\perp \, , \, \Phi(\delta_x)^\perp \right\rangle_{k_0} + c_f = f^\perp(x) + c_f \mathbb{1}(x) = f(x) \, . \tag{13}$$

So by defining $k(x,y) := \langle \Phi(\delta_y) \, , \, \Phi(\delta_x) \rangle$ and applying (13) to $f = \Phi(\delta_y)$, we see that $\Phi(\delta_y) = k(.,y)$. Thus (13) may be rewritten as

$$\forall f \in \mathscr{H}, \forall x \in \mathscr{X}, \qquad \langle f \, , \, k(.,x) \rangle = f(x).$$

Thus $\mathscr{H}$ is an RKHS with reproducing kernel and Φ is its associated KME.

(d) As k_0 is characteristic to $\mathcal{M}_f^0$, Φ is injective over $\mathcal{M}_f^0$. And $\Phi(\nu_0) \in \mathcal{H}\backslash\Phi(\mathcal{M}_f^0)$. Thus Φ is injective over $\mathcal{M}_f$, so k is characteristic to $\mathcal{M}_f$. To conclude, (12) shows that

$$\begin{aligned}\langle \delta_y - \nu_0 \, , \, \delta_x - \nu_0 \rangle &= \langle \Phi_{k_0}(\delta_y) + (\delta_y - \nu_0)(\mathbb{1})\mathbb{1} \, , \, \Phi_{k_0}(\delta_x) + (\delta_x - \nu_0)(\mathbb{1})\mathbb{1} \rangle \\ &= \langle \Phi_{k_0}(\delta_y) + 0 \, , \, \Phi_{k_0}(\delta_x) + 0 \rangle \\ &= k_0(x, y) \ .\end{aligned}$$

Equivalence of (v) with (iii) & (iv): First, notice that the kernel k constructed in the proof of (v) $\Rightarrow$ (ii) verifies:

$$\begin{aligned}k(x, y) &= \langle \Phi(\delta_x) \, , \, \Phi(\delta_y) \rangle \\ &= \langle \Phi_{k_0}(\delta_x) + \delta_x(\mathbb{1})\mathbb{1} \, , \, \Phi_{k_0}(\delta_y) + \delta_y(\mathbb{1})\mathbb{1} \rangle \\ &= \langle \Phi_{k_0}(\delta_x) \, , \, \Phi_{k_0}(\delta_y) \rangle + \|\mathbb{1}\|^2 \\ &= k_0(x, y) + 1 \ ,\end{aligned}$$

where we used (10), (12) and the fact that by construction $\langle . \, , \, . \rangle$ coincides with $\langle . \, , \, . \rangle_{k_0}$ on $\mathcal{M}_f^0$. Thus the proof of (v) $\Rightarrow$ (ii) shows that, if k_0 characteristic to $\mathcal{M}_f^0$, then the kernel $k_0(x, y) + 1$ is characteristic to $\mathcal{M}_f$, thus $\int$s.p.d. (Thm. 6). $k(x, y) := k_0(x, y) + 1$ is $\int$s.p.d.. More generally, if instead of fixing $\|\mathbb{1}\|_k = 1$ in (10) we fixed $\|\mathbb{1}\|_k = \epsilon$ for some real $\epsilon > 0$, then we would have ended up with an $\int$s.p.d. kernel k verifying $k(x, y) := k_0(x, y) + \epsilon^2$. Thus (ii) implies (iii) and (iv). Conversely, given any kernel k of the previous form, the inner products defined by k and k_0 coincide on $\mathcal{M}_f^0$. So if k is characteristic to $\mathcal{M}_f^0$, then so is k_0. Thus (iii) or (iv) implies (ii). ■

A.4. Proof of Theorem 12 Continued

The proof of Theorem 12 used the following lemma.

