On Integral Probability Metrics, ϕ-Divergences and Binary Classification

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf and Gert R. G. Lanckriet

***Abstract*—ϕ-divergences are a widely studied class of distance measures between probabilities. In this paper, a different class of distance measures on probabilities, called the integral probability metrics (IPMs) is considered. IPMs, for example, the Wasserstein distance and Dudley metric have, thus far, only been used in a limited setting, as theoretical tools in mass transportation problems, in metrizing the weak topology (on the set of all Borel probability measures defined on a metric space), etc., and their practical applicability have not been well investigated. In this paper, novel properties of IPMs are presented by exploring their relation to ϕ-divergences and binary classification, which we believe would make IPMs as widely and practically applicable as ϕ-divergences.**

Firstly, to understand the relation between IPMs and ϕ-divergences, the necessary and sufficient conditions under which these classes intersect are derived, using which the total variation distance is shown to be the only non-trivial ϕ-divergence that is also an IPM. This shows that IPMs are essentially different from ϕ-divergences. Secondly, since the empirical estimation of ϕ-divergences, especially the KL-divergence is well-studied, the empirical estimation of IPMs from finite i.i.d. samples is then considered and their consistency and convergence rates are analyzed. The empirical estimators of the Wasserstein distance and Dudley metric are shown to be *strongly consistent*. Thirdly, similar to the relation between ϕ-divergences and binary classification, IPMs are related to binary classification by showing that IPMs between the class-conditional distributions are the negative of the optimal risk associated with a binary classifier. In particular, the Wasserstein distance is shown to be related to the Lipschitz classifier, the Dudley metric to the bounded Lipschitz classifier and the maximum mean discrepancy (also an IPM) to the Parzen window classifier.

***Index Terms*—Integral probability metrics, ϕ-divergences, Wasserstein distance, Dudley metric, Maximum mean discrepancy, Reproducing kernel Hilbert space, Rademacher average, Lipschitz classifier, nearest neighbor classifier, Parzen window classifier, support vector machine.**

Bharath K. Sriperumbudur and Gert R. G. Lanckriet are with the Department of Electrical and Computer Engineering, University of California, San Diego, CA, 92093, USA. E-mail: bharathsv@ucsd.edu, gert@ece.ucsd.edu.

Kenji Fukumizu is with the Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan. E-mail: fukumizu@ism.ac.jp.

Arthur Gretton is with the Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA; supported by grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747, and ARO MURI W911NF0810242. He is also affiliated with the Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany. E-mail: arthur.gretton@gmail.com.

Bernhard Schölkopf is with the Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany. E-mail: bernhard.schoelkopf@tuebingen.mpg.de.

Bharath K. Sriperumbudur wishes to acknowledge the support from the Max Planck Institute (MPI) for Biological Cybernetics, National Science Foundation (grant dms-mspa 0625409), the Fair Isaac Corporation and the University of California MICRO program. Part of this work was done while the author was visiting MPI.

I. Introduction

The notion of distance between probability measures [1] has found many applications in probability theory, mathematical statistics and information theory. Popular applications include homogeneity tests (the two-sample problem), independence tests, goodness-of-fit tests, establishing central limit theorems, density estimation, signal detection, channel and source coding, etc. One of the widely studied and well understood families of distances/divergences between probability measures is the *Ali-Silvey distance* [2], also called the *Csiszár's ϕ-divergence* [3], which is defined as

$$D_\phi(\mathbb{P},\mathbb{Q}) := \begin{cases} \int_M \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}, & \mathbb{P} \ll \mathbb{Q} \\ +\infty, & \text{otherwise} \end{cases}, \quad (1)$$

where $(M,\mathcal{A})$ is a measurable space and $\phi : [0,\infty) \to (-\infty,\infty]$ is a convex function.[1] Let $\mathscr{P}$ be the set of all probability measures defined on M. Some of the well-known distance/divergence measures on $\mathscr{P}$ are obtained by appropriately choosing ϕ: Kullback-Liebler (KL) divergence ($\phi(t) = t\log t$), Hellinger distance ($\phi(t) = (\sqrt{t}-1)^2$), total variation distance ($\phi(t) = |t-1|$), χ^2-divergence ($\phi(t) = (t-1)^2$), etc. See [4] and references therein for some of the statistical and information theoretic applications where ϕ-divergences are used.

On the other hand, another popular family (particularly in probability theory and mathematical statistics) of distance measures on $\mathscr{P}$ is the *integral probability metric* (IPM) [5] defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_M f\, d\mathbb{P} - \int_M f\, d\mathbb{Q} \right|, \quad (2)$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on M. This definition of IPMs is motivated from the notion of *weak convergence* of probability measures on metric spaces [6, Section 9.3, Lemma 9.3.2]. In probability theory, IPMs appear in the context of proving central limit theorems using Stein's method [7], [8]. These are also the fundamental quantities that appear in empirical process theory [9], where $\mathbb{Q}$ is replaced by the *empirical distribution* of $\mathbb{P}$.

Some popular distance measures in probability theory and statistics can be obtained by appropriately choosing $\mathcal{F}$. Suppose (M,ρ) is a metric space, where $\mathcal{A}$ is the Borel σ-algebra induced by the metric topology and $\mathscr{P}$ is the set of all Borel probability measures on $\mathcal{A}$. Choosing $\mathcal{F} = \{f : \|f\|_{BL} \le 1\}$ in (2) yields the *Dudley metric* [10, Chapter 19, Definition

[1] Usually, the condition $\phi(1) = 0$ is used in the definition of ϕ-divergence. Here, we do not enforce this condition.

2.2], which metrizes the *weak topology* on $\mathscr{P}$ when M is separable [6, Theorem 11.3.3]. Therefore, the Dudley metric is very useful in the context of proving the convergence of probability measures with respect to the weak topology. Here $\|f\|_{BL} := \|f\|_\infty + \|f\|_L$, $\|f\|_\infty := \sup\{|f(x)| : x \in M\}$ and $\|f\|_L := \sup\{|f(x) - f(y)|/\rho(x,y) : x \neq y \text{ in } M\}$, where $\|f\|_L$ is called the Lipschitz semi-norm of a real-valued function f on M. Choosing $\mathcal{F} = \{f : \|f\|_L \leq 1\}$ yields the *Kantorovich metric*. The famous Kantorovich-Rubinstein theorem [6, Theorem 11.8.2] shows that when M is separable, the Kantorovich metric is the dual representation of the so called *Wasserstein distance* defined as

$$W_1(\mathbb{P},\mathbb{Q}) := \inf_{\mu \in \mathcal{L}(\mathbb{P},\mathbb{Q})} \int \rho(x,y)\, d\mu(x,y), \tag{3}$$

where $\mathbb{P},\mathbb{Q} \in \mathscr{P}_1 := \{\mathbb{P} : \int \rho(x,y)\, d\mathbb{P}(x) < \infty, \forall\, y \in M\}$ and $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $M \times M$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Due to this duality, in this paper, we refer to the Kantorovich metric as the Wasserstein distance and denote it as W when M is separable. The Wasserstein distance has found applications in information theory [11], mathematical statistics [12], [13], *mass transportation problems* [14] and is also called as the *earth mover's distance* in engineering applications (see [15] and references therein for a list of engineering applications where the Wasserstein distance is used). We refer the interested reader to [16, Chapter 5] for the generalizations of W_1. $\gamma_{\mathcal{F}}$ is the *total variation metric* when $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$. Note that the classical central limit theorem and the Berry-Esséen theorem in $\mathbb{R}^d$ use the Kolmogorov distance. The Kolmogorov distance also appears in hypothesis testing as the Kolmogorov-Smirnov statistic [10]. Recently, [17], [18] considered $\mathcal{F}$ to be a unit ball in a reproducing kernel Hilbert space (RKHS) [19], [20], i.e., $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ and obtained a *Hilbertian metric* on $\mathscr{P}$, called the *maximum mean discrepancy* (MMD). Here, $\mathcal{H}$ represents an RKHS with k as its reproducing kernel (r.k.).[2] MMD is used in statistical applications like homogeneity testing [17], independence testing [21], testing for conditional independence [22], etc.

Some of the previously mentioned IPMs, e.g., the Kantorovich distance and Dudley metric are mainly studied and used as tools of theoretical interest in probability theory. However, their practical applicability is not well studied. As mentioned before, the Dudley metric is proposed and used only in the context of metrizing the weak topology on $\mathscr{P}$ [6, Chapter 11]. On the other hand, the Kantorovich distance is more popular and well-studied in its primal form in (3) as the Wasserstein distance than as an IPM [14], [23]. In this work, we present novel properties of IPMs (that have not been explored before) by studying their relation to ϕ-divergences and binary classification. We believe the results presented would provide a better understanding of IPMs and make them as widely and practically applicable as ϕ-divergences. Our contributions in this paper are three-fold and explained in detail below.

1) IPMs and ϕ-divergences: Since the properties of ϕ-divergences are widely studied (see [4], [16] and references therein), in this work, we first investigate the relation between IPMs and ϕ-divergences, the motivation being that if some of the IPMs can be realized as ϕ-divergences, then the properties of ϕ-divergences will carry over to those IPMs also. To this end, in Section II, we first show that $\gamma_{\mathcal{F}}$ is closely related to the variational form of D_ϕ [24]–[26] and is "trivially" a ϕ-divergence if $\mathcal{F}$ is chosen to be the set of all real-valued measurable functions on M (see Theorem 1). Next, we generalize this result by determining the necessary and sufficient conditions on $\mathcal{F}$ and ϕ for which $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}), \forall\, \mathbb{P},\mathbb{Q} \in \mathscr{P}_0 \subset \mathscr{P}$, where $\mathscr{P}_0$ is some subset of $\mathscr{P}$. This leads to our first contribution in this paper, answering the question, "Given a set of distance/divergence measures, $\{\gamma_{\mathcal{F}} : \mathcal{F}\}$ (indexed by $\mathcal{F}$) and $\{D_\phi : \phi\}$ (indexed by ϕ) defined on $\mathscr{P}$, is there a set of distance measures that is common to both these families?" We show that the classes $\{\gamma_{\mathcal{F}} : \mathcal{F}\}$ and $\{D_\phi : \phi\}$ of distance measures intersect *non-trivially only* at the *total variation distance*, which in turn shows that these classes are essentially different and therefore the properties of ϕ-divergences will not carry over to IPMs.

2) Estimation of IPMs: Though IPMs like the Wasserstein distance and Dudley metric are far-reaching as theoretical tools, they have a definite drawback: explicit calculation is difficult for most concrete examples. The same issue also arises with MMD and ϕ-divergences where the exact computation is not straightforward for certain distributions. Therefore, given two probability measures $\mathbb{P}$ and $\mathbb{Q}$, one approach to compute the distance (say the Wasserstein distance) between them is to estimate it based on finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ and hope that the estimate converges to the true distance between $\mathbb{P}$ and $\mathbb{Q}$ given a large number of samples. This situation also arises in statistical inference applications where $\mathbb{P}$ and $\mathbb{Q}$ are known only through finite samples drawn i.i.d. from them and one would like to estimate the distance between $\mathbb{P}$ and $\mathbb{Q}$.

The empirical estimation of ϕ-divergences, especially the KL-divergence is a well-studied problem (see [27], [28] and references therein). Wang *et al.* [27] used a data-dependent space partitioning scheme and showed that the non-parametric estimator of KL-divergence is *strongly consistent*. However, the rate of convergence of this estimator can be arbitrarily slow depending on the distributions. On the other hand, by exploiting the variational representation of ϕ-divergences, Nguyen *et al.* [28] provide an estimate of a "lower bound" of the KL-divergence by solving a convex program. Therefore, as our second contribution, in Section III, we consider the non-parametric estimation of some IPMs, in particular the Wasserstein distance, Dudley metric and MMD based on finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$. The estimates of the Wasserstein distance and Dudley metric are obtained by solving linear programs while an estimator of MMD is computed in closed form (see Section III-A). In Section III-B, we then show that these estimators are strongly consistent and

[2]A function $k : M \times M \to \mathbb{R}$, $(x,y) \mapsto k(x,y)$ is a *reproducing kernel* of the Hilbert space $\mathcal{H}$ if and only if the following hold: *(i)* $\forall\, y \in M,\ k(.,y) \in \mathcal{H}$ and *(ii)* $\forall\, y \in M, \forall\, f \in \mathcal{H},\ \langle f, k(.,y)\rangle_{\mathcal{H}} = f(y)$. $\mathcal{H}$ is called a reproducing kernel Hilbert space.

also provide their rates of convergence. We use tools from concentration inequalities and empirical process theory [9] to establish these results on consistency and rate of convergence. In Section III-C, we describe simulation results that demonstrate the practical viability of these estimators. Since the total variation distance is also an IPM, in Section III-D, we discuss its empirical estimation and show that the empirical estimator is not strongly consistent. Because of the latter, we provide lower bounds for the total variation distance in terms of the Wasserstein distance, Dudley metric and MMD, which can be consistently estimated. These bounds also translate as lower bounds on the KL-divergence through Pinsker's inequality [29].

Our study shows that estimating IPMs (especially the Wasserstein distance, Dudley metric and MMD) is much simpler than estimating ϕ-divergences, and that the estimators are strongly consistent while exhibiting good rates of convergence. In addition, it has to be noted that IPMs also consider the properties of the underlying space M (metric property determined by ρ in the case of Wasserstein and Dudley metrics; similarity property determined by the kernel k [30] in the case of MMD) while computing the distance between $\mathbb{P}$ and $\mathbb{Q}$, which is not the case with ϕ-divergences. This property is useful when $\mathbb{P}$ and $\mathbb{Q}$ have disjoint support.[3] With these advantages, we believe that IPMs can find many applications in information theory, image processing, machine learning and other areas.

3) IPMs and Binary Classification: Finally, as our third contribution, we show how IPMs naturally appear in binary classification. Many previous works [4], [25], [31], [32] related ϕ-divergences (between $\mathbb{P}$ and $\mathbb{Q}$) to binary classification (where $\mathbb{P}$ and $\mathbb{Q}$ are the class conditional distributions) as the negative of the optimal risk associated with a loss function (see [33, Section 1.3] for a detailed list of references). In Section IV, we present a series of results that relate IPMs to binary classification. First, in Section IV-A, we provide a result (similar to that for ϕ-divergences), which shows $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ as the negative of the optimal risk associated with a binary classifier that separates the class conditional distributions, $\mathbb{P}$ and $\mathbb{Q}$, wherein the classification rule is restricted to $\mathcal{F}$. Therefore, the Dudley metric, Wasserstein distance, total variation distance and MMD can be understood as the negative of the optimal risk associated with a classifier wherein the classification rule is restricted to $\{f : \|f\|_{BL} \le 1\}$, $\{f : \|f\|_L \le 1\}$, $\{f : \|f\|_\infty \le 1\}$ and $\{f : \|f\|_{\mathcal{H}} \le 1\}$ respectively. Next, in Sections IV-B and IV-C, we present a second result that relates the empirical estimators studied in Section III to the binary classification setting, by relating the Wasserstein distance and Dudley metric to the margins of the Lipschitz [34] and bounded Lipschitz classifiers respectively and MMD to the Parzen window classifier [30], [35] (see *kernel classification rule* [36, Chapter 10]).

We believe the results presented in this paper address properties of IPMs that have not been explored before and help to understand them from a more practical perspective, which opens up many opportunities for applications. For convenience, in the following, we introduce some notation that we use throughout the paper. Supplementary results used in proofs are collected in the Appendix.

[3] When $\mathbb{P}$ and $\mathbb{Q}$ have disjoint support, $D_\phi(\mathbb{P},\mathbb{Q}) = +\infty$ irrespective of the properties of M, while $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ varies with the properties of M. Therefore, in such cases, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ provides a better notion of distance between $\mathbb{P}$ and $\mathbb{Q}$.

A. Notation

For a measurable function f and a signed measure $\mathbb{P}$, $\mathbb{P}f := \int f\,d\mathbb{P}$ denotes the expectation of f under $\mathbb{P}$. $[\![A]\!]$ represents the indicator function for set A. Given an i.i.d. sample $X_1,\dots,X_n$ drawn from $\mathbb{P}$, $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ represents the empirical distribution, where δ_x represents the Dirac measure at x. We use $\mathbb{P}_n f$ to represent the empirical expectation $\frac{1}{n}\sum_{i=1}^n f(X_i)$. We define $\text{Lip}(M,\rho) := \{f : M \to \mathbb{R}\,|\,\|f\|_L < \infty\}$ and $BL(M,\rho) := \{f : M \to \mathbb{R}\,|\,\|f\|_{BL} < \infty\}$. We also define $W(\mathbb{P},\mathbb{Q}) := \gamma_{\mathcal{F}_W}(\mathbb{P},\mathbb{Q})$, $\beta(\mathbb{P},\mathbb{Q}) := \gamma_{\mathcal{F}_\beta}(\mathbb{P},\mathbb{Q})$, $\gamma_k(\mathbb{P},\mathbb{Q}) := \gamma_{\mathcal{F}_k}(\mathbb{P},\mathbb{Q})$ and $TV(\mathbb{P},\mathbb{Q}) := \gamma_{\mathcal{F}_{TV}}(\mathbb{P},\mathbb{Q})$, where $\mathcal{F}_W := \{f : \|f\|_L \le 1\}$, $\mathcal{F}_\beta := \{f : \|f\|_{BL} \le 1\}$, $\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \le 1\}$ and $\mathcal{F}_{TV} := \{f : \|f\|_\infty \le 1\}$.

