Non-parametric Estimation of Integral Probability Metrics

Bharath K. Sriperumbudur[1], Kenji Fukumizu[2], Arthur Gretton[3,4], Bernhard Schölkopf[4] and Gert R. G. Lanckriet[1]

[1]Department of ECE, UC San Diego, USA.
[2]The Institute of Statistical Mathematics, Tokyo, Japan.
[3]Machine Learning Department, Carnegie Mellon University, USA.
[4]MPI for Biological Cybernetics, Tübingen, Germany.
bharathsv@ucsd.edu, fukumizu@ism.ac.jp, arthur.gretton@gmail.com,
bernhard.schoelkopf@tuebingen.mpg.de, gert@ece.ucsd.edu

***Abstract*—In this paper, we develop and analyze a non-parametric method for estimating the class of integral probability metrics (IPMs), examples of which include the Wasserstein distance, Dudley metric, and maximum mean discrepancy (MMD). We show that these distances can be estimated efficiently by solving a linear program in the case of Wasserstein distance and Dudley metric, while MMD is computable in a closed form. All these estimators are shown to be *strongly consistent* and their convergence rates are analyzed. Based on these results, we show that IPMs are simple to estimate and the estimators exhibit good convergence behavior compared to ϕ-divergence estimators.**

I. Introduction

Given samples from two probability measures, $\mathbb{P}$ and $\mathbb{Q}$, it is often of interest (especially in inference problems in statistics) to estimate the distance/divergence between unknown $\mathbb{P}$ and $\mathbb{Q}$. The commonly and popularly used distance/divergence measure between probabilities is the *Ali-Silvey distance* [1], also called the *Csiszár's ϕ-divergence* [2], which is defined as

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

where M is a measurable space and $\phi : [0,\infty) \to (-\infty,\infty]$ is a convex function. $\mathbb{P} \ll \mathbb{Q}$ denotes that $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$. Well-known distance/divergence measures obtained by appropriately choosing ϕ include the Kullback-Liebler (KL) divergence ($\phi(t) = t\log t$), Hellinger distance ($\phi(t) = (\sqrt{t}-1)^2$), and total variation distance ($\phi(t) = |t-1|$). The non-parametric estimation of ϕ-divergence, especially the KL-divergence has recently been studied in depth [3]–[5].

The goal of this paper is to study the non-parametric estimation of another popular family (particularly in probability theory and mathematical statistics) of distance measures on probabilities, the *integral probability metrics* (IPM) [6], defined as

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

where $\mathcal{F}$ in (1) is a class of real-valued bounded measurable functions on M. Mostly, IPMs have been studied as tools of theoretical interest in probability theory [7, Chapter 11], with applications in mass transportation problems [8], empirical process theory [9], etc. By appropriately choosing $\mathcal{F}$, various popular distance measures can be obtained:

(a) *Dudley metric:* Choose $\mathcal{F} = \{f : \|f\|_{BL} \le 1\}$ in (1), where $\|f\|_{BL} := \|f\|_\infty + \|f\|_L$, $\|f\|_\infty := \sup\{|f(x)| : x \in M\}$ and $\|f\|_L := \sup\{|f(x) - f(y)|/\rho(x,y) : x \neq y \text{ in } M\}$. $\|f\|_L$ is called the Lipschitz semi-norm of a real-valued function f on a metric space, (M,ρ). The Dudley metric is popularly used in proving the convergence of probability measures with respect to the weak topology [7, Chapter 11].

(b) *Kantorovich metric and Wasserstein distance:* Choosing $\mathcal{F} = \{f : \|f\|_L \le 1\}$ in (1) yields the *Kantorovich metric*. The famous Kantorovich-Rubinstein theorem [7, Theorem 11.8.2] shows that when M is separable, the Kantorovich metric is the dual representation of *Wasserstein distance* [7, p. 420]. Due to this duality, in this paper, we refer to the Kantorovich metric as the Wasserstein distance. The Wasserstein distance has found application in information theory [10], mathematical statistics [11], and mass transportation problems [8].

(c) *Total variation metric and Kolmogorov distance:* $\gamma_{\mathcal{F}}$ is the *total variation metric* when $\mathcal{F} = \{f : \|f\|_\infty \le 1\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$. The Kolmogorov distance is popularly used in proving the classical central limit theorem in $\mathbb{R}^d$, and also appears as the Kolmogorov-Smirnov statistic in hypothesis testing [12].

