Journal of Machine Learning Research 18 (2018) 1-29 Submitted 8/17; Revised 6/18; Published 7/18

# Characteristic and Universal Tensor Product Kernels

**Zoltán Szabó** ZOLTAN.SZABO@POLYTECHNIQUE.EDU
*CMAP, École Polytechnique*
*Route de Saclay, 91128 Palaiseau, France*

**Bharath K. Sriperumbudur** BKS18@PSU.EDU
*Department of Statistics*
*Pennsylvania State University*
*314 Thomas Building*
*University Park, PA 16802*

**Editor:** Francis Bach

## Abstract

Maximum mean discrepancy (MMD), also called energy distance or N-distance in statistics and Hilbert-Schmidt independence criterion (HSIC), specifically distance covariance in statistics, are among the most popular and successful approaches to quantify the difference and independence of random variables, respectively. Thanks to their kernel-based foundations, MMD and HSIC are applicable on a wide variety of domains. Despite their tremendous success, quite little is known about when HSIC characterizes independence and when MMD with tensor product kernel can discriminate probability distributions. In this paper, we answer these questions by studying various notions of characteristic property of the tensor product kernel.

**Keywords:** tensor product kernel, kernel mean embedding, characteristic kernel, $\mathcal{I}$-characteristic kernel, universality, maximum mean discrepancy, Hilbert-Schmidt independence criterion

## 1. Introduction

Kernel methods (Schölkopf and Smola, 2002) are among the most flexible and influential tools in machine learning and statistics, with superior performance demonstrated in a large number of areas and applications. The key idea in these methods is to map the data samples into a possibly infinite-dimensional feature space—precisely, a reproducing kernel Hilbert space (RKHS; Aronszajn, 1950)—and apply linear methods in the feature space, without the explicit need to compute the map. A generalization of this idea to probability measures, i.e., mapping probability measures into an RKHS (Berlinet and Thomas-Agnan, 2004, Chapter 4; Smola et al., 2007) has found novel applications in nonparametric statistics and machine learning. Formally, given a probability measure $\mathbb{P}$ defined on a measurable space $\mathcal{X}$ and an RKHS $\mathcal{H}_k$ with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel (which is symmetric and positive definite), $\mathbb{P}$ is embedded into $\mathcal{H}_k$ as

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, \mathrm{d}\mathbb{P}(x) =: \mu_k(\mathbb{P}), \tag{1}$$

where $\mu_k(\mathbb{P})$ is called the *mean element* or *kernel mean embedding* of $\mathbb{P}$. The *mean embedding* of $\mathbb{P}$ has lead to a new generation of solutions in two-sample testing (Baringhaus and Franz, 2004; Székely and Rizzo, 2004, 2005; Borgwardt et al., 2006; Harchaoui et al., 2007; Gretton et al., 2012), goodness-of-fit testing (Chwialkowski et al., 2016; Liu et al., 2016; Jitkrittum et al., 2017b; Balasubramanian et al., 2017), domain adaptation (Zhang et al., 2013) and generalization (Blanchard et al., 2017), kernel belief propagation (Song et al., 2011), kernel Bayes' rule (Fukumizu et al., 2013), model criticism (Lloyd et al., 2014; Kim et al., 2016), approximate Bayesian computation (Park et al., 2016), probabilistic programming (Schölkopf et al., 2015), distribution classification (Muandet et al., 2011; Zaheer et al., 2017), distribution regression (Szabó et al., 2016; Law et al., 2018) and topological data analysis (Kusano et al., 2016). A recent survey on the topic is provided by Muandet et al. (2017).

Crucial to the success of the mean embedding based representation is whether it encodes all the information about the distribution, in other words whether the map in (1) is injective in which case the kernel is referred to as *characteristic* (Fukumizu et al., 2008; Sriperumbudur et al., 2010). Various characterizations for the characteristic property of $k$ is known in the literature (Fukumizu et al., 2008, 2009; Sriperumbudur et al., 2010; Gretton et al., 2012) using which the popular kernels on $\mathbb{R}^d$ such as Gaussian, Laplacian, B-spline, inverse multiquadrics, and the Matérn class are shown to be characteristic. The characteristic property is closely related to the notion of *universality* (Steinwart, 2001; Micchelli et al., 2006; Carmeli et al., 2010; Sriperumbudur et al., 2011)—$k$ is said to be universal if the corresponding RKHS $\mathcal{H}_k$ is dense in a certain target function class, for example, the class of continuous functions on compact domains—and the relation between these notions has recently been explored by Sriperumbudur et al. (2011); Simon-Gabriel and Schölkopf (2016).

Based on the mean embedding in (1), Smola et al. (2007) and Gretton et al. (2012) defined a semi-metric, called the maximum mean discrepancy (MMD) on the space of probability measures:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k},$$

which is a metric iff $k$ is characteristic. A fundamental application of MMD is in non-parametric hypothesis testing that includes two-sample (Gretton et al., 2012) and independence tests (Gretton et al., 2008). Particularly in independence testing, as a measure of independence, MMD measures the distance between the joint distribution $\mathbb{P}_{XY}$ and the product of marginals $\mathbb{P}_X \otimes \mathbb{P}_Y$ of two random variables $X$ and $Y$ which are respectively defined on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, with the kernel $k$ being defined on $\mathcal{X}\times\mathcal{Y}$. As aforementioned, if $k$ is characteristic, then $\mathrm{MMD}_k(\mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y) = 0$ implies $\mathbb{P}_{XY} = \mathbb{P}_X \otimes \mathbb{P}_Y$, i.e., $X$ and $Y$ are independent. A simple way to define a kernel on $\mathcal{X}\times\mathcal{Y}$ is through the tensor product of kernels $k_X$ and $k_Y$ defined on $\mathcal{X}$ and $\mathcal{Y}$ respectively: $k = k_X \otimes k_Y$, i.e., $k\left((x,y),(x',y')\right) = k_X(x,x')k_Y(y,y')$, $x,x' \in \mathcal{X}$, $y,y' \in \mathcal{Y}$, with the corresponding RKHS $\mathcal{H}_k = \mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}$ being the tensor product space generated by $\mathcal{H}_{k_X}$ and $\mathcal{H}_{k_Y}$. This means, when $k = k_X \otimes k_Y$,

$$\mathrm{MMD}_k(\mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y) = \|\mu_{k_X\otimes k_Y}(\mathbb{P}_{XY}) - \mu_{k_X\otimes k_Y}(\mathbb{P}_X \otimes \mathbb{P}_Y)\|_{\mathcal{H}_{k_X}\otimes\mathcal{H}_{k_Y}}. \tag{2}$$

In addition to the simplicity of defining a joint kernel $k$ on $\mathcal{X}\times\mathcal{Y}$, the tensor product kernel offers a principled way of combining inner products ($k_X$ and $k_Y$) on domains that can correspond to different modalities (say images, texts, audio). By exploiting the isomorphism between tensor product Hilbert spaces and the space of Hilbert-Schmidt operators[1], it follows from (2) that

$$\mathrm{MMD}_k(\mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y) = \|C_{XY}\|_{\mathrm{HS}} =: \mathrm{HSIC}_k(\mathbb{P}_{XY}), \tag{3}$$

which is the Hilbert-Schmidt norm of the cross-covariance operator $C_{XY} := \mu_{k_X \otimes k_Y}(\mathbb{P}_{XY}) - \mu_{k_X}(\mathbb{P}_X) \otimes \mu_{k_Y}(\mathbb{P}_Y)$ and is known as the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton et al., 2005a). HSIC has enjoyed tremendous success in a variety of applications such as independent component analysis (Gretton et al., 2005a), feature selection (Song et al., 2012), independence testing (Gretton et al., 2008; Jitkrittum et al., 2017a), post selection inference (Yamada et al., 2018) and causal detection (Mooij et al., 2016; Pfister et al., 2017; Strobl et al., 2017). Recently, MMD and HSIC (as defined in (3) for two components) have been shown by Sejdinovic et al. (2013b) to be equivalent to other popular statistical measures such as the energy distance (Baringhaus and Franz, 2004; Székely and Rizzo, 2004, 2005)—also known as N-distance (Zinger et al., 1992; Klebanov, 2005)—and distance covariance (Székely et al., 2007; Székely and Rizzo, 2009; Lyons, 2013) respectively. HSIC has been generalized to $M \geq 2$ components (Quadrianto et al., 2009; Sejdinovic et al., 2013a) to measure the joint independence of $M$ random variables

$$\mathrm{HSIC}_k(\mathbb{P}) = \left\| \mu_{\otimes_{m=1}^M k_m}(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m) \right\|_{\otimes_{m=1}^M \mathcal{H}_{k_m}}, \tag{4}$$

where $\mathbb{P}$ is a joint measure on the product space $\mathcal{X} := \times_{m=1}^M \mathcal{X}_m$ and $(\mathbb{P}_m)_{m=1}^M$ are the marginal measures of $\mathbb{P}$ defined on $(\mathcal{X}_m)_{m=1}^M$ respectively. The extended HSIC measure has recently been analyzed in the context of independence testing (Pfister et al., 2017). In addition to testing, the extended HSIC measure is also useful in the problem of independent subspace analysis (ISA; Cardoso, 1998), wherein the latent sources are separated by maximizing the degree of independence among them. In all the applications of HSIC, the key requirement is that $k = \otimes_{m=1}^M k_m$ captures the joint independence of $M$ random variables (with joint distribution $\mathbb{P}$)—we call this property as $\mathcal{I}$-characteristic—, which is guaranteed if $k$ is characteristic. Since $k$ is defined in terms of $(k_m)_{m=1}^M$, it is of fundamental importance to understand the characteristic and $\mathcal{I}$-characteristic properties of $k$ in terms of the characteristic property of $(k_m)_{m=1}^M$, which is one of the main goals of this work.

For $M = 2$, the characterization of independence, i.e., the $\mathcal{I}$-characteristic property of $k$, is studied by Blanchard et al. (2011) and Gretton (2015) where it has been shown that if $k_1$ and $k_2$ are universal, then $k$ is universal[2] and therefore HSIC captures independence. A stronger version of this result can be obtained by combining (Lyons, 2013, Theorem 3.11) and (Sejdinovic et al., 2013b, Proposition 29): if $k_1$ and $k_2$ are characteristic, then the HSIC associated with $k = k_1 \otimes k_2$ characterizes independence. Apart from these results, not much is known about the characteristic/$\mathcal{I}$-characteristic/universality properties of $k$ in

1. In the equivalence one assumes that $\mathcal{H}_{k_X}$, $\mathcal{H}_{k_Y}$ are separable; this holds under mild conditions, for example if $\mathcal{X}$ and $\mathcal{Y}$ are separable topological domains and $k_X$, $k_Y$ are continuous (Steinwart and Christmann, 2008, Lemma 4.33).
2. Blanchard et al. (2011) deal with $c$-universal kernels while Gretton (2015) deals with $c_0$-universal kernels. A brief description of these notions are given in Section 3. Carmeli et al. (2010); Sriperumbudur et al. (2010) provide further details on these notions of universality.

terms of the individual kernels. Our **goal** is to resolve this question and understand the characteristic, $\mathcal{I}$-characteristic and universal property of the product kernel ($\otimes_{m=1}^{M} k_m$) in terms of the kernel components ($(k_m)_{m=1}^{M}$) for $M \ge 2$. Because of the relatedness of MMD and HSIC to energy distance and distance covariance, our results also contribute to the better understanding of these other measures that are popular in the statistical literature.

Specifically, our results shed light on the following **surprising phenomena** of the $\mathcal{I}$-characteristic property of $\otimes_{m=1}^{M} k_m$ for $M \ge 3$:

1. characteristic property of $(k_m)_{m=1}^{M}$ is not sufficient but necessary for $\otimes_{m=1}^{M} k_m$ to be $\mathcal{I}$-characteristic;
2. universality of $(k_m)_{m=1}^{M}$ is sufficient for $\otimes_{m=1}^{M} k_m$ to be $\mathcal{I}$-characteristic, and
3. if at least one of $(k_m)_{m=1}^{M}$ is only characteristic and not universal, then $\otimes_{m=1}^{M} k_m$ need not be $\mathcal{I}$-characteristic.

The paper is organized as follows. In Section 3, we conduct a comprehensive analysis about the above mentioned properties of $k$ and $(k_m)_{m=1}^{M}$ for any positive integer $M$. To this end, we define various notions of characteristic property on the product space $\mathcal{X}$ (see Definition 1 and Figure 2(a) in Section 3) and explore the relation between them. In order to keep our presentation in this section to be non-technical, we relegate the problem formulation to Section 3, with the main results of the paper being presented in Section 4. A summary of the results is captured in Figure 1 while the proofs are provided in Section 5. Various definitions and notation that are used throughout the paper are collected in Section 2.

## 2. Definitions and Notation

$\mathbb{N} := \{1, 2, \ldots\}$ and $\mathbb{R}$ denotes the set of natural numbers and real numbers respectively. For $M \in \mathbb{N}$, $[M] := \{1, \ldots, M\}$. $\mathbf{1}_d := (1, 1, \ldots, 1) \in \mathbb{R}^d$ and $\mathbf{0}$ denotes the matrix of zeros. For $a := (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $b := (b_1, \ldots, b_d) \in \mathbb{R}^d$, $\langle a, b\rangle = \sum_{i=1}^{d} a_i b_i$ is the Euclidean inner product. For sets $A$ and $B$, $A\backslash B = \{a \in A : a \notin B\}$ is their difference, $|A|$ is the cardinality of $A$ and $\times_{m=1}^{M} A_m = \{(a_1, \ldots, a_M) : a_m \in A_m \, m \in [M]\}$ is the Descartes product of sets $(A_m)_{m=1}^{M}$. $\mathcal{P}(\mathcal{X})$ denotes the power set of a set $\mathcal{X}$, i.e., all subsets of $\mathcal{X}$ (including the empty set and $\mathcal{X}$). The Kronecker delta is defined as $\delta_{a,b} = 1$ if $a = b$, and zero otherwise. $\chi_A$ is the indicator function of set $A$: $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ otherwise. $\mathbb{R}^{d_1 \times \ldots \times d_M}$ is the set of $d_1 \times \ldots \times d_M$-sized tensors.