II. IPMs and ϕ-divergences

In this section, we consider $\{\gamma_{\mathcal{F}} : \mathcal{F}\}$ and $\{D_\phi : \phi\}$, which are classes of IPMs and ϕ-divergences on $\mathscr{P}$ indexed by $\mathcal{F}$ and ϕ, respectively. We derive conditions on $\mathcal{F}$ and ϕ such that $\forall\,\mathbb{P},\mathbb{Q} \in \mathscr{P}_0 \subset \mathscr{P}$, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for some chosen $\mathscr{P}_0$. This would show whether the class of IPMs and the class of ϕ-divergences are the same or not and under what conditions.

Consider the variational form of D_ϕ [24], [26], [28] given by

$$\begin{aligned} D_\phi(\mathbb{P},\mathbb{Q}) &= \sup_{f:M\to\mathbb{R}} \left[\int_M f\,d\mathbb{P} - \int_M \phi^*(f)\,d\mathbb{Q}\right] \\ &= \sup_{f:M\to\mathbb{R}} (\mathbb{P}f - \mathbb{Q}\phi^*(f)), \end{aligned} \tag{4}$$

where $\phi^*(t) = \sup\{tu - \phi(u) : u \in \mathbb{R}\}$ is the *convex conjugate* of ϕ. Suppose $\mathcal{F}$ is such that $f \in \mathcal{F} \Rightarrow -f \in \mathcal{F}$. Then,

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{Q}f| = \sup_{f\in\mathcal{F}} (\mathbb{P}f - \mathbb{Q}f). \tag{5}$$

Recently, Reid and Williamson [33, Section 8.2] considered the generalization of D_ϕ by modifying its variational form as

$$D_{\phi,\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} (\mathbb{P}f - \mathbb{Q}\phi^*(f)). \tag{6}$$

Let $\mathcal{F}_\star$ be the set of all real-valued measurable functions on M and let $\phi_\star$ be the convex function defined as in (7). It is easy to show that $\phi_\star^*(u) = u$. Comparing $\gamma_{\mathcal{F}}$ in (5) to D_ϕ in (4) through $D_{\phi,\mathcal{F}}$ in (6), we see that $\gamma_{\mathcal{F}} = D_{\phi_\star,\mathcal{F}}$ and $D_\phi = D_{\phi,\mathcal{F}_\star}$. This means $\gamma_{\mathcal{F}}$ is obtained by fixing ϕ to $\phi_\star$ in $D_{\phi,\mathcal{F}}$ with $\mathcal{F}$ as the variable and D_ϕ is obtained by fixing $\mathcal{F}$ to $\mathcal{F}_\star$ in $D_{\phi,\mathcal{F}}$ with ϕ as the variable. This provides a nice relation between $\gamma_{\mathcal{F}}$ and D_ϕ, leading to the following simple result which shows that $\gamma_{\mathcal{F}_\star}$ is "trivially" a ϕ-divergence.

Theorem 1 ($\gamma_{\mathcal{F}_\star}$ is a ϕ-divergence): Let $\mathcal{F}_\star$ be the set of all real-valued measurable functions on M and let

$$\phi_\star(t) = 0[\![t=1]\!] + \infty[\![t \neq 1]\!]. \tag{7}$$

Then

$$\gamma_{\mathcal{F}_\star}(\mathbb{P},\mathbb{Q}) = D_{\phi_\star}(\mathbb{P},\mathbb{Q}) = 0[\![\mathbb{P}=\mathbb{Q}]\!] + \infty[\![\mathbb{P}\neq\mathbb{Q}]\!]. \tag{8}$$

Conversely, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q}) = 0[\![\mathbb{P}=\mathbb{Q}]\!] + \infty[\![\mathbb{P}\neq\mathbb{Q}]\!]$ implies $\mathcal{F}=\mathcal{F}_\star$ and $\phi=\phi_\star$.

Proof: (8) simply follows by using $\mathcal{F}_\star$ and $\phi_\star$ in $\gamma_{\mathcal{F}}$ and D_ϕ or by using $\phi^*_\star(u) = u$ in (4). For the converse, note that $D_\phi(\mathbb{P},\mathbb{Q}) = 0[\![\mathbb{P}=\mathbb{Q}]\!] + \infty[\![\mathbb{P}\neq\mathbb{Q}]\!]$ implies $\phi(1)=0$ and $\int \phi(d\mathbb{P}/d\mathbb{Q})\,d\mathbb{Q} = \infty,\ \forall\,\mathbb{P}\neq\mathbb{Q}$, which means $\phi(x)=\infty,\ \forall\, x\neq 1$ and so $\phi=\phi_\star$. Consider $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \gamma_{\mathcal{F}_\star}(\mathbb{P},\mathbb{Q}) = \sup\{\mathbb{P}f-\mathbb{Q}f : f\in\mathcal{F}_\star\},\ \forall\,\mathbb{P},\mathbb{Q}\in\mathscr{P}$. Suppose $\mathcal{F}\subsetneq\mathcal{F}_\star$. Then it is easy to see that $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) < \gamma_{\mathcal{F}_\star}(\mathbb{P},\mathbb{Q})$ for some $\mathbb{P},\mathbb{Q}\in\mathscr{P}$, which leads to a contradiction. Therefore, $\mathcal{F}=\mathcal{F}_\star$. ∎

From (8), it is clear that $\gamma_{\mathcal{F}_\star}(\mathbb{P},\mathbb{Q})$ is the strongest way to measure the distance between probability measures, and is not a very useful metric in practice.[4] We therefore consider a more restricted function class than $\mathcal{F}_\star$ resulting in a variety of more interesting IPMs, including the Dudley metric, Wasserstein metric, total variation distance, etc. Now, the question is for what other, more restricted function classes $\mathcal{F}$ does there exist a ϕ such that $\gamma_{\mathcal{F}}$ is a ϕ-divergence? We answer this in the following theorem. To this end, we introduce some notation. Let us define $\mathscr{P}_\lambda$ as the set of all probability measures, $\mathbb{P}$ that are absolutely continuous with respect to some σ-finite measure, λ. For $\mathbb{P}\in\mathscr{P}_\lambda$, let $p=\frac{d\mathbb{P}}{d\lambda}$ be the Radon-Nikodym derivative of $\mathbb{P}$ with respect to λ. Let Φ be the class of all convex functions $\phi:[0,\infty)\to(-\infty,\infty]$ continuous at 0 and finite on $(0,\infty)$.

Theorem 2 (Necessary and sufficient conditions): Let $\mathcal{F}\subset\mathcal{F}_\star$ and $\phi\in\Phi$. Then for any $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ if and only if any one of the following hold:

(i) $\mathcal{F} = \{f : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$,
$\phi(u) = \alpha(u-1)[\![0\le u\le 1]\!] + \beta(u-1)[\![u\ge 1]\!]$ for some $\alpha<\beta<\infty$.

(ii) $\mathcal{F} = \{f : f=c,\ c\in\mathbb{R}\}$,
$\phi(u) = \alpha(u-1)[\![u\ge 0]\!]$, $\alpha\in\mathbb{R}$.

The proof idea is as follows. First note that $\gamma_{\mathcal{F}}$ in (2) is a pseudometric[5] on $\mathscr{P}_\lambda$ for any $\mathcal{F}$. Since we want to prove $\gamma_{\mathcal{F}} = D_\phi$, this suggests that we first study the conditions on ϕ for which D_ϕ is a pseudometric. This is answered by Lemma 3, which is a simple modification of a result in [37, Theorem 2].

[4]Unless $\mathbb{P}$ and $\mathbb{Q}$ are exactly the same, $\gamma_{\mathcal{F}_\star}(\mathbb{P},\mathbb{Q}) = +\infty$ and therefore is a trivial and useless metric in practice.

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

Lemma 3: For $\phi\in\Phi$, D_ϕ is a pseudometric on $\mathscr{P}_\lambda$ if and only if ϕ is of the form

$$\phi(u) = \alpha(u-1)[\![0\le u\le 1]\!] + \beta(u-1)[\![u\ge 1]\!], \tag{9}$$

for some $\beta\ge\alpha$.

Before we prove Lemma 3, we need the following lemma from [37], which is quite easy to prove. Lemma 4 shows that $D_\phi(\mathbb{P},\mathbb{Q})$ in (10) associated with ϕ in (9) is proportional to the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$. Note that the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ can be written as $\int_M |p-q|\,d\lambda$, where p and q are defined as in Lemma 4.

Lemma 4 ([37]): For ϕ in (9),

$$D_\phi(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_M |p-q|\,d\lambda, \tag{10}$$

for any $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$, where p and q are the Radon-Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$ with respect to λ.

Proof of Lemma 3: ($\Rightarrow$) If ϕ is of the form in (9), then by Lemma 4, we have $D_\phi(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_M|p-q|\,d\lambda$, which is a metric on $\mathscr{P}_\lambda$ if $\beta>\alpha$ and therefore is a pseudometric on $\mathscr{P}_\lambda$. If $\beta=\alpha$, $D_\phi(\mathbb{P},\mathbb{Q})=0$ for all $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$ and therefore is a pseudometric on $\mathscr{P}_\lambda$.

($\Leftarrow$) If D_ϕ is a pseudometric on $\mathscr{P}_\lambda$, then it satisfies the triangle inequality and ($\mathbb{P}=\mathbb{Q}\Rightarrow D_\phi(\mathbb{P},\mathbb{Q})=0$) and therefore by [37, Theorem 2], ϕ is of the form in (9). ∎

Remark 5: (on the proof of Lemma 3) Theorem 2 in [37] says that D_ϕ satisfies ($\mathbb{P}=\mathbb{Q}\Leftrightarrow D_\phi(\mathbb{P},\mathbb{Q})=0$) and the triangle inequality if and only if ϕ is of the form in (9) for some $\beta>\alpha$. In that case, the strict inequality between α and β is needed such that ϕ is strictly convex for ($\mathbb{P}=\mathbb{Q}\Leftrightarrow D_\phi(\mathbb{P},\mathbb{Q})=0$) to hold. In Lemma 3, we only need ($\mathbb{P}=\mathbb{Q}\Rightarrow D_\phi(\mathbb{P},\mathbb{Q})=0$) to characterize ϕ for D_ϕ to be a pseudometric and so a trivial change in the proof of Theorem 2 in [37] yields ϕ in (9) with an inequality that is not strict between α and β.

Proof of Theorem 2: ($\Rightarrow$) Suppose *(i)* holds. Then for any $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$, we have

$$\begin{aligned}\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) &= \sup\left\{|\mathbb{P}f-\mathbb{Q}f| : \|f\|_\infty\le\frac{\beta-\alpha}{2}\right\}\\ &= \frac{\beta-\alpha}{2}\sup\{|\mathbb{P}f-\mathbb{Q}f| : \|f\|_\infty\le 1\}\\ &= \frac{\beta-\alpha}{2}\int_M|p-q|\,d\lambda \stackrel{(a)}{=} D_\phi(\mathbb{P},\mathbb{Q}),\end{aligned}$$

where (a) follows from Lemma 4.

Suppose *(ii)* holds. Then $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})=0$ and $D_\phi(\mathbb{P},\mathbb{Q}) = \alpha\int_M q\phi(p/q)\,d\lambda = \alpha\int_M(p-q)\,d\lambda = 0$.

($\Leftarrow$) Suppose $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for any $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$. Since $\gamma_{\mathcal{F}}$ is a pseudometric on $\mathscr{P}_\lambda$ (irrespective of $\mathcal{F}$), D_ϕ is a pseudometric on $\mathscr{P}_\lambda$. Therefore, by Lemma 3, $\phi(u) = \alpha(u-1)[\![0\le u\le 1]\!] + \beta(u-1)[\![u\ge 1]\!]$ for some $\beta\ge\alpha$. Now, let us consider two cases.

Case 1: $\beta>\alpha$
By Lemma 4, $D_\phi(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_M|p-q|\,d\lambda$. Since $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = D_\phi(\mathbb{P},\mathbb{Q})$ for all $\mathbb{P},\mathbb{Q}\in\mathscr{P}_\lambda$, we have

$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \frac{\beta-\alpha}{2}\int_M |p-q|\,d\lambda = \frac{\beta-\alpha}{2}\sup\{|\mathbb{P}f-\mathbb{Q}f| : \|f\|_\infty \le 1\} = \sup\{|\mathbb{P}f-\mathbb{Q}f| : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$ and therefore $\mathcal{F} = \{f : \|f\|_\infty \le \frac{\beta-\alpha}{2}\}$.

Case 2: $\beta = \alpha$
$\phi(u) = \alpha(u-1), u \ge 0, \alpha < \infty$. Now, $D_\phi(\mathbb{P},\mathbb{Q}) = \int_M q\phi(p/q)\,d\lambda = \alpha\int_M(p-q) = 0$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda$. Therefore, $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}}|\mathbb{P}f-\mathbb{Q}f| = 0$ for all $\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda$, which means $\forall\,\mathbb{P},\mathbb{Q} \in \mathscr{P}_\lambda$, $\forall\, f \in \mathcal{F}$, $\mathbb{P}f = \mathbb{Q}f$. This, in turn, means f is a constant on M, i.e., $\mathcal{F} = \{f : f = c,\, c \in \mathbb{R}\}$. ■

Note that in Theorem 2, the cases *(i)* and *(ii)* are disjoint as $\alpha < \beta$ in case *(i)* and $\alpha = \beta$ in case *(ii)*. Case *(i)* shows that the family of ϕ-divergences and the family of IPMs intersect only at the total variation distance, which follows from Lemma 4. Case *(ii)* is trivial as the distance between any two probability measures is zero. This result shows that IPMs and ϕ-divergences are essentially different. Theorem 2 also addresses the open question posed by Reid and Williamson [33, pp. 56] of "whether there exist $\mathcal{F}$ such that $\gamma_{\mathcal{F}}$ is not a metric but equals D_ϕ for some $\phi \ne t \mapsto |t-1|$?" This is answered affirmatively by case *(ii)* in Theorem 2 as $\gamma_{\mathcal{F}}$ with $\mathcal{F} = \{f : f = c,\, c \in \mathbb{R}\}$ is a pseudometric (not a metric) on $\mathscr{P}_\lambda$ but equals D_ϕ for $\phi(u) = \alpha(u-1)[\![u \ge 0]\!] \ne u \mapsto |u-1|$.

III. Non-parametric Estimation of IPMs

Suppose one wishes to compute the Wasserstein distance or Dudley metric between $\mathbb{P}$ and $\mathbb{Q}$. This is not straightforward as the explicit calculation is difficult for most concrete examples.[6] Similar is the case with MMD and ϕ-divergences for certain distributions, where the explicit calculation is not straightforward. Therefore, one approach to compute the distance between $\mathbb{P}$ and $\mathbb{Q}$ is to estimate it from finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$ and hope that the estimate converges to the true distance with large sample sizes. The same problem of estimating the distance between $\mathbb{P}$ and $\mathbb{Q}$ appears in statistical inference applications, e.g., the two sample problem, where $\mathbb{P}$ and $\mathbb{Q}$ are known only through random i.i.d. samples.

To this end, the non-parametric estimation of ϕ-divergences, especially the KL-divergence is well studied (see [27], [28], [39] and references therein). Since IPMs and ϕ-divergences are essentially different classes of distance measures on $\mathscr{P}$, in Section III-A, we consider the non-parametric estimation of IPMs, especially the Wasserstein distance, Dudley metric and MMD. We show that the Wasserstein and Dudley metrics can be estimated by solving linear programs (see Theorems 6 and 7) whereas an estimator for MMD can be obtained in closed form (see Theorem 8). These results are significant because to our knowledge, statistical applications (e.g. hypothesis tests) involving the Wasserstein distance in (3) are restricted only to $\mathbb{R}$ [40] as the closed form expression for the Wasserstein distance is known only for $\mathbb{R}$ (see footnote 6). In our case, the results in Section III-A show that the estimation of above mentioned IPMs is possible without any difficulty even in $\mathbb{R}^d$ ($d > 1$) and therefore can be used in testing applications.

In Section III-B, we present the consistency and convergence rate analysis of these estimators. To this end, in Theorem 9, we present a general result on the statistical consistency of the estimators of IPMs by using tools from empirical process theory [9]. As a special case, in Corollary 10, we show that the estimators of Wasserstein distance and Dudley metrics are *strongly consistent*, i.e., suppose $\{\theta_l\}$ is a sequence of estimators of θ, then θ_l is strongly consistent if θ_l converges a.s. to θ as $l \to \infty$. Then, in Theorem 12, we provide a probabilistic bound on the deviation between $\gamma_{\mathcal{F}}$ and its estimate for any $\mathcal{F}$ in terms of the Rademacher complexity (see Definition 11), which is then used to derive the rates of convergence for the estimators of Wasserstein distance, Dudley metric and MMD in Corollary 13. Using the Borel-Cantelli lemma, we then show that MMD is also strongly consistent. The results in this section show that IPMs (especially the Wasserstein distance, Dudley metric and MMD) are easier to estimate than the KL-divergence and the estimators exhibit better convergence behavior than the estimators of the KL-divergence [27], [28]. In Section III-C, we present simulation results to demonstrate the performance of these estimators.

