Contents

1 Introduction

2 Hilbert Space Embedding of Probability Measures

3 Conditions for Characteristic Kernels

4 Metrization of the Weak Topology
In probability theory, the distance between probability measures is used in studying limit theorems, the popular example being the central limit theorem. In the present work, a particular pseudometric on probability distributions is considered, an integral probability metric (IPM).

Denoting $\mathcal{P}$ the set of all Borel probability measures on $(M, A)$, the IPM between $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is defined as

$$d_\mathcal{F}(P, Q) = \sup_{f \in \mathcal{F}} \left| \int_M f \, dP - \int_M f \, dQ \right|$$

Where $\mathcal{F}$ is a class of real-valued bounded measurable functions on $M$. 
Common choices for $\mathcal{F}$ include bounded continuous functions, bounded uniformly continuous functions, the unit ball in $L^\infty$ (which produces the total variation distance), etcetera. In this work, we consider the class of functions to be the unit ball in a Reproducing Kernel Hilbert Space $\mathcal{H}$, this is

$$\mathcal{F} = \{ f \in \mathcal{H} : < f, f >_{\mathcal{H}} \leq 1 \}$$

And denote $d_{\mathcal{F}} = d_k$
Three questions are addressed in this paper

- When is $d_k$ a metric between probability measures?
- How to deal, in practice, with dissimilar distributions who are really close in distance?
- When does $d_k$ metrize the weak topology on $\mathcal{P}$
Let $k$ be a symmetric positive definite Kernel and $\mathcal{H}$ the Hilbert space generated by it. We consider the following set of probability measures

$$\mathcal{P}_k = \{ P \in \mathcal{P} : \int_M \sqrt{k(x,x)} dP(x) < \infty \}$$

The IPM can be viewed as an inner product of the embedded measures.

**Theorem 2.1**

For $P, Q \in \mathcal{P}_k$

$$d_k(P, Q) = || \int_M k(\cdot, x) dP(x) - \int_M k(\cdot, x) dQ(x)||_\mathcal{H} := ||Pk - Qk||_\mathcal{H}$$
For a proof, define $T_P : \mathcal{H} \to \mathbb{R}$ as $T_P(f) = \int_M f dP$ and note that

$$|T_P(f)| \leq \int_M |< f, k(\cdot, x) >|\ dP(x) \leq ||f||_{\mathcal{H}} \int_M \sqrt{k(x, x)} dP(x)$$

So $T_P$ is linear continuous. Denote by $t_P$ the element given by the Riesz representation theorem so that $T_P(f) = < f, t_P >$. It is easy to check that

$$t_P = \int_M k(\cdot, x) dP(x) := Pk$$

$$|Pf - Qf| = |< f, t_P - t_Q >|$$

And taking supremum over $||f|| = 1$, the result follows.
This result is valid for any kernel $k$. However, especially in statistical inference applications, it is not possible to check whether $P \in \mathcal{P}_k$ as $P$ is not known. Therefore, one would prefer to have a kernel such that

$$\int_{\mathcal{M}} \sqrt{k(x, x)} dP(x) < \infty \forall P \in \mathcal{P}_k$$

The following result shows that this is equivalent to $k$ being bounded.
Theorem 2.2

For $f$ measurable, $\int_M f \, dP < \infty \forall P \in \mathcal{P}$ if and only if $f$ is bounded.

The if part is obvious. For the only if, assume $f$ is not bounded and find a sequence $x_n$ such that $f(x_n) > n^2$, so that $C = \sum_n \frac{1}{f(x_n)} < \infty$. Defining the probability measure

$$P = \sum_n \frac{\delta_{x_n}}{Cf(x_n)}$$

It is clear that

$$\int_M f \, dP = \sum_n \frac{f(x_n)}{Cf(x_n)} = \infty$$
Combining both of the preceding theorems, it shows that if $k$ is measurable and bounded, then $d_k(P, Q) = ||Pk - Qk||_H$ for any $P, Q \in \mathcal{P}$, which yields the following embedding