For a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$, $\mathcal{B}(\mathcal{X}) := \mathcal{B}(\tau_{\mathcal{X}})$ is the Borel sigma-algebra on $\mathcal{X}$ induced by the topology $\tau_{\mathcal{X}}$. Probability and finite signed measures in the paper are meant w.r.t. the measurable space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Given $\{(\mathcal{X}_i, \tau_i)\}_{i \in I}$ topological spaces, their product $\times_{i \in I} \mathcal{X}_i$ is enriched with the product topology; it is the coarsest topology for which the canonical projections $\pi_i : \times_{i \in I} \mathcal{X}_i \to (\mathcal{X}_i, \tau_i)$ are continuous for all $i \in I$. A topological space $(\mathcal{X}, \tau_{\mathcal{X}})$ is called second-countable if $\tau_{\mathcal{X}}$ has a countable basis.[3] $C(\mathcal{X})$ denotes the space of continuous functions on $\mathcal{X}$. $C_0(\mathcal{X})$ denotes the class of real-valued functions vanishing at infinity on a locally compact Hausdorff (LCH) space[4] $\mathcal{X}$, i.e., for any $\epsilon > 0$, the set $\{x \in \mathcal{X} : |f(x)| \ge \epsilon\}$

3. Second-countability implies separability; in metric spaces the two notions coincide (Dudley, 2004, Proposition 2.1.4). By the Urysohn's theorem, a topological space is separable and metrizable if and only if it is regular, Hausdorff and second-countable. Any uncountable discrete space is *not* second-countable.
4. LCH spaces include $\mathbb{R}^d$, discrete spaces, and topological manifolds. Open or closed subsets, finite products of LCH spaces are LCH. Infinite-dimensional Hilbert spaces are *not* LCH.

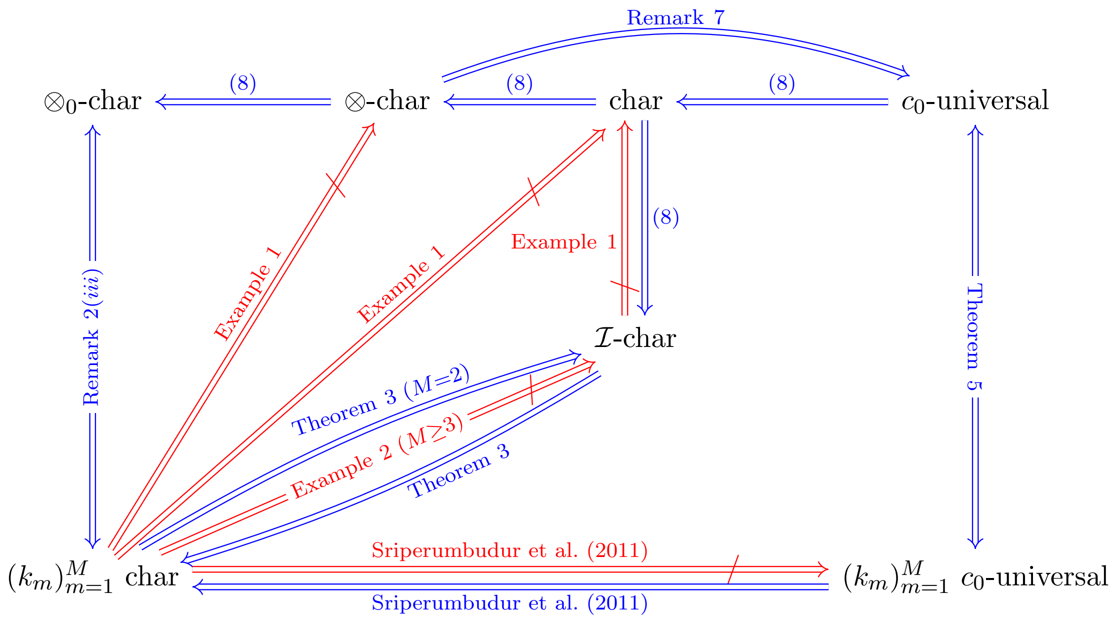


Figure 1: Summary of results: "char" denotes characteristic. In addition to the usual characteristic property, three new notions $\otimes_0$-characteristic, $\otimes$-characteristic and $\mathcal{I}$-characteristic are introduced in Definition 1 which along with $c_0$-universal (in the top right corner) correspond to the property of the tensor product kernel $\otimes_{m=1}^M k_m$, while the bottom part of the picture corresponds to the individual kernels $(k_m)_{m=1}^M$ being characteristic or $c_0$-universal. If $(k_m)_{m=1}^M$-s are continuous, bounded and translation invariant kernels on $\mathbb{R}^{d_m}$, $m \in [M]$, all the notions are equivalent (see Theorem 4).

is compact. $C_0(\mathcal{X})$ is endowed with the uniform norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. $\mathcal{M}_b(\mathcal{X})$ and $\mathcal{M}_1^+(\mathcal{X})$ are the space of finite signed measures and probability measures on $\mathcal{X}$, respectively. For $\mathbb{P}_m \in \mathcal{M}_1^+(\mathcal{X}_m)$, $\otimes_{m=1}^M \mathbb{P}_m$ denotes the product probability measure on the product space $\times_{m=1}^M \mathcal{X}_m$, i.e., $\otimes_{m=1}^M \mathbb{P}_m \in \mathcal{M}_1^+(\times_{m=1}^M \mathcal{X}_m)$. $\delta_x$ is the Dirac measure supported on $x \in \mathcal{X}$. For $\mathbb{F} \in \mathcal{M}_b\left(\times_{m=1}^M \mathcal{X}_m\right)$, the finite signed measure $\mathbb{F}_m$ denotes its marginal on $\mathcal{X}_m$. $\mathcal{H}_{k_m}$ is the reproducing kernel Hilbert space (RKHS) associated with the reproducing kernel $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$, which in this paper is assumed to be measurable and bounded. The tensor product of $(k_m)_{m=1}^M$ is a kernel, defined as

$$\otimes_{m=1}^M k_m \left(\left(x_1, \ldots, x_M\right), \left(x_1', \ldots, x_M'\right)\right) = \prod_{m=1}^M k_m\left(x_m, x_m'\right), \quad x_m, x_m' \in \mathcal{X}_m,$$

whose associated RKHS is denoted as $\mathcal{H}_{\otimes_{m=1}^M k_m} = \otimes_{m=1}^M \mathcal{H}_{k_m}$ (Berlinet and Thomas-Agnan, 2004, Theorem 13), where the r.h.s. is the tensor product of RKHSs $(\mathcal{H}_{k_m})_{m=1}^M$. For $h_m \in \mathcal{H}_m$, $m \in [M]$, the multi-linear operator $\otimes_{m=1}^M h_m \in \otimes_{m=1}^M \mathcal{H}_m$ is defined as

$$\left(\otimes_{m=1}^M h_m\right)(v_1, \ldots, v_M) = \prod_{m=1}^M \langle h_m, v_m \rangle_{\mathcal{H}_m}, \quad v_m \in \mathcal{H}_m.$$

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on a LCH space $\mathcal{X}$ is called a $c_0$-kernel if $k(\cdot, x) \in C_0(\mathcal{X})$ for all $x \in \mathcal{X}$. $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be a translation invariant kernel on $\mathbb{R}^d$ if $k(x, y) = \psi(x - y)$, $x, y \in \mathbb{R}^d$ for a positive definite function $\psi : \mathbb{R}^d \to \mathbb{R}$. $\mu_k(\mathbb{F})$ denotes the kernel mean embedding of $\mathbb{F} \in \mathcal{M}_b(\mathcal{X})$ to $\mathcal{H}_k$ which is defined as $\mu_k(\mathbb{F}) = \int_{\mathcal{X}} k(\cdot, x)\,\mathrm{d}\mathbb{F}(x)$, where the integral is meant in the Bochner sense.

## 3. Problem Formulation

In this section, we formally introduce the goal of the paper. To this end, we start with a definition. For simplicity, throughout the paper, we assume that all kernels are bounded. The definition is based on the observation (Sriperumbudur et al., 2010, Lemma 8) that a bounded kernel $k$ on a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$ is characteristic if and only if

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') > 0,\ \forall\, \mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\} \text{ such that } \mathbb{F}(\mathcal{X}) = 0.$$

In other words, characteristic kernels are integrally strictly positive definite (ispd; see Sriperumbudur et al., 2010, p. 1523) w.r.t. the class of finite signed measures that assign zero measure to $\mathcal{X}$. The following definition extends this observation to tensor product kernels on product spaces.

**Definition 1 ($\mathcal{F}$-ispd tensor product kernel)** *Suppose $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ is a bounded kernel on a topological space $(\mathcal{X}_m, \tau_{\mathcal{X}_m})$, $m \in [M]$. Let $\mathcal{F} \subseteq \mathcal{M}_b(\mathcal{X})$ be such that $0 \in \mathcal{F}$ where $\mathcal{X} := \times_{m=1}^M \mathcal{X}_m$. $k := \otimes_{m=1}^M k_m$ is said to be* $\mathcal{F}$-ispd *if*

$$\mu_k(\mathbb{F}) = 0 \Rightarrow \mathbb{F} = 0 \quad (\mathbb{F} \in \mathcal{F}),\ \textit{or equivalently}$$
$$\|\mu_k(\mathbb{F})\|_{\mathcal{H}_k}^2 = \int_{\times_{m=1}^M \mathcal{X}_m} \int_{\times_{m=1}^M \mathcal{X}_m} \left(\otimes_{m=1}^M k_m\right)\left(x, x'\right)\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') > 0, \quad \forall\, \mathbb{F} \in \mathcal{F}\backslash\{0\}. \tag{5}$$

*Specifically,*

- *if $k_m$-s are $c_0$-kernels on locally compact Polish (LCP)* [5] *spaces $\mathcal{X}_m$-s and $\mathcal{F} = \mathcal{M}_b(\mathcal{X})$, then $k$ is called $c_0$-universal.*
- *if*

$$\begin{aligned}
\mathcal{F} &= [\mathcal{M}_b(\mathcal{X})]^0 &&:= \{\mathbb{F} \in \mathcal{M}_b(\mathcal{X}) : \mathbb{F}(\mathcal{X}) = 0\},\\
\mathcal{F} &= \left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0 &&:= \left\{\mathbb{F} \in \otimes_{m=1}^M \mathcal{M}_b\left(\mathcal{X}_m\right), \mathbb{F}(\mathcal{X}) = 0\right\},\\
\mathcal{F} &= \mathcal{I} &&:= \left\{\mathbb{P} - \otimes_{m=1}^M \mathbb{P}_m : \mathbb{P} \in \mathcal{M}_1^+\left(\times_{m=1}^M \mathcal{X}_m\right)\right\}, \quad (M \geq 2)\\
\mathcal{F} &= \otimes_{m=1}^M \mathcal{M}_b^0(\mathcal{X}_m) &&:= \left\{\mathbb{F} = \otimes_{m=1}^M \mathbb{F}_m \,:\, \mathbb{F}_m \in \mathcal{M}_b\left(\mathcal{X}_m\right), \mathbb{F}_m(\mathcal{X}_m) = 0,\ \forall\, m \in [M]\right\},
\end{aligned}$$

  *then $k$ is called* characteristic, $\otimes$-characteristic, $\mathcal{I}$-characteristic *and* $\otimes_0$-characteristic, *respectively.*

5. A topological space is called Polish if it is complete, separable and metrizable. For example, $\mathbb{R}^d$ and countable discrete spaces are Polish. Open and closed subsets, products and disjoint unions of countably many Polish spaces are Polish. Every second-countable LCH space is Polish.

In Definition 1, $k$ being characteristic matches the usual notion of characteristic kernels on a product space, i.e., there are no two distinct probability measures on $\mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$ such that the MMD between them is zero. The other notions such as $\otimes$-characteristic, $\mathcal{I}$-characteristic and $\otimes_0$-characteristic are typically weaker than the usual characteristic property since

$$\otimes_{m=1}^{M} \mathcal{M}_b^0(\mathcal{X}_m) \subseteq \left[\otimes_{m=1}^{M} \mathcal{M}_b(\mathcal{X}_m)\right]^0 \subseteq \underset{\substack{\cup\!\mid \\ \mathcal{I}}}{\left[\mathcal{M}_b\left(\times_{m=1}^{M} \mathcal{X}_m\right)\right]^0} \subseteq \mathcal{M}_b\left(\times_{m=1}^{M} \mathcal{X}_m\right). \tag{6}$$

Below we provide further intuition on the $\mathcal{F}$ measure classes enlisted in Definition 1.