(d) *Maximum mean discrepancy:* $\gamma_{\mathcal{F}}$ is called the *maximum mean discrepancy (MMD)* [13] when $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$. Here, $\mathcal{H}$ represents a reproducing kernel Hilbert space (RKHS) [14] with k as its reproducing kernel (r.k.). MMD is used in statistical applications including homogeneity testing [13], independence testing [15], and testing for conditional independence [16].

Having briefly mentioned different IPMs and their applications, we now consider the problem of non-parametrically estimating $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$, where $\mathbb{P}$ and $\mathbb{Q}$ are known only through random samples drawn from them. Our focus is non-parametric estimation as we do not want to impose any strong assumptions on $\mathbb{P}$ and $\mathbb{Q}$. The key properties that any estimator should satisfy are (a) *consistency* (resp. *strong consistency*), i.e., suppose $\{\theta_l\}$ is a sequence of estimators of θ, then θ_l is consistent (*resp.* strongly consistent) if θ_l converges in probability (*resp.* a.s.) to θ as $l \to \infty$, (b) *fast* rate of convergence and (c) a simple implementation.

Before presenting our results on the estimation of IPMs, we will briefly discuss prior work on the estimation of ϕ-divergences, which we hope will help the reader to appreciate the advantages involved in the estimation of IPMs. As mentioned earlier, the non-parametric estimation of ϕ-divergences, especially the KL-divergence, is well studied (see [3]–[5] and references therein). Wang *et al.* [3] propose a simple histogram-based KL estimator, using a data-dependent space partitioning scheme, and show 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. In addition, for increasing dimensionality of the data (in $\mathbb{R}^d$), the method is inefficient both in statistical and computational terms. Nguyen *et al.* [5] provide a consistent estimator of the KL-divergence by solving a convex program (specifically, a quadratic program [17, Chapter 4]). Although this approach is efficient and the dimensionality of data is not an issue, the rate of convergence of this estimator can also be arbitrarily slow, depending on the distributions. One should therefore bear in mind the difficulty in empirically estimating ϕ-divergences, as compared with our estimates of integral probability metrics.

In Section II, we consider the non-parametric estimation of IPMs, in particular the Wasserstein distance (W), Dudley metric (β) and MMD (γ_k). The estimates of W and β are obtained by solving linear programs,while an estimator of γ_k is computed in closed form, which means that these distances are computationally simpler to estimate than the KL-divergence (the KL-divergence estimator due to [5] solves a quadratic program). In addition, an increase in the dimensionality of data has only a mild effect on the complexity of estimating these metrics, unlike in the case of KL-divergence, where space partitioning schemes [3] become increasingly difficult to implement as the number of dimensions grows. Next, in Section III, we show that these estimators are strongly consistent, and provide their rates of convergence, using concentration inequalities and tools from empirical process theory [9]. Based on these results, it will be clear that all these estimators 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 [3], [5]. Our experimental results in [18] confirm the convergence theory discussed in Section III and therefore demonstrate the practical viability of these estimators.

Since the total variation distance is also an IPM, in Section IV, we briefly discuss its empirical estimation and show that the empirical estimator is not strongly consistent. Because of this, we provide new lower bounds for the total variation distance in terms of W, β and γ_k, which can be consistently estimated. These bounds also translate as lower bounds on the KL-divergence through Pinsker's inequality [19].

Due to space limitations, complete proofs of the results are not provided. We refer the reader to [18] for complete proofs and additional results, including experimental results.

II. Non-parametric Estimation of Wasserstein Distance, Dudley Metric and MMD

In this section, we show that the Wasserstein and Dudley metrics can be estimated by solving linear programs (see Theorems 1 and 2) whereas an estimator for MMD can be obtained in closed form (Theorem 3; proved in [13]).

Given $\{X^{(1)}_1, X^{(1)}_2, \dots, X^{(1)}_m\}$ and $\{X^{(2)}_1, X^{(2)}_2, \dots, X^{(2)}_n\}$, which are i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$ respectively, we propose to estimate $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ by the following estimator,

$$\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n) = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{N} \widetilde{Y}_i f(X_i) \right|, \tag{2}$$

where $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X^{(1)}_i}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X^{(2)}_i}$ represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$ respectively, $N = m+n$, $\widetilde{Y}_i = \frac{1}{m}$ when $X_i = X^{(1)}_i$ for $i = 1,\dots,m$ and $\widetilde{Y}_{m+i} = -\frac{1}{n}$ when $X_{m+i} = X^{(2)}_i$ for $i = 1,\dots,n$. Here, δ_x represents the Dirac measure at x. The computation of $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ in (2) is not straightforward for any arbitrary $\mathcal{F}$. To obtain meaningful results, in the following, we restrict ourselves to $\mathcal{F}_W := \{f : \|f\|_L \le 1\}$, $\mathcal{F}_\beta := \{f : \|f\|_{BL} \le 1\}$ and $\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \le 1\}$ and compute (2). Let us denote $W := \gamma_{\mathcal{F}_W}$, $\beta := \gamma_{\mathcal{F}_\beta}$ and $\gamma_k := \gamma_{\mathcal{F}_k}$.