Since the total variation distance is also an IPM, we discuss its empirical estimation and consistency in Section III-D. By citing earlier work [41], we show that the empirical estimator of the total variation distance is not consistent. Since the total variation distance cannot be estimated consistently, in Theorem 15, we provide two lower bounds on the total variation distance, one involving the Wasserstein distance and Dudley metric and the other involving MMD. These bounds can be estimated consistently based on the results in Section III-B and, moreover, they translate to lower bounds on the KL-divergence through Pinsker's inequality (see [29] and references therein for more lower bounds on the KL-divergence in terms of the total variation distance).

A. Non-parametric estimation of Wasserstein distance, Dudley metric and MMD

Let $\{X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}\}$ and $\{X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}\}$ be i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$ respectively. The empirical estimate of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ is given by

$$\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n) = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^N \widetilde{Y}_i f(X_i)\right|, \tag{11}$$

where $\mathbb{P}_m$ and $\mathbb{Q}_n$ represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$, $N = m+n$ and

$$\widetilde{Y}_i = \begin{cases} \frac{1}{m}, & X_i = X^{(1)}_. \\ -\frac{1}{n}, & X_i = X^{(2)}_. \end{cases}. \tag{12}$$

The computation of $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ in (11) is not straightforward for any arbitrary $\mathcal{F}$. In the following, we restrict ourselves to $\mathcal{F}_W := \{f : \|f\|_L \le 1\}$, $\mathcal{F}_\beta := \{f : \|f\|_{BL} \le 1\}$ and

[6] The explicit form for the Wasserstein distance in (3) is known for $(M,\rho(x,y)) = (\mathbb{R},|x-y|)$ [16], [38], which is given as $W_1(\mathbb{P},\mathbb{Q}) = \int_{(0,1)}|F_{\mathbb{P}}^{-1}(u) - F_{\mathbb{Q}}^{-1}(u)|\,du = \int_{\mathbb{R}}|F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)|\,dx$, where $F_{\mathbb{P}}(x) = \mathbb{P}((-\infty,x])$ and $F_{\mathbb{P}}^{-1}(u) = \inf\{x \in \mathbb{R}|F_{\mathbb{P}}(x) \ge u\}$, $0 < u < 1$. It is easy to show that this explicit form can be extended to $(\mathbb{R}^d, \|\cdot\|_1)$. However, the exact computation of $W_1(\mathbb{P},\mathbb{Q})$ is not straightforwardfor all $\mathbb{P}$ and $\mathbb{Q}$. See Section III-C for some examples where $W_1(\mathbb{P},\mathbb{Q})$ can be computed exactly. Note that since $\mathbb{R}^d$ is separable, by the Kantorovich-Rubinstein theorem, $W(\mathbb{P},\mathbb{Q}) = W_1(\mathbb{P},\mathbb{Q}), \forall\,\mathbb{P},\mathbb{Q}$.

$\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \le 1\}$ and compute (11). Let us denote $W := \gamma_{\mathcal{F}_W}$, $\beta := \gamma_{\mathcal{F}_\beta}$ and $\gamma_k := \gamma_{\mathcal{F}_k}$.

Theorem 6 (Estimator of Wasserstein distance): For all $\alpha \in [0,1]$, the following function solves (11) for $\mathcal{F} = \mathcal{F}_W$:

$$f_\alpha(x) := \alpha \min_{i=1,\dots,N}(a_i^\star + \rho(x, X_i)) + (1-\alpha)\max_{i=1,\dots,N}(a_i^\star - \rho(x, X_i)), \quad (13)$$

where

$$W(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star, \quad (14)$$

and $\{a_i^\star\}_{i=1}^N$ solve the following linear program,

$$\begin{aligned} &\max_{a_1,\dots,a_N} \sum_{i=1}^N \widetilde{Y}_i a_i \\ &\text{s.t. } -\rho(X_i, X_j) \le a_i - a_j \le \rho(X_i, X_j), \forall\, i, j. \end{aligned} \quad (15)$$

Proof: Consider $W(\mathbb{P}_m, \mathbb{Q}_n) = \sup\{\sum_{i=1}^N \widetilde{Y}_i f(X_i) : \|f\|_L \le 1\}$. Note that

$$1 \ge \|f\|_L = \sup_{x \ne x'} \frac{|f(x) - f(x')|}{\rho(x, x')} \ge \max_{X_i \ne X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)},$$

which means

$$\begin{aligned} W(\mathbb{P}_m, \mathbb{Q}_n) \le \sup &\sum_{i=1}^N \widetilde{Y}_i f(X_i) \\ &\text{s.t. } \max_{X_i \ne X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)} \le 1. \end{aligned} \quad (16)$$

The right hand side of (16) can be equivalently written as

$$\begin{aligned} &\sup \sum_{i=1}^N \widetilde{Y}_i f(X_i) \\ &\text{s.t. } -\rho(X_i, X_j) \le f(X_i) - f(X_j) \le \rho(X_i, X_j), \forall\, i, j. \end{aligned}$$

Let $a_i := f(X_i)$. Therefore, we have $W(\mathbb{P}_m, \mathbb{Q}_n) \le \sum_{i=1}^N \widetilde{Y}_i a_i^\star$, where $\{a_i^\star\}_{i=1}^N$ solve the linear program in (15). Note that the objective in (15) is linear in $\{a_i\}_{i=1}^N$ with linear inequality constraints and therefore by Theorem 25 (see the Appendix), the optimum lies on the boundary of the constraint set, which means $\max_{X_i \ne X_j} \frac{|a_i^\star - a_j^\star|}{\rho(X_i, X_j)} = 1$. Therefore, by Lemma 20 (see the Appendix), f on $\{X_1, \dots, X_N\}$ can be extended to a function f_α (on M) defined in (13) where $f_\alpha(X_i) = f(X_i) = a_i^\star$ and $\|f_\alpha\|_L = \|f\|_L = 1$, which means f_α is a maximizer of (11) and $W(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star$. ■

Theorem 7 (Estimator of Dudley metric): For all $\alpha \in [0,1]$, the following function solves (11) for $\mathcal{F} = \mathcal{F}_\beta$:

$$g_\alpha(x) := \max\left(-\max_{i=1,\dots,N}|a_i^\star|, \min\left(h_\alpha(x), \max_{i=1,\dots,N}|a_i^\star|\right)\right) \quad (17)$$

where

$$h_\alpha(x) := \alpha \min_{i=1,\dots,N}(a_i^\star + L^\star \rho(x, X_i)) + (1-\alpha)\max_{i=1,\dots,N}(a_i^\star - L^\star \rho(x, X_i)), \quad (18)$$

$$\beta(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star, \quad (19)$$

$$L^\star = \max_{X_i \ne X_j} \frac{|a_i^\star - a_j^\star|}{\rho(X_i, X_j)}, \quad (20)$$

and $\{a_i^\star\}_{i=1}^N$ solve the following linear program,

$$\begin{aligned} &\max_{a_1,\dots,a_N,b,c} \sum_{i=1}^N \widetilde{Y}_i a_i \\ &\text{s.t. } -b\,\rho(X_i, X_j) \le a_i - a_j \le b\,\rho(X_i, X_j), \forall\, i, j \\ &\quad\; -c \le a_i \le c, \forall\, i \\ &\quad\; b + c \le 1. \end{aligned} \quad (21)$$

Proof: The proof is similar to that of Theorem 6. Note that

$$\begin{aligned} 1 \ge \|f\|_L + \|f\|_\infty &= \sup_{x \ne y} \frac{|f(x) - f(y)|}{\rho(x, y)} + \sup_{x \in M} |f(x)| \\ &\ge \max_{X_i \ne X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)} + \max_i |f(X_i)|, \end{aligned}$$

which means

$$\beta(\mathbb{P}_m, \mathbb{Q}_n) \le \sup\left\{\sum_{i=1}^N \widetilde{Y}_i f(X_i) : \max_i |f(X_i)| + \max_{X_i \ne X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)} \le 1\right\}. \quad (22)$$

Let $a_i := f(X_i)$. Therefore, $\beta(\mathbb{P}_m, \mathbb{Q}_n) \le \sum_{i=1}^N \widetilde{Y}_i a_i^\star$, where $\{a_i^\star\}_{i=1}^N$ solve

$$\begin{aligned} &\max_{a_1,\dots,a_N} \sum_{i=1}^N \widetilde{Y}_i a_i \\ &\text{s.t. } \max_{X_i \ne X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} + \max_i |a_i| \le 1. \end{aligned} \quad (23)$$

Introducing variables b and c such that $\max_{X_i \ne X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} \le b$ and $\max_i |a_i| \le c$ reduces the program in (23) to (21). In addition, it is easy to see that the optimum occurs at the boundary of the constraint set and therefore $\max_{X_i \ne X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} + \max_i |a_i| = 1$. Hence, by Lemma 21 (see the Appendix), g_α in (17) extends f defined on $\{X_1, \dots, X_n\}$ to M, i.e., $g_\alpha(X_i) = f(X_i)$ and $\|g_\alpha\|_{BL} = \|f\|_{BL} = 1$. Note that h_α in (18) is the Lipschitz extension of g to M (by Lemma 20). Therefore, g_α is a solution to (11) and (19) holds. ■

Theorem 8 (Estimator of MMD): For $\mathcal{F} = \mathcal{F}_k$, the following function is the unique solution to (11):

$$f = \frac{1}{\|\sum_{i=1}^N \widetilde{Y}_i k(.,X_i)\|_{\mathcal{H}}} \sum_{i=1}^N \widetilde{Y}_i k(., X_i), \quad (24)$$

and

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \left\|\sum_{i=1}^N \widetilde{Y}_i k(., X_i)\right\|_{\mathcal{H}} = \sqrt{\sum_{i,j=1}^N \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)}. \quad (25)$$

Proof: Consider $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n) := \sup\{\sum_{i=1}^N \widetilde{Y}_i f(X_i) : \|f\|_{\mathcal{H}} \le 1\}$, which can be written as

$$\gamma_k(\mathbb{P}_m,\mathbb{Q}_n) = \sup_{\|f\|_{\mathcal{H}}\le 1} \langle f, \sum_{i=1}^N \widetilde{Y}_i k(.,X_i)\rangle_{\mathcal{H}}, \qquad (26)$$

where we have used the reproducing property of $\mathcal{H}$, i.e., $\forall\, f \in \mathcal{H},\, \forall\, x \in M,\, f(x) = \langle f, k(.,x)\rangle_{\mathcal{H}}$. It is easy to see that the objective function in (26) is linear in f and the constraint set is convex in f and therefore by Theorem 25 (see the Appendix), the optimum occurs on the boundary of the constraint set, i.e., $\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} = 1\}$. The Lagrangian function associated with (26) is given by

$$J(f,\lambda) = \langle f, \sum_{i=1}^N \widetilde{Y}_i k(.,X_i)\rangle_{\mathcal{H}} - \lambda(\langle f,f\rangle_{\mathcal{H}}^{1/2} - 1), \qquad (27)$$

where $\lambda \ge 0$. Minimizing J with respect to f gives

$$f = \frac{\|f\|_{\mathcal{H}}}{\lambda} \sum_{i=1}^N \widetilde{Y}_i k(.,X_i) = \frac{1}{\lambda} \sum_{i=1}^N \widetilde{Y}_i k(.,X_i),$$

which implies $\lambda = \|\sum_{i=1}^N \widetilde{Y}_i k(.,X_i)\|_{\mathcal{H}}$, therefore resulting in f as in (24). Using this f in (26) with $k(X_i,X_j) = \langle k(.,X_i), k(.,X_j)\rangle_{\mathcal{H}}$ yields (25). ■

The above result related to MMD also appears in [17]. We presented here for completeness.

B. Consistency and rate of convergence

In Section III-A, we presented the empirical estimators of W, β and γ_k. For these estimators to be reliable, we need them to converge to the population values as $m,n \to \infty$. Even if this holds, we would like to have a fast rate of convergence such that in practice, fewer samples are sufficient to obtain reliable estimates. We address these issues in this subsection.

Before we start presenting the results, we briefly introduce some terminology and notation from empirical process theory. For any $r \ge 1$ and probability measure $\mathbb{Q}$, define the L_r norm $\|f\|_{\mathbb{Q},r} := (\int |f|^r \, d\mathbb{Q})^{1/r}$ and let $L_r(\mathbb{Q})$ denote the metric space induced by this norm. The *covering number* $\mathcal{N}(\varepsilon, \mathcal{F}, L_r(\mathbb{Q}))$ is the minimal number of $L_r(\mathbb{Q})$ balls of radius ε needed to cover $\mathcal{F}$. $\mathcal{H}(\varepsilon, \mathcal{F}, L_r(\mathbb{Q})) := \log \mathcal{N}(\varepsilon, \mathcal{F}, L_r(\mathbb{Q}))$ is called the *entropy* of $\mathcal{F}$ using the $L_r(\mathbb{Q})$ metric. Define the minimal envelope function: $F(x) := \sup_{f\in\mathcal{F}} |f(x)|$.

We now present a general result on the consistency of $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$, which simply follows from Theorem 22 (see the Appendix).

Theorem 9: Suppose the following conditions hold:

(i) $\int F \, d\mathbb{P} < \infty$.

(ii) $\int F \, d\mathbb{Q} < \infty$.

(iii) $\forall \varepsilon > 0$, $\frac{1}{m}\mathcal{H}(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_m)) \xrightarrow{\mathbb{P}} 0$ as $m \to \infty$.

(iv) $\forall \varepsilon > 0$, $\frac{1}{n}\mathcal{H}(\varepsilon, \mathcal{F}, L_1(\mathbb{Q}_n)) \xrightarrow{\mathbb{Q}} 0$ as $n \to \infty$.

Then, $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$ as $m,n \to \infty$.

Proof: Consider $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| = \left|\sup_{f\in\mathcal{F}} |\mathbb{P}_m f - \mathbb{Q}_n f| - \sup_{f\in\mathcal{F}} |\mathbb{P} f - \mathbb{Q} f|\right| \le \sup_{f\in\mathcal{F}} ||\mathbb{P}_m f - \mathbb{Q}_n f| - |\mathbb{P} f - \mathbb{Q} f|| \le \sup_{f\in\mathcal{F}} |\mathbb{P}_m f - \mathbb{Q}_n f - \mathbb{P} f + \mathbb{Q} f| \le \sup_{f\in\mathcal{F}} [|\mathbb{P}_m f - \mathbb{P} f| + |\mathbb{Q}_n f - \mathbb{Q} f|] \le \sup_{f\in\mathcal{F}} |\mathbb{P}_m f - \mathbb{P} f| + \sup_{f\in\mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q} f|$. Therefore, by Theorem 22, $\sup_{f\in\mathcal{F}} |\mathbb{P}_m f - \mathbb{P} f| \xrightarrow{a.s.} 0$, $\sup_{f\in\mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q} f| \xrightarrow{a.s.} 0$ and the result follows. ■

The following corollary to Theorem 9 shows that $W(\mathbb{P}_m,\mathbb{Q}_n)$ and $\beta(\mathbb{P}_m,\mathbb{Q}_n)$ are strongly consistent.

Corollary 10 (Consistency of W and β): Let (M,ρ) be a totally bounded metric space. Then, $|W(\mathbb{P}_m,\mathbb{Q}_n) - W(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$ and $|\beta(\mathbb{P}_m,\mathbb{Q}_n) - \beta(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$ as $m,n \to \infty$.

Proof: For any $f \in \mathcal{F}_W$,

$$f(x) \le \sup_{x\in M} |f(x)| \le \sup_{x,y} |f(x) - f(y)| \le \|f\|_L \sup_{x,y} \rho(x,y) \le \|f\|_L \text{diam}(M) \le \text{diam}(M) < \infty,$$

where $\text{diam}(M)$ represents the diameter of M. Therefore, $\forall\, x \in M$, $F(x) \le \text{diam}(M) < \infty$, which satisfies *(i)* and *(ii)* in Theorem 9. Kolmogorov and Tihomirov [42] have shown that

$$\mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty) \le \mathcal{N}(\frac{\varepsilon}{4}, M, \rho) \log\left(2\left\lceil \frac{2\text{diam}(M)}{\varepsilon}\right\rceil + 1\right). \qquad (28)$$

Since $\mathcal{H}(\varepsilon, \mathcal{F}_W, L_1(\mathbb{P}_m)) \le \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty)$, the conditions *(iii)* and *(iv)* in Theorem 9 are satisfied and therefore, $|W(\mathbb{P}_m,\mathbb{Q}_n) - W(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$ as $m,n \to \infty$. Since $\mathcal{F}_\beta \subset \mathcal{F}_W$, the envelope function associated with $\mathcal{F}_\beta$ is upper bounded by the envelope function associated with $\mathcal{F}_W$ and $\mathcal{H}(\varepsilon, \mathcal{F}_\beta, \|\cdot\|_\infty) \le \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty)$. Therefore, the result for β follows. ■

Similar to Corollary 10, a consistency result for γ_k can be provided by estimating the entropy number of $\mathcal{F}_k$. See Cucker and Zhou [43, Chapter 5] for the estimates of entropy numbers for various $\mathcal{H}$. However, in the following, we adopt a different approach to prove the strong consistency of γ_k. To this end, we first provide a general result on the rate of convergence of $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ and then, as a special case, obtain the rates of convergence of the estimators of W, β and γ_k. Using this result, we then prove the strong consistency of γ_k. We start with the following definition.