**Remark 2** *(i)* $\boldsymbol{\mathcal{F} = \mathcal{M}_b(\mathcal{X})}$ **:** *If $k_m$-s are $c_0$-kernels on LCH spaces $\mathcal{X}_m$ for all $m \in [M]$, then $k$ is also a $c_0$-kernel on LCH space $\mathcal{X}$ implying that if $k$ satisfies* (5)*, then $k$ is $c_0$-universal (Sriperumbudur et al., 2010, Proposition 2). It is well known (Sriperumbudur et al., 2010) that $c_0$-universality reduces to c-universality (i.e., the notion of universality proposed by Steinwart, 2001) if $\mathcal{X}$ is compact which is guaranteed if and only if each $\mathcal{X}_m$, $m \in [M]$ is compact.*

*(ii)* $\boldsymbol{\mathcal{F} = \mathcal{I}}$ **:** *This family is useful to describe the joint* independence *of $M$ random variables—hence the name $\mathcal{I}$-characteristic—defined on kernel-endowed domains $(\mathcal{X}_m)_{m=1}^{M}$: If $\mathbb{P}$ denotes the joint distribution of random variables $(X_m)_{m=1}^{M}$ and $(\mathbb{P}_m)_{m=1}^{M}$ are the associated marginals on $(\mathcal{X}_m)_{m=1}^{M}$, then by definition $k = \otimes_{m=1}^{M} k_m$ is $\mathcal{I}$-characteristic iff*

$$\mathrm{HSIC}_k(\mathbb{P}) = 0 \Longleftrightarrow \mathbb{P} = \otimes_{m=1}^{M} \mathbb{P}_m.$$

*In other words, HSIC captures joint independence exactly with $\mathcal{I}$-characteristic kernels. Similarly, the $\mathcal{I}$-characteristic property ensures that COCO (constrained covariance; Gretton et al., 2005b) is a joint independence measure as COCO is defined by replacing the Hilbert-Schmidt norm of the cross-covariance operator (see* (3) *and* (4)*) with its spectral norm.*

*(iii)* $\boldsymbol{\mathcal{F} = \otimes_{m=1}^{M} \mathcal{M}_b^0(\mathcal{X}_m)}$ **:** *In this case $\mathcal{F}$ is chosen to be the product of finite signed measures on $\mathcal{X}$ such that each marginal measure $\mathbb{F}_m$ assigns zero to the corresponding space $\mathcal{X}_m$. This choice is relevant as the characteristic property of individual kernels $(k_m)_{m=1}^{M}$ need* not *imply the characteristic property of $\otimes_{m=1}^{M} k_m$, but is equivalent to the $\otimes_0$-characteristic property of $\otimes_{m=1}^{M} k_m$. The equivalence holds for bounded kernels $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ on topological spaces $\mathcal{X}_m$ ($m \in [M]$) since for any $\mathbb{F} = \otimes_{m=1}^{M} \mathbb{F}_m \in \otimes_{m=1}^{M} \mathcal{M}_b(\mathcal{X}_m)$, $\mathbb{F}_m(\mathcal{X}_m) = 0$ ($\forall\, m \in [M]$)*

$$\|\mu_k(\mathbb{F})\|^2_{\mathcal{H}_{\otimes_{m=1}^{M} k_m}} = \prod_{m=1}^{M} \|\mu_{k_m}(\mathbb{F}_m)\|^2_{\mathcal{H}_{k_m}}, \tag{7}$$

*and the l.h.s. is positive iff each term on the r.h.s. is positive.*

*(iv)* $\boldsymbol{\mathcal{F} = \left[\otimes_{m=1}^{M} \mathcal{M}_b(\mathcal{X}_m)\right]^0}$ **:** *This class is similar to the one discussed in (iii) above—i.e., class of product measures—with the slight difference that the joint measure $\mathbb{F}$ is restricted to assign zero measure to $\mathcal{X}$ without requiring all the marginal measures $\mathbb{F}_m$*

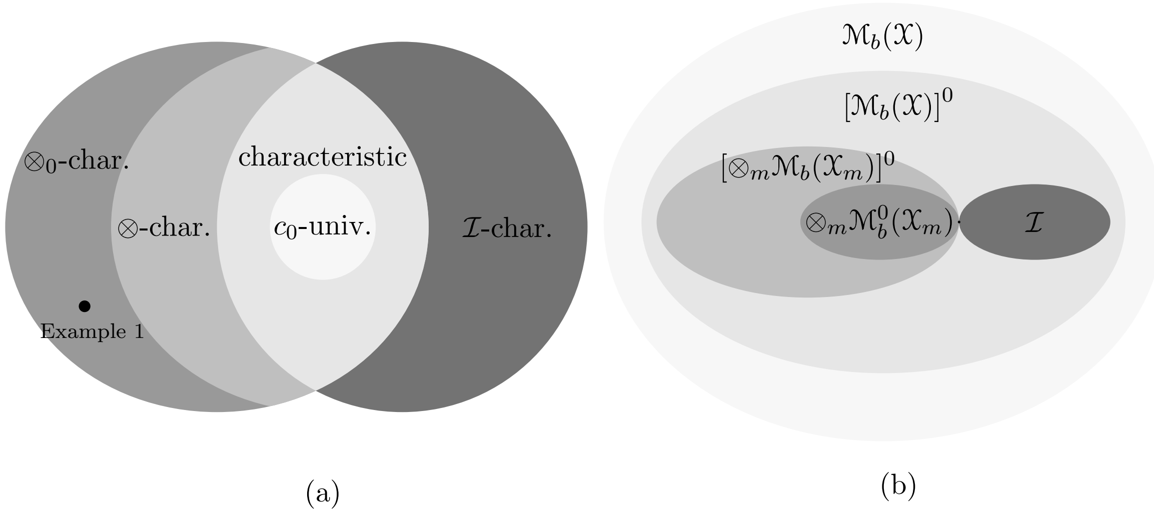


Figure 2: (a) $\mathcal{F}$-ispd $\otimes_{m=1}^M k_m$ kernels (see (8)); (b) $\mathcal{F} \subseteq \mathcal{M}_b(\mathcal{X})$, $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$. Example 1: $\otimes_{m=1}^M k_m$ is $\otimes_0$-characteristic but not $\otimes$-characteristic and therefore not characteristic.

*to assign zero measure to the corresponding space $\mathcal{X}_m$. While the need for considering such a measure class may not be clear at this juncture, however, based on* (7)*, it turns out that this choice of $\mathcal{F}$ has quite surprising connections to the characteristic property and $c_0$-universality of the product kernel; for details see Remark 7.*

*(v)* ***$\mathcal{F}$-ispd relations:*** *Given the relations in* (6)*, it immediately follows that $k = \otimes_{m=1}^M k_m$ satisfies*

$$\begin{array}{ccccccc} \otimes_0\text{-}\textit{characteristic} & \Longleftarrow & \otimes\text{-}\textit{characteristic} & \Longleftarrow & \textit{characteristic} & \Longleftarrow & c_0\text{-}\textit{universal} \\ & & & & \Downarrow & & \\ & & & & \mathcal{I}\text{-}\textit{characteristic} & & \end{array} \tag{8}$$

*when $\mathcal{X}_m$ for all $m \in [M]$ are LCP. A visual illustration of* (6) *and* (8) *is provided in Figure 2.*

*(vi)* $\left[\boldsymbol{\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)}\right]^{\mathbf{0}} \boldsymbol{\cap \mathcal{I} = \{0\}}$ **:** *While it is clear that $\left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0$ and $\mathcal{I}$ are subsets of $\left[\mathcal{M}_b(\times_{m=1}^M \mathcal{X}_m)\right]^0$, it is interesting to note that $\left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0$ and $\mathcal{I}$ have a trivial intersection with* 0 *being the measure common to each of them, assuming that $\mathcal{X}_m$-s are second-countable for all $m \in [M]$; see Section 5.1.*

Having defined the $\mathcal{F}$-ispd property, our **goal** is to investigate whether the characteristic or $c_0$-universal property of $k_m$-s ($m \in [M]$) imply different $\mathcal{F}$-ispd properties of $\otimes_{m=1}^M k_m$, and vice versa.

## 4. Main Results

In this section, we present our main results related to the $\mathcal{F}$-ispd property of tensor product kernels, which are summarized in Figure 1. The results in this section will deal with various assumptions on $\mathcal{X}_m$, such as second-countability, Hausdorff, locally compact Hausdorff

(LCH) and locally compact Polish (LCP), so that they are presented in more generality. However, for simplicity, all these assumptions can be unified by simply assuming a stronger condition that $\mathcal{X}_m$'s are LCP.

Our first example illustrates that the characteristic property of $k_m$-s does not imply the characteristic property of the tensor product kernel. In light of Remark 2(iv) of Section 3, it follows that the class of $\otimes_0$-characteristic tensor product kernels form a *strictly* larger class than characteristic tensor product kernels; see also Figure 2.

**Example 1** *Let $\mathcal{X}_1 = \mathcal{X}_2 = \{1, 2\}$, $\tau_{\mathcal{X}_1} = \tau_{\mathcal{X}_2} = \mathcal{P}(\{1, 2\})$, $k_1(x, x') = k_2(x, x') = 2\delta_{x,x'} - 1$. It is easy to verify that $k_1$ and $k_2$ are characteristic. However, it can be proved that $k_1 \otimes k_2$ is* not *$\otimes$-characteristic and therefore not characteristic. On the hand, interestingly, $k_1 \otimes k_2$ is $\mathcal{I}$-characteristic. We refer the reader to Section 5.2 for details.*

In the above example, we showed that the tensor product of $k_1$ and $k_2$ (which are characteristic kernels) is $\mathcal{I}$-characteristic. The following result generalizes this behavior for any bounded characteristic kernels. In addition, under a mild assumption, it shows the converse to be true for any $M$.

**Theorem 3** *Let $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ be bounded kernels on topological spaces $\mathcal{X}_m$ for all $m \in [M]$, $M \geq 2$. Then the following holds.*

- *(i) Suppose $\mathcal{X}_m$ is second-countable for all $m \in [M]$ with $M = 2$. If $k_1$ and $k_2$ are characteristic, then $k_1 \otimes k_2$ is $\mathcal{I}$-characteristic.*
- *(ii) Suppose $\mathcal{X}_m$ is Hausdorff and $|\mathcal{X}_m| \geq 2$ for all $m \in [M]$. If $\otimes_{m=1}^M k_m$ is $\mathcal{I}$-characteristic, then $k_1, \ldots, k_M$ are characteristic.*

Lyons (2013) has showed an analogous result to Theorem 3(i) for distance covariances ($M = 2$) on metric spaces of negative type (Lyons, 2013, Theorem 3.11), which by Sejdinovic et al. (2013b, Proposition 29) holds for HSIC yielding the $\mathcal{I}$-characteristic property of $k_1 \otimes k_2$. Recently, Gretton (2015) presented a direct proof showing that HSIC corresponding to $k_1 \otimes k_2$ captures independence if $k_1$ and $k_2$ are translation invariant characteristic kernels on $\mathbb{R}^d$ (which is equivalent to $c_0$-universality). Blanchard et al. (2011) proved a result similar to Theorem 3(i) assuming that $\mathcal{X}_m$'s are compact and $k_1$, $k_2$ being $c$-universal. In contrast, Theorem 3(i) establishes the result for bounded kernels on general second-countable topological spaces. In fact, the results of Gretton (2015); Blanchard et al. (2011) are special cases of Theorems 4 and 5 below. Theorem 3(i) raises a pertinent question: whether $\otimes_{m=1}^M k_m$ is $\mathcal{I}$-characteristic if $k_m$-s are characteristic for all $m \in [M]$ where $M > 2$? The following example provides a negative answer to this question. On a positive side, however, we will see in Theorem 5 that the $\mathcal{I}$-characteristic property of $\otimes_{m=1}^M k_m$ can be guaranteed for any $M \geq 2$ if a stronger condition is imposed on $k_m$-s (and $\mathcal{X}_m$-s). Theorem 3(ii) generalizes Proposition 3.15 of Lyons (2013) for any $M > 2$, which states that every kernel $k_m$, $m \in [M]$ being characteristic is necessary for the tensor kernel $\otimes_{m=1}^M k_m$ to be $\mathcal{I}$-characteristic.

**Example 2** *Let $M = 3$ and $\mathcal{X}_m := \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\mathcal{X}_m)$, $k_m(x, x') = 2\delta_{x,x'} - 1$ ($m = 1, 2, 3$). As mentioned in Example 1, $(k_m)_{m=1}^3$ are characteristic. However, it can be shown that $\otimes_{m=1}^3 k_m$ is not $\mathcal{I}$-characteristic. See Section 5.4 for details.*

In Remark 2(iii) and Example 1, we showed that in general, only the $\otimes_0$-characteristic property of $\otimes_{m=1}^M k_m$ is equivalent to the characteristic property of $k_m$-s. Our next result shows that all the various notions of characteristic property of $\otimes_{m=1}^M k_m$ *coincide* if $k_m$-s are translation-invariant, continuous bounded kernels on $\mathbb{R}^d$.

**Theorem 4** *Suppose $k_m : \mathbb{R}^{d_m} \times \mathbb{R}^{d_m} \to \mathbb{R}$ are continuous, bounded and translation-invariant kernels for all $m \in [M]$. Then the following statements are equivalent:*

*(i) $k_m$-s are characteristic for all $m \in [M]$;*
*(ii) $\otimes_{m=1}^M k_m$ is $\otimes_0$-characteristic;*
*(iii) $\otimes_{m=1}^M k_m$ is $\otimes$-characteristic;*
*(iv) $\otimes_{m=1}^M k_m$ is $\mathcal{I}$-characteristic;*
*(v) $\otimes_{m=1}^M k_m$ is characteristic.*

The following result shows that on LCP spaces, the tensor product of $M \geq 2$ $c_0$-universal kernels is also $c_0$-universal, and vice versa.

**Theorem 5** *Suppose $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ are $c_0$-kernels on LCP spaces $\mathcal{X}_m$ ($m \in [M]$). Then $\otimes_{m=1}^M k_m$ is $c_0$-universal iff $k_m$-s are $c_0$-universal for all $m \in [M]$.*

**Remark 6** *(i) A special case of Theorem 5 for $M = 2$ is proved by Lyons (2013, Lemma 3.8) in the context of distance covariance which reduces to Theorem 5 through the equivalence established by Sejdinovic et al. (2013b). Another special case of Theorem 5 is proved by Blanchard et al. (2011, Lemma 5.2) for c-universality with $M = 2$ using the Stone-Weierstrass theorem: if $k_1$ and $k_2$ are c-universal then $k_1 \otimes k_2$ is c-universal.*

*(ii) Since the notions of $c_0$-universality and characteristic property are equivalent for translation invariant $c_0$-kernels on $\mathbb{R}^d$ (Carmeli et al., 2010, Prop. 5.16, Sriperumbudur et al., 2010, Theorem 9), Theorem 4 can be considered as a special case of Theorem 5. In other words, requiring $(k_m)_{m=1}^M$ to be also $c_0$-kernels in Theorem 4(i)-(iv) is equivalent to*

*(v) $k_m$-s are $c_0$-universal for all $m \in [M]$;*
*(vi) $\otimes_{m=1}^M k_m$ is $c_0$-universal.*

*(iii) Since the $c_0$-universality of $\otimes_{m=1}^M k_m$ implies its $\mathcal{I}$-characteristic property (see* (8)*), Theorem 5 also provides a generalization of Theorem 3(i) to $M \geq 2$ under additional assumptions on $k_m$-s, while constraining $\mathcal{X}_m$-s to LCP-s instead of second-countable topological spaces.*

In Example 2 and Theorem 5, we showed that for $M \geq 3$ components while the characteristic property of $(k_m)_{m=1}^M$ is not sufficient, their universality is enough to guarantee the $\mathcal{I}$-characteristic property of $\otimes_{m=1}^M k_m$. The next example demonstrates that these results are tight: If at least one $k_m$ is not universal but only characteristic, then $\otimes_{m=1}^M k_m$ might not be $\mathcal{I}$-characteristic.