Theorem 1 (Estimator of Wasserstein distance): For all $\alpha \in [0,1]$, the following function solves (2) for $\mathcal{F} = \mathcal{F}_W$:

$$\begin{aligned} f_\alpha(x) \quad := \quad & \alpha \min_{i=1,\dots,N} (a^\star_i + \rho(x,X_i)) \\ & +(1-\alpha) \max_{i=1,\dots,N} (a^\star_i - \rho(x,X_i)), \end{aligned}$$

where

$$W(\mathbb{P}_m,\mathbb{Q}_n) = \sum_{i=1}^{N} \widetilde{Y}_i a^\star_i, \tag{3}$$

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

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

Theorem 2 (Estimator of Dudley metric): For all $\alpha \in [0,1]$, the following function solves (2) for $\mathcal{F} = \mathcal{F}_\beta$:

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

where

$$\begin{aligned} h_\alpha(x) := \alpha & \min_{i=1,\dots,N} (a^\star_i + L^\star \rho(x,X_i)) \\ & +(1-\alpha) \max_{i=1,\dots,N} (a^\star_i - L^\star \rho(x,X_i)), \end{aligned}$$

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

$$\beta(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^{N} \widetilde{Y}_i a_i^\star, \tag{4}$$

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

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

Theorem 3 (Estimator of MMD [13]): For $\mathcal{F} = \mathcal{F}_k$, the following function is the unique solution to (2):

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

and

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \sqrt{\sum_{i,j=1}^{N} \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)}. \tag{5}$$

The following observations can be made about the estimators in Theorems 1–3.

(a) Since W and β are estimated by solving a linear program and γ_k is obtained in closed form, it is easy to see that these estimators have a computational advantage over KL-divergence estimators in [5].

(b) Note that the estimators in (3) and (4) depend on $\{X_i\}_{i=1}^N$ only through ρ, while the one in (5) depends on $\{X_i\}_{i=1}^N$ only through k. This means, once $\{\rho(X_i, X_j)\}_{i,j=1}^N$ or $\{k(X_i, X_j)\}_{i,j=1}^N$ is known, the complexity of the corresponding estimators is independent of d (when $M = \mathbb{R}^d$), unlike in the estimation of KL-divergence [3].

(c) Because these estimators depend on M only through ρ or k, the domain M is immaterial as long as ρ or k is defined on M. Therefore, these estimators extend to arbitrary domains, unlike the KL-divergence, where the domain is usually chosen to be $\mathbb{R}^d$ [3].

(d) Unlike with the KL-divergence, the estimators of W, β and γ_k account for the properties of the underlying space M. This is useful when $\mathbb{P}$ and $\mathbb{Q}$ have disjoint support. When $\mathbb{P}$ and $\mathbb{Q}$ have disjoint support, $D_\phi(\mathbb{P}, \mathbb{Q}) = +\infty$ irrespective of M, while W, β and γ_k vary with the properties of M. Therefore, in such cases, these IPMs provides a better notion of distance between $\mathbb{P}$ and $\mathbb{Q}$, compared to ϕ-divergences.

III. Consistency and rate of convergence

In Section II, we presented 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 so that in practice, fewer samples are sufficient to obtain reliable estimates. We address these issues in this section.

Before we present our results, we briefly introduce some terminology and notation from empirical process theory. For any $r \geq 1$ 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 strong consistency of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$, using which we prove the consistency of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ in Corollary 5.

Theorem 4: 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 $m, n \to \infty$.

The following corollary to Theorem 4 shows that $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ are strongly consistent.

Corollary 5 (Consistency of W and β): Let (M, ρ) be a totally bounded metric space. Then, as $m, n \to \infty$,

(i) $|W(\mathbb{P}_m, \mathbb{Q}_n) - W(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$.

(ii) $|\beta(\mathbb{P}_m, \mathbb{Q}_n) - \beta(\mathbb{P}, \mathbb{Q})| \xrightarrow{a.s.} 0$.