Definition 11 (Rademacher complexity): Let $\mathcal{F}$ be a class of functions on M and $\{\sigma_i\}_{i=1}^m$ be independent Rademacher random variables, i.e., $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = \frac{1}{2}$. The Rademacher process is defined as $\{\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i) : f \in \mathcal{F}\}$ for some $\{x_i\}_{i=1}^m \subset M$. The Rademacher complexity over $\mathcal{F}$ is defined as

$$R_m(\mathcal{F}; \{x_i\}_{i=1}^m) := \mathbb{E} \sup_{f\in\mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right|. \qquad (29)$$

We now present a general result that provides a probabilistic bound on the deviation of $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ from $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$.

Theorem 12: For any $\mathcal{F}$ such that $\nu := \sup_{x\in M} F(x) < \infty$, with probability at least $1-\delta$, the following holds:

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sqrt{18\nu^2\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right) + 2R_m(\mathcal{F};\{X_i^{(1)}\}) + 2R_n(\mathcal{F};\{X_i^{(2)}\}). \quad (30)$$

Proof: From the proof of Theorem 9, we have $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f| + \sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f|$. We now bound the terms $\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f|$ and $\sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f|$, which are the fundamental quantities that appear in empirical process theory.

Note that $\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f|$ satisfies (74) (see the Appendix) with $c_i=\frac{2\nu}{m}$. Therefore, by McDiarmid's inequality in (75) (see the Appendix), we have that with probability at least $1-\frac{\delta}{4}$, the following holds:

$$\begin{aligned}\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f| &\le \mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f| + \sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}}\\ &\stackrel{(a)}{\le} 2\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})\Big| + \sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}}, \quad (31)\end{aligned}$$

where (a) follows from bounding $\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f|$ by using the symmetrization inequality in (76) (see the Appendix). Note that the expectation in the second line of (31) is taken jointly over $\{\sigma_i\}_{i=1}^m$ and $\{X_i^{(1)}\}_{i=1}^m$. $\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})\Big|$ can be written as $\mathbb{E}\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\Big|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})\Big|$, where the inner expectation, which we denote as $\mathbb{E}_\sigma$, is taken with respect to $\{\sigma_i\}_{i=1}^m$ conditioned on $\{X_i^{(1)}\}_{i=1}^m$ and the outer expectation is taken with respect to $\{X_i^{(1)}\}_{i=1}^m$. Since $\mathbb{E}_\sigma\sup_{f\in\mathcal{F}}|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})|$ satisfies (74) (see the Appendix) with $c_i=\frac{2\nu}{m}$, by McDiarmid's inequality in (75) (see the Appendix), with probability at least $1-\frac{\delta}{4}$, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})\right| \le \mathbb{E}_\sigma\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m\sigma_i f(X_i^{(1)})\right| + \sqrt{\frac{2\nu^2}{m}\log\frac{4}{\delta}}. \quad (32)$$

Tying (31) and (32), we have that with probability at least $1-\frac{\delta}{2}$, the following holds:

$$\sup_{f\in\mathcal{F}}|\mathbb{P}_m f-\mathbb{P}f| \le 2R_m(\mathcal{F};\{X_i^{(1)}\}_{i=1}^m) + \sqrt{\frac{18\nu^2}{m}\log\frac{4}{\delta}}. \quad (33)$$

Performing similar analysis for $\sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f|$, we have that with probability at least $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}|\mathbb{Q}_n f-\mathbb{Q}f| \le 2R_n(\mathcal{F};\{X_i^{(2)}\}_{i=1}^n) + \sqrt{\frac{18\nu^2}{n}\log\frac{4}{\delta}}. \quad (34)$$

The result follows by adding (33) and (34). ■

Theorem 12 holds for any $\mathcal{F}$ for which ν is finite. However, to obtain the rate of convergence for $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$, one requires an estimate of $R_m(\mathcal{F};\{X_i^{(1)}\}_{i=1}^m)$ and $R_n(\mathcal{F};\{X_i^{(2)}\}_{i=1}^n)$. Note that if $R_m(\mathcal{F};\{X_i^{(1)}\}_{i=1}^m) \xrightarrow{\mathbb{P}} 0$ as $m\to\infty$ and $R_n(\mathcal{F};\{X_i^{(2)}\}_{i=1}^n) \xrightarrow{\mathbb{Q}} 0$ as $n\to\infty$, then $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \xrightarrow{\mathbb{P},\mathbb{Q}} 0$ as $m,n\to\infty$. Also note that if $R_m(\mathcal{F};\{X_i^{(1)}\}_{i=1}^m)=O_{\mathbb{P}}(r_m)$ and $R_n(\mathcal{F};\{X_i^{(2)}\}_{i=1}^n)=O_{\mathbb{Q}}(r_n)$, then from (33) and (34), $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m\vee m^{-1/2}+r_n\vee n^{-1/2})$ as $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{P})=O_{\mathbb{P}}(r_m\vee m^{-1/2})$ and $\gamma_{\mathcal{F}}(\mathbb{Q}_n,\mathbb{Q})=O_{\mathbb{Q}}(r_n\vee n^{-1/2})$, where $a\vee b:=\max(a,b)$. The following corollary to Theorem 12 provides the rate of convergence for W, β and γ_k (for a fixed k), by choosing a specific $\mathcal{F}$.

Corollary 13 (Rates of convergence for W, β and γ_k): *(i)* Let M be a bounded subset of $(\mathbb{R}^d,\|\cdot\|_s)$ for some $1\le s\le\infty$. Then, $|W(\mathbb{P}_m,\mathbb{Q}_n)-W(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m+r_n)$ and $|\beta(\mathbb{P}_m,\mathbb{Q}_n)-\beta(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m+r_n)$, where

$$r_m = \begin{cases} m^{-1/2}\log m, & d=1\\ m^{-1/(d+1)}, & d\ge 2\end{cases}. \quad (35)$$

In addition if M is a bounded, convex subset of $(\mathbb{R}^d,\|\cdot\|_s)$ with non-empty interior, then

$$r_m = \begin{cases} m^{-1/2}, & d=1\\ m^{-1/2}\log m, & d=2\\ m^{-1/d}, & d>2\end{cases}. \quad (36)$$

(ii) Let M be a measurable space. Suppose k is measurable and $\sup_{x\in M}k(x,x)\le C<\infty$. Then, $|\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_k(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(m^{-1/2}+n^{-1/2})$. In addition, $|\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_k(\mathbb{P},\mathbb{Q})| \xrightarrow{a.s.} 0$ as $m,n\to\infty$, i.e., the estimator of MMD is strongly consistent.

Proof: *(i)* Define $R_m^1(\mathcal{F}) := R_m(\mathcal{F};\{X_i^{(1)}\}_{i=1}^m)$. The generalized entropy bound [34, Theorem 16] gives that for every $\varepsilon>0$,

$$R_m^1(\mathcal{F}) \le 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^\infty \sqrt{\mathcal{H}(\tau,\mathcal{F},L_2(\mathbb{P}_m))}\,d\tau. \quad (37)$$

Let $\mathcal{F}=\mathcal{F}_W$. Since M is a bounded subset of $\mathbb{R}^d$, it is totally bounded and therefore the entropy number in (37) can be bounded through (28) by noting that

$$\mathcal{H}(\tau,\mathcal{F}_W,L_2(\mathbb{P}_m)) \le \mathcal{H}(\tau,\mathcal{F}_W,\|\cdot\|_\infty) \le \frac{C_1}{\tau^{d+1}}+\frac{C_2}{\tau^d}, \quad (38)$$

where we have used the fact that $\mathcal{N}(\varepsilon,M,\|\cdot\|_s)=O(\varepsilon^{-d})$, $1\le s\le\infty$ and $\log(\lceil x\rceil+1)\le x+1$.[7] The constants C_1 and C_2 depend only on the properties of M and are independent of τ. Substituting (38) in (37), we have

$$\begin{aligned}R_m^1(\mathcal{F}_W) &\le \inf_{\varepsilon>0}\left[2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^\infty\sqrt{\mathcal{H}(\tau,\mathcal{F}_W,L_2(\mathbb{P}_m))}\,d\tau\right]\\ &\le \inf_{\varepsilon>0}\left[2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{4R}\left(\frac{\sqrt{C_1}}{\tau^{(d+1)/2}}+\frac{\sqrt{C_2}}{\tau^{d/2}}\right)d\tau\right],\end{aligned}$$

where $R:=\mathrm{diam}(M)$. Note the change in upper limits of the integral from ∞ to $4R$. This is because M is totally bounded and $\mathcal{H}(\tau,\mathcal{F}_W,\|\cdot\|_\infty)$ depends on $\mathcal{N}(\tau/4,M,\rho)$. The rates in

[7] Note that for any $x\in M\subset\mathbb{R}^d$, $\|x\|_\infty\le\cdots\le\|x\|_s\le\cdots\le\|x\|_2\le\|x\|_1\le\sqrt{d}\|x\|_2$. Therefore, $\forall\, s\ge 2$, $\mathcal{N}(\varepsilon,M,\|\cdot\|_s)\le\mathcal{N}(\varepsilon,M,\|\cdot\|_2)$ and $\forall\, 1\le s\le 2$, $\mathcal{N}(\varepsilon,M,\|\cdot\|_s)\le\mathcal{N}(\varepsilon,M,\sqrt{d}\|\cdot\|_2)=\mathcal{N}(\varepsilon/\sqrt{d},M,\|\cdot\|_2)$. Use $\mathcal{N}(\varepsilon,M,\|\cdot\|_2)=O(\varepsilon^{-d})$ [44, Lemma 2.5].

(35) are simply obtained by solving the right hand side of the above inequality. As mentioned in the paragraph preceding the statement of Corollary 13, we have $r_m \vee m^{-1/2} = r_m$ and so the result for $W(\mathbb{P}_m, \mathbb{Q}_n)$ follows.

Suppose M is convex. Then M is connected. It is easy to see that M is also centered, i.e., for all subsets $A \subset M$ with diam$(A) \le 2r$ there exists a point $x \in M$ such that $\|x - a\|_s \le r$ for all $a \in A$. Since M is connected and centered, we have from [42] that

$$\begin{aligned}\mathcal{H}(\tau, \mathcal{F}_W, L_2(\mathbb{P}_m)) &\le \mathcal{H}(\tau, \mathcal{F}_W, \|\cdot\|_\infty) \le \\ \mathcal{N}(\frac{\tau}{2}, M, \|\cdot\|_s)\log 2 &+ \log\left(2\left\lceil\frac{2\,\mathrm{diam}(M)}{\tau}\right\rceil + 1\right) \\ &\le C_3\tau^{-d} + C_4\tau^{-1} + C_5,\end{aligned} \tag{39}$$

where we used the fact that $\mathcal{N}(\varepsilon, M, \|\cdot\|_s) = O(\varepsilon^{-d})$. C_3, C_4 and C_5 are constants that depend only on the properties of M and are independent of τ. Substituting (39) in (37), we have,

$$R^1_m(\mathcal{F}_W) \le \inf_{\varepsilon > 0}\left[2\varepsilon + \frac{4\sqrt{2}}{\sqrt{m}}\int_{\varepsilon/4}^{2R}\frac{\sqrt{C_3}}{\tau^{d/2}}\,d\tau\right] + O(m^{-1/2}).$$

Again note the change in upper limits of the integral from ∞ to $2R$. This is because $\mathcal{H}(\tau, \mathcal{F}_W, \|\cdot\|_\infty)$ depends on $\mathcal{N}(\tau/2, M, \rho)$. The rates in (36) are obtained by solving the right hand side of the above inequality. Since $r_m \vee m^{-1/2} = r_m$, the result for $W(\mathbb{P}_m, \mathbb{Q}_n)$ follows.

Since $\mathcal{F}_\beta \subset \mathcal{F}_W$, we have $R^1_m(\mathcal{F}_\beta) \le R^1_m(\mathcal{F}_W)$ and therefore, the result for $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ follows. The rates in (36) can also be directly obtained for β by using the entropy number of $\mathcal{F}_\beta$, i.e., $\mathcal{H}(\varepsilon, \mathcal{F}_\beta, \|\cdot\|_\infty) = O(\varepsilon^{-d})$ [9, Theorem 2.7.1] in (37).

(ii) $R^1_m(\mathcal{F}_k)$ can be bounded as

$$\begin{aligned}&\mathbb{E}_\sigma \sup_{f\in\mathcal{F}_k}\Big|\sum_{i=1}^m \frac{\sigma_i f(X_i^{(1)})}{m}\Big| = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}_k}\Big|\langle f, \sum_{i=1}^m\frac{\sigma_i k(.,X_i^{(1)})}{m}\rangle_{\mathcal{H}}\Big| \\ &= \mathbb{E}_\sigma\Big\|\sum_{i=1}^m\frac{\sigma_i k(.,X_i^{(1)})}{m}\Big\|_{\mathcal{H}} = \mathbb{E}_\sigma\sqrt{\sum_{i,j=1}^m\frac{\sigma_i\sigma_j k(X_i^{(1)},X_j^{(1)})}{m^2}} \\ &\le \sqrt{\mathbb{E}_\sigma\sum_{i,j=1}^m\frac{\sigma_i\sigma_j k(X_i^{(1)},X_j^{(1)})}{m^2}} = \frac{1}{m}\sqrt{\sum_{i=1}^m k(X_i^{(1)},X_i^{(1)})} \\ &\le \frac{\sqrt{C}}{\sqrt{m}}.\end{aligned} \tag{40}$$

By using (40) in (33), it is easy to see that $\gamma_k(\mathbb{P}_m, \mathbb{P}) = O_{\mathbb{P}}(m^{-1/2})$. In addition, by the Borel-Cantelli lemma, $\gamma_k(\mathbb{P}_m, \mathbb{P}) \xrightarrow{a.s.} 0$ as $m \to \infty$. Performing similar analysis for $\gamma_k(\mathbb{Q}_n, \mathbb{Q})$ and adding (33) and (34) yields the result. ■

Remark 14: (i) Note that the rate of convergence of W and β is dependent on the dimension, d, which means that in large dimensions, more samples are needed to obtain useful estimates of W and β. Also note that the rates are independent of the metric, $\|\cdot\|_s$, $1 \le s \le \infty$.

(ii) Note that when M is a bounded, convex subset of $(\mathbb{R}^d, \|\cdot\|_s)$, faster rates are obtained than for the case where M is just a bounded (but not convex) subset of $(\mathbb{R}^d, \|\cdot\|_s)$.

(iii) In the case of MMD, we have not made any assumptions on M except it being a measurable space. This means in the case of $\mathbb{R}^d$, the rate is independent of d, which is a very useful property. The condition of the kernel being bounded is satisfied by a host of kernels, the examples of which include the Gaussian kernel, $k(x,y) = \exp(-\sigma\|x-y\|_2^2)$, $\sigma > 0$, Laplacian kernel, $k(x,y) = \exp(-\sigma\|x-y\|_1)$, $\sigma > 0$, inverse multiquadrics, $k(x,y) = (c^2 + \|x-y\|_2^2)^{-t}$, $c > 0$, $t > d/2$, etc. on $\mathbb{R}^d$. See Wendland [45] for more examples. The result in *(ii)* of Corollary 13 also appears in [17]. As mentioned before, the estimates for $R_m(\mathcal{F}_k; \{X_i^{(1)}\}_{i=1}^m)$ can be directly obtained by using the entropy numbers of $\mathcal{F}_k$. See Cucker and Zhou [43, Chapter 5] for the estimates of entropy numbers for various $\mathcal{H}$.

C. Simulation results

So far, in Sections III-A and III-B, we have presented the empirical estimation of W, β and γ_k and their convergence analysis. Now, the question is how good are these estimators in practice? In this subsection, we demonstrate the performance of these estimators through simulations.

As we have mentioned before, given $\mathbb{P}$ and $\mathbb{Q}$, it is usually difficult to exactly compute W, β and γ_k. However, in order to test the performance of their estimators, in the following, we consider some examples where W, β and γ_k can be computed exactly.

1) Estimator of W: For the ease of computation, let us consider $\mathbb{P}$ and $\mathbb{Q}$ (defined on the Borel σ-algebra of $\mathbb{R}^d$) as product measures, $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}^{(i)}$ and $\mathbb{Q} = \otimes_{i=1}^d \mathbb{Q}^{(i)}$, where $\mathbb{P}^{(i)}$ and $\mathbb{Q}^{(i)}$ are defined on the Borel σ-algebra of $\mathbb{R}$. In this setting, when $\rho(x,y) = \|x-y\|_1$, it is easy to show that

$$W(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^d W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}), \tag{41}$$

where

$$W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}) = \int_{\mathbb{R}}\left|F_{\mathbb{P}^{(i)}}(x) - F_{\mathbb{Q}^{(i)}}(x)\right|\,dx, \tag{42}$$

and $F_{\mathbb{P}^{(i)}}(x) = \mathbb{P}^{(i)}((-\infty, x])$ [38] (see footnote 6). Therefore, the computation of $W(\mathbb{P}, \mathbb{Q})$ reduces to d computations of the form of (42). Now, in the following, we consider two examples where W in (42) can be computed in closed form. Note that we need M to be a bounded subset of $\mathbb{R}^d$ such that the consistency of $W(\mathbb{P}_m, \mathbb{Q}_n)$ is guaranteed by Corollary 13.