**Example 3** *Let $M = 3$ and $\mathcal{X}_m := \{1, 2\}$, $\tau_{\mathcal{X}_m} = \mathcal{P}(\mathcal{X}_m)$, for all $m \in$ [3], $k_1(x, x') = 2\delta_{x,x'} - 1$, and $k_m(x, x') = \delta_{x,x'}$ ($m = 2, 3$). $k_1$ is characteristic (Example 1), $k_2$ and $k_3$ are universal since the associated Gram matrix $\mathbf{G} = [k_m(x, x')]_{x,x' \in \mathcal{X}_m}$ is an identity matrix,*

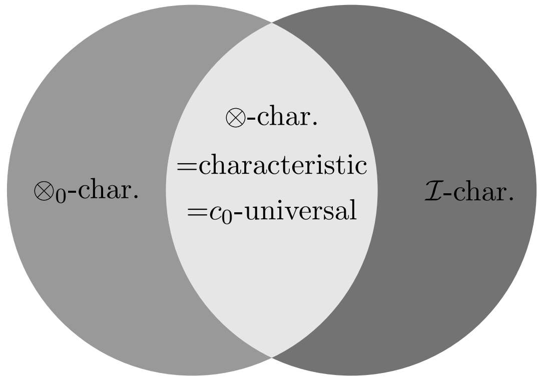


Figure 3: Simplification of the $\mathcal{F}$-ispd property of tensor product kernels; see Remark 7.

*which is strictly positive definite ($m = 2, 3$). However, $\otimes_{m=1}^3 k_m$ is* not *$\mathcal{I}$-characteristic. See Section 5.7 for details.*

**Remark 7** *Note that the l.h.s. in* (7) *is positive if and only if each term on the r.h.s. is positive, i.e., if $k = \otimes_{m=1}^M k_m$ is $\otimes$-characteristic with $k_m$-s being $c_0$-kernels on LCP $\mathcal{X}_m$-s, then all $k_m$-s are $c_0$-universal. A similar result was also proved by Steinwart and Ziegel (2017, Lemma 3.4). Combining this with Theorem 5 yields that for tensor product $c_0$-kernels, the notions of $\otimes$-characteristic, characteristic and $c_0$-universality are equivalent, which is quite surprising as for a joint kernel $k$ (that is not of product type), these notions need not necessarily coincide. In light of this discussion, Figure 2(a) can be simplified to Figure 3.*

## 5. Proofs

In this section, we provide the proofs of our results presented in Section 4.

### 5.1 Proof of Remark 2(iv)

By the second-countability of $\mathcal{X}_m$-s, $\mathcal{B}\left(\times_{m=1}^M \mathcal{X}_m\right) = \otimes_{m=1}^M \mathcal{B}(\mathcal{X}_m)$, where the r.h.s. is defined as the $\sigma$-field generated by the cylinder sets $A_m \times_{n\neq m} \mathcal{X}_n$ where $m \in [M]$ and $A_m \in \mathcal{B}(\mathcal{X}_m)$. Suppose there exists $\mathbb{F} \in \left[\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\right]^0 \cap \mathcal{I}$ such that $\mathbb{F} \neq 0$. This means there exists $\mathbb{P} \in \mathcal{M}_1^+\left(\times_{m=1}^M \mathcal{X}_m\right)$ with $(\mathbb{P}_m)_{m=1}^M$ being the marginals of $\mathbb{P}$ such that $\mathbb{F} = \otimes_{m=1}^M \mathbb{F}_m = \mathbb{P} - \otimes_{m=1}^M \mathbb{P}_m$. Since $\mathbb{F} \neq 0$ there exists $A_m \times_{n\neq m} \mathcal{X}_n$ for some $m \in [M]$ and $A_m \in \mathcal{B}(\mathcal{X}_m)$ such that $0 \neq \mathbb{F}(A_m \times_{n\neq m} \mathcal{X}_n) = \mathbb{F}_m(A_m)\prod_{n\neq m} \mathbb{F}_n(\mathcal{X}_n) = \mathbb{P}\left(A_m \times_{n\neq m} \mathcal{X}_n\right) - \mathbb{P}_m(A_m)\prod_{n\neq m} \mathbb{P}_n(\mathcal{X}_n) = \mathbb{P}_m(A_m) - \mathbb{P}_m(A_m) = 0$, leading to a contradiction.

### 5.2 Proof of Example 1

The proof is structured as follows.

1. First we show that $k := k_1 = k_2$ is a kernel and it is characteristic.

2. Next it is proved that $k_1 \otimes k_2$ is not $\otimes$-characteristic, which implies $k_1 \otimes k_2$ is not characteristic.
3. Finally, the $\mathcal{I}$-characteristic property of $k_1 \otimes k_2$ is established.

The individual steps are as follows:

$k$ **is a kernel.** Assume w.l.o.g. that $x_1 = \ldots = x_N = 1$, $x_{N+1} = \ldots = x_n = 2$. Then it is easy to verify that the Gram matrix $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n = \mathbf{a}\mathbf{a}^\top$ where $\mathbf{a} := \left(\mathbf{1}_N^\top, -\mathbf{1}_{n-N}^\top\right)^\top$ and $\mathbf{a}^\top$ is the transpose of $\mathbf{a}$. Clearly $\mathbf{G}$ is positive semidefinite and so $k$ is a kernel.

$k$ **is characteristic**. We will show that $k$ satisfies (5). On $\mathcal{X} = \{1, 2\}$ a finite signed measure $\mathbb{F}$ takes the form $\mathbb{F} = a_1\delta_1 + a_2\delta_2$ for some $a_1, a_2 \in \mathbb{R}$. Thus,

$$\mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\} \Leftrightarrow (a_1, a_2) \neq \mathbf{0} \qquad \text{and} \qquad \mathbb{F}(\mathcal{X}) = 0 \Leftrightarrow a_1 + a_2 = 0. \tag{9}$$

Consider

$$\begin{aligned} \int_{\mathcal{X}}\int_{\mathcal{X}} k(x, x')\,\mathrm{d}\mathbb{F}(x)\,\mathrm{d}\mathbb{F}(x') &= a_1^2 k(1,1) + a_2^2 k(2,2) + 2a_1a_2k(1,2) \\ &= a_1^2 + a_2^2 - 2a_1a_2 = (a_1 - a_2)^2 = 4a_1^2 > 0, \end{aligned} \tag{10}$$

where we used (9) and the facts that $k(1,1) = k(2,2) = 1$, $k(1,2) = -1$.

$k_1 \otimes k_2$ **is not $\otimes$-characteristic**. We construct a witness $\mathbb{F} = \mathbb{F}_1 \otimes \mathbb{F}_2 \in \otimes_{m=1}^2 \mathcal{M}_b(\mathcal{X}_m)\backslash\{0\}$ such that

$$\mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = \mathbb{F}_1(\mathcal{X}_1)\mathbb{F}_2(\mathcal{X}_2) = 0, \tag{11}$$

and

$$\begin{aligned} 0 &= \int_{\mathcal{X}_1\times\mathcal{X}_2}\int_{\mathcal{X}_1\times\mathcal{X}_2} \underbrace{(k_1 \otimes k_2)((i_1, i_2), (i_1', i_2'))}_{k_1(i_1,i_1')k_2(i_2,i_2')}\,\mathrm{d}\mathbb{F}(i_1, i_2)\,\mathrm{d}\mathbb{F}(i_1', i_2') \\ &= \prod_{m=1}^2 \int_{\mathcal{X}_m}\int_{\mathcal{X}_m} k_m(i_m, i_m')\,\mathrm{d}\mathbb{F}_m(i_m)\,\mathrm{d}\mathbb{F}_m(i_m'). \end{aligned} \tag{12}$$

Finite signed measures on $\{1, 2\}$ take the form $\mathbb{F}_1 = \mathbb{F}_1(\mathbf{a}) = a_1\delta_1 + a_2\delta_2$, $\mathbb{F}_2 = \mathbb{F}_2(\mathbf{b}) = b_1\delta_1 + b_2\delta_2$ form, where $\mathbf{a} = (a_1, a_2) \in \mathbb{R}^2$, $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$. With these notations, (11) and (12) can be rewritten as

$$\begin{aligned} 0 &= (a_1 + a_2)(b_1 + b_2), \\ 0 &= \left[\sum_{i,i'=1}^2 k_1(i, i')a_ia_{i'}\right]\left[\sum_{j,j'=1}^2 k_2(j, j')b_jb_{j'}\right] = (a_1 - a_2)^2(b_1 - b_2)^2. \end{aligned}$$

Keeping the solutions where neither $\mathbf{a}$ nor $\mathbf{b}$ is the zero vector, there are 2 (symmetric) possibilities: (i) $a_1 + a_2 = 0$, $b_1 = b_2$ and (ii) $a_1 = a_2$, $b_1 + b_2 = 0$. In other words, for any $a, b \neq 0$, the possibilities are (i) $\mathbf{a} = (a, -a)$, $\mathbf{b} = (b, b)$ and (ii) $\mathbf{a} = (a, a)$, $\mathbf{b} = (b, -b)$. This establishes the non-$\left[\otimes_{m=1}^2 \mathcal{M}_b(\mathcal{X}_m)\right]^0$-ispd property of $k_1 \otimes k_2$.

$k_1 \otimes k_2$ **is $\mathcal{I}$-characteristic**. Our goal is to show that $k_1 \otimes k_2$ is $\mathcal{I}$-characteristic, i.e., for any $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X}_1 \times \mathcal{X}_2)$, $\mu_{k_1 \otimes k_2}(\mathbb{F}) = 0$ implies $\mathbb{F} = 0$, where $\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$. We divide the proof into two parts:

1. First we derive the equations of
$$\mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = 0 \quad \text{and} \quad \int\int_{(\mathcal{X}_1\times\mathcal{X}_2)^2} (k_1 \otimes k_2)\left((i,j),(r,s)\right) \mathrm{d}\mathbb{F}(i,j)\, \mathrm{d}\mathbb{F}(r,s) = 0 \tag{13}$$
for general finite signed measures $\mathbb{F} = \sum_{i,j=1}^2 a_{ij}\delta_{(i,j)}$ on $\mathcal{X}_1 \times \mathcal{X}_2$.
2. Then, we apply the $\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$ parameterization and solve for $\mathbb{P}$ that satisfies (13) to conclude that $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$, i.e., $\mathbb{F} = 0$. Note that in the chosen parametrization for $\mathbb{F}$, $\mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) = 0$ holds automatically.

The details are as follows.

**Step 1**.

$$\begin{aligned}
0 &= \mathbb{F}(\mathcal{X}_1 \times \mathcal{X}_2) \Leftrightarrow 0 = a_{11} + a_{12} + a_{21} + a_{22}, && (14)\\
0 &= \int_{\mathcal{X}_1\times\mathcal{X}_2}\int_{\mathcal{X}_1\times\mathcal{X}_2} \underbrace{(k_1 \otimes k_2)\left((i,j),(r,s)\right)}_{k_1(i,r)k_2(j,s)} \mathrm{d}\mathbb{F}(i,j)\, \mathrm{d}\mathbb{F}(r,s)\\
&= \sum_{i,j=1}^{2}\sum_{r,s=1}^{2} k_1(i,r)k_2(j,s)a_{ij}a_{rs} = \sum_{i,r=1}^{2} k_1(i,r) \sum_{j,s=1}^{2} k_2(j,s)a_{ij}a_{rs}\\
&= k_1(1,1)\left[k_2(1,1)a_{11}a_{11} + k_2(1,2)a_{11}a_{12} + k_2(2,1)a_{12}a_{11} + k_2(2,2)a_{12}a_{12}\right]\\
&\quad + k_1(1,2)\left[k_2(1,1)a_{11}a_{21} + k_2(1,2)a_{11}a_{22} + k_2(2,1)a_{12}a_{21} + k_2(2,2)a_{12}a_{22}\right]\\
&\quad + k_1(2,1)\left[k_2(1,1)a_{21}a_{11} + k_2(1,2)a_{21}a_{12} + k_2(2,1)a_{22}a_{11} + k_2(2,2)a_{22}a_{12}\right]\\
&\quad + k_1(2,2)\left[k_2(1,1)a_{21}a_{21} + k_2(1,2)a_{21}a_{22} + k_2(2,1)a_{22}a_{21} + k_2(2,2)a_{22}a_{22}\right]\\
&= \underbrace{\left(a_{11}^2 - 2a_{11}a_{12} + a_{12}^2\right)}_{(a_{11}-a_{12})^2} + \underbrace{\left(a_{21}^2 - 2a_{21}a_{22} + a_{22}^2\right)}_{(a_{21}-a_{22})^2} - 2\underbrace{\left(a_{11}a_{21} - a_{11}a_{22} - a_{12}a_{21} + a_{12}a_{22}\right)}_{(a_{11}-a_{12})(a_{21}-a_{22})}\\
&= (a_{11} - a_{12} - a_{21} + a_{22})^2. && (15)
\end{aligned}$$

Solving (14) and (15) yields

$$a_{11} + a_{22} = 0 \quad \text{and} \quad a_{12} + a_{21} = 0. \tag{16}$$

**Step 2**. Any $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X}_1 \times \mathcal{X}_2)$ can be parametrized as

$$\mathbb{P} = \sum_{i,j=1}^{2} p_{ij}\delta_{(i,j)}, \quad p_{ij} \geq 0, \forall\,(i,j) \quad \text{and} \quad \sum_{i,j=1}^{2} p_{ij} = 1. \tag{17}$$

Let $\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 = \sum_{i,j=1}^2 a_{ij}\delta_{(i,j)}$; for illustration see Table 1. It follows from step 1 that $\mathbb{F}$ satisfying (16) is equivalent to satisfying (13). Therefore, for the choice of $\mathbb{F} := \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$, we obtain

$$p_{11} - (p_{11} + p_{12})(p_{11} + p_{21}) + p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) = 0, \tag{18}$$

| $\mathbb{P}$: $y\backslash x$ | 1 | 2 | $\mathbb{P}_2$ |
|---|---|---|---|
| 1 | $p_{11}$ | $p_{21}$ | $q_1 = p_{11} + p_{21}$ |
| 2 | $p_{12}$ | $p_{22}$ | $q_2 = p_{12} + p_{22}$ |
| $\mathbb{P}_1$ | $p_1 = p_{11} + p_{12}$ | $p_2 = p_{21} + p_{22}$ | |