$$\Pi : \mathcal{P} \rightarrow \mathcal{H}$$

$$\Pi(P) = Pk$$

Since $d_k(P, Q) = ||\Pi(P) - \Pi(Q)||$, the question of when is $d_k$ a metric can be formulated as when is $\Pi$ injective. An useful characterization is provided for this purpose
\[ d_k^2(P, Q) = \langle \int_M k(\cdot, x) dP(x), \int_M k(\cdot, y) dP(y) \rangle + \langle \int_M k(\cdot, x) dQ(x), \int_M k(\cdot, y) dQ(y) \rangle - 2 \langle \int_M k(\cdot, x) dP(x), \int_M k(\cdot, y) dQ(y) \rangle \]
\[ = \int_M \int_M k(x, y) dP(x) dP(y) + \int_M \int_M k(x, y) dQ(x) dQ(y) - 2 \int_M \int_M k(x, y) dP(x) dQ(y) \]
\[ = \int_M \int_M k(x, y) d(P - Q)(x) d(P - Q)(y) \]
A final useful characterization, now on the kernel, is presented in Bochner’s theorem

**Theorem 2.3**

A continuous function $g : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure $\mu$ on $\mathbb{R}^d$, that is

$$g(x) = \int_{\mathbb{R}^d} e^{-ix^T y} d\mu(y)$$
Contents

1. Introduction
2. Hilbert Space Embedding of Probability Measures
3. Conditions for Characteristic Kernels
4. Metrization of the Weak Topology
In this section, we address the question of when is \( d_k \) a metric on \( \mathcal{P} \)? In other words, when is \( \Pi \) injective? or under what conditions is \( k \) characteristic?. Although some characterizations are available for \( k \) so that \( d_k \) is a metric, they are difficult to check in practice.

**Definition 3.1**

A bounded measurable positive definite kernel \( k \) is characteristic to a set \( Q \subset \mathcal{P} \) of probability measures defined on \((M, A)\) if for \( P, Q \in Q \),

\[
d_k(P, Q) = 0 \iff P = Q
\]

\( k \) is simply said to be characteristic if it is characteristic to \( \mathcal{P} \). The RKHS \( \mathcal{H} \) induced by such \( k \) is called a characteristic RKHS.
Before proceeding, let’s see a couple of common examples for non-characteristic kernels.

Example 3.2

The trivial kernel given by a nonnegative constant \( k(x, y) = C \) is not characteristic, as

\[
d_k^2(P, Q) = C + C - 2C = 0 \forall P, Q
\]
Example 3.3

The dot product kernel in $\mathcal{R}^d$ given by $k(x, y) = x^T y$ is not characteristic since

$$d_k^2(P, Q) = \mu_P^2 + \mu_Q^2 - 2\mu_P\mu_Q = ||\mu_P - \mu_Q||^2$$

Where $\mu_P = \int_{\mathcal{R}^d} x dP$ is the mean of the probability measure. So

$$d_k^2(P, Q) = 0 \Rightarrow \mu_P = \mu_Q \not\Rightarrow P = Q$$
Example 3.4

The polynomial kernel of order two in $\mathcal{R}^d$ given by $k(x, y) = (1 + x^T y)^2$ is also not characteristic since

$$d_k^2(P, Q) = \int \int (1 + 2x^T y + x^T y y^T x)d(P - Q)(x)d(P - Q)(y)$$

$$= 2||\mu_P - \mu_Q||^2 + ||\Sigma_P - \Sigma_Q + \mu_P \mu_P^T - \mu_Q \mu_Q^T||$$

Where $\mu_P = \int_{\mathcal{R}^d} x dP$ is the mean of the probability measure and $\Sigma_P$ its covariance matrix. So

$$d_k^2(P, Q) = 0 \Rightarrow \mu_P = \mu_Q \land \Sigma_P = \Sigma_Q \nRightarrow P = Q$$
The authors present a condition on the kernel that is sufficient for it to be characteristic.

**Definition 3.5**

A measurable and bounded kernel $k$ is said to be integrally strictly positive definite if

$$\int \int k(x, y)dm(x)dm(y) > 0$$

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

**Theorem 3.6**

Let $k$ be an integrally strictly positive definite kernel on $M$. Then $k$ is characteristic to $\mathcal{P}$.
For the proof, the following lemma is established, and the theorem is a direct consequence of it.