Example 1: Let $M = \times_{i=1}^d [a_i, s_i]$. Suppose $\mathbb{P}^{(i)} = U[a_i, b_i]$ and $\mathbb{Q}^{(i)} = U[r_i, s_i]$, which are uniform distributions on $[a_i, b_i]$ and $[r_i, s_i]$ respectively, where $-\infty < a_i \le r_i \le b_i \le s_i < \infty$. Then, it is easy to verify that $W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}) = (s_i + r_i - a_i - b_i)/2$ and $W(\mathbb{P}, \mathbb{Q})$ follows from (41).

Figures 1(a-c) show the empirical estimates of W (shown in thick dotted lines) for $d = 1$, $d = 2$ and $d = 5$ respectively. Here, we chose $a_i = -\frac{1}{2}$, $b_i = \frac{1}{2}$, $r_i = 0$ and $s_i = 1$ for all $i = 1, \ldots, d$ such that $W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}) = \frac{1}{2}$, $\forall i$ and $W(\mathbb{P}, \mathbb{Q}) = \frac{d}{2}$, shown in thin dotted lines in Figures 1(a-c). Note that the

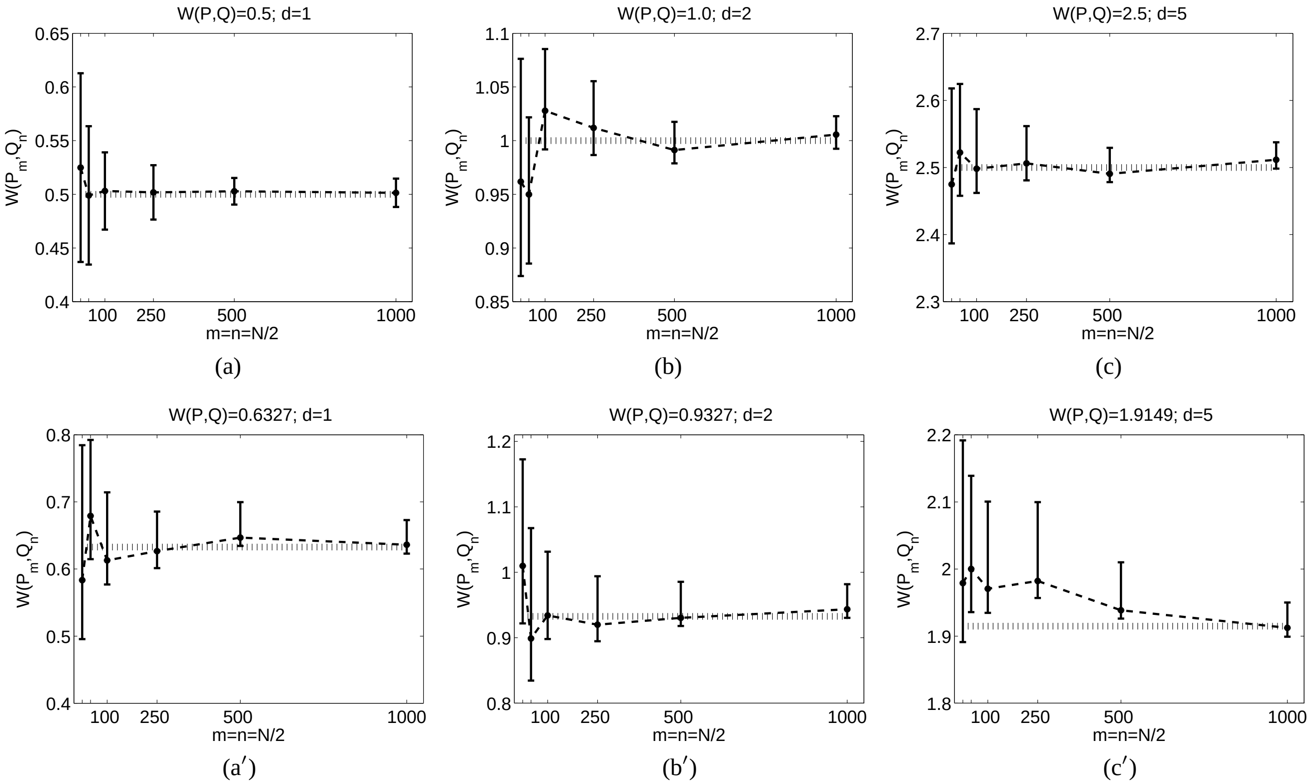

Fig. 1. (a-c) represent the empirical estimates of the Wasserstein distance (shown in thick dotted lines) between $\mathbb{P} = U[-\frac{1}{2}, \frac{1}{2}]^d$ and $\mathbb{Q} = U[0,1]^d$ with $\rho(x,y) = \|x-y\|_1$, for increasing sample size N, where $d=1$ in (a), $d=2$ in (b) and $d=5$ in (c). Here $U[l_1,l_2]^d$ represents a uniform distribution on $[l_1,l_2]^d$ (see Example 1 for details). Similarly, (a′-c′) represent the empirical estimates of the Wasserstein distance (shown in thick dotted lines) between $\mathbb{P}$ and $\mathbb{Q}$, which are truncated exponential distributions on $\mathbb{R}^d_+$ (see Example 2 for details), for increasing sample size N. Here $d=1$ in (a′), $d=2$ in (b′) and $d=5$ in (c′) with $\rho(x,y) = \|x-y\|_1$. The population values of the Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ are shown in thin dotted lines in (a-c, a′-c′). Error bars are obtained by replicating the experiment 20 times.

present choice of $\mathbb{P}$ and $\mathbb{Q}$ would result in a KL-divergence of $+\infty$. ■

Example 2: Let $M = \times_{i=1}^d [0,c_i]$. Suppose $\mathbb{P}^{(i)}$ and $\mathbb{Q}^{(i)}$ have densities $p_i(x) = \frac{d\mathbb{P}^{(i)}}{dx} = \frac{\lambda_i e^{-\lambda_i x}}{1-e^{-\lambda_i c_i}}$ and $q_i(x) = \frac{d\mathbb{Q}^{(i)}}{dx} = \frac{\mu_i e^{-\mu_i x}}{1-e^{-\mu_i c_i}}$ respectively, where $\lambda_i > 0$, $\mu_i > 0$. Note that $\mathbb{P}^{(i)}$ and $\mathbb{Q}^{(i)}$ are exponential distributions supported on $[0,c_i]$ with rate parameters λ_i and μ_i. Then, it can be shown that

$$W(\mathbb{P}^{(i)},\mathbb{Q}^{(i)}) = \left|\frac{1}{\lambda_i} - \frac{1}{\mu_i} - \frac{c_i(e^{-\lambda_i c_i} - e^{-\mu_i c_i})}{(1-e^{-\lambda_i c_i})(1-e^{-\mu_i c_i})}\right|,$$

and $W(\mathbb{P},\mathbb{Q})$ follows from (41).

Figures 1(a′-c′) show the empirical estimates of W (shown in thick dotted lines) for $d=1$, $d=2$ and $d=5$ respectively. Let $\lambda = (\lambda_1, \overset{d}{\ldots}, \lambda_d)$, $\mu = (\mu_1, \overset{d}{\ldots}, \mu_d)$ and $c = (c_1, \overset{d}{\ldots}, c_d)$. In Figure 1(a′), we chose $\lambda = (3)$, $\mu = (1)$ and $c = (5)$ which gives $W(\mathbb{P},\mathbb{Q}) = 0.6327$. In Figure 1(b′), we chose $\lambda = (3,2)$, $\mu = (1,5)$ and $c = (5,6)$, which gives $W(\mathbb{P},\mathbb{Q}) = 0.9327$. In Figure 1(c′), we chose $\lambda = (3,2,1/2,2,7)$, $\mu = (1,5,5/2,1,8)$ and $c = (5,6,3,2,10)$, which gives $W(\mathbb{P},\mathbb{Q}) = 1.9149$. The population values $W(\mathbb{P},\mathbb{Q})$ are shown in thin dotted lines in Figures 1(a′-c′). ■

The empirical estimates in Figure 1 are obtained by drawing N i.i.d. samples (with $m = n = N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and then solving the linear program in (15). It is easy to see that the estimate of $W(\mathbb{P},\mathbb{Q})$ improves with increasing sample size and that $W(\mathbb{P}_m,\mathbb{Q}_n)$ estimates $W(\mathbb{P},\mathbb{Q})$ correctly, which therefore demonstrates the efficacy of the estimator. Error bars are obtained by replicating the experiment 20 times.

2) Estimator of γ_k: We now consider the performance of $\gamma_k(\mathbb{P},\mathbb{Q})$. [17], [18] have shown that when k is measurable and bounded,

$$\begin{aligned}\gamma_k(\mathbb{P},\mathbb{Q}) &= \left\|\int_M k(.,x)\,d\mathbb{P}(x) - \int_M k(.,x)\,d\mathbb{Q}(x)\right\|_{\mathcal{H}} \\ &= \Big[\int k(x,y)\,d\mathbb{P}(x)\,d\mathbb{P}(y) \\ &\quad + \int k(x,y)\,d\mathbb{Q}(x)\,d\mathbb{Q}(y) \\ &\quad -2\int k(x,y)\,d\mathbb{P}(x)\,d\mathbb{Q}(y)\Big]^{\frac{1}{2}}. \end{aligned} \tag{43}$$

Note that, although $\gamma_k(\mathbb{P},\mathbb{Q})$ has a closed form in (43), exact computation is not always possible for all choices of k, $\mathbb{P}$ and $\mathbb{Q}$. In such cases, one has to resort to numerical techniques to compute the integrals in (43). In the following, we present two examples where we choose $\mathbb{P}$ and $\mathbb{Q}$ such that $\gamma_k(\mathbb{P},\mathbb{Q})$ can be computed exactly, which is then used to verify the performance of $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$. Also note that for the consistency of $\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)$, by Corollary 10, we just need the kernel, k to be measurable and bounded and no assumptions on M are required.

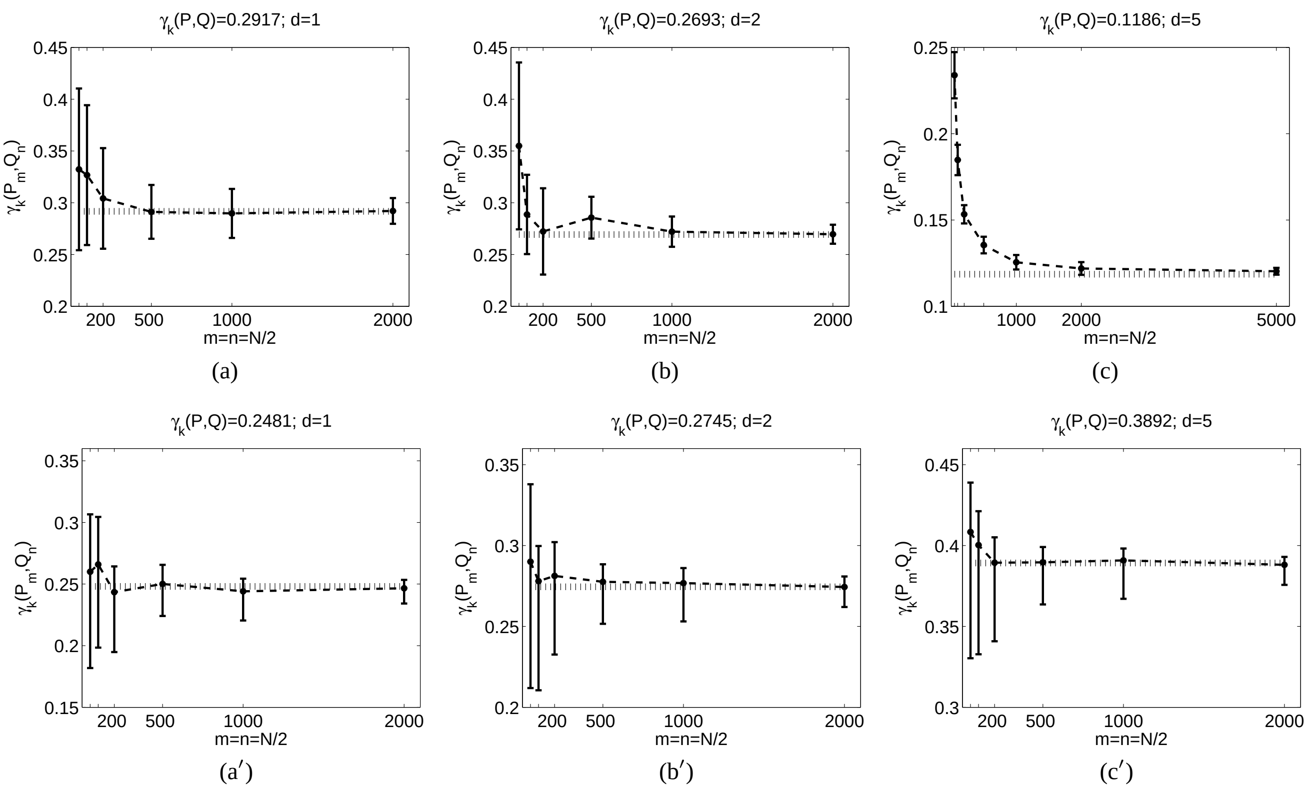

Fig. 2. (a-c) represent the empirical estimates of MMD (shown in thick dotted lines) between $\mathbb{P} = N(0, 2I_d)$ and $\mathbb{Q} = N(1, 2I_d)$ with $k(x, y) = \exp(-\frac{1}{2}\|x-y\|_2^2)$, for increasing sample size N, where $d = 1$ in (a), $d = 2$ in (b) and $d = 5$ in (c) (see Example 3 for details). Here $N(\mu, \sigma^2 I_d)$ represents a normal distribution with mean vector $(\mu_1, \overset{d}{\ldots}, \mu_d)$ and covariance matrix $\sigma^2 I_d$. I_d represents the $d \times d$ identity matrix. Similarly, (a′-c′) represent the empirical estimates of MMD (shown in thick dotted lines) between $\mathbb{P}$ and $\mathbb{Q}$, which are exponential distributions on $\mathbb{R}^d_+$ (see Example 4 for details), for increasing sample size N. Here $d = 1$ in (a′), $d = 2$ in (b′) and $d = 5$ in (c′) with $k(x, y) = \exp(-\frac{1}{4}\|x - y\|_1)$. The population values of MMD are shown in thin dotted lines in (a-c,a′-c′). Error bars are obtained by replicating the experiment 20 times.

Example 3: Let $M = \mathbb{R}^d$, $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}^{(i)}$ and $\mathbb{Q} = \otimes_{i=1}^d \mathbb{Q}^{(i)}$. Suppose $\mathbb{P}^{(i)} = N(\mu_i, \sigma_i^2)$ and $\mathbb{Q}^{(i)} = N(\lambda_i, \theta_i^2)$, where $N(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. Let $k(x, y) = \exp(-\|x - y\|_2^2/2\tau^2)$. Clearly k is measurable and bounded. With this choice of k, $\mathbb{P}$ and $\mathbb{Q}$, γ_k in (43) can be computed exactly as

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \prod_{i=1}^d \frac{\tau}{\sqrt{2\sigma_i^2 + \tau^2}} + \prod_{i=1}^d \frac{\tau}{\sqrt{2\theta_i^2 + \tau^2}} - 2\prod_{i=1}^d \frac{\tau e^{-\frac{(\mu_i - \lambda_i)^2}{2(\sigma_i^2 + \theta_i^2 + \tau^2)}}}{\sqrt{\sigma_i^2 + \theta_i^2 + \tau^2}},$$

as the integrals in (43) simply involve the convolution of Gaussian distributions.