$\Rightarrow$

| $\mathbb{F} := \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$ | 1 | 2 |
|---|---|---|
| 1 | $a_{11} = p_{11} - (p_{11} + p_{12})(p_{11} + p_{21})$ | $a_{21} = p_{21} - (p_{21} + p_{22})(p_{11} + p_{21})$ |
| 2 | $a_{12} = p_{12} - (p_{11} + p_{12})(p_{12} + p_{22})$ | $a_{22} = p_{22} - (p_{21} + p_{22})(p_{12} + p_{22})$ |

Table 1: Joint ($\mathbb{P}$), joint minus product of the marginals ($\mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2$).

| $\mathbb{P}$: $y\backslash x$ | 1 | 2 | $\mathbb{P}_2$ |
|---|---|---|---|
| 1 | $p_{11} = \frac{a[1-(a+b)]}{a+b}$ | $p_{21} = a$ | $q_1 = \frac{a}{a+b}$ |
| 2 | $p_{12} = \frac{b[1-(a+b)]}{a+b}$ | $p_{22} = b$ | $q_2 = \frac{b}{a+b}$ |
| $\mathbb{P}_1$ | $p_1 = 1 - (a+b)$ | $p_2 = a + b$ | |

Table 2: Family of probability distributions solving (17)–(19).

$$p_{12} - (p_{11} + p_{12})(p_{12} + p_{22}) + p_{21} - (p_{21} + p_{22})(p_{11} + p_{21}) = 0, \tag{19}$$

where $(p_{ij})_{i,j\in[2]}$ satisfy (17). Solving (17)–(19), we obtain

$$p_{11} = \frac{a[1-(a+b)]}{a+b}, \quad p_{12} = \frac{b[1-(a+b)]}{a+b}, \quad p_{21} = a \quad \text{and} \quad p_{22} = b,$$

with $0 \le a, b \le 1$, $a + b \le 1$ and $(a, b) \neq \mathbf{0}$. The resulting distribution family with its marginals is summarized in Table 2. It can be seen that each member of this family (any $a$, $b$ in the constraint set) factorizes: $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$. In other words, $\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 = 0$; hence $k_1 \otimes k_2$ is $\mathcal{I}$-characteristic.

**Remark.** We would like to mention that while $k_1$ and $k_2$ are characteristic, they are not universal. Since $\mathcal{X}$ is finite, the usual notion of universality (also called $c$-universality) matches with $c_0$-universality. Therefore, from (10), we have $\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x')\, \mathrm{d}\mathbb{F}(x)\, \mathrm{d}\mathbb{F}(x) = (a_1 - a_2)^2$ where $\mathbb{F} = a_1\delta_1 + a_2\delta_2$ for some $a_1, a_2 \in \mathbb{R}\backslash\{0\}$. Clearly, the choice of $a_1 = a_2$ establishes that there exists $\mathbb{F} \in \mathcal{M}_b(\mathcal{X})\backslash\{0\}$ such that $\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x')\, \mathrm{d}\mathbb{F}(x)\, \mathrm{d}\mathbb{F}(x) = 0$. Hence $k$ is not universal. Note that the constraint in (9), which is needed to verify the characteristic property of $k$ is not needed to verify its universality.

### 5.3 Proof of Theorem 3

Define $\mathcal{H}_m := \mathcal{H}_{k_m}$.

$(i)$ Suppose $k_1$ and $k_2$ are characteristic and that for some $\mathbb{F} = \mathbb{P} - \mathbb{P}_1 \otimes \mathbb{P}_2 \in \mathcal{I}$,

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \ni \int_{\mathcal{X}_1 \times \mathcal{X}_2} (k_1 \otimes k_2)(\cdot, x) \,\mathrm{d}\mathbb{F}(x) = \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(\cdot, x_1) \otimes k_2(\cdot, x_2) \,\mathrm{d}\mathbb{F}(x) = 0, \tag{20}$$

where $x = (x_1, x_2)$. We want to show that $\mathbb{F} = 0$. By the second-countability of $\mathcal{X}_m$-s, the product $\sigma$-field, i.e., $\otimes_{m=1}^{2} \mathcal{B}(\mathcal{X}_m)$ generated by the cylinder sets $B_1 \times \mathcal{X}_2$ and $\mathcal{X}_1 \times B_2$ ($B_m \in \mathcal{B}(\mathcal{X}_m), m = 1, 2$), coincides with the Borel $\sigma$-field $\mathcal{B}(\mathcal{X}_1 \times \mathcal{X}_2)$ on the product space (Dudley, 2004, Lemma 4.1.7):

$$\otimes_{m=1}^{2} \mathcal{B}(\mathcal{X}_m) = \mathcal{B}(\mathcal{X}_1 \times \mathcal{X}_2).$$

Hence, it is sufficient to prove that $\mathbb{F}(B_1 \times B_2) = 0$, $\forall\, B_m \in \mathcal{B}(\mathcal{X}_m)$, $m = 1, 2$. To this end, it follows from (20) that for all $h_2 \in \mathcal{H}_2$,

$$\mathcal{H}_1 \ni \int_{\mathcal{X}_1 \times \mathcal{X}_2} k_1(\cdot, x_1) h_2(x_2) \,\mathrm{d}\mathbb{F}(x) = \int_{\mathcal{X}_1} k_1(\cdot, x_1) \,\mathrm{d}\nu(x_1) = 0, \tag{21}$$

where

$$\nu(B_1) := \nu_{h_2}(B_1) = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \chi_{B_1}(x_1) h_2(x_2) \,\mathrm{d}\mathbb{F}(x), \quad B_1 \in \mathcal{B}(\mathcal{X}_1).$$

Since $k_1$ is characteristic, (21) implies $\nu = 0$, provided that $|\nu|(\mathcal{X}_1) < \infty$ and $\nu(\mathcal{X}_1) = 0$. These two requirements hold:

$$\begin{aligned}
\nu(\mathcal{X}_1) &= \int_{\mathcal{X}_1 \times \mathcal{X}_2} h_2(x_2) \,\mathrm{d}\mathbb{F}(x) = \int_{\mathcal{X}_2} h_2(x_2) \,\mathrm{d}[\mathbb{P}_2 - \mathbb{P}_2](x_2) = 0,\\
|\nu|(\mathcal{X}_1) &\le \int_{\mathcal{X}_1 \times \mathcal{X}_2} \underbrace{|h_2(x_2)|}_{\left|\langle h_2, k_2(\cdot, x_2)\rangle_{\mathcal{H}_2}\right|} \,\mathrm{d}[\mathbb{P} + \mathbb{P}_1 \otimes \mathbb{P}_2](x_1, x_2)\\
&\le \|h_2\|_{\mathcal{H}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_2} \sqrt{k_2(x_2, x_2)} \,\mathrm{d}[\mathbb{P} + \mathbb{P}_1 \otimes \mathbb{P}_2](x_1, x_2)\\
&\le 2\, \|h_2\|_{\mathcal{H}_2} \int_{\mathcal{X}_2} \sqrt{k_2(x_2, x_2)} \,\mathrm{d}\mathbb{P}_2(x_2) < \infty,
\end{aligned}$$

where the last inequality follows from the boundedness of $k_2$. The established $\nu = 0$ implies that for $\forall\, B_1 \in \mathcal{B}(\mathcal{X}_1)$ and $\forall\, h_2 \in \mathcal{H}_2$,

$$0 = \nu(B_1) = \left\langle h_2, \int_{\mathcal{X}_1 \times \mathcal{X}_2} \chi_{B_1}(x_1) k_2(\cdot, x_2) \,\mathrm{d}\mathbb{F}(x) \right\rangle_{\mathcal{H}_2},$$

and hence

$$0 = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \chi_{B_1}(x_1) k_2(\cdot, x_2) \,\mathrm{d}\mathbb{F}(x) = \int_{\mathcal{X}_2} k_2(\cdot, x_2) \,\mathrm{d}\theta_{B_1}(x_2), \tag{22}$$

where

$$\theta_{B_1}(B_2) = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \chi_{B_1}(x_1) \chi_{B_2}(x_2) \,\mathrm{d}\mathbb{F}(x), \quad B_2 \in \mathcal{B}(\mathcal{X}_2).$$

Using the characteristic property of $k_2$, it follows from (22) that $\theta_{B_1} = 0$ for $\forall B_1 \in \mathcal{B}(\mathcal{X}_1)$, i.e.,

$$0 = \theta_{B_1}(B_2) = \mathbb{F}(B_1 \times B_2), \quad \forall\, B_1 \in \mathcal{B}(\mathcal{X}_1), \forall\, B_2 \in \mathcal{B}(\mathcal{X}_2)$$

provided that $\theta_{B_1}(\mathcal{X}_2) = 0$ and $|\theta_{B_1}|(\mathcal{X}_2) < \infty$. Indeed, both these conditions hold:

$$\theta_{B_1}(\mathcal{X}_2) = \int_{\mathcal{X}_1 \times \mathcal{X}_2} \chi_{B_1}(x_1)\, \mathrm{d}\mathbb{F}(x) = \int_{\mathcal{X}_1} \chi_{B_1}(x_1)\, \mathrm{d}[\mathbb{P}_1 - \mathbb{P}_1](x_1) = 0,$$

$$|\theta_{B_1}|(\mathcal{X}_2) \leq \int_{\mathcal{X}_1 \times \mathcal{X}_2} \mathrm{d}[\mathbb{P} + \mathbb{P}_1 \otimes \mathbb{P}_2](x) = 2.$$

$(ii)$ Assume w.l.o.g. that $k_1$ is not characteristic. This means there exists $\mathbb{P}_1 \neq \mathbb{P}_1' \in \mathcal{M}_1^+(\mathcal{X}_1)$ such that $\mu_{k_1}(\mathbb{P}_1) = \mu_{k_1}(\mathbb{P}_1')$. Our goal is to construct an $\mathbb{F} \in \mathcal{M}_1^+\left(\times_{m=1}^M \mathcal{X}_m\right)$ such that

$$\mu_{\otimes_{m=1}^M k_m}\left(\mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m\right) = \int_{\times_{m=1}^M} \otimes_{m=1}^M k_m(\cdot, x_m)\, \mathrm{d}\big[\mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m\big] = 0, \text{ but } \mathbb{F} \neq \otimes_{m=1}^M \mathbb{F}_m.$$

Define $\mathbb{I} := \mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m \in \mathcal{I}$. In other words we want to get a witness $\mathbb{I} \in \mathcal{I}$ proving that $\otimes_{m=1}^M k_m$ is not $\mathcal{I}$-characteristic. Let us take $z \neq z' \in \mathcal{X}_2$, which is possible since $|\mathcal{X}_2| \geq 2$. Let us define $\mathbb{F}$ as[6]

$$\mathbb{F} = \frac{\mathbb{P}_1 \otimes \delta_z \otimes (\otimes_{m=3}^M \mathbb{Q}_m) + \mathbb{P}_1' \otimes \delta_{z'} \otimes (\otimes_{m=3}^M \mathbb{Q}_m)}{2} \in \mathcal{M}_1^+\left(\times_{m=1}^M \mathcal{X}_m\right).$$

It is easy to verify that

$$\mathbb{F}_1 = \frac{\mathbb{P}_1 + \mathbb{P}_1'}{2},\ \mathbb{F}_2 = \frac{\delta_z + \delta_{z'}}{2} \text{ and } \mathbb{F}_m = \mathbb{Q}_m \quad (m = 3, \ldots, M),$$

where $\mathbb{Q}_3, \ldots, \mathbb{Q}_M$ are arbitrary probability measures on $\mathcal{X}_3, \ldots, \mathcal{X}_M$, respectively. First we check that $\mathbb{I} \neq 0$. Indeed it is the case since

- $z \neq z'$ and $\mathcal{X}_2$ is a Hausdorff space, there exists $B_2 \in \mathcal{B}(\mathcal{X}_2)$ such that $z \in B_2$, $z' \notin B_2$.
- $\mathbb{P}_1 \neq \mathbb{P}_1'$, $\mathbb{P}_1(B_1) \neq \mathbb{P}_1'(B_1)$ for some $B_1 \in \mathcal{B}(\mathcal{X}_1)$.