Lemma 20 *Let k be a continuous, $\int$s.p.d. kernel and let $(\mu_\alpha)_\alpha$ be bounded in $\mathcal{M}_+$ (meaning $\sup_\alpha \|\mu_\alpha\|_{TV} < \infty$). Then $\mu_\alpha \xrightarrow{w-k} \mu \ \Rightarrow \ \mu_\alpha \xrightarrow{\sigma} \mu$. Consequently: $\mu_\alpha \xrightarrow{\|\cdot\|_k} \mu \ \Rightarrow \ \mu_\alpha \xrightarrow{\sigma} \mu$.*

Proof We will show that $\mu_\alpha(f) \to \mu(f)$ for any $f \in \mathscr{C}_c$. As $\mathscr{C}_c$ is a dense subset of $\mathscr{C}_0$ and μ_α is bounded, combining Prop. 32.5 and Thm. 33.2 of Treves (1967) then shows that $\mu_\alpha(f) \to \mu(f)$ for any $f \in \mathscr{C}_0$ (weak-$*$ convergence), which implies weak-convergence, $\mu_\alpha \xrightarrow{\sigma} \mu$ (Berg et al., 1984, Chap. 2, Cor. 4.3), and thus concludes.

Let K be a compact subset of $\mathcal{X}$. First, we show that there exists a function $h \in \mathcal{H}_k$ such that $h(x) > 0$ for any $x \in K$. To do so, let $f \in \mathscr{C}_b$ such that $f \geq 1$ on K. k being $\int$s.p.d. and $\mathcal{M}_f$ being the dual of $(\mathscr{C}_b)_c$, $\mathcal{H}_k$ is dense in $(\mathscr{C}_b)_c$ (Thm. 6). So we can find a sequence of functions $f_n \in \mathcal{H}_k$ that converges to f for the topology of $(\mathscr{C}_b)_c$. By definition of the topology of $(\mathscr{C}_b)_c$, this implies in particular that the restrictions of f_n to K converge in infinity norm, meaning: $\sup_{x \in K} |f_n(x) - f(x)| \to 0$. Thus, for a sufficiently large n, $f_n > 0$ on K, so we can take $h = f_n$.

Now, let us define the measures $h.\mu_\alpha$ as $[h.\mu_\alpha](f) = \mu_\alpha(hf)$ for any $f \in \mathscr{C}_b$. Then $\|h.\mu_\alpha\|_{TV} \leq \|h\|_\infty \|\mu_\alpha\|_{TV}$, so the new net $(h.\mu_\alpha)_\alpha$ is bounded. But bounded sets are relatively compact for the weak-$*$ topology $w(\mathcal{M}_f, \mathscr{C}_0)$. (Treves 1967, Thm. 33.2, or Banach-Alaoglu theorem). So we can extract a subnet $h.\mu_\beta$ of $h.\mu_\alpha$ that converges in weak-$*$ topology. Then $h.\mu_\beta$ is also a Cauchy-net for the weak-$*$ topology, meaning that for any $\epsilon > 0$ and any sufficiently large β, β':

$$|\mu_\beta(hf) - \mu_{\beta'}(hf)| \leq \epsilon, \quad \forall f \in \mathscr{C}_0 \ .$$

This inequality holds in particular for functions f whose support is contained in K, which we denote $f \in \mathscr{C}_c(K)$. But the mapping $f \longmapsto g := hf$ is a bijective map from $\mathscr{C}_c(K)$ to itself (because $h > 0$ on K), so we actually have $|\mu_\beta(g) - \mu_{\beta'}(g)| \leq \epsilon$ for any $g \in \mathscr{C}_c(K)$. But this holds for any compact subset K of $\mathcal{X}$. So the inequality also holds for any function $g \in \mathscr{C}_c(\mathcal{X})$, which shows that μ_β is a Cauchy-net for the topology of pointwise convergence in $\mathscr{C}_c(\mathcal{X})$, also known as the *vague* topology. But $\mathcal{M}_+$ is vaguely complete (Bourbaki, 2007, Chap.III, §1, n.9, Prop.14), so μ_β converges to a measure $\mu' \in \mathcal{M}_+$. But for any $f \in \mathscr{C}_c(\mathcal{X})$, $\mu'(f) = \lim_\beta \mu_\beta(f) = \lim_\alpha \mu_\alpha(f) = \mu(f)$, thus μ' and μ coincide on $\mathscr{C}_c(\mathcal{X})$, which is a dense subset of $\mathscr{C}_0$. Thus $\mu' = \mu$, and $\mu_\alpha(f) \to \mu(f)$ for any $f \in \mathscr{C}_c$. ■

Note that if we additionally supposed that $\mathscr{H}_k \hookrightarrow \mathscr{C}_0$ (meaning that k is c_0-universal), then Lemma 20 is a simple consequence of Lemma 11 and the fact that weak-$*$ and weak convergence coincide on $\mathscr{P}$.