Proof: The proof idea is to check the conditions *(i)–(iv)* in Theorem 4. Since M is a totally bounded metric space, it can be shown that $\forall\, x \in M$, $F(x) < \infty$ for $\mathcal{F} = \mathcal{F}_W$ and $\mathcal{F} = \mathcal{F}_\beta$, which therefore satisfies *(i)* and *(ii)* in Theorem 4. It can be shown that $\mathcal{H}(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_m))$ and $\mathcal{H}(\varepsilon, \mathcal{F}, L_1(\mathbb{Q}_n))$ are independent of m and n for $\mathcal{F} = \mathcal{F}_W$ and $\mathcal{F} = \mathcal{F}_\beta$, therefore satisfying *(iii)* and *(iv)* in Theorem 4. For details, refer to [18, Corollary 9]. ■

Similar to Corollary 5, a strong consistency result for γ_k can be provided by estimating the entropy number of $\mathcal{F}_k$. However, in the following, we adopt a different and simpler approach. To this end, we first provide a general result on the rate of convergence of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$, expanding on the proof strategy in [20, Appendix A.2]. As a special case, obtain the rates of convergence of the estimators of W, β and γ_k. Using this result, we recover the strong consistency of γ_k obtained in [13, Theorem 4]. We start with the following definition.

Definition 6 (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

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

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

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| \le \sqrt{18\nu^2\log\frac{4}{\delta}}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)$$
$$+2R_m(\mathcal{F};\{X_i^{(1)}\})+2R_n(\mathcal{F};\{X_i^{(2)}\}). \quad (6)$$

Theorem 7 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 (6), $|\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)-\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_m\vee m^{-1/2}+r_n\vee n^{-1/2})$, where $a\vee b:=\max(a,b)$. The following corollary to Theorem 7 provides the rate of convergence for W, β and γ_k.

Corollary 8 (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

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

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

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

(ii) [13, Theorem 4]: 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: The proof involves the estimation of $R_m(\mathcal{F};\{X_i^{(1)}\})$ and $R_n(\mathcal{F};\{X_i^{(2)}\})$ for $\mathcal{F}=\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$, which is then used in (6). For details, see [18, Corollary 12]. ■

Several observations follow Corollary 8:

(a) 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, $\|\cdot\|_s$, $1\le s\le\infty$.

(b) 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 $(\mathbb{R}^d,\|\cdot\|_s)$.

(c) In the case of MMD, we have not made any assumptions on M except that it be 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 numerous kernels, including 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 [21] for more examples.

The results derived in this section show that the estimators of the Wasserstein distance, Dudley metric and MMD exhibit good convergence behavior, irrespective of the distributions, unlike estimators of the ϕ-divergence [3], [5].

IV. Non-parametric estimation of total variation distance

So far, the results in Sections II and III show that the estimators of IPMs (specifically, W, β and γ_k) exhibit nice properties compared to those of ϕ-divergences. As shown in Section I, since the total variation distance,

$$TV(\mathbb{P},\mathbb{Q}) := \sup\left\{\int_M f\,d(\mathbb{P}-\mathbb{Q}) : \|f\|_\infty \le 1\right\},$$

is both an IPM and a ϕ-divergence, we consider here its empirical estimation and investigate consistency. Suppose M is a metric space. Let $TV(\mathbb{P}_m,\mathbb{Q}_n)$ be an empirical estimator of $TV(\mathbb{P},\mathbb{Q})$, which can be shown to be

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

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

$$\begin{array}{cl} \max\limits_{a_1,\ldots,a_N} & \sum\limits_{i=1}^N \widetilde{Y}_i a_i \\ \text{s.t.} & -1\le a_i\le 1,\ \forall\, i. \end{array}$$

The question is whether this estimator is consistent. First note that $a_i^\star = \text{sign}(\widetilde{Y}_i)$ and therefore, $TV(\mathbb{P}_m,\mathbb{Q}_n)=2$ for any m,n. This means for any $\mathbb{P},\mathbb{Q}$ such that $TV(\mathbb{P},\mathbb{Q})<2$, $TV(\mathbb{P}_m,\mathbb{Q}_n)$ is not a consistent estimator of $TV(\mathbb{P},\mathbb{Q})$. Indeed $a_i^\star,\ \forall\, i$ are indepedent of the actual samples, $\{X_i\}_{i=1}^N$ drawn from $\mathbb{P}$ and $\mathbb{Q}$, unlike in the estimation of the Wasserstein and Dudley metrics, and therefore it is not surprising that $TV(\mathbb{P}_m,\mathbb{Q}_n)$ is not a consistent estimator of $TV(\mathbb{P},\mathbb{Q})$.