Let $S = B_1 \times B_2 \times \left(\times_{m=3}^M \mathcal{X}_m\right)$, and compare its measure under $\mathbb{F}$ and $\otimes_{m=1}^M \mathbb{F}_m$:

$$\begin{aligned}
\mathbb{F}(S) &= \frac{\mathbb{P}_1(B_1)\ \overbrace{\delta_z(B_2)}^{=1\,(z \in B_2)}\ \prod_{m=3}^M \overbrace{\mathbb{Q}_m(\mathcal{X}_m)}^{=1} + \mathbb{P}_1'(B_1)\ \overbrace{\delta_{z'}(B_2)}^{=0\,(z' \notin B_2)}\ \prod_{m=3}^M \overbrace{\mathbb{Q}_m(\mathcal{X}_m)}^{=1}}{2} \\
&= \frac{\mathbb{P}_1(B_1)}{2},
\end{aligned}$$

$$\left(\otimes_{m=1}^M \mathbb{F}_m\right)(S) = \prod_{m=1}^M \mathbb{F}_m(B_m) = \frac{\mathbb{P}_1(B_1) + \mathbb{P}_1'(B_1)}{2}\, \frac{\overbrace{\delta_z(B_2)}^{=1} + \overbrace{\delta_{z'}(B_2)}^{=0}}{2} \prod_{m=3}^M \overbrace{\mathbb{Q}_m(\mathcal{X}_m)}^{=1}$$

6. The $\mathbb{F}$ construction specializes to that of Lyons (2013, Proposition 3.15) in the $M = 2$ case; Lyons used it for distance covariances, which is known to be equivalent to HSIC (Sejdinovic et al., 2013b).

$$= \frac{\mathbb{P}_1(B_1) + \mathbb{P}'_1(B_1)}{4} \neq \frac{\mathbb{P}_1(B_1)}{2},$$

where the last equality holds since $\mathbb{P}_1(B_1) \neq \mathbb{P}'_1(B_1)$. This shows that $\mathbb{I} = \mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m \neq 0$ since $\mathbb{I}(S) \neq 0$.
Next we prove that $\mu_{\otimes_{m=1}^M k_m}\left(\mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m\right) = 0$. Indeed,

$$\begin{aligned}
\mu_{\otimes_{m=1}^M k_m}(\mathbb{I}) &= \int_{\times_{m=1}^M \mathcal{X}_m} \otimes_{m=1}^M k_m(\cdot, x_m)\, \mathrm{d}\left[\mathbb{F} - \otimes_{m=1}^M \mathbb{F}_m\right](x_1, \ldots, x_M) \\
&= \int_{\times_{m=1}^M \mathcal{X}_m} \otimes_{m=1}^M k_m(\cdot, x_m)\, \mathrm{d}\left(\left[\frac{\mathbb{P}_1 \otimes \delta_z + \mathbb{P}'_1 \otimes \delta_{z'}}{2} - \frac{\mathbb{P}_1 + \mathbb{P}'_1}{2} \otimes \frac{\delta_z + \delta_{z'}}{2}\right]\right. \\
&\qquad\qquad \left. \otimes \left(\otimes_{m=3}^M \mathbb{Q}_m\right)\right)(x_1, \ldots, x_M) \\
&= \int_{\times_{m=1}^M \mathcal{X}_m} \otimes_{m=1}^M k_m(\cdot, x_m)\, \mathrm{d}\left(\left[\frac{\mathbb{P}_1(x_1) \otimes \delta_z(x_2) + \mathbb{P}'_1(x_1) \otimes \delta_{z'}(x_2)}{2}\right.\right. \\
&\qquad - \frac{\mathbb{P}_1(x_1) \otimes \delta_z(x_2) + \mathbb{P}_1(x_1) \otimes \delta_{z'}(x_2)}{4} \\
&\qquad\quad \left.\left. - \frac{\mathbb{P}'_1(x_1) \otimes \delta_z(x_2) + \mathbb{P}'_1(x_1) \otimes \delta_{z'}(x_2)}{4}\right] \otimes \left(\otimes_{m=3}^M \mathbb{Q}_m(x_m)\right)\right) \\
&\overset{(*)}{=} \left[\frac{\mu_{k_1}(\mathbb{P}_1) \otimes k_2(\cdot, z) + \mu_{k_1}(\mathbb{P}'_1) \otimes k_2(\cdot, z')}{2}\right. \\
&\qquad - \frac{\mu_{k_1}(\mathbb{P}_1) \otimes k_2(\cdot, z) + \mu_{k_1}(\mathbb{P}_1) \otimes k_2(\cdot, z')}{4} \\
&\qquad\quad \left. - \frac{\mu_{k_1}(\mathbb{P}'_1) \otimes k_2(\cdot, z) + \mu_{k_1}(\mathbb{P}'_1) \otimes k_2(\cdot, z')}{4}\right] \otimes \left[\otimes_{m=3}^M \mu_{k_m}(\mathbb{Q}_m)\right] \\
&= \underbrace{0}_{\in \mathcal{H}_{k_1 \otimes k_2}} \otimes \left[\otimes_{m=3}^M \mu_{k_m}(\mathbb{Q}_m)\right] = 0,
\end{aligned}$$

where we used $\mu_{k_1}(\mathbb{P}_1) = \mu_{k_1}(\mathbb{P}'_1)$ in $(*)$.

### 5.4 Proof of Example 2

Let $M = 3$, $\times_{m=1}^M \mathcal{X}_m = \{(i_1, i_2, i_3) : i_m \in \{1, 2\},\, m \in [3]\}$, $k_m(x, x') = 2\delta_{x,x'} - 1$. Our goal is to show that $\otimes_{m=1}^3 k_m$ is *not* $\mathcal{I}$-characteristic. The structure of the proof is as follows:

1. First we describe the equations of the non-characteristic property of $\otimes_{m=1}^3 k_m$ with a general finite signed measure $\mathbb{F} = \sum_{i_1,i_2,i_3=1}^2 a_{i_1,i_2,i_3} \delta_{(i_1,i_2,i_3)}$ on $\times_{m=1}^3 \mathcal{X}_m$ where $a_{i_1,i_2,i_3} \in \mathbb{R}$ $(\forall\, i_1, i_2, i_3)$.
2. Next, we apply the $\mathbb{F} = \mathbb{P} - \otimes_{m=1}^3 \mathbb{P}_m$ parameterization and show that there exists $\mathbb{P}$ that satisfies the equations of step 1 to conclude that $\otimes_{m=1}^3 k_m$ is not $\mathcal{I}$-characteristic.

The details are as follows.

**Step 1**. The equations of non-characteristic property in terms of $\mathbf{A} = [a_{i_1,i_2,i_3}]_{(i_m)_{m=1}^3 \in [2]^3} \in \mathbb{R}^{2\times2\times2}$ are

$$\mathbb{F} \in \mathcal{M}_b\left(\times_{m=1}^3 \mathcal{X}_m\right) \setminus \{0\} \Leftrightarrow \mathbf{A} \neq \mathbf{0},$$

$$0 = \mathbb{F}(\times_{m=1}^{3} \mathcal{X}_m) \Leftrightarrow 0 = \sum_{i_1,i_2,i_3=1}^{2} a_{i_1,i_2,i_3}, \tag{23}$$

$$0 = \int_{\times_{m=1}^{3} \mathcal{X}_m} \int_{\times_{m=1}^{3} \mathcal{X}_m} \underbrace{(\otimes_{m=1}^{3} k_m)\left((i_1,i_2,i_3),(i_1',i_2',i_3')\right)}_{\prod_{m=1}^{3} k_m(i_m,i_m')} \mathrm{d}\mathbb{F}(i_1,i_2,i_3)\, \mathrm{d}\mathbb{F}(i_1',i_2',i_3')$$

$$= \sum_{i_1,i_2,i_3=1}^{2} \sum_{i_1',i_2',i_3'=1}^{2} \prod_{m=1}^{3} k_m(i_m,i_m') a_{i_1,i_2,i_3} a_{i_1',i_2',i_3'}. \tag{24}$$

Solving (23) and (24) yields

$$a_{1,1,1} + a_{1,2,2} + a_{2,1,2} + a_{2,2,1} = 0 \quad \text{and} \quad a_{1,1,2} + a_{1,2,1} + a_{2,1,1} + a_{2,2,2} = 0.$$

**Step 2**. The equations of non $\mathcal{I}$-characteristic property can be obtained from step 1 by choosing $\mathbb{F} = \mathbb{P} - \otimes_{m=1}^{M} \mathbb{P}_m$, where

$$\mathbb{P} = \sum_{i_1,i_2,i_3=1}^{2} p_{i_1,i_2,i_3} \delta_{(i_1,i_2,i_3)} \quad \text{and} \quad \mathbf{P} = [p_{i_1,i_2,i_3}]_{(i_m)_{m=1}^{3} \in [2]^3} \in \mathbb{R}^{2\times 2\times 2}.$$

In other words, it is sufficient to obtain a $\mathbf{P}$ that solves the following system of equations for which $\mathbf{A} = \mathbf{A}(\mathbf{P}) \neq \mathbf{0}$:

$$\sum_{i_1,i_2,i_3=1}^{2} p_{i_1,i_2,i_3} = 1, \tag{25}$$

$$p_{i_1,i_2,i_3} \geq 0,\ \forall\, (i_1,i_2,i_3) \in [2]^3, \tag{26}$$

$$a_{1,1,1} + a_{1,2,2} + a_{2,1,2} + a_{2,2,1} = 0, \tag{27}$$

$$a_{1,1,2} + a_{1,2,1} + a_{2,1,1} + a_{2,2,2} = 0, \tag{28}$$

where

$$a_{i_1,i_2,i_3} = p_{i_1,i_2,i_3} - p_{1,i_1} p_{2,i_2} p_{3,i_3}, \tag{29}$$

and

$$p_{1,i_1} = \sum_{i_2,i_3=1}^{2} p_{i_1,i_2,i_3}, \quad p_{2,i_2} = \sum_{i_1,i_3=1}^{2} p_{i_1,i_2,i_3}, \quad p_{3,i_3} = \sum_{i_1,i_2=1}^{2} p_{i_1,i_2,i_3}. \tag{30}$$

One can get an analytical description for the solution of (25)–(30), where the solution $\mathbf{P}(\mathbf{z})$ is parameterized by $\mathbf{z} = (z_0, \ldots, z_5) \in \mathbb{R}^6$. For explicit expressions, we refer the reader to Appendix A. In the following, we present two examples of $\mathbf{P}$ that satisfy (25)–(30) such that $\mathbf{A} \neq \mathbf{0}$, thereby establishing the non $\mathcal{I}$-characteristic property of $\otimes_{m=1}^{3} k_m$.

1. $\mathbf{P}$:

$$p_{1,1,1} = \frac{1}{5}, \qquad p_{1,1,2} = \frac{1}{10}, \qquad p_{1,2,1} = \frac{1}{10}, \qquad p_{1,2,2} = \frac{1}{10},$$
$$p_{2,1,1} = \frac{1}{5}, \qquad p_{2,1,2} = \frac{1}{10}, \qquad p_{2,2,1} = \frac{1}{10}, \qquad p_{2,2,2} = \frac{1}{10},$$

and $\mathbf{A}$:

$$a_{1,1,1} = \frac{1}{50}, \qquad a_{1,1,2} = -\frac{1}{50}, \qquad a_{1,2,1} = -\frac{1}{50}, \qquad a_{1,2,2} = \frac{1}{50}, \tag{31}$$

$$a_{2,1,1} = \frac{1}{50}, \qquad a_{2,1,2} = -\frac{1}{50}, \qquad a_{2,2,1} = -\frac{1}{50}, \qquad a_{2,2,2} = \frac{1}{50}. \tag{32}$$

2. $\mathbf{P}$:

$$p_{1,1,1} = 0, \qquad p_{1,1,2} = \frac{1}{10}, \qquad p_{1,2,1} = \frac{1}{10}, \qquad p_{1,2,2} = \frac{1}{10},$$
$$p_{2,1,1} = \frac{1}{10}, \qquad p_{2,1,2} = \frac{1}{10}, \qquad p_{2,2,1} = \frac{3}{10}, \qquad p_{2,2,2} = \frac{1}{5},$$

and $\mathbf{A}$:

$$a_{1,1,1} = -\frac{9}{200}, \qquad a_{1,1,2} = \frac{11}{200}, \qquad a_{1,2,1} = -\frac{1}{200}, \qquad a_{1,2,2} = -\frac{1}{200},$$
$$a_{2,1,1} = -\frac{1}{200}, \qquad a_{2,1,2} = -\frac{1}{200}, \qquad a_{2,2,1} = \frac{11}{200}, \qquad a_{2,2,2} = -\frac{9}{200}.$$

In fact these examples are obtained with the choices $\mathbf{z} = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)$ and $\mathbf{z} = \left(\frac{3}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{2}{10}\right)$ respectively. See Appendix A for details.

### 5.5 Proof of Theorem 4

It follows from (8) and Remark 2(iii) that $(v) \Rightarrow (iii) \Rightarrow (ii) \Leftrightarrow (i)$. It also follows from (8) and Theorem 3(ii) that $(v) \Rightarrow (iv) \Rightarrow (i)$. We now show that $(i) \Rightarrow (v)$ which establishes the equivalence of $(i)$–$(v)$. Suppose $(i)$ holds. Then by Bochner's theorem (Wendland, 2005, Theorem 6.6), we have that for all $m \in [M]$,

$$k_m(x_m, y_m) = \int_{\mathbb{R}^{d_m}} e^{-\sqrt{-1}\langle \omega_m, x_m - y_m\rangle} \,\mathrm{d}\Lambda_m(\omega_m), \ x_m, y_m \in \mathbb{R}^{d_m},$$

where $(\Lambda_m)_{m=1}^M$ are finite non-negative Borel measures on $(\mathbb{R}^{d_m})_{m=1}^M$ respectively. This implies

$$\otimes_{m=1}^M k_m(x_m, y_m) = \otimes_{m=1}^M \int_{\mathbb{R}^{d_m}} e^{-\sqrt{-1}\langle \omega_m, x_m - y_m\rangle} \,\mathrm{d}\Lambda_m(\omega_m) = \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle \omega, x - y\rangle} \,\mathrm{d}\Lambda(\omega),$$

where $x = (x_1, \ldots, x_M) \in \mathbb{R}^d$, $y = (y_1, \ldots, y_M) \in \mathbb{R}^d$, $\omega = (\omega_1, \ldots, \omega_M) \in \mathbb{R}^d$, $d = \sum_{m=1}^M d_m$ and $\Lambda := \otimes_{m=1}^M \Lambda_m$. Sriperumbudur et al. (2010, Theorem 9) showed that $k_m$ is characteristic iff $\mathrm{supp}\,(\Lambda_m) = \mathbb{R}^{d_m}$, where $\mathrm{supp}(\cdot)$ denotes the support of its argument. Since $\mathrm{supp}(\Lambda) = \mathrm{supp}\left(\otimes_{m=1}^M \Lambda_m\right) = \times_{m=1}^M \mathrm{supp}\,(\Lambda_m) = \times_{m=1}^M \mathbb{R}^{d_m} = \mathbb{R}^d$, it follows that $\otimes_{m=1}^M k_m$ is characteristic.

### 5.6 Proof of Theorem 5

The $c_0$-kernel property of $k_m$-s ($m = 1, \ldots, M$) implies that of $\otimes_{m=1}^M k_m$. Moreover, $\mathcal{X}_m$-s are LCP spaces, hence $\times_{m=1}^M \mathcal{X}_m$ is also LCP.