A.5. Proof of Theorem 17 Continued

Proof We are left with proving (a) and (b). To do so, we will use the decomposition $D = \sum_{|p| \leq m} \partial^p \mu_p$ of Lemma 15. Indeed, k being in $\mathscr{C}_b^{(m,m)}$, by Corollary 4, $\partial^p \mu_p$ embeds into $\mathscr{H}_k$ for any $|p| \leq m$ and $\mu_p \in \mathcal{M}_f$. Thus

$$\begin{aligned}\langle \partial^p \mu_p \, , \, \partial^q \mu_q \rangle_k &= \langle \Phi_{\partial^{(0,p)}k}(\mu_p) \, , \, \Phi_{\partial^{(0,q)}k}(\mu_q) \rangle_k \\ &= \iint \left\langle \partial^{(0,p)} k(.,y) \, , \, \partial^{(0,q)} k(.,x) \right\rangle_k \mathrm{d}\bar{\mu}_q(x) \, \mathrm{d}\mu_p(y) \\ &= \iint \partial^{(q,p)} k(x,y) \, \mathrm{d}\bar{\mu}_q(x) \, \mathrm{d}\mu_p(y) \\ &= \iiint i^{|p+q|} \xi^{p+q} e^{i(x-y)\cdot\xi} \, \mathrm{d}\Lambda(\xi) \, \mathrm{d}\bar{\mu}_q(x) \, \mathrm{d}\mu_p(x) \ ,\end{aligned}$$

where for $\xi = (\xi_1, \ldots \xi_d) \in \mathbb{R}^d$, we defined $\xi^p := \xi_1^{p_1} \xi_2^{p_2} \cdots \xi_d^{p_d}$. The first line uses Proposition 14, the second line uses twice the definition of a weak integral (1), the third uses (9) from Lemma 19 and the fourth line uses the fact that $\partial^{(q,p)} k(x,y) = (-1)^{|p|} \partial^{p+q} \psi(x-y)$ and $\mathscr{F} \partial^{p+q} \psi = i^{|p+q|} \xi^{p+q} \mathscr{F} \psi = i^{|p+q|} \xi^{p+q} \Lambda$.

Let us denote $\xi^p \Lambda$ the measure defined by $\xi^p \Lambda(A) := \int_A \xi^p \, \mathrm{d}\Lambda(\xi)$. We will now show that $\xi^{p+q}\Lambda$ is finite, so that we can apply the usual Bochner theorem and permute the order of integrations. To do so, notice that $\partial^{(p,p)} k(x,y) = (-1)^{|p|} \partial^{2p} \psi(x-y)$ is a continuous kernel, thus, by Bochner's theorem, its associated measure Λ_∂ is finite and verifies $\mathscr{F} \Lambda_\partial = \partial^{2p} \psi$. But the usual calculus rules with Fourier transforms show that $\partial^{2p} \psi = (-i)^{|2p|} \xi^{2p} \Lambda$. Thus $\Lambda_\partial = i^{|p|} \xi^{2p} \Lambda$, showing that $\tilde{\Lambda}$ is a finite measure. Noting now that $2|\xi^{p+q}| \leq \xi^{2p} + \xi^{2q}$, this also implies that $\xi^{p+q}\Lambda$ is a finite measure. Consequently:

$$\langle \partial^p \mu_p \, , \, \partial^q \mu_q \rangle_k = \iiint i^{|p+q|} e^{i(x-y)} \, \mathrm{d}[\xi^{p+q}\Lambda](\xi) \, \mathrm{d}\bar{\mu}_q(x) \, \mathrm{d}\mu_p(x)$$

$$
\begin{aligned}
&= \iiint i^{|p+q|} e^{i(x-y)} \,\mathrm{d}\bar{\mu}_q(x) \,\mathrm{d}\mu_p(x) \,\mathrm{d}\tilde{\Lambda}(\xi) \\
&= \int i^{|p+q|} \xi^{p+q} \, \mathscr{F}\mu_q(\xi) \overline{\mathscr{F}\mu_p(\xi)} \,\mathrm{d}\Lambda(\xi) \\
&= \int [\mathscr{F}(\, \partial^p \mu_p)](\xi) \overline{[\mathscr{F}(\, \partial^q \mu_q)](\xi)} \,\mathrm{d}\Lambda(\xi).
\end{aligned}
$$