Figures 2(a-c) show the empirical estimates of γ_k (shown in thick dotted lines) for $d = 1$, $d = 2$ and $d = 5$ respectively. Here we chose $\mu_i = 0$, $\lambda_i = 1$, $\sigma_i = \sqrt{2}$, $\theta_i = \sqrt{2}$ for all $i = 1, \ldots, d$ and $\tau = 1$. Using these values in (44), it is easy to check that $\gamma_k(\mathbb{P}, \mathbb{Q}) = 5^{-d/4}(2 - 2e^{-d/10})^{1/2}$, which is shown in thin dotted lines in Figures 2(a-c). ■

Example 4: Let $M = \mathbb{R}^d_+$, $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}^{(i)}$ and $\mathbb{Q} = \otimes_{i=1}^d \mathbb{Q}^{(i)}$. Suppose $\mathbb{P}^{(i)} = \text{Exp}(1/\lambda_i)$ and $\mathbb{Q}^{(i)} = \text{Exp}(1/\mu_i)$, which are exponential distributions on $\mathbb{R}_+$ with rate parameters $\lambda_i > 0$ and $\mu_i > 0$ respectively. Suppose $k(x, y) = \exp(-\alpha\|x - y\|_1)$, $\alpha > 0$, which is a Laplacian kernel on $\mathbb{R}^d$. Then, it is easy to verify that $\gamma_k(\mathbb{P}, \mathbb{Q})$ in (43) reduces to

$$\gamma_k^2(\mathbb{P}, \mathbb{Q}) = \prod_{i=1}^d \frac{\lambda_i}{\lambda_i + \alpha} + \prod_{i=1}^d \frac{\mu_i}{\mu_i + \alpha} - 2\prod_{i=1}^d \frac{\lambda_i \mu_i (\lambda_i + \mu_i + 2\alpha)}{(\lambda_i + \alpha)(\mu_i + \alpha)(\lambda_i + \mu_i)}.$$

Figures 2(a′-c′) show the empirical estimates of γ_k (shown in thick dotted lines) for $d = 1$, $d = 2$ and $d = 5$ respectively. Here, we chose $\{\lambda_i\}_{i=1}^d$ and $\{\mu_i\}_{i=1}^d$ as in Example 2 with $\alpha = \frac{1}{4}$, which gives $\gamma_k(\mathbb{P}, \mathbb{Q}) = 0.2481$ for $d = 1$, 0.2745 for $d = 2$ and 0.3592 for $d = 5$, shown in thin dotted lines in Figures 2(a′-c′). ■

As in the case of W, the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is verified by drawing N i.i.d. samples (with $m = n = N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and computing $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ in (25). Figure 2 shows the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ for various sample sizes and d. It is easy to see that the quality of the estimate improves with increasing sample size and that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ estimates $\gamma_k(\mathbb{P}, \mathbb{Q})$ correctly. As in the case of W, the error bars are obtained by replicating the experiment 20 times.

3) Estimator of β: In the case of W and γ_k, we have some closed form expression to start with (see (42) and (43)), which can be solved by numerical methods. The resulting value is then used as the baseline to test the performance of the estimators of W and γ_k. On the other hand, in the case of β, we are not aware of any such closed form expression to compute the baseline. However, it is possible to compute $\beta(\mathbb{P},\mathbb{Q})$ when $\mathbb{P}$ and $\mathbb{Q}$ are discrete distributions on M, i.e., $\mathbb{P}=\sum_{i=1}^r \lambda_i\delta_{X_i}$, $\mathbb{Q}=\sum_{i=1}^s \mu_i\delta_{Z_i}$, where $\sum_{i=1}^r \lambda_i=1$, $\sum_{i=1}^s \mu_i=1$, $\lambda_i\ge 0,\,\forall\, i$, $\mu_i\ge 0,\,\forall\, i$, and $X_i, Z_i\in M$. This is because, for this choice of $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\beta(\mathbb{P},\mathbb{Q}) = \sup\Big\{\sum_{i=1}^r \lambda_i f(X_i) - \sum_{i=1}^s \mu_i f(Z_i) : \|f\|_{BL}\le 1\Big\}$$
$$= \sup\Big\{\sum_{i=1}^{r+s} \theta_i f(V_i) : \|f\|_{BL}\le 1\Big\}, \qquad (44)$$

where $\theta = (\lambda_1,\ldots,\lambda_r,-\mu_1,\ldots,-\mu_s)$, $V = (X_1,\ldots,X_r,Z_1,\ldots,Z_s)$ with $\theta_i := (\theta)_i$ and $V_i := (V)_i$. Now, (44) is of the form of (11) and therefore, by Theorem 7, $\beta(\mathbb{P},\mathbb{Q})=\sum_{i=1}^{r+s}\theta_i a_i^\star$, where $\{a_i^\star\}$ solve the following linear program,

$$\begin{aligned}\max_{a_1,\ldots,a_{r+s},b,c}\ & \sum_{i=1}^{r+s}\theta_i a_i\\ \text{s.t.}\ & -b\,\rho(V_i,V_j)\le a_i-a_j\le b\,\rho(V_i,V_j),\,\forall\, i,j\\ & -c\le a_i\le c,\,\forall\, i\\ & b+c\le 1.\end{aligned} \qquad (45)$$

Therefore, for these distributions, one can compute the baseline which can then be used to verify the performance of $\beta(\mathbb{P}_m,\mathbb{Q}_n)$. In the following, we consider a simple example to demonstrate the performance of $\beta(\mathbb{P}_m,\mathbb{Q}_n)$.

Example 5: Let $M=\{0,1,2,3,4,5\}\subset\mathbb{R}$, $\lambda = (\frac{1}{3},\frac{1}{6},\frac{1}{8},\frac{1}{4},\frac{1}{8})$, $\mu=(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$, $X=(0,1,2,3,4)$ and $Z=(2,3,4,5)$. With this choice, $\mathbb{P}$ and $\mathbb{Q}$ are defined as $\mathbb{P}=\sum_{i=1}^5\lambda_i\delta_{X_i}$ and $\mathbb{Q}=\sum_{i=1}^4\mu_i\delta_{Z_i}$. By solving (45) with $\rho(x,y)=|x-y|$, we get $\beta(\mathbb{P},\mathbb{Q})=0.5278$. Note that the KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is $+\infty$.

Figure 3 shows the empirical estimates of $\beta(\mathbb{P},\mathbb{Q})$ (shown in a thick dotted line) which are computed by drawing N i.i.d. samples (with $m=n=N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and solving the linear program in (21). It can be seen that $\beta(\mathbb{P}_m,\mathbb{Q}_n)$ estimates $\beta(\mathbb{P},\mathbb{Q})$ correctly. ■

Since we do not know how to compute $\beta(\mathbb{P},\mathbb{Q})$ for $\mathbb{P}$ and $\mathbb{Q}$ other than the ones we discussed here, we do not provide any other non-trivial examples to test the performance of $\beta(\mathbb{P}_m,\mathbb{Q}_n)$.

D. Non-parametric estimation of total variation distance

In this subsection, we consider the empirical estimation of total variation distance,

$$TV(\mathbb{P},\mathbb{Q}) := \sup\{\mathbb{P}f-\mathbb{Q}f : \|f\|_\infty\le 1\}, \qquad (46)$$

and the statistical consistency of the estimator. Let $TV(\mathbb{P}_m,\mathbb{Q}_n)$ be an empirical estimator of $TV(\mathbb{P},\mathbb{Q})$. Using similar arguments as in Theorems 6 and 7, it can be shown that

$$TV(\mathbb{P}_m,\mathbb{Q}_n)=\sum_{i=1}^N \widetilde{Y}_i a_i^\star, \qquad (47)$$

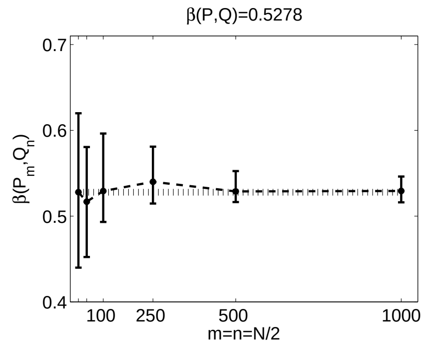

Fig. 3. Empirical estimates of the Dudley metric (shown in a thick dotted line) between discrete distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}$ (see Example 5 for details), for increasing sample size N. The population value of the Dudley metric is shown in a thin dotted line. Error bars are obtained by replicating the experiment 20 times.

where $\{a_i^\star\}_{i=1}^N$ solve the following linear program,

$$\begin{aligned}\max_{a_1,\ldots,a_N}\ & \sum_{i=1}^N \widetilde{Y}_i a_i\\ \text{s.t.}\ & -1\le a_i\le 1,\,\forall\, i.\end{aligned} \qquad (48)$$

Now, the question is whether this estimator consistent. To answer this question, we consider an equivalent representation of TV given as

$$TV(\mathbb{P},\mathbb{Q}) = 2\sup_{A\in\mathcal{A}}|\mathbb{P}(A)-\mathbb{Q}(A)|, \qquad (49)$$

where the supremum is taken over all measurable subsets A of M [41]. Note that $|TV(\mathbb{P}_m,\mathbb{Q}_n)-TV(\mathbb{P},\mathbb{Q})|\le TV(\mathbb{P}_m,\mathbb{P})+TV(\mathbb{Q}_n,\mathbb{Q})$. It is easy to see that $TV(\mathbb{P}_m,\mathbb{P})\overset{a.s.}{\nrightarrow}0$ as $m\to\infty$ for all $\mathbb{P}$ and therefore, the estimator in (47) is not strongly consistent. This is because if $\mathbb{P}$ is absolutely continuous, then $TV(\mathbb{P}_m,\mathbb{P})=2$, where we have considered the set A that is the finite support of $\mathbb{P}_m$ such that $\mathbb{P}_m(A)=1$ and $\mathbb{P}(A)=0$. In fact, Devroye and Györfi [41] have proved that for any empirical measure, $\widehat{\mathbb{P}}_m$ (a function depending on $\{X_i^{(1)}\}_{i=1}^m$ assigning a nonnegative number to any measurable set), there exists a distribution, $\mathbb{P}$ such that for all m,

$$\sup_{A\in\mathcal{A}}|\widehat{\mathbb{P}}_m(A)-\mathbb{P}(A)|>\frac{1}{4}\ \text{a.s.} \qquad (50)$$

This indicates that, for the strong consistency of distribution estimates in total variation, the set of probability measures has to be restricted. Barron *et al.* [46] have studied the classes of distributions that can be estimated consistently in total variation. Therefore, for such distributions, the total variation distance between them can be estimated by an estimator that is strongly consistent.

The issue in the estimation of $TV(\mathbb{P},\mathbb{Q})$ is that the set $\mathfrak{F}_{TV}:=\{f:\|f\|_\infty\le 1\}$ is too large to obtain meaningful results if no assumptions on distributions are made. On the other hand, one can choose a more manageable subset $\mathfrak{F}$ of $\mathfrak{F}_{TV}$ such that $\gamma_{\mathfrak{F}}(\mathbb{P},\mathbb{Q})\le TV(\mathbb{P},\mathbb{Q}),\,\forall\,\mathbb{P},\mathbb{Q}\in\mathscr{P}$ and $\gamma_{\mathfrak{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ is a consistent estimator of $\gamma_{\mathfrak{F}}(\mathbb{P},\mathbb{Q})$. Examples of such choice of $\mathfrak{F}$ include $\mathfrak{F}_\beta$ and $\{\mathbb{1}_{(-\infty,t]} : t\in\mathbb{R}^d\}$, where the former yields the Dudley metric while the latter results in the Kolmogorov distance. The empirical estimator of the Dudley metric and its consistency have been presented in Sections III-A and III-B. The empirical estimator of the Kolmogorov distance between $\mathbb{P}$ and $\mathbb{Q}$ is well studied and is strongly consistent, which simply follows from the famous Glivenko-Cantelli theorem [36, Theorem 12.4].

Since the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$ cannot be estimated consistently for all $\mathbb{P},\mathbb{Q}\in\mathscr{P}$, in the following, we present two lower bounds on TV, one involving W and β and the other involving γ_k, which can be estimated consistently.

Theorem 15 (Lower bounds on TV): *(i)* For all $\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P},\mathbb{Q}\in\mathscr{P}$, we have

$$TV(\mathbb{P},\mathbb{Q})\ge\frac{W(\mathbb{P},\mathbb{Q})\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})-\beta(\mathbb{P},\mathbb{Q})}. \tag{51}$$

(ii) Suppose $C:=\sup_{x\in M}k(x,x)<\infty$. Then

$$TV(\mathbb{P},\mathbb{Q})\ge\frac{\gamma_k(\mathbb{P},\mathbb{Q})}{\sqrt{C}}. \tag{52}$$

Before, we prove Theorem 15, we present a simple lemma.

Lemma 16: Let $\theta:V\to\mathbb{R}$ and $\psi:V\to\mathbb{R}$ be convex functions on a real vector space V. Suppose

$$a=\sup\{\theta(x):\psi(x)\le b\}, \tag{53}$$

where θ is not constant on $\{x:\psi(x)\le b\}$ and $a<\infty$. Then

$$b=\inf\{\psi(x):\theta(x)\ge a\}. \tag{54}$$

Proof: Note that $A:=\{x:\psi(x)\le b\}$ is a convex subset of V. Since θ is not constant on A, by Theorem 25 (see the Appendix), θ attains its supremum on the boundary of A. Therefore, any solution, x_* to (53) satisfies $\theta(x_*)=a$ and $\psi(x_*)=b$. Let $G:=\{x:\theta(x)>a\}$. For any $x\in G$, $\psi(x)>b$. If this were not the case, then x_* is not a solution to (53). Let $H:=\{x:\theta(x)=a\}$. Clearly, $x_*\in H$ and so there exists an $x\in H$ for which $\psi(x)=b$. Suppose $\inf\{\psi(x):x\in H\}=c<b$, which means for some $x^*\in H$, $x^*\in A$. From (53), this implies θ attains its supremum relative to A at some point of relative interior of A. By Theorem 25, this implies θ is constant on A leading to a contradiction. Therefore, $\inf\{\psi(x):x\in H\}=b$ and the result in (54) follows. ∎

Proof of Theorem 15: *(i)* Note that $\|f\|_L$, $\|f\|_{BL}$ and $\|f\|_\infty$ are convex functionals on the vector spaces Lip(M,ρ), $BL(M,\rho)$ and $U(M):=\{f:M\to\mathbb{R}\,|\,\|f\|_\infty<\infty\}$ respectively. Similarly, $\mathbb{P}f-\mathbb{Q}f$ is a convex functional on Lip(M,ρ), $BL(M,\rho)$ and $U(M)$. Since $\mathbb{P}\neq\mathbb{Q}$, $\mathbb{P}f-\mathbb{Q}f$ is not constant on $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_{TV}$. Therefore, by appropriately choosing ψ, θ, V and b in Lemma 16, the following sequence of inequalities are obtained. Define $\beta:=\beta(\mathbb{P},\mathbb{Q})$, $W:=W(\mathbb{P},\mathbb{Q})$, $TV:=TV(\mathbb{P},\mathbb{Q})$.

$$\begin{aligned}1&=\inf\{\|f\|_{BL}:\mathbb{P}f-\mathbb{Q}f\ge\beta,\,f\in BL(M,\rho)\}\\&\ge\inf\{\|f\|_L:\mathbb{P}f-\mathbb{Q}f\ge\beta,\,f\in BL(M,\rho)\}\\&\quad+\inf\{\|f\|_\infty:\mathbb{P}f-\mathbb{Q}f\ge\beta,\,f\in BL(M,\rho)\}\\&=\frac{\beta}{W}\inf\{\|f\|_L:\mathbb{P}f-\mathbb{Q}f\ge W,\,f\in BL(M,\rho)\}\\&\quad+\frac{\beta}{TV}\inf\{\|f\|_\infty:\mathbb{P}f-\mathbb{Q}f\ge TV,\,f\in BL(M,\rho)\}\\&\ge\frac{\beta}{W}\inf\{\|f\|_L:\mathbb{P}f-\mathbb{Q}f\ge W,\,f\in \text{Lip}(M,\rho)\}\\&\quad+\frac{\beta}{TV}\inf\{\|f\|_\infty:\mathbb{P}f-\mathbb{Q}f\ge TV,\,f\in U(M)\}\\&=\frac{\beta}{W}+\frac{\beta}{TV},\end{aligned}$$

which gives (51).

(ii) To prove (52), we use the coupling formulation for TV [47, p. 19] given by

$$TV(\mathbb{P},\mathbb{Q})=2\inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})}\mu(X\neq Y), \tag{55}$$

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

$$\begin{aligned}\left|\int_M f\,d(\mathbb{P}-\mathbb{Q})\right|&=\left|\int(f(x)-f(y))\,d\lambda(x,y)\right|\\&\le\int|f(x)-f(y)|\,d\lambda(x,y)\\&\overset{(a)}{=}\int|\langle f,k(.,x)-k(.,y)\rangle_{\mathcal{H}}|\,d\lambda(x,y)\\&\overset{(b)}{\le}\|f\|_{\mathcal{H}}\int\|k(.,x)-k(.,y)\|_{\mathcal{H}}\,d\lambda(x,y),\end{aligned}$$

where we have used the reproducing property of $\mathcal{H}$ in (a) and the Cauchy-Schwartz inequality in (b). Taking the supremum over $f\in\mathcal{F}_k$ and the infimum over $\lambda\in\mathcal{L}(\mathbb{P},\mathbb{Q})$ gives

$$\gamma_k(\mathbb{P},\mathbb{Q})\le\inf_{\lambda\in\mathcal{L}(\mathbb{P},\mathbb{Q})}\int\|k(.,x)-k(.,y)\|_{\mathcal{H}}\,d\lambda(x,y). \tag{56}$$

Consider

$$\begin{aligned}\|k(.,x)-k(.,y)\|_{\mathcal{H}}&\le\mathbb{1}_{x\neq y}\|k(.,x)-k(.,y)\|_{\mathcal{H}}\\&\le\mathbb{1}_{x\neq y}\left[\|k(.,x)\|_{\mathcal{H}}+\|k(.,y)\|_{\mathcal{H}}\right]\\&=\mathbb{1}_{x\neq y}\left[\sqrt{k(x,x)}+\sqrt{k(y,y)}\right]\\&\le2\sqrt{C}\mathbb{1}_{x\neq y}.\end{aligned} \tag{57}$$

Using (57) in (56) yields (52). ∎

Remark 17: *(i)* As mentioned before, a simple lower bound on TV can be obtained as $TV(\mathbb{P},\mathbb{Q})\ge\beta(\mathbb{P},\mathbb{Q}),\,\forall\,\mathbb{P},\mathbb{Q}\in\mathscr{P}$. It is easy to see that the bound in (51) is tighter as $\frac{W(\mathbb{P},\mathbb{Q})\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})-\beta(\mathbb{P},\mathbb{Q})}\ge\beta(\mathbb{P},\mathbb{Q})$ with equality if and only if $\mathbb{P}=\mathbb{Q}$.