($\Leftarrow$) Assume that $\otimes_{m=1}^M k_m$ is $c_0$-universal. Since $\otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m) \subseteq \mathcal{M}_b\left(\times_{m=1}^M \mathcal{X}_m\right)$, we have that for all $\mathbb{F} = \otimes_{m=1}^M \mathbb{F}_m \in \otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\backslash\{0\}$,

$$0 < \int_{\times_{m=1}^M \mathcal{X}_m} \int_{\times_{m=1}^M \mathcal{X}_m} \underbrace{\left(\otimes_{m=1}^M k_m\right)(x, x')}_{\prod_{m=1}^M k_m(x_m, x'_m)} \mathrm{d}\mathbb{F}(x)\, \mathrm{d}\mathbb{F}(x')$$
$$= \prod_{m=1}^M \int_{\mathcal{X}_m \times \mathcal{X}_m} k_m(x_m, x'_m)\, \mathrm{d}\mathbb{F}_m(x_m)\, \mathrm{d}\mathbb{F}_m\left(x'_m\right),$$

where $x = (x_1, \ldots, x_M)$ and $x' = (x'_1, \ldots, x'_M)$. The above inequality implies

$$\int_{\mathcal{X}_m \times \mathcal{X}_m} k_m(x_m, x'_m)\, \mathrm{d}\mathbb{F}_m(x_m)\, \mathrm{d}\mathbb{F}_m\left(x'_m\right) > 0, \ \forall\, m \in [M].$$

Since $\mathbb{F} \in \otimes_{m=1}^M \mathcal{M}_b(\mathcal{X}_m)\backslash\{0\}$ iff $\mathbb{F}_m \in \mathcal{M}_b(\mathcal{X}_m)\backslash\{0\}$ for all $m \in [M]$, the result follows.

($\Rightarrow$) Assume that $k_m$-s are $c_0$-universal. By the note above $\otimes_{m=1}^M k_m$ is $c_0$-kernel; its $c_0$-universality is equivalent to the injectivity of $\mu = \mu_{\otimes_{m=1}^M k_m}$ on $\mathcal{M}_b\left(\times_{m=1}^M \mathcal{X}_m\right)$. In other words, we want to prove that $\mu(\mathbb{F}) = 0$ implies $\mathbb{F} = 0$, where $\mathbb{F} \in \mathcal{M}_b\left(\times_{m=1}^M \mathcal{X}_m\right)$. We will use the shorthand $\mathcal{H}_m = \mathcal{H}_{k_m}$ below.

Suppose there exists $\mathbb{F} \in \mathcal{M}_b\left(\times_{m=1}^M \mathcal{X}_m\right)$ such that

$$\mu_{\mathbb{F}} = \int_{\times_{m=1}^M \mathcal{X}_m} \underbrace{\left(\otimes_{m=1}^M k_m\right)(\cdot, x)}_{\otimes_{m=1}^M k_m(\cdot, x_m)} \mathrm{d}\mathbb{F}(x) = 0 \ (\in \otimes_{m=1}^M \mathcal{H}_m). \tag{33}$$

Since $\mathcal{X}_m$-s are LCP, $\otimes_{m=1}^M \mathcal{B}(\mathcal{X}_m) = \mathcal{B}\left(\times_{m=1}^M \mathcal{X}_m\right)$ (Steinwart and Christmann, 2008, page 480). Hence, in order to get $\mathbb{F} = 0$ it is sufficient to prove that

$$\mathbb{F}\left(\times_{m=1}^M B_m\right) = 0, \quad \forall\, B_m \in \mathcal{B}(\mathcal{X}_m), m \in [M].$$

We will prove by induction that for $m = 0, \ldots, M$

$$\left(\otimes_{j=m+1}^M \mathcal{H}_j \ni\right) 0 = \int_{\times_{j=1}^M \mathcal{X}_j} \prod_{j=1}^m \chi_{B_j}(x_j) \otimes_{j=m+1}^M k_j(\cdot, x_j)\, \mathrm{d}\mathbb{F}(x)$$
$$=: o(B_1, \ldots, B_m, k_{m+1}, \ldots, k_M), \forall\, B_j \in \mathcal{B}(\mathcal{X}_j),\ j \in [m], \tag{34}$$

which

- ($*$) reduces to (33) when $m = 0$ by defining $\prod_{j=1}^0 \chi_{B_j}(x_j) := 1$;
- ($\dagger$) for $m = M$, $\otimes_{m=M+1}^M \mathcal{H}_m$ is defined to be equal to $\mathbb{R}$ and $\otimes_{m=M+1}^M k_m(\cdot, x_m) := 1$, in which case $o(B_1, \ldots, B_M) = \mathbb{F}\left(\times_{j=1}^M B_j\right) = 0 \Rightarrow \mathbb{F} = 0$, the result we want to prove.

From the above, it is clear that (34) holds for $m = 0$. Assuming (34) holds for some $m$, we now prove that it holds for $m + 1$. To this end, it follows from (34) that $\forall\, h_{m+2} \in \mathcal{H}_{m+2}, \ldots, \forall\, h_M \in \mathcal{H}_M$,

$$(\mathcal{H}_{m+1} \ni)\, 0 = o(B_1, \ldots, B_m, k_{m+1}, \ldots, k_M)\, (h_{m+2}, \ldots, h_M)$$

$$
\begin{aligned}
&= \left[\int_{\times_{j=1}^M \mathcal{X}_j} \left(\prod_{j=1}^m \chi_{B_j}(x_j)\right) \otimes_{j=m+1}^M k_j(\cdot, x_j)\,\mathrm{d}\mathbb{F}(x)\right](h_{m+2},\ldots,h_M)\\
&= \int_{\times_{j=1}^M \mathcal{X}_j} k_{m+1}(\cdot, x_{m+1}) \prod_{j=1}^m \chi_{B_j}(x_j) \prod_{j=m+2}^M h_j(x_j)\,\mathrm{d}\mathbb{F}(x)\\
&= \int_{\mathcal{X}_{m+1}} k_{m+1}(\cdot, x_{m+1})\,\mathrm{d}\nu(x_{m+1}),
\end{aligned}
$$

where

$$
\begin{aligned}
\nu(B) &:= \nu_{B_1,\ldots,B_m,h_{m+2},\ldots,h_M}(B)\\
&= \int_{\times_{j=1}^M \mathcal{X}_j} \left[\prod_{j=1}^m \chi_{B_j}(x_j)\right] \chi_B(x_{m+1}) \left[\prod_{j=m+2}^M h_j(x_j)\right] \mathrm{d}\mathbb{F}(x),\ B \in \mathcal{B}(\mathcal{X}_{m+1}).
\end{aligned}
$$

By the $c_0$-universality of $k_{m+1}$,

$$
\nu = 0 \text{ for } \forall\, h_{m+2} \in \mathcal{H}_{m+2},\ldots,\forall\, h_M \in \mathcal{H}_M \tag{35}
$$

provided that $\nu \in \mathcal{M}_b(\mathcal{X}_{m+1})$, in other words if $|\nu|(\mathcal{X}_{m+1}) < \infty$. This condition is met:

$$
\begin{aligned}
|\nu|(\mathcal{X}_{m+1}) &\le \int_{\times_{j=1}^M \mathcal{X}_j} \prod_{j=m+2}^M \underbrace{\left|\langle h_j, k_j(\cdot, x_j)\rangle_{\mathcal{H}_j}\right|}_{\le \|h_j\|_{\mathcal{H}_j}\sqrt{k_j(x_j,x_j)}} \mathrm{d}|\mathbb{F}|(x)\\
&\le |\mathbb{F}|\left(\times_{m=1}^M \mathcal{X}_m\right) \prod_{j=m+2}^M \|h_j\|_{\mathcal{H}_j} \sup_{x\in\mathcal{X}_j, x'\in\mathcal{X}_j} \sqrt{k_j(x,x')} < \infty,
\end{aligned}
$$

where we used the boundedness of $k_m$-s in the last inequality. (35) implies that for $\forall\, B_1 \in \mathcal{B}(\mathcal{X}_1),\ldots,\forall\, B_{m+1} \in \mathcal{B}(\mathcal{X}_{m+1})$ and $\forall\, h_{m+2} \in \mathcal{H}_{m+2},\ldots,\forall\, h_M \in \mathcal{H}_M$

$$
\begin{aligned}
0 = \nu(B_{m+1}) &= \int_{\times_{j=1}^M \mathcal{X}_j} \left[\prod_{j=1}^{m+1} \chi_{B_j}(x_j)\right] \left[\prod_{j=m+2}^M h_j(x_j)\right] \mathrm{d}\mathbb{F}(x)\\
&= \left\langle \otimes_{j=m+2}^M h_j, \int_{\times_{j=1}^M \mathcal{X}_j} \left[\prod_{j=1}^{m+1} \chi_{B_j}(x_j)\right] \otimes_{j=m+2}^M k_j(\cdot, x_j)\,\mathrm{d}\mathbb{F}(x) \right\rangle_{\otimes_{j=m+2}^M \mathcal{H}_j},
\end{aligned}
$$

and therefore

$$
\begin{aligned}
o(B_1,\ldots,B_{m+1},k_{m+2},\ldots,k_M) &= \int_{\times_{j=1}^M \mathcal{X}_j} \left[\prod_{j=1}^{m+1} \chi_{B_j}(x_j)\right] \otimes_{j=m+2}^M k(\cdot, x_j)\,\mathrm{d}\mathbb{F}(x)\\
&= 0 \left(\in \otimes_{j=m+2}^M \mathcal{H}_j\right)
\end{aligned}
$$

for $\forall\, B_1 \in \mathcal{B}(\mathcal{X}_1), \ldots, \forall\, B_{m+1} \in \mathcal{B}(\mathcal{X}_{m+1})$, i.e., (34) holds for $m+1$. Therefore, by induction, (34) holds for $m = M$ and the result follows from (†). To justify the convention in (†), consider the case of $m = M-1$ in which case (34) can be written as

$$\int_{\mathcal{X}_M} k_M(\cdot, x_M)\, \mathrm{d}\nu(x_M) = 0,$$

where

$$\nu(B) = \int_{\times_{j=1}^M \mathcal{X}_j} \left[\prod_{j=1}^{M-1} \chi_{B_j}(x_j)\right] \chi_B(x_M)\, \mathrm{d}\mathbb{F}(x),\ B \in \mathcal{B}(\mathcal{X}_M).$$

Then by the $c_0$-universal property of $k_M$, since

$$|\nu|(\mathcal{X}_M) \le \int_{\times_{j=1}^M \mathcal{X}_j} 1\, \mathrm{d}|\mathbb{F}|(x) = |\mathbb{F}|\left(\times_{j=1}^M \mathcal{X}_j\right) < \infty$$

we obtain

$$\int_{\times_{j=1}^M \mathcal{X}_j} \prod_{j=1}^M \chi_{B_j}(x_j)\, \mathrm{d}\mathbb{F}(x) = \mathbb{F}\left(\times_{j=1}^M B_j\right) = 0, \forall\, B_1 \in \mathcal{B}(\mathcal{X}_1), \ldots, \forall\, B_M \in \mathcal{B}(\mathcal{X}_M).$$

### 5.7 Proof of Example 3

The proof follows by a simple modification of that of Example 2 (Section 5.4). The equations of a witness $\mathbf{A} = [a_{i_1,i_2,i_3}]_{(i_m)_{m=1}^3 \in [2]^3}$ (and corresponding $\mathbf{P} = [p_{i_1,i_2,i_3}]_{(i_m)_{m=1}^3 \in [2]^3}$) for the non-$\mathcal{I}$-characteristic property of $\otimes_{m=1}^3 k_m$ take the form:

$$\begin{aligned}
&\mathbf{A} \neq \mathbf{0},\\
&0 = \sum_{i_1,i_2,i_3=1}^{2} a_{i_1,i_2,i_3}, &(36)\\
&0 = \sum_{i_1,i_2,i_3=1}^{2} \sum_{i_1',i_2',i_3'=1}^{2} \prod_{m=1}^{3} k_m(i_m, i_m') a_{i_1,i_2,i_3} a_{i_1',i_2',i_3'}\\
&\quad = (a_{1,1,1} - a_{2,1,1})^2 + (a_{1,1,2} - a_{2,1,2})^2 + (a_{1,2,1} - a_{2,2,1})^2 + (a_{1,2,2} - a_{2,2,2})^2, &(37)
\end{aligned}$$

where (36) and (37) are equivalent to

$$0 = \sum_{i_1,i_2,i_3=1}^{2} a_{i_1,i_2,i_3}, \quad a_{1,1,1} = a_{2,1,1}, \quad a_{1,1,2} = a_{2,1,2}, \quad a_{1,2,1} = a_{2,2,1}, \quad a_{1,2,2} = a_{2,2,2}. \quad (38)$$

While (38) is more restrictive than (27) and (28) (hence its solution set might even be empty), one can immediately see that the example of $\mathbf{A} \neq \mathbf{0}$ given in (31) and (32) fulfills (38) proving the non-$\mathcal{I}$-characteristic property of $\otimes_{m=1}^3 k_m$.

## Acknowledgments

The authors profusely thank Ingo Steinwart for fascinating discussions on topics related to the paper and for contributing to Remark 7. The authors also thank the anonymous reviewers for their constructive comments that improved the manuscript. A part of the work was carried out while BKS was visiting ZSz at CMAP, École Polytechnique. BKS is supported by NSF-DMS-1713011 and also thanks CMAP and DSI for their generous support. ZSz is highly grateful for the Greek hospitality around the Aegean Sea; it greatly contributed to the development of the induction arguments.


