Electronic Journal of Statistics
Vol. 6 (2012) 1550–1599
ISSN: 1935-7524
DOI: 10.1214/12-EJS722

On the empirical estimation of integral probability metrics

Bharath K. Sriperumbudur

Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London, WC1N 3AR, U.K.
e-mail: bharath@gatsby.ucl.ac.uk

Kenji Fukumizu

The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: fukumizu@ism.ac.jp

Arthur Gretton

Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London, WC1N 3AR, U.K.
e-mail: arthur.gretton@gmail.com
Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany

Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany
e-mail: bs@tuebingen.mpg.de

and

Gert R. G. Lanckriet

Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0407, U.S.A.
e-mail: gert@ece.ucsd.edu

Abstract: Given two probability measures, $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable space, S, the integral probability metric (IPM) is defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup\left\{\left|\int_S f\,d\mathbb{P} - \int_S f\,d\mathbb{Q}\right| : f \in \mathcal{F}\right\},$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on S. By appropriately choosing $\mathcal{F}$, various popular distances between $\mathbb{P}$ and $\mathbb{Q}$, including the Kantorovich metric, Fortet-Mourier metric, dual-bounded Lipschitz distance (also called the Dudley metric), total variation distance, and *kernel distance*, can be obtained.

In this paper, we consider the problem of estimating $\gamma_{\mathcal{F}}$ from finite random samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$. Although the above mentioned distances cannot be computed in closed form for every $\mathbb{P}$ and $\mathbb{Q}$, we show their empirical estimators to be easily computable, and strongly consistent (except for the total-variation distance). We further analyze their rates of convergence. Based on these results, we discuss the advantages of certain choices of $\mathcal{F}$ (and therefore the corresponding IPMs) over others—in particular, the *kernel distance* is shown to have three favorable properties compared with the other mentioned distances: it is computationally cheaper,

the empirical estimate converges at a faster rate to the population value, and the rate of convergence is independent of the dimension d of the space (for $S = \mathbb{R}^d$). We also provide a novel interpretation of IPMs and their empirical estimators by relating them to the problem of binary classification: while the IPM between class-conditional distributions is the negative of the optimal risk associated with a binary classifier, the smoothness of an appropriate binary classifier (e.g., support vector machine, Lipschitz classifier, etc.) is inversely related to the empirical estimator of the IPM between these class-conditional distributions.

AMS 2000 subject classifications: Primary 62G05.
Keywords and phrases: Integral probability metrics, empirical estimation, Kantorovich metric, dual-bounded Lipschitz distance (Dudley metric), kernel distance, reproducing kernel Hilbert space, Rademacher average, Lipschitz classifier, support vector machine.

Received July 2011.

Contents

1. Introduction

Given samples from two unknown probability measures, $\mathbb{P}$ and $\mathbb{Q}$, it is often of interest (in applications such as two-sample and independence testing) to

estimate the distance (or divergence) between them. The goal of this paper is to study the empirical estimation of a popular family of distance measures on probabilities, the *integral probability metrics* (IPM) [29]—also called *probability metrics with a ζ-structure* [55]—defined as

$$\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int_S f\, d\mathbb{P} - \int_S f\, d\mathbb{Q} \right|, \tag{1.1}$$

where $\mathcal{F}$ in (1.1) is a class of real-valued bounded measurable functions on S (the choice of functions being the crucial distinction between different IPMs). IPMs have been employed as tools of theoretical interest in probability theory [15, Chapter 11], with applications including mass transportation problems [34] and empirical process theory [50]. In statistics, IPMs have been used in nonparametric two-sample testing, including the Kolmogorov-Smirnov test [7, 39, 40] and the kernel test [20, 21]; as well as in independence testing [20, Section 7.4], [22]. By appropriately choosing $\mathcal{F}$ in (1.1), various popular distance measures can be obtained:

(a) *Kantorovich metric, Wasserstein distance and Fortet-Mourier metric:* Setting $\mathcal{F} = \{f : \|f\|_L \le 1\}$ in (1.1) yields the *Kantorovich metric*, where $\|f\|_L$ is the Lipschitz semi-norm of a bounded continuous real-valued function f on a metric space, (S,ρ),

$$\|f\|_L := \sup\left\{ \frac{|f(x)-f(y)|}{\rho(x,y)} \,:\, x\neq y \text{ in } S \right\}.$$

The famous Kantorovich-Rubinstein theorem [15, Theorem 11.8.2] shows that when S is separable, the Kantorovich metric is the dual representation of the *Wasserstein distance* [15, p. 420]—more specifically, the *L_1-Wasserstein distance*—defined as

$$W_1(\mathbb{P},\mathbb{Q}) := \inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})} \int \rho(x,y)\, d\mu(x,y), \tag{1.2}$$

where $\mathbb{P},\mathbb{Q} \in \{\mathbb{P} : \int \rho(x,y)\, d\mathbb{P}(x) < \infty, \forall\, y \in S\}$ and $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $S\times S$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. The L_1-Wasserstein distance (and therefore the Kantorovich metric) has found applications in information theory [19], mathematical statistics [33, 55] and mass transportation problems [34].

The Fortet-Mourier metric [35, p. 17] is a generalization of the Kantorovich metric, with $\mathcal{F} := \{\|f\|_c \le 1\}$, where

$$\|f\|_c := \sup\left\{ \frac{|f(x)-f(y)|}{c(x,y)} \,:\, x\neq y \text{ in } S \right\},$$

and $c(x,y) = \rho(x,y)\max(1,\rho(x,a)^{p-1},\rho(y,a)^{p-1})$, $p\ge 1$ for some $a\in S$. Note that when $p=1$, the Fortet-Mourier metric is the same as the Kantorovich metric.

(b) *Dudley metric:* Choosing $\mathcal{F} = \{f : \|f\|_{BL} \leq 1\}$ in (1.1) yields the *dual-bounded Lipschitz distance*—also called the *Dudley metric*—where

$$\|f\|_{BL} := \|f\|_\infty + \|f\|_L,$$

with $\|f\|_\infty := \sup\{|f(x)| : x \in S\}$. The Dudley metric is used in proving the convergence of probability measures with respect to the weak topology [15, Chapter 11].

(c) *Total variation metric and Kolmogorov distance:* $\gamma_\mathcal{F