(ii) From (51), it is easy to see that $TV(\mathbb{P},\mathbb{Q})=0$ or $W(\mathbb{P},\mathbb{Q})=0$ implies $\beta(\mathbb{P},\mathbb{Q})=0$ while the converse is not true. This shows that the topology induced by β on $\mathscr{P}$ is coarser than the topology induced by either W or TV.

(iii) The bounds in (51) and (52) translate as lower bounds on the KL-divergence through Pinsker's inequality: $TV^2(\mathbb{P},\mathbb{Q})\le 2\,KL(\mathbb{P},\mathbb{Q}),\,\forall\,\mathbb{P},\mathbb{Q}\in\mathscr{P}$. See Fedotov *et al.* [29] and references therein for more refined bounds between TV and KL. Therefore, using these bounds, one can obtain a consistent estimate of a lower bound on TV and KL. The bounds in (51) and (52) also translate to lower bounds on other distance

measures on $\mathscr{P}$. See [48] for a detailed discussion on the relation between various metrics.

To summarize, in this section, we have considered the empirical estimation of IPMs along with their convergence rate analysis. We have shown that IPMs such as the Wasserstein distance, Dudley metric and MMD are simpler to estimate than the KL-divergence. This is because the Wasserstein distance and Dudley metric are estimated by solving a linear program while estimating the KL-divergence involves solving a quadratic program [28]. Even more, the estimator of MMD has a simple closed form expression. On the other hand, space partitioning schemes like in [27], to estimate the KL-divergence, become increasingly difficult to implement as the number of dimensions increases whereas an increased number of dimensions has only a mild effect on the complexity of estimating W, β and γ_k. In addition, the estimators of IPMs, especially the Wasserstein distance, Dudley metric and MMD, exhibit good convergence behavior compared to KL-divergence estimators as the latter can have an arbitrarily slow rate of convergence depending on the probability distributions [27], [28]. With these advantages, we believe that IPMs can find applications in information theory, detection theory, image processing, machine learning and other areas. As an example, in the following section, we show how IPMs are related to binary classification.

IV. IPMs and Binary Classification

Many previous works, e.g., [4], [25], [32] have studied the problem of relating the risk (expected loss) in binary classification problems to ϕ-divergences (see [33, Section 1.3] for a list of detailed references). Since IPMs are essentially different from ϕ-divergences (see Section II), we are interested to study the relation between IPMs and binary classification. In this section, we show how IPMs, measuring the distance between class conditional distributions, appear naturally in binary classification problems. First, in Section IV-A we provide a novel interpretation for β, W, TV and γ_k (see Theorem 18), as the optimal L-risk of a binary classification problem. Second, in Section IV-B, we relate W and β to the margin of the Lipschitz classifier [34] and the bounded Lipschitz classifier respectively. Third, in Section IV-C, we discuss the relation between γ_k and the Parzen window classifier [30], [35] (also called the kernel classification rule [36, Chapter 10]).

A. Interpretation of β, W, TV and γ_k as the optimal L-risk of a binary classification problem

Let us consider the binary classification problem with X being a M-valued random variable, Y being a $\{-1,+1\}$-valued random variable and the product space, $M\times\{-1,+1\}$, being endowed with a Borel probability measure μ. A discriminant function, f is a real valued measurable function on M, whose sign is used to make a classification decision. The goal is to choose an f that minimizes the probability of making the incorrect classification, also known as the *Bayes risk*. Formally, the Bayes decision rule involves solving

$$\begin{aligned}\widetilde{f} &= \arg\inf_{f\in\mathcal{F}_\star}\mu(Y\neq \text{sign}(f(X)))\\ &= \arg\inf_{f\in\mathcal{F}_\star}\int_M [\![y\neq\text{sign}(f(x))]\!]\,d\mu(x,y),\end{aligned}\tag{58}$$

where $(a,b)\mapsto [\![a\neq b]\!]$ is the $0-1$ loss function and $\mu(Y\neq \text{sign}(\widetilde{f}(X)))$ is the Bayes risk. (58) can be generalized by replacing the $0-1$ loss function with some loss function, $L:\{-1,+1\}\times\mathbb{R}\to\mathbb{R}$, for example the squared-loss function, $L(a,b)=(a-b)^2$. Given L, the associated optimal L-risk is defined as

$$\begin{aligned}R^L_{\mathcal{F}_\star} &= \inf_{f\in\mathcal{F}_\star}\int_M L(y,f(x))\,d\mu(x,y)\\ &= \inf_{f\in\mathcal{F}_\star}\Big\{\varepsilon\int_M L_1(f(x))\,d\mathbb{P}(x)\\ &+ \quad (1-\varepsilon)\int_M L_{-1}(f(x))\,d\mathbb{Q}(x)\Big\},\end{aligned}\tag{59}$$

where $L_1(\alpha):=L(1,\alpha)$, $L_{-1}(\alpha):=L(-1,\alpha)$, $\mathbb{P}(X):=\mu(X|Y=+1)$, $\mathbb{Q}(X):=\mu(X|Y=-1)$, $\varepsilon:=\mu(M,Y=+1)$. Here, $\mathbb{P}$ and $\mathbb{Q}$ represent the class-conditional distributions and ε is the prior distribution of class $+1$.

Nguyen *et al.* [25] have shown an equivalence between ϕ-divergences (between $\mathbb{P}$ and $\mathbb{Q}$) and the optimal L-risk associated with a loss-function, L, that satisfies $L_1(\alpha)=L_{-1}(-\alpha)$. They showed that for each loss function, L, there exists exactly one corresponding ϕ-divergence such that the optimal L-risk is equal to the negative ϕ-divergence between $\mathbb{P}$ and $\mathbb{Q}$.[8] For example, the total-variation distance, Hellinger distance and χ^2-divergence are shown to be related to the optimal L-risk where L is the hinge loss ($L(y,\alpha)=\max(0,1-y\alpha)$), exponential loss ($L(y,\alpha)=\exp(-y\alpha)$) and logistic loss ($L(y,\alpha)=\log(1+\exp(-y\alpha))$) respectively.[9] In statistical machine learning, these losses are well-studied and are shown to result in various binary classification algorithms like support vector machines, Adaboost and logistic regression. See [30], [49] for details.

Since IPMs and ϕ-divergences are essentially different, we present and prove the following result that relates IPMs (between the class conditional distributions) and the optimal L-risk of a binary classification problem.

Theorem 18 ($\gamma_\mathcal{F}$ and associated L): Let $L_1(\alpha)=-\frac{\alpha}{\varepsilon}$ and $L_{-1}(\alpha)=\frac{\alpha}{1-\varepsilon}$. Let $\mathcal{F}\subset\mathcal{F}_\star$ be such that $f\in\mathcal{F}\Rightarrow -f\in\mathcal{F}$. Then, $\gamma_\mathcal{F}(\mathbb{P},\mathbb{Q})=-R^L_\mathcal{F}$.

Proof: From (59), we have

$$\begin{aligned}&\varepsilon\int_M L_1(f)\,d\mathbb{P}+(1-\varepsilon)\int_M L_{-1}(f)\,d\mathbb{Q}\\ &= \int_M f\,d\mathbb{Q}-\int_M f\,d\mathbb{P}=\mathbb{Q}f-\mathbb{P}f.\end{aligned}\tag{60}$$

[8] This result holds even if L does not satisfy the property $L_1(\alpha)=L_{-1}(-\alpha)$, $\forall\alpha\in\mathbb{R}$. However, such an assumption is made to completely analyze the relation between L and ϕ.

[9] By choosing $L_1(\alpha)=-\frac{\alpha}{\varepsilon}$ and $L_{-1}(\alpha)=\frac{e^{\alpha-1}}{1-\varepsilon}$, it can be shown that the associated optimal L-risk is the negative of the KL-divergence between $\mathbb{P}$ and $\mathbb{Q}$.

Therefore,

$$R_{\mathcal{F}}^{L} = \inf_{f\in\mathcal{F}}(\mathbb{Q}f-\mathbb{P}f) = -\sup_{f\in\mathcal{F}}(\mathbb{P}f-\mathbb{Q}f) \stackrel{(a)}{=} -\sup_{f\in\mathcal{F}}|\mathbb{P}f-\mathbb{Q}f| = -\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}), \quad (61)$$

where (a) follows from the fact that $\mathcal{F}$ is symmetric around zero, i.e., $f\in\mathcal{F}\Rightarrow -f\in\mathcal{F}$. ■

Theorem 18 shows that $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ is the negative of the optimal L-risk that is associated with a binary classifier that classifies the class conditional distributions $\mathbb{P}$ and $\mathbb{Q}$ using the loss function, L, in Theorem 18, when the discriminant function is restricted to $\mathcal{F}$. Therefore, Theorem 18 provides a novel interpretation for the total variation distance, Dudley metric, Wasserstein distance and MMD, which can be understood as the optimal L-risk associated with binary classifiers where the discriminant function, f is restricted to $\mathcal{F}_{TV}$, $\mathcal{F}_{\beta}$, $\mathcal{F}_W$ and $\mathcal{F}_k$ respectively.

Suppose, we are given a finite number of samples $\{(X_i,Y_i)\}_{i=1}^N$, $X_i\in M$, $Y_i\in\{-1,+1\},\forall\, i$ drawn i.i.d. from μ and we would like to build a classifier, $f\in\mathcal{F}$ that minimizes the expected loss (with L as in Theorem 18) based on this finite number of samples. This is usually carried out by solving an empirical equivalent of (59), which reduces to (11), i.e., $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)=\sup\{|\sum_{i=1}^N \widetilde{Y}_i f(X_i)| : f\in\mathcal{F}\}$ by noting that $X_{\cdot}^{(1)} := X_i$ when $Y_i=1$, $X_{\cdot}^{(2)} := X_i$ when $Y_i=-1$, and $f\in\mathcal{F}\Rightarrow -f\in\mathcal{F}$. This means the sign of $f\in\mathcal{F}$ that solves (11) is the classifier we are looking for.

B. Wasserstein distance and Dudley metric: Relation to Lipschitz and bounded Lipschitz classifiers

The Lipschitz classifier is defined as the solution, f_{lip} to the following program:

$$\inf_{f\in \text{Lip}(M,\rho)} \|f\|_L \quad \text{s.t. } Y_i f(X_i)\ge 1,\, i=1,\dots,N, \quad (62)$$

which is a large margin classifier with margin[10] $\frac{1}{\|f_{\text{lip}}\|_L}$. The program in (62) computes a *smooth* function, f that classifies the training sequence, $\{(X_i,Y_i)\}_{i=1}^N$ correctly (note that the constraints in (62) are such that $\text{sign}(f(X_i))=Y_i$, which means f classifies the training sequence correctly, assuming the training sequence is separable). The smoothness is controlled by $\|f\|_L$ (the smaller the value of $\|f\|_L$, the smoother f and vice-versa). See [34] for a detailed study on the Lipschitz classifier. Replacing $\|f\|_L$ by $\|f\|_{BL}$ in (62) gives the bounded Lipschitz classifier, f_{BL} which is the solution to the following program:

$$\inf_{f\in BL(M,\rho)} \|f\|_{BL} \quad \text{s.t. } Y_i f(X_i)\ge 1,\, i=1,\dots,N. \quad (63)$$

Note that replacing $\|f\|_L$ by $\|f\|_{\mathcal{H}}$ in (62), taking the infimum over $f\in\mathcal{H}$, yields the hard-margin support vector machine (SVM) [50]. We now show how the empirical estimates of W and β appear as upper bounds on the margins of the Lipschitz and bounded Lipschitz classifiers, respectively.

Theorem 19: The Wasserstein distance and Dudley metric are related to the margins of Lipschitz and bounded Lipschitz classifiers as

$$\frac{1}{\|f_{\text{lip}}\|_L} \le \frac{W(\mathbb{P}_m,\mathbb{Q}_n)}{2}, \quad (64)$$

$$\frac{1}{\|f_{\text{BL}}\|_{BL}} \le \frac{\beta(\mathbb{P}_m,\mathbb{Q}_n)}{2}. \quad (65)$$

In addition, there exists $f^\star\in \text{Lip}(M,\rho)$, $f_\star\in BL(M,\rho)$ such that $\|f^\star\|_L=\|f_{\text{lip}}\|_L$, $\|f_\star\|_{BL}=\|f_{\text{BL}}\|_{BL}$ and $\text{sign}(f^\star)$, $\text{sign}(f_\star)$ are 1-nearest neighbor (NN) classifiers.[11]

Proof: Define $W_{mn} := W(\mathbb{P}_m,\mathbb{Q}_n)$. By Lemma 16, we have

$$1=\inf\Big\{\|f\|_L : \sum_{i=1}^N \widetilde{Y}_i f(X_i)\ge W_{mn},\, f\in \text{Lip}(M,\rho)\Big\},$$

which can be written as

$$\frac{2}{W_{mn}}=\inf\Big\{\|f\|_L : \sum_{i=1}^N \widetilde{Y}_i f(X_i)\ge 2,\, f\in \text{Lip}(M,\rho)\Big\}.$$

Note that $\{f\in \text{Lip}(M,\rho) : Y_i f(X_i)\ge 1,\forall\, i\}\subset\{f\in \text{Lip}(M,\rho) : \sum_{i=1}^N \widetilde{Y}_i f(X_i)\ge 2\}$, and therefore

$$\frac{2}{W_{mn}}\le\inf\Big\{\|f\|_L : Y_i f(X_i)\ge 1,\forall\, i,\, f\in \text{Lip}(M,\rho)\Big\},$$

hence proving (64). Similar analysis for β yields (65).

The existence of $f^\star\in \text{Lip}(M,\rho)$ such that $\|f^\star\|_L = \|f_{\text{Lip}}\|_L$ and $\text{sign}(f^\star)$ being a 1-NN classifier follows from [34, Lemmas 11-13]. However, for completeness, we provide the proof of these results. To this end, let us define $\rho(a,B) := \inf_{b\in B}\rho(a,b)$, $\rho(A,B) := \inf_{a\in A, b\in B}\rho(a,b)$ for any $A,B\subset M$, $X^+ := \{X_i : Y_i=1\}$ and $X^- := \{X_i : Y_i=-1\}$.

Consider the Lipschitz classifier in (62). Then,

$$\|f\|_L = \sup_{x\ne y\in M}\frac{|f(x)-f(y)|}{\rho(x,y)} \ge \max_{X_i\ne X_j}\frac{|f(X_i)-f(X_j)|}{\rho(X_i,X_j)} \ge \max_{X_i\ne X_j}\frac{|Y_i-Y_j|}{\rho(X_i,X_j)} = \frac{2}{\rho(X^+,X^-)} =: L^\star. \quad (66)$$

So, any solution, f_{lip} to (62) satisfies $\|f_{\text{lip}}\|_L = L^\star$. Let $f_{\text{lip}}(X_i)=Y_i$. Clearly, f_{lip} satisfies the constraints in (62). Therefore, by Lemma 20 (see the Appendix),

$$f_\alpha(x) := \alpha\min_{i=1,\dots,N}(Y_i+L^\star\rho(x,X_i)) +(1-\alpha)\max_{i=1,\dots,N}(Y_i-L^\star\rho(x,X_i)), \quad (67)$$

is a solution to (62) for any $\alpha\in[0,1]$. Consider $f_{\frac{1}{2}}$, which can be rewritten as:

$$f_{\frac{1}{2}}(x) = \frac{1}{2}\min(L^\star\rho(x,X^+)+1, L^\star\rho(x,X^-)-1) -\frac{1}{2}\min(L^\star\rho(x,X^-)+1, L^\star\rho(x,X^+)-1).$$

[10]The margin is a technical term used in statistical machine learning. See [30] for details.

[11]The 1-nearest neighbor rule, f_{NN} is defined as: $f_{\text{NN}}(x)=1$ if $\inf\{\rho(x,X_i) : Y_i=1\}\le\inf\{\rho(x,X_i) : Y_i=-1\}$ and $f_{\text{NN}}(x)=-1$ otherwise.

Consider x such that $\rho(x, X^+) \geq \rho(x, X^-)$. Then, we have

$$\begin{aligned} f_{\frac{1}{2}}(x) &= -\frac{1}{2}\min(L^\star\rho(x,X^-)+1, L^\star\rho(x,X^+)-1) \\ &\qquad +\frac{1}{2}(L^\star\rho(x,X^-)-1) \\ &= \frac{1}{2}\max(L^\star\rho(x,X^-)-L^\star\rho(x,X^+),-2) \leq 0. \end{aligned}$$

Similarly, let us consider x such that $\rho(x, X^+) \leq \rho(x, X^-)$. Then

$$\begin{aligned} f_{\frac{1}{2}}(x) &= \frac{1}{2}\min(L^\star\rho(x,X^+)+1, L^\star\rho(x,X^-)-1) \\ &\qquad -\frac{1}{2}(L^\star\rho(x,X^+)-1) \\ &= \frac{1}{2}\min(L^\star\rho(x,X^-)-L^\star\rho(x,X^+),2) \geq 0. \end{aligned}$$

Therefore, sign$(f_{\frac{1}{2}})$ represents a 1-NN classifier. Choose $f^\star = f_{\frac{1}{2}}$.