Thus, with the decomposition $D = \sum_{|p| \leq m} \partial^p \mu_p$, we get

$$
\begin{aligned}
\|D\|_k^2 = \left\| \sum_{|p| \leq m} \partial^p \mu_p \right\|_k^2 &= \int \sum_{|p|,|q| \leq m} [\mathscr{F}(\, \partial^p \mu_p)](\xi) \overline{[\mathscr{F}(\, \partial^q \mu_q)](\xi)} \,\mathrm{d}\Lambda(\xi) \\
&= \int | \sum_{|p| \leq m} [\mathscr{F}(\, \partial^p \mu_p)](\xi)|^2 \,\mathrm{d}\Lambda(\xi) \\
&= \int | \, \mathscr{F}D(\xi)|^2 \,\mathrm{d}\Lambda(\xi) \ ,
\end{aligned}
$$

where we used the linearity of the Fourier operator on the last line. ■

Appendix B. Short Introduction to Schwartz-Distributions

To introduce Schwartz-distributions, the first step is to notice that any continuous function f is uniquely characterized by the values taken by $f(\varphi) := \int \varphi(x) f(x) \,\mathrm{d}x$ when φ goes through $\mathscr{C}_c$. Rather than seeing f as a function that acts on points x in $\mathscr{X}$, we could thus equivalently see f as a linear functional that acts on other functions φ in $\mathscr{C}_c$ and takes its values in $\mathbb{C}$. Such functionals are called *linear forms*. We could do the same for measures: a signed measure μ is also characterized by the values of $\mu(\varphi) := \int \varphi(x) \,\mathrm{d}\mu(x)$. So we could also see it as a linear functional that acts on functions φ in $\mathscr{C}_c$. Doing so effectively identifies f with the signed measure μ_f that has density f, because both define the same linear form $\varphi \longmapsto \int \varphi(x) f(x) \,\mathrm{d}x$. So from this perspective, a function f becomes a particular kind of measure, and a measure μ a sort of 'generalized function'. Moreover, seen as linear forms over $\mathscr{C}_c$, f and μ are continuous in the sense that if φ_α converges to φ, then $\mu(\varphi_\alpha)$ converges to $\mu(\varphi)$. Thus, by definition, we just identified f and μ with elements of the dual of $\mathscr{C}_c$.

We may now ask whether there are other continuous linear forms over $\mathscr{C}_c$. The answer is negative and is given by the Riesz-Markov-Kakutani representer theorem (see Appendix C). It states that the dual of $\mathscr{C}_c$ is exactly the set of signed regular Borel measures $\mathscr{M}_r$, meaning that any continuous linear form over $\mathscr{C}_c$ can be written as $\varphi \longmapsto \int \varphi \,\mathrm{d}\mu(x)$ for some $\mu \in \mathscr{M}_r$, and can thus be identified with a measure μ. So it seems that our generalization of functions to measures using continuous linear forms is as general as it can get. But this is forgetting the following detail. To distinguish a measure μ from all the others in $\mathscr{M}_r$, we do not need to know the values $\mu(\varphi)$ for *all* functions φ of $\mathscr{C}_c$. Actually, it suffices to know them for all φ in $\mathscr{C}_c^\infty$. This is because $\mathscr{C}_c^\infty$ is a dense subset of $\mathscr{C}_c$. Thus for any $\varphi \in \mathscr{C}_c$, even if $\varphi \notin \mathscr{C}_c^\infty$, we can reconstruct the value $\mu(\varphi)$ by taking a sequence φ_α in $\mathscr{C}_c^\infty$ that converges to φ and noticing that, by continuity, $\mu(\varphi)$ is the limit of $\mu(\varphi_\alpha)$. So instead of