**Lemma 3.7**

Let $k$ be measurable and bounded kernel on $M$. Then $\exists P \neq Q \in \mathcal{P}$ such that $d_k(P, Q) = 0$ if and only if there exists a finite non-zero signed Borel measure $m$ that satisfies:

- $\int \int k(x, y)dm(x)dm(y) = 0$
- $m(M) = 0$
Although the condition for characteristic $k$ in Theorem 3.6 is easy to understand compared to other characterizations in literature, it is not always easy to check for integral strict positive definiteness of $k$. Next, the authors present a sufficient condition for characteristic $k$ which is simple to check in the case $M = \mathbb{R}^d$ and radial kernels, this is, kernels of the form

$$k(x, y) = g(x - y)$$

For measurable $g$
Theorem 3.8

Assume $k(x, y) = g(x - y)$ for measurable $g$ and let $\Lambda$ be the Fourier Transform of $g$. Then $k$ is characteristic if and only if $\text{Supp}(\Lambda) = \mathbb{R}^d$.

A (lengthy and dense) proof is presented at the end of the paper. This is, however, a very useful characterization for radial kernels.
Example 3.9

For the gaussian kernel $k(x, y) = e^{-\frac{(x-y)^2}{2\sigma^2}}$, the shape of the Fourier Transform is

$$\Lambda(z) = \sigma e^{-\frac{\sigma^2 z^2}{2}}$$

Which is supported in the entire space. Thus this kernel defines a characteristic RKHS.

The following corollaries helps with the characterization of the characteristic property even further

Corollary 3.10

In the same framework, if supp$(g)$ is compact, then $k$ is characteristic.
Corollary 3.11

*If $k$ is a radial characteristic kernel and $k_1$ is a radial kernel, then $k + k_1$ and $kk_1$ are also characteristic.*

Proof: Both of them are obviously radial. The antitransform of a sum is the sum of the antitransforms and the antitransform of a product is the convolution of the antitransforms. Note that in the above result, we do not need $k_1$ to be characteristic. Therefore, one can generate all sorts of kernels that are characteristic by starting with a characteristic kernel, $k$. 
Contents

1 Introduction

2 Hilbert Space Embedding of Probability Measures

3 Conditions for Characteristic Kernels

4 Metrization of the Weak Topology
So far, we have studied different characterizations for the kernel $k$ such that $d_k$ is a metric. The metric property of $d_k$ is crucial in many statistical inference applications like hypothesis testing. Therefore, in practice, it is important to use characteristic kernels. However, characteristic kernels, while guaranteeing $d_k$ to be a metric, may nonetheless have difficulty in distinguishing certain distributions on the basis of finite samples. A fair question to ask, then, is how weak or strong $d_k$ is when compared to other metrics.
Recall that we defined $d_{\mathcal{F}}$ in terms of a functional space $\mathcal{F}$. Usual choices for $\mathcal{F}$ are

- $\mathcal{F} = \{ f : \|f\|_\infty \leq 1 \}$ induces the total variation distance $d_{TV}$
- $\mathcal{F} = \{ f : \|f\|_L \leq 1 \}, \|f\|_L = \sup \frac{|f(x) - f(y)|}{\rho(x,y)}$ where $\rho$ is a metric in $M$ induces the Wasserstein distance $d_W$
- $\mathcal{F} = \{ f : \|f\|_\infty + \|f\|_L \leq 1 \}$ induces the Dudley distance $d_D$

We now compare the performance of $d_k$ to those induced metrics.
Theorem 4.1

Assume $k(x, x) \leq C \forall x$ and define

$$b(x, y) = \|k(\cdot, x) - k(\cdot, y)\|$$

Then, for any $P, Q \in \mathcal{P}$

1. $d_k(P, Q) \leq d_W(P, Q) \leq \sqrt{d_k^2(P, Q) + 4C}$ if $(M, b)$ is separable.
2. $\frac{d_k(P, Q)}{1+\sqrt{C}} \leq d_W(P, Q) \leq 2^{\frac{3}{2}}\sqrt{d_k^2(P, Q) + 4C}$ if $(M, b)$ is separable.
3. $d_k(P, Q) \leq d_{TV}(P, Q)$