The proof for the existence of $f_\star$ is similar to that of $f^\star$. Consider,

$$\begin{aligned} \|f\|_{BL} &= \sup_{x\neq y\in M}\frac{|f(x)-f(y)|}{\rho(x,y)} + \sup_{x\in M}|f(x)| \\ &\geq \max_{X_i\neq X_j}\frac{|f(X_i)-f(X_j)|}{\rho(X_i,X_j)} + \max_i |f(X_i)| \\ &\geq \max_{X_i\neq X_j}\frac{|Y_i-Y_j|}{\rho(X_i,X_j)} + \max_i |Y_i| \\ &= \frac{2}{\rho(X^+,X^-)} + 1 =: L^\star + 1. \end{aligned} \tag{68}$$

So f_{BL} satisfies $\|f_{\text{BL}}\|_{BL} = L^\star + 1$. Let $f_{\text{BL}}(X_i) = Y_i$, which satisfies the constraints of the program in (63). Therefore, by Lemma 21 (see the Appendix), $h_\alpha = \max(-1, \min(f_\alpha, 1))$, is a solution to (63) for any $\alpha \in [0, 1]$, where f_α is defined in (67). Now, based on the proof of sign$(f_{\frac{1}{2}})$ being a 1-NN classifier, it is easy to check that sign$(h_{\frac{1}{2}})$ is a 1-NN classifier. Choose $f_\star = h_{\frac{1}{2}}$. ∎

The significance of this result is as follows. (64) shows that $\|f_{\text{lip}}\|_L \geq \frac{2}{W(\mathbb{P}_m,\mathbb{Q}_n)}$, which means the smoothness of the classifier, f_{lip}, computed as $\|f_{\text{lip}}\|_L$ is bounded by the inverse of the Wasserstein distance between $\mathbb{P}_m$ and $\mathbb{Q}_n$. So, if the distance between the class-conditionals $\mathbb{P}$ and $\mathbb{Q}$ is "small" (in terms of W), then the resulting Lipschitz classifier is less smooth, i.e., a "complex" classifier is required to classify the distributions $\mathbb{P}$ and $\mathbb{Q}$. A similar explanation holds for the bounded Lipschitz classifier.

C. Maximum mean discrepancy: Relation to Parzen window classifier and support vector machine

Consider the maximizer f, for the empirical estimator of MMD, in (24). Computing $y = \text{sign}(f(x))$ gives

$$y = \begin{cases} +1, & \frac{1}{m}\sum_{Y_i=1}k(x,X_i) > \frac{1}{n}\sum_{Y_i=-1}k(x,X_i) \\ -1, & \frac{1}{m}\sum_{Y_i=1}k(x,X_i) \leq \frac{1}{n}\sum_{Y_i=-1}k(x,X_i) \end{cases}, \tag{69}$$

which is exactly the classification function of a Parzen window classifier [30], [35]. It is easy to see that (69) can be rewritten as

$$y = \text{sign}(\langle w, k(.,x)\rangle_{\mathcal{H}}), \tag{70}$$

where $w = \mu^+ - \mu^-$, $\mu^+ := \frac{1}{m}\sum_{Y_i=1}k(.,X_i)$ and $\mu^- := \frac{1}{n}\sum_{Y_i=-1}k(.,X_i)$. μ^+ and μ^- represent the class means associated with $X^+ := \{X_i : Y_i = 1\}$ and $X^- := \{X_i : Y_i = -1\}$ respectively.

The Parzen window classification rule in (70) can be interpreted as a *mean classifier* in $\mathcal{H}$: $\langle w, k(.,x)\rangle_{\mathcal{H}}$ represents a hyperplane in $\mathcal{H}$ passing through the origin with w being its normal along the direction that joins the means, μ^+ and μ^- in $\mathcal{H}$. From (25), we can see that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is the RKHS distance between the mean functions, μ^+ and μ^-.

Suppose $\|\mu^+\|_{\mathcal{H}} = \|\mu^-\|_{\mathcal{H}}$, i.e., μ^+ and μ^- are equidistant from the origin in $\mathcal{H}$. Then, the rule in (70) can be equivalently written as

$$y = \text{sign}\left(\|k(.,x)-\mu^-\|^2_{\mathcal{H}} - \|k(.,x)-\mu^+\|^2_{\mathcal{H}}\right). \tag{71}$$

(71) provides another interpretation of the rule in (69), i.e., as a nearest-neighbor rule: assign to x the label associated with the mean μ^+ or μ^-, depending on which mean function is closest to $k(.,x)$ in $\mathcal{H}$.

The classification rule in (69) differs from the "classical" Parzen window classifier in two respects. (i) Usually, the kernel (called the smoothing kernel) in the Parzen window rule is translation invariant in $\mathbb{R}^d$. In our case, M need not be $\mathbb{R}^d$ and k need not be translation invariant. So, the rule in (69) can be seen as a generalization of the classical Parzen window rule. (ii) The kernel in (69) is positive definite unlike in the classical Parzen window rule where k need not have to be so.

Recently, Reid and Williamson *et al.* [33, Section 8, Appendix E] have related MMD to Fisher Discriminant analysis [36, Section 4.3] in $\mathcal{H}$ and SVM [50]. One approachto relate MMD to SVM is along the lines of Theorem 19, where it is easy to see that the margin of an SVM, computed as $\frac{1}{\|f\|_{\mathcal{H}}}$, can be upper bounded by $\frac{\gamma_k(\mathbb{P}_m,\mathbb{Q}_n)}{2}$, which says that the smoothness of an SVM classifier is bounded by the inverse of the MMD between $\mathbb{P}$ and $\mathbb{Q}$.

V. Conclusion & Discussion

In this work, we present integral probability metrics (IPMs) from a more practical perspective and prove several novel properties. We first relate IPMs to ϕ-divergences and show that they are essentially different. More specifically, we prove that total variation distance is the only "non-trivial" ϕ-divergence that is also an IPM. We then study the consistency and convergence rates of the empirical estimators of IPMs and show that the empirical estimators of Wasserstein distance, Dudley metric and maximum mean discrepancy are strongly consistent and have a good convergence behavior. We illustrate how IPMs naturally appear in a binary classification setting, first by relating them to the optimal L-risk of a binary classifier and, second, by relating the Wasserstein distance to the margin of a Lipschitz classifier, the Dudley metric to the margin of a bounded Lipschitz classifier and the maximum mean discrepancy to the Parzen window classifier. With many IPMs being used only as theoretical tools, we believe that this study highlights properties of IPMs that have not been explored before and would improve their practical applicability.

De Groot [51], [52] introduced the concept of *statistical information* that is widely studied in information theory and statistics. [4], [31] have shown that every statistical information is a ϕ-divergence and every ϕ-divergence is a statistical information. Since an IPM is trivially a ϕ-divergence (see Theorems 1 and 2), it can be related to statistical information (see Eq. (77) in [4]).

Another question one can ask involves the relation between IPMs and Bregman divergences. It can be shown that IPMs and Bregman divergences do not intersect because Bregman divergences do not satisfy the triangle inequality, whereas IPMs satisfy the triangle inequality since they are pseudo-metrics on $\mathscr{P}$. However, recently Chen *et al.* [53], [54] have studied "square-root metrics" based on Bregman divergences. One could investigate conditions on $\mathcal{F}$ for which $\gamma_\mathcal{F}$ coincides with such a family.

Similarly, in the case of ϕ-divergences, some functions of D_ϕ are shown to be metrics on $\mathscr{P}_\lambda$ (see Theorem 2 for the notation), for example, the square root of variational distance, the square root of Hellinger's distance, the square root of Jensen-Shannon divergence [55]–[57], etc. Also, Österreicher and Vajda [58, Theorem 1] have shown that certain powers of D_ϕ are metrics on $\mathscr{P}_\lambda$. Therefore, one could investigate conditions on $\mathcal{F}$ for which $\gamma_\mathcal{F}$ equals such functions of D_ϕ.

Appendix
Supplementary Results

In this section, we collect results that are used to prove results in Sections III and IV.

We quote the following result on Lipschitz extensions from [34] (see also [59], [60]).

Lemma 20 (Lipschitz extension): Given a function f defined on a finite subset $x_1,\ldots,x_n$ of M, there exists a function $\widetilde{f}$ which coincides with f on $x_1,\ldots,x_n$, is defined on the whole space M, and has the same Lipschitz constant as f. Additionally, it is possible to explicitly construct $\widetilde{f}$ in the form

$$\widetilde{f}(x) = \alpha \min_{i=1,\ldots,n}(f(x_i)+L(f)\rho(x,x_i)) \\ +(1-\alpha)\max_{i=1,\ldots,n}(f(x_i)-L(f)\rho(x,x_i)), \quad (72)$$

for any $\alpha\in[0,1]$, with $L(f)=\max_{x_i\neq x_j}\frac{f(x_i)-f(x_j)}{\rho(x_i,x_j)}$.

The following result on bounded Lipschitz extensions is quoted from [6, Proposition 11.2.3].

Lemma 21 (Bounded Lipschitz extension): If $A\subset M$ and $f\in BL(A,\rho)$, then f can be extended to a function $h\in BL(M,\rho)$ with $h=f$ on A and $\|h\|_{BL}=\|f\|_{BL}$. Additionally, it is possible to explicitly construct h as

$$h=\max\left(-\|f\|_\infty,\min\left(g,\|f\|_\infty\right)\right), \quad (73)$$

where g is a function on M such that $g=f$ on A and $\|g\|_L=\|f\|_L$.

The following result is quoted from [44, Theorem 3.7].

Theorem 22: Let $F(x)=\sup_{f\in\mathcal{F}}|f(x)|$ be the envelope function for $\mathcal{F}$. Assume that $\int F\,d\mathbb{P}<\infty$, and suppose moreover that for any $\varepsilon>0$, $\frac{1}{m}\mathcal{H}(\varepsilon,\mathcal{F},L_1(\mathbb{P}_m))\xrightarrow{\mathbb{P}}0$. Then $\sup_{f\in\mathcal{F}}(\mathbb{P}_mf-\mathbb{P}f)\xrightarrow{a.s.}0$.

Theorem 23 ([61] McDiarmid's Inequality): Let $X_1,\ldots,X_n,X_1',\ldots,X_n'$ be independent random variables taking values in a set M, and assume that $f:M^n\to\mathbb{R}$ satisfies

$$|f(x_1,\ldots,x_n)-f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|\le c_i, \quad (74)$$

$\forall\,x_1,\ldots,x_n,x_1',\ldots,x_n'\in M$. Then for every $\epsilon>0$,

$$\Pr\left(f(X_1,\ldots,X_n)-\mathbb{E}f(X_1,\ldots,X_n)\ge\epsilon\right)\le e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}. \quad (75)$$

Lemma 24 ([9] Symmetrization): Let $\sigma_1,\ldots,\sigma_N$ be i.i.d. Rademacher random variables. Then,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\mathbb{E}f-\frac{1}{N}\sum_{i=1}^N f(x_i)\right|\le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{i=1}^N\sigma_if(x_i)\right|. \quad (76)$$

The following result is quoted from [62, Theorem 32.1].

Theorem 25: Let f be a convex function, and let C be a convex set contained in the domain of f. If f attains its supremum relative to C at some point of relative interior of C, then f is actually constant throughout C.

Acknowledgment

We wish to thank Robert Williamson and Mark Reid for helpful conversations.

References

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

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

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

[4] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[5] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, pp. 429–443, 1997.

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

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

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

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

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

[11] R. M. Gray, D. L. Neuhoff, and P. C. Shields, "A generalization of Ornstein's $\overline{d}$ distance with applications to information theory," *Annals of Probability*, vol. 3, pp. 315–328, 1975.

[12] V. M. Zolotarev, "Probability metrics," *Theory of Probability and its Applications*, vol. 28, pp. 278–302, 1983.

[13] S. T. Rachev, "On a class of minimum functionals in a space of probability measures," *Theory of Probability and its Applications*, vol. 29, pp. 41–48, 1984.

[14] ——, "The Monge-Kantorovich mass transference problem and its stochastic applications," *Theory of Probability and its Applications*, vol. 29, pp. 647–676, 1985.

[15] E. Levina and P. Bickel, "The earth mover's distance is the Mallows distance: Some insights from statistics," in *Proc. of Intl. Conf. Computer Vision*, 2001, pp. 251–256.

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

[17] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two sample problem," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 513–520.

[18] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. of the 21^{st} Annual Conference on Learning Theory*, R. Servedio and T. Zhang, Eds., 2008, pp. 111–122.

[19] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[20] S. Saitoh, *Theory of Reproducing Kernels and its Applications*. Harlow, UK: Longman, 1988.

[21] 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 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. MIT Press, 2008, pp. 585–592.

[22] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 489–496.

[23] S. T. Rachev and L. Rüschendorf, *Mass transportation problems. Vol. I Theory, Vol. II Applications*, ser. Probability and its Applications. Berlin: Springer-Verlag, 1998.

[24] A. Keziou, "Dual representation of ϕ-divergences and applications," *Comptes Rendus Mathematique*, vol. 336, pp. 857–862, 2003.

[25] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *Annals of Statistics*, vol. 37, no. 2, pp. 876–904, 2009.

[26] M. Broniatowski and A. Keziou, "Parametric estimation and tests through divergences and the duality technique," *Journal of Multivariate Analysis*, vol. 100, pp. 16–36, 2009.

[27] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Information Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[28] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," Department of Statistics, University of California, Berkeley, Tech. Rep. 764, 2008.

[29] A. A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Information Theory*, vol. 49, no. 6, pp. 1491–1498, 2003.

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

[31] F. Österreicher and I. Vajda, "Statistical information and discrimination," *IEEE Trans. on Information Theory*, vol. 39, pp. 1036–1039, 1993.

[32] A. Buja, W. Stuetzle, and Y. Shen, "Loss functions for binary class probability estimation and classification: Structure and applications," University of Pennsylvania, Tech. Rep., November 2005.

[33] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *http://arxiv.org/abs/0901.0356v1*, January 2009.

[34] U. von Luxburg and O. Bousquet, "Distance-based classification with Lipschitz functions," *Journal for Machine Learning Research*, vol. 5, pp. 669–695, 2004.

[35] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. UK: Cambridge University Press, 2004.

[36] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[37] M. Khosravifard, D. Fooladivanda, and T. A. Gulliver, "Confliction of the convexity and metric properties in f-divergences," *IEICE Trans. Fundamentals*, vol. E90-A, no. 9, pp. 1848–1853, 2007.

[38] S. S. Vallander, "Calculation of the Wasserstein distance between probability distributions on the line," *Theory Probab. Appl.*, vol. 18, pp. 784–786, 1973.

[39] Q. Wang, S. R. Kulkarni, and S. Verdú, "A nearest-neighbor approach to estimating divergence between continuous random vectors," in *IEEE Symposium on Information Theory*, 2006.

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

[41] L. Devroye and L. Györfi, "No empirical probability measure can converge in the total variation sense for all distributions," *Annals of Statistics*, vol. 18, no. 3, pp. 1496–1499, 1990.

[42] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in functional space," *American Mathematical Society Translations*, vol. 2, no. 17, pp. 277–364, 1961.

[43] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, UK: Cambridge University Press, 2007.

[44] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge, UK: Cambridge University Press, 2000.

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

[46] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Information Theory*, vol. 38, no. 5, pp. 1437–1454, 1992.

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

[48] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics." *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.

[49] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Advances in Computational Mathematics*, vol. 13, no. 1, pp. 1–50, 2000.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[51] M. H. D. Groot, "Uncertainty, information and sequential experiments," *Annals of Mathematical Statistics*, vol. 33, pp. 404–419, 1962.

[52] ——, *Optimal Statistical Decisions*. New York: McGraw Hill, 1970.

[53] P. Chen, Y. Chen, and M. Rao, "Metrics defined by Bregman divergences," *Communications in Mathematical Sciences*, vol. 6, no. 4, pp. 915–926, 2008.

[54] ——, "Metrics defined by Bregman divergences: Part 2," *Communications in Mathematical Sciences*, vol. 6, no. 4, pp. 927–948, 2008.

[55] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information Theory*, vol. 49, pp. 1858–1860, 2003.

[56] B. Fuglede and F. Topsøe, "Jensen-Shannon divergence and Hilbert space embedding," 2003, preprint.

[57] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures," in *AISTATS*, 2005.

[58] F. Österreicher and I. Vajda, "A new class of metric divergences on probability spaces and its applicability in statistics," *Ann. Inst. Statist. Math.*, vol. 55, pp. 639–653, 2003.

[59] E. J. McShane, "Extension of range of functions," *Bulletin of the American Mathematical Society*, vol. 40, pp. 837–842, 1934.

[60] H. Whitney, "Analytic extensions of differentiable functions defined in closed sets," *Transactions of the American Mathematical Society*, vol. 36, pp. 63–89, 1934.

[61] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, pp. 148–188, 1989.

[62] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton University Press, 1970.