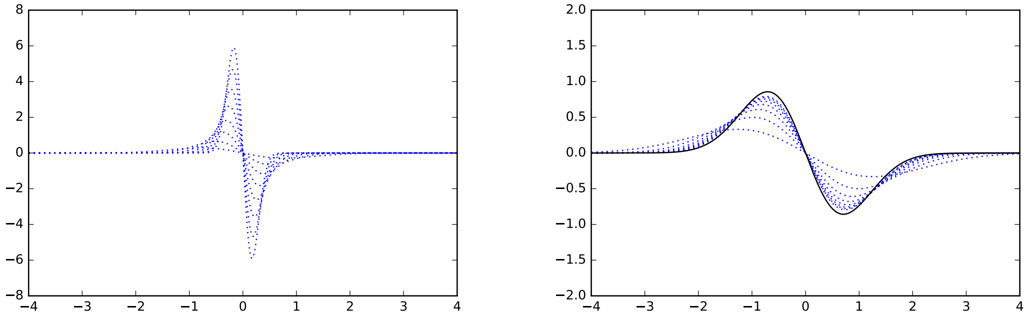

Figure 2: Left: the difference f_σ of two Gaussians that get closer and closer and more and more peaked with decreasing σ. Right: the KMEs of f_σ. Note the difference in the y-axis scale. f_σ converges to a *dipole*, which is not a measure, but a Schwartz-distribution. It cannot be represented as a function, but its KME can (black solid line). Note that the KMEs of f_σ seem to converge to the KME of the dipole.

seeing a function or a measure as an element of $(\mathscr{C}_c)'$, we could also see it as an element of $(\mathscr{C}_c^\infty)'$.

But do we gain anything from it? Yes indeed, because now, we can define linear functionals over $\mathscr{C}_c^\infty$ that we could not define over $\mathscr{C}_c$. For example, suppose that $\mathscr{X} = \mathbb{R}$ and consider the linear form d_x that, to each function φ associates its derivative $\partial\varphi(x)$ evaluated at x. This is a valid (continuous) linear form over $\mathscr{C}_c^\infty$—called a dipole in x—but it cannot be defined over $\mathscr{C}_c$, because not all continuous functions are differentiable. This example shows that, although each measure in $(\mathscr{C}_c)'$ can be seen as an element of $(\mathscr{C}_c^\infty)'$, the latter space contains many more linear forms which do not correspond to a signed measure. This bigger set of linear forms, which we denote $\mathscr{D}^\infty$, is called the set of Schwartz-distributions.

Now, why are distributions useful? First of all, because they can all be seen as limits of functions (Schwartz, 1978)[Theo. XV, Chap. III]. As an example, consider the sequence of functions

$$f_\sigma : \quad x \quad \longmapsto \quad \tfrac{1}{\sigma} g(\tfrac{x+\sigma}{\sigma}) - \tfrac{1}{\sigma} g(\tfrac{x-\sigma}{\sigma}) \ ,$$

where g is a Gaussian (see Figure 2). f_σ is the difference of two Gaussians that get closer and closer and more and more peaked with decreasing σ. Now, applying f_σ to a function $\varphi \in \mathscr{C}_c^\infty$, it is not difficult to see that $f_\sigma(\varphi)$ converges to $\partial\varphi(0) = d_0(\varphi)$ when $\sigma \to 0$. The dipole d_0 can thus be seen as a weak limit of the functions f_σ, although it is itself neither a function nor even a signed measure.

Another reason to use distributions is that many common linear operations can be extended to them (or to big subsets of them), such as differentiation, Fourier transformation and convolution. Let us show for example how to extend differentiation. If we want the distributional derivative ∂ to be an extension of the usual derivative, then of course we should require that $\partial\mu_f = \mu_{f'}$ whenever f is a continuously differentiable function over