Suppose $M=\mathbb{R}^d$ and let $\mathbb{P}$, $\mathbb{Q}$ be absolutely continuous w.r.t. the Lebesgue measure. Then $TV(\mathbb{P},\mathbb{Q})$ can be consistently estimated in a strong sense using the total variation distance between the kernel density estimators of $\mathbb{P}$ and $\mathbb{Q}$. This is because if $\widetilde{\mathbb{P}}_m$ and $\widetilde{\mathbb{Q}}_n$ represent the kernel density estimators associated with $\mathbb{P}$ and $\mathbb{Q}$ respectively, then $|TV(\widetilde{\mathbb{P}}_m,\widetilde{\mathbb{Q}}_n)-TV(\mathbb{P},\mathbb{Q})| \le TV(\widetilde{\mathbb{P}}_m,\mathbb{P})+TV(\widetilde{\mathbb{Q}}_n,Q) \xrightarrow{a.s.} 0$ as $m,n\to\infty$ (see [22, Chapter 6] and references therein).

The issue in the estimation of $TV(\mathbb{P},\mathbb{Q})$ is that the set $\mathcal{F}_{TV} := \{f : \|f\|_\infty\le 1\}$ 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 $\mathcal{F}$ of $\mathcal{F}_{TV}$ such that $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})\le TV(\mathbb{P},\mathbb{Q}),\ \forall\,\mathbb{P},\mathbb{Q}$ and $\gamma_{\mathcal{F}}(\mathbb{P}_m,\mathbb{Q}_n)$ is a consistent estimator of $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$. Possible choices for $\mathcal{F}$ include $\mathcal{F}_\beta$ 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 II and III. 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 Glivenko-Cantelli theorem [23, 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 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 9 (Lower bounds on TV): (i) For all $\mathbb{P} \neq \mathbb{Q}$, we have

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{W(\mathbb{P}, \mathbb{Q})\beta(\mathbb{P}, \mathbb{Q})}{W(\mathbb{P}, \mathbb{Q}) - \beta(\mathbb{P}, \mathbb{Q})}. \tag{7}$$

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

$$TV(\mathbb{P}, \mathbb{Q}) \geq \frac{\gamma_k(\mathbb{P}, \mathbb{Q})}{\sqrt{C}}. \tag{8}$$

Based on the above result, the following observations can be made:

(a) A simple lower bound on TV can be obtained as $TV(\mathbb{P}, \mathbb{Q}) \geq \beta(\mathbb{P}, \mathbb{Q}), \forall\, \mathbb{P}, \mathbb{Q}$. It is easy to see that the bound in (7) is tighter as $\frac{W(\mathbb{P},\mathbb{Q})\beta(\mathbb{P},\mathbb{Q})}{W(\mathbb{P},\mathbb{Q})-\beta(\mathbb{P},\mathbb{Q})} \geq \beta(\mathbb{P}, \mathbb{Q})$ with equality if and only if $\mathbb{P} = \mathbb{Q}$.

(b) The bounds in (7) and (8) translate as lower bounds on the KL-divergence through Pinsker's inequality: $TV^2(\mathbb{P}, \mathbb{Q}) \leq 2\, KL(\mathbb{P}, \mathbb{Q}), \forall\, \mathbb{P}, \mathbb{Q}$. See Fedotov *et al.* [19] and references therein for more refined bounds relating TV and KL. Therefore, using these bounds, one can obtain a consistent estimate of a lower bound on TV and KL.

V. Conclusion & Discussion

In this work, we have studied the non-parametric estimation of integral probability metrics and showed that the empirical estimators of the Wasserstein distance (W), Dudley metric (β) and maximum mean discrepancy (γ_k) are simple to compute, strongly consistent and have a good convergence behavior, compared to those of ϕ-divergences. In addition, we provided two lower bounds on the total variation distance in terms of these IPMs, which then translate to lower bounds on KL-divergence through Pinsker's inequality. Our experimental results [18] demonstrate the practical viability of these estimators, which are not reported here due to space limitations.

One interesting problem yet to be explored in connection with this work is: What is the minimax rate for estimating W, β and γ_k, and do the proposed estimators achieve this rate?

Acknowledgments

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

References

[1] 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.

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

[3] 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.

[4] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric estimation of the likelihood ratio and divergence functionals," in *IEEE International Symposium on Information Theory*, 2007.

[5] ——, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," Department of Statistics, University of California, Berkeley, Tech. Rep. 764, 2008.

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

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

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

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

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

[11] 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.

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

[13] 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.

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

[15] 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.

[16] 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.

[17] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

[19] 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.

[20] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two sample problem," MPI for Biological Cybernetics, Tech. Rep. 157, 2008.

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

[22] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. New York: Wiley, 1985.

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