## Appendix A. Analytical Solution to (25)–(30) in Example 2

The solution of (25)–(30) takes the form

$$
\begin{aligned}
p_{1,1,1} &= -\frac{\begin{array}{l} z_2+z_1+z_4+z_5-3z_2z_1-4z_2z_4-4z_1z_4-z_2z_3-2z_2z_0-2z_1z_3-3z_2z_5\\ -2z_4z_3-z_1z_0-3z_1z_5-2z_4z_0-4z_4z_5-z_3z_0-z_3z_5-z_0z_5+2z_2z_1^2+2z_2^2z_1\\ +4z_2z_4^2+2z_2^2z_4+4z_1z_4^2+2z_1^2z_4+2z_2^2z_0+2z_1^2z_3+2z_2z_5^2+2z_2^2z_5+2z_4^2z_3\\ +2z_1z_5^2+2z_1^2z_5+2z_4^2z_0+2z_4z_5^2+4z_4^2z_5-z_2^2-z_1^2-3z_4^2+2z_4^3-z_5^2\\ +6z_2z_1z_4+2z_2z_1z_3+2z_2z_4z_3+2z_2z_1z_0+4z_2z_1z_5+4z_2z_4z_0+4z_1z_4z_3\\ +6z_2z_4z_5+2z_1z_4z_0+6z_1z_4z_5+2z_2z_3z_0+2z_2z_3z_5+2z_1z_3z_0+2z_2z_0z_5\\ +2z_1z_3z_5+2z_4z_3z_0+2z_4z_3z_5+2z_1z_0z_5+2z_4z_0z_5 \end{array}}{\begin{array}{l} 2z_2z_1-z_1-2z_4-z_3-z_0-2z_5-z_2+2z_2z_4+2z_1z_4+2z_2z_0+2z_1z_3+2z_2z_5\\ +2z_4z_3+2z_1z_5+2z_4z_0+4z_4z_5+2z_3z_0+2z_3z_5+2z_0z_5+2z_4^2+2z_5^2 \end{array}},\\
p_{1,1,2} &= z_2,\\
p_{1,2,1} &= z_1,\\
p_{1,2,2} &= z_4,\\
p_{2,1,1} &= -\frac{\begin{array}{l} z_4+z_3+z_0+z_5-z_2z_1-z_2z_4-z_1z_4-z_2z_3-2z_2z_0-2z_1z_3-2z_2z_5\\ -3z_4z_3-z_1z_0-2z_1z_5-3z_4z_0-4z_4z_5-3z_3z_0-4z_3z_5-4z_0z_5+2z_2z_0^2\\ +2z_1z_3^2+2z_2z_5^2+2z_4z_3^2+2z_4^2z_3+2z_1z_5^2+2z_4z_0^2+2z_4^2z_0+4z_4z_5^2+2z_4^2z_5\\ +2z_3z_0^2+2z_3^2z_0+4z_3z_5^2+2z_3^2z_5+4z_0z_5^2+2z_0^2z_5-z_4^2-z_3^2-z_0^2-3z_5^2\\ +2z_5^3+2z_2z_1z_3+2z_2z_4z_3+2z_2z_1z_0+2z_2z_1z_5+2z_2z_4z_0+2z_1z_4z_3\\ +2z_2z_4z_5+2z_1z_4z_0+2z_1z_4z_5+2z_2z_3z_0+2z_2z_3z_5+2z_1z_3z_0+4z_2z_0z_5\\ +4z_1z_3z_5+4z_4z_3z_0+6z_4z_3z_5+2z_1z_0z_5+6z_4z_0z_5+6z_3z_0z_5 \end{array}}{\begin{array}{l} 2z_2z_1-z_1-2z_4-z_3-z_0-2z_5-z_2+2z_2z_4+2z_1z_4+2z_2z_0+2z_1z_3+2z_2z_5\\ +2z_4z_3+2z_1z_5+2z_4z_0+4z_4z_5+2z_3z_0+2z_3z_5+2z_0z_5+2z_4^2+2z_5^2 \end{array}},\\
p_{2,1,2} &= z_3,\\
p_{2,2,1} &= z_0,\\
p_{2,2,2} &= z_5,
\end{aligned}
$$

form, where $\mathbf{z} = (z_0, z_1, \ldots, z_5) \in \mathbb{R}^6$ satisfies

$$
\begin{aligned}
0 \leq\ &(2z_0z_2 - z_1 - z_2 - z_3 - 2z_4 - 2z_5 - z_0 + 2z_0z_3 + 2z_1z_2 + 2z_0z_4 + 2z_1z_3 + 2z_0z_5 \\
&+2z_1z_4 + 2z_1z_5 + 2z_2z_4 + 2z_2z_5 + 2z_3z_4 + 2z_3z_5 + 4z_4z_5 + 2z_4^2 + 2z_5^2) \times \\
&(z_0z_3 - z_3 - z_4 - z_5 - z_0z_1 - z_0 - z_1z_2 + z_0z_5 - 2z_1z_4 - z_2z_3 - z_1z_5 - 2z_2z_4 - z_2z_5 \\
&+z_3z_5 + 2z_0z_2^2 + 2z_1z_2^2 + 2z_1^2z_2 + 2z_0z_4^2 + 2z_1^2z_3 + 4z_1z_4^2 + 2z_1^2z_4 + 2z_1z_5^2 + 4z_2z_4^2 \\
&+2z_1^2z_5 + 2z_2^2z_4 + 2z_2z_5^2 + 2z_3z_4^2 + 2z_2^2z_5 + 2z_4z_5^2 + 4z_4^2z_5 - z_1^2 - z_2^2 - z_4^2 + 2z_4^3 + z_5^2 \\
&+2z_0z_1z_2 + 2z_0z_1z_3 + 2z_0z_1z_4 + 2z_0z_2z_3 + 2z_0z_1z_5 + 4z_0z_2z_4 + 2z_1z_2z_3 + 2z_0z_2z_5 \\
&+2z_0z_3z_4 + 6z_1z_2z_4 + 4z_1z_2z_5 + 4z_1z_3z_4 + 2z_0z_4z_5 + 2z_1z_3z_5 + 2z_2z_3z_4 + 6z_1z_4z_5 \\
&+2z_2z_3z_5 + 6z_2z_4z_5 + 2z_3z_4z_5)\,,
\end{aligned}
$$

$$
\begin{aligned}
0 \leq\ &(2z_0z_2 - z_1 - z_2 - z_3 - 2z_4 - 2z_5 - z_0 + 2z_0z_3 + 2z_1z_2 + 2z_0z_4 + 2z_1z_3 + 2z_0z_5 \\
&+2z_1z_4 + 2z_1z_5 + 2z_2z_4 + 2z_2z_5 + 2z_3z_4 + 2z_3z_5 + 4z_4z_5 + 2z_4^2 + 2z_5^2) \times \\
&(z_1z_2 - z_2 - z_4 - z_5 - z_0z_1 - z_0z_3 - z_1 - z_0z_4 - 2z_0z_5 + z_1z_4 - z_2z_3 + z_2z_4 \\
&-z_3z_4 - 2z_3z_5 + 2z_0^2z_2 + 2z_0z_3^2 + 2z_0^2z_3 + 2z_0z_4^2 + 2z_1z_3^2 + 2z_0^2z_4 + 4z_0z_5^2 + 2z_0^2z_5 \\
&+2z_1z_5^2 + 2z_2z_5^2 + 2z_3z_4^2 + 2z_3^2z_4 + 4z_3z_5^2 + 2z_3^2z_5 + 4z_4z_5^2 + 2z_4^2z_5 - z_0^2 - z_3^2 + z_4^2 \\
&-z_5^2 + 2z_5^3 + 2z_0z_1z_2 + 2z_0z_1z_3 + 2z_0z_1z_4 + 2z_0z_2z_3 + 2z_0z_1z_5 + 2z_0z_2z_4 + 2z_1z_2z_3 \\
&+4z_0z_2z_5 + 4z_0z_3z_4 + 6z_0z_3z_5 + 2z_1z_2z_5 + 2z_1z_3z_4 + 6z_0z_4z_5 + 4z_1z_3z_5 + 2z_2z_3z_4 \\
&+2z_1z_4z_5 + 2z_2z_3z_5 + 2z_2z_4z_5 + 6z_3z_4z_5)\,,
\end{aligned}
$$

$$
\begin{aligned}
&2z_0z_2 + 2z_0z_3 + 2z_1z_2 + 2z_0z_4 + 2z_1z_3 + 2z_0z_5 + 2z_1z_4 + 2z_1z_5 + 2z_2z_4 + 2z_2z_5 \\
&\quad + 2z_3z_4 + 2z_3z_5 + 4z_4z_5 + 2z_4^2 + 2z_5^2 \neq z_0 + z_1 + z_2 + z_3 + 2z_4 + 2z_5,
\end{aligned}
$$

$$
\begin{aligned}
&(2z_0z_2 - z_1 - z_2 - z_3 - 2z_4 - 2z_5 - z_0 + 2z_0z_3 + 2z_1z_2 + 2z_0z_4 + 2z_1z_3 + 2z_0z_5 \\
&\quad+2z_1z_4 + 2z_1z_5 + 2z_2z_4 + 2z_2z_5 + 2z_3z_4 + 2z_3z_5 + 4z_4z_5 + 2z_4^2 + 2z_5^2) \times \\
&\quad(z_1 + z_2 + z_4 + z_5 - z_0z_1 - 2z_0z_2 - z_0z_3 - 3z_1z_2 - 2z_0z_4 - 2z_1z_3 - z_0z_5 - 4z_1z_4 \\
&\quad-z_2z_3 - 3z_1z_5 - 4z_2z_4 - 3z_2z_5 - 2z_3z_4 - z_3z_5 - 4z_4z_5 + 2z_0z_2^2 + 2z_1z_2^2 + 2z_1^2z_2 \\
&\quad+2z_0z_4^2 + 2z_1^2z_3 + 4z_1z_4^2 + 2z_1^2z_4 + 2z_1z_5^2 + 4z_2z_4^2 + 2z_1^2z_5 + 2z_2^2z_4 + 2z_2z_5^2 \\
&\quad+2z_3z_4^2 + 2z_2^2z_5 + 2z_4z_5^2 + 4z_4^2z_5 - z_1^2 - z_2^2 - 3z_4^2 + 2z_4^3 - z_5^2 + 2z_0z_1z_2 \\
&\quad+2z_0z_1z_3 + 2z_0z_1z_4 + 2z_0z_2z_3 + 2z_0z_1z_5 + 4z_0z_2z_4 + 2z_1z_2z_3 + 2z_0z_2z_5 \\
&\quad+2z_0z_3z_4 + 6z_1z_2z_4 + 4z_1z_2z_5 + 4z_1z_3z_4 + 2z_0z_4z_5 + 2z_1z_3z_5 + 2z_2z_3z_4 + 6z_1z_4z_5 \\
&\quad+ 2z_2z_3z_5 + 6z_2z_4z_5 + 2z_3z_4z_5) \leq 0,
\end{aligned}
$$

$$
\begin{aligned}
&(2z_0z_2 - z_1 - z_2 - z_3 - 2z_4 - 2z_5 - z_0 + 2z_0z_3 + 2z_1z_2 + 2z_0z_4 + 2z_1z_3 + 2z_0z_5 \\
&\quad+2z_1z_4 + 2z_1z_5 + 2z_2z_4 + 2z_2z_5 + 2z_3z_4 + 2z_3z_5 + 4z_4z_5 + 2z_4^2 + 2z_5^2) \times \\
&\quad(z_0 + z_3 + z_4 + z_5 - z_0z_1 - 2z_0z_2 - 3z_0z_3 - z_1z_2 - 3z_0z_4 - 2z_1z_3 - 4z_0z_5 \\
&\quad-z_1z_4 - z_2z_3 - 2z_1z_5 - z_2z_4 - 2z_2z_5 - 3z_3z_4 - 4z_3z_5 - 4z_4z_5 + 2z_0^2z_2
\end{aligned}
$$

$$
\begin{aligned}
&+2z_0z_3^2+2z_0^2z_3+2z_0z_4^2+2z_1z_3^2+2z_0^2z_4+4z_0z_5^2+2z_0^2z_5+2z_1z_5^2+2z_2z_5^2\\
&+2z_3z_4^2+2z_3^2z_4+4z_3z_5^2+2z_3^2z_5+4z_4z_5^2+2z_4^2z_5-z_0^2-z_3^2-z_4^2-3z_5^2+2z_5^3\\
&+2z_0z_1z_2+2z_0z_1z_3+2z_0z_1z_4+2z_0z_2z_3+2z_0z_1z_5+2z_0z_2z_4+2z_1z_2z_3\\
&+4z_0z_2z_5+4z_0z_3z_4+6z_0z_3z_5+2z_1z_2z_5+2z_1z_3z_4+6z_0z_4z_5+4z_1z_3z_5\\
&+2z_2z_3z_4+2z_1z_4z_5+2z_2z_3z_5+2z_2z_4z_5+6z_3z_4z_5)\le 0,
\end{aligned}
$$

and $\quad 0\le z_0,z_1,z_2,z_3,z_4,z_5\le 1.$

The above analytic solution to (25)–(30) is obtained by symbolic math programming in MATLAB.

## References


N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. Technical report, 2017. (`https://arxiv.org/abs/1709.08148`).

L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

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

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186, 2011.

G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee, and C. Scott. Domain generalization by marginal transfer learning. Technical report, 2017. (`https://arxiv.org/abs/1711.07910`).

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

J.-F. Cardoso. Multidimensional independent component analysis. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1941–1944, 1998.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML; PMLR)*, volume 48, pages 2606–2615, 2016.

R. M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496, 2008.

K. Fukumizu, F. Bach, and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

A. Gretton. A simpler condition for consistency of a kernel independence test. Technical report, University College London, 2015. (`http://arxiv.org/abs/1501.06103`).

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005a.

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

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592, 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.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616, 2007.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning (ICML; PMLR)*, volume 70, pages 1742–1751, 2017a.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems (NIPS)*, pages 261–270, 2017b.

B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2280–2288, 2016.

L. Klebanov. *N-Distances and Their Applications*. Charles University, Prague, 2005.

G. Kusano, K. Fukumizu, and Y. Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. In *International Conference on Machine Learning (ICML)*, pages 2004–2013, 2016.

H. C. L. Law, D. J. Sutherland, D. Sejdinovic, and S. Flaxman. Bayesian approaches to distribution regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 84, pages 1167–1176, 2018.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, pages 276–284, 2016.

J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI Conference on Artificial Intelligence*, pages 1242–1250, 2014.

R. Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

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

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18, 2011.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10 (1-2):1–141, 2017.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 51, pages 51:398–407, 2016.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 5–31, 2017.

N. Quadrianto, L. Song, and A. Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2009.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132, 2013a.

D. Sejdinovic, B. K. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013b.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. Technical report, Max Planck Institute for Intelligent Systems, 2016. (`https://arxiv.org/abs/1604.05251`).

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715, 2011.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

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, 2010.

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*, 6(3):67–93, 2001.

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

I. Steinwart and J. F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. Technical report, Faculty for Mathematics and Physics, University of Stuttgart, 2017. (`https://arxiv.org/abs/1712.05279`).

E. V. Strobl, S. Visweswaran, and K. Zhang. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. Technical report, 2017. (`https://arxiv.org/abs/1702.03877`).

Z. Szabó, B. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

H. Wendland. *Scattered Data Approximation.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

M. Yamada, Y. Umezu, K. Fukumizu, and I. Takeuchi. Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics (AISTATS; PMLR)*, volume 84, pages 152–160, 2018.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3394–3404, 2017.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. *Journal of Machine Learning Research*, 28(3):819–827, 2013.

A. A. Zinger, A. V. Kakosyan, and L. B. Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 1992.