$\mathcal{X} = \mathbb{R}$ whose usual derivative is f'. Now, by integration by part, we get, for any $\varphi \in \mathscr{C}_c^\infty$:

$$\mu_{f'}(\varphi) = \int f'\varphi = -\int f\varphi' = -\mu_f(\varphi') \ .$$

This suggests to define the derivative of any $D \in \mathscr{D}^\infty$ as $\partial^p D(\varphi) := (-1)^{|p|} D(\partial^p \varphi)$ for any $\varphi \in \mathscr{C}_c^\infty$. Doing so, we just defined a notion of differentiation that is compatible with the usual differentiation and makes *any* distribution infinitely many times differentiable. In particular, any function and any measure is infinitely differentiable in this distributional sense. Moreover, if a sequence of differentiable functions f_n converges to a distribution D (in the sense that $f_n(\varphi)$ converges to $D(\varphi)$ for any φ), then their usual derivatives f_n' converges to ∂D (in the same *distributional* sense). All this makes distributions extremely useful for solving linear differential equations and more generally for physicists. Last but not least, note that, by construction, if Q is a probability measure with smooth density q, then $\partial^p Q$ is the signed measure with density $\partial^p q$.

Appendix C. Other Background Material

Formally, a topological vector space (TVS) $\mathscr{E}$ is a vector space equipped with a topology that is *compatible* with its linear structure, in the sense that the addition $\mathscr{E} \times \mathscr{E} \longrightarrow \mathscr{E}$ and scalar multiplication $\mathbb{C} \times \mathscr{E} \longrightarrow \mathscr{E}$ become continuous for this topology (when their domains are equipped with the product topology). This makes the topology translation-invariant and hence completely defined by the neighborhoods of the origin. A TVS is locally convex (loc. cv.) if there exists a basis of (origin-) neighborhoods consisting of convex sets only. Obviously, the origin-centered balls of any semi-norm are convex. But interestingly, one can show that a TVS is loc. cv. iff its topology can be defined by a family of (continuous) semi-norms. So we can think of loc. cv. TVSs as "multi-normed" spaces, i.e. where convergence is given by a family of possibly multiple semi-norms $(\|.\|_\alpha)_{\alpha\in\mathscr{I}}$ (where the index set $\mathscr{I}$ can be uncountable). If this family contains only a single norm, $\mathscr{E}$ is a normed space. The origin-centered balls of these semi-norms are actually not only convex, they are *barrels*.

Definition 21 (Barrel) *A subset T of a TVS $\mathscr{E}$ is called a* barrel *if it is*

(i) absorbing*: for any $f \in \mathscr{E}$, there exists $c_f > 0$ such that $f \in c_f T$;*
(ii) balanced*: for any $f \in \mathscr{E}$, if $f \in T$ then $\lambda f \in T$ for any $\lambda \in \mathbb{C}$ with $|\lambda| \leq 1$;*
(iii) convex *;*
(iv) closed*.*

Given that the topology of loc. cv. TVS can be defined by a family of semi-norms, it is not surprising that in loc. cv. spaces there always exists a basis of origin-neighborhoods consisting only of barrels. However, there might be barrels that are not a neighborhood of 0. This leads to

Definition 22 (Barreled spaces) *A TVS is* barreled *if any barrel is a neighborhood of the origin.*

Although many authors include local convexity in the definition, in general, a barreled space need not be loc. cv. Barreled spaces were introduced by Bourbaki, because they were well-suited for the following generalization of the celebrated *Banach-Steinhaus* theorem.

Theorem 23 (Banach-Steinhaus) *Let $\mathscr{E}$ be a barreled TVS, $\mathscr{F}$ be a loc. cv. TVS, and let $L(\mathscr{E}, \mathscr{F})$ be the set of continuous linear maps form $\mathscr{E}$ to $\mathscr{F}$. For any $H \subset L(\mathscr{E}, \mathscr{F})$ the following properties are equivalent:*

(i) H is equicontinuous.
(ii) H is bounded for the topology of pointwise convergence.
(iii) H is bounded for the topology of bounded convergence.

When $\mathscr{E}$ is a normed space and $\mathscr{F} = \mathbb{C}$, then $L(\mathscr{E}, \mathscr{F})$ is by definition $\mathscr{E}'$. With $\|.\|_{\mathscr{E}'}$ being the dual norm in $\mathscr{E}'$, the equivalence of (ii) and (iii) states that

$$\left(\forall f \in \mathscr{E}, \ \sup_{h \in H} |h(f)| < \infty\right) \quad \Longleftrightarrow \quad \sup_{h \in H} \|h\|_{\mathscr{E}'} < \infty \,.$$

Obviously, to understand the content of the Banach-Steinhaus theorem, one needs the definition of a bounded set. Let us define them now.

When $\mathscr{E}$ is a normed space, then a subset B of $\mathscr{E}$ is called *bounded* if $\sup_{f \in B} \|f\|_{\mathscr{E}} < \infty$. In a more general loc. cv. TVS $\mathscr{E}$, where the topology is given by a family of semi-norms $(\|.\|_\alpha)_{\alpha \in \mathscr{I}}$, a subset B of $\mathscr{E}$ is called *bounded* if, for any $\alpha \in \mathscr{I}$, $\sup_{f \in B} \|f\|_\alpha < \infty$. This can be shown equivalent to the following, more usual definition.

Definition 24 (Bounded Sets in a TVS) *A subset B of a TVS $\mathscr{E}$ is* bounded*, if, for any neighborhood $U \subset \mathscr{E}$ of the origin, there exists a real $c_B > 0$ such that $B \subset c_B U$.*

Note that the notion of boundedness depends on the underlying topology. By default, a bounded set of some dual space $\mathscr{E} = \mathscr{F}'$ designates a set that is bounded for the strong dual topology. We now move on to an unrelated topic: the Riesz Representation theorem for Hilbert spaces. Most of this paper relies on this one theorem.

Theorem 25 (Riesz Representation Theorem for Hilbert Spaces) *A Hilbert space $\mathscr{H}$ and its topological dual $\mathscr{H}'$ are isometrically (anti-) isomorphic via the Riesz representer map*

$$\begin{array}{rccl} \imath : & \mathscr{H} & \longrightarrow & \mathscr{H}' \\ & f & \longmapsto & D_f := \left\{ \begin{array}{ccc} \mathscr{H} & \longrightarrow & \mathbb{C} \\ g & \longmapsto & \langle g \, , \, f \rangle \end{array} \right. \end{array} .$$

In particular, for any continuous linear form $D \in \mathscr{H}'$, there exists a unique element $f \in \mathscr{H}$, called the Riesz representer of D, *such that*

$$\forall g \in \mathscr{H}, \qquad D(g) = \langle g \, , \, f \rangle \ .$$

Note that "anti" in "anti-isomorphic" simply means that, instead of being linear, $\imath$ is anti-linear: for any $\lambda \in \mathbb{C}$ and $f \in \mathscr{H}$, $\imath(\lambda f) = \bar{\lambda} \imath(f)$. Often, we prefer to say that $\mathscr{H}$ is isometrically isomorphic to $\overline{\mathscr{H}}'$, where $\overline{\mathscr{H}}'$ denotes the conjugate of $\mathscr{H}$, where the scalar multiplication is replaced by $(\lambda, f) \longmapsto \bar{\lambda} f$. $\mathcal{H}_k'$ and $\overline{\mathcal{H}_k}'$ are obviously isomorphic via the complex conjugation map $D \longmapsto \bar{D}$.

The Riesz representation theorem for Hilbert spaces is not to be confounded with the following theorem, also known as the Riesz—or Riesz-Markov-Kakutani—representation theorem. In this paper, we always refer to the latter as the Riesz-Markov-Kakutani representation theorem. This theorem has numerous variants, depending on which dual pair $(\mathcal{E}, \mathcal{E}')$ one uses. Here we state it for $\mathcal{E} = \mathscr{C}_0$.

Theorem 26 (Riesz-Markov-Kakutani) *Let $\mathcal{X}$ be a locally compact Hausdorff space. The spaces $\mathcal{M}_f(\mathcal{X})$ and $(\mathscr{C}_0(\mathcal{X}))'$ are isomorphic, both algebraically and topologically via the map*

$$\begin{array}{rccl} \imath : & \mathcal{M}_f(\mathcal{X}) & \longrightarrow & (\mathscr{C}_0(\mathcal{X}))' \\ & \mu & \longmapsto & D_\mu := \left\{ \begin{array}{ccc} \mathscr{C}_0 & \longrightarrow & \mathbb{C} \\ \varphi & \longmapsto & \int \varphi \, \mathrm{d}\mu \end{array} \right. \end{array} .$$

In other words, for any continuous linear form D over $\mathscr{C}_0(\mathcal{X})$, there exists a unique finite Borel measure $\mu \in \mathcal{M}_f$ such that, for any test function $\varphi \in \mathscr{C}_0(\mathcal{X})$, $D(\varphi) = \int \varphi \, \mathrm{d}\mu$. Moreover, $\sup_{\|\varphi\|_\infty \leq 1} D(\varphi) = |\mu|(\mathcal{X})$, or in short: $\|D\|_{(\mathscr{C}_0)'} = \|\mu\|_{TV}$, where $\|\mu\|_{TV}$ denotes the total variation norm of μ. This is why, in this paper, we identify $\mathcal{M}_f$—a space of σ-additive set functions—with $\mathcal{M}_f$—a space of linear functionals.

In this paper, to embed a space of measures into an RKHS $\mathcal{H}_k$ we successively apply both Riesz representation theorems: If $\mathcal{H}_k$ embeds continuously into $\mathscr{C}_0$, then $(\mathscr{C}_0)'$ embeds continuously into $\overline{\mathcal{H}_k}'$, via the embedding map Φ_k. But $(\mathscr{C}_0)' = \mathcal{M}_f$ (Riesz-Markov-Kakutani Representation) and $\overline{\mathcal{H}_k}' = \mathcal{H}_k$ (Riesz Representation). Thus Φ_k may also be seen as an embedding of $\mathcal{M}_f$ into $\mathcal{H}_k$.

For a further introduction to TVSs and the theorems mentioned here, we suggest Treves (1967).

References

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups Theory of Positive Definite and Related Functions*. Springer, 1984.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

N. Bourbaki. *Intégration - Chapitres 1-4*. Springer, reprint of the 1965 original edition, 2007.

A. Caponnetto, C. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(7):1615–1646, 2008.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

M. Duc-Jacquet. *Approximation des Fonctionnelles Linéaires sur les Espaces Hilbertiens Autoreproduisants*. PhD thesis, Université Joseph-Fourier - Grenoble I, 1973.

G.K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

D. H. Fremlin, D. J. H. Garling, and R. G. Haydon. Bounded measures on topological spaces. *Proceedings of the London Mathematical Society*, s3-25(1):115–136, 1972.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimensionality reduction for supervised learning. *Journal of Machine Learning Research*, 5(12):73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Neural Information Processing Systems*, 2008.

K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Neural Information Processing Systems*, 2009a.

K. Fukumizu, A. Gretton, B. Schölkopf, and B. Sriperumbudur. Characteristic kernels on groups and semigroups. In *Neural Information Processing Systems*, 2009b.

A. Gretton and L. Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, 2005.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Neural Information Processing Systems*, 2007.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Neural Information Processing Systems*, 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

C. Guilbart. *Etude des Produits Scalaires sur l'Espace des Mesures: Estimation par Projections*. PhD thesis, Université des Sciences et Techniques de Lille, 1978.

O. Lehtö. Some remarks on the kernel function in Hilbert function space. *Annales Academiae Scientiarum Fennicae*, 109:6, 1952.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, 2015.

D. Lopez-Paz, P. Hennig, and B. Schölkopf. The randomized dependence coefficient. In *Neural Information Processing Systems*, 2013.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12):2651–2667, 2006.

Š. Schwabik. *Topics in Banach Space Integration*. Number 10 in Series in Real Analysis. World Scientific, 2005.

L. Schwartz. Espaces de fonctions différentiables à valeurs vectorielles. *Journal d'Analyse Mathématique*, 4(1):88–148, 1954.

L. Schwartz. *Théorie des Distributions*. Hermann, 1978.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions – arXiv version. arXiv: 1604.05251, 2016.

B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Conference On Learning Theory*, 2008.

B. K. Sriperumbudur, K. Fukumizu, and G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics*, 2010a.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010b.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(12):67–93, 2001.

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

F. Treves. *Topological Vector Spaces, Distributions and Kernels*. Academic Press, 1967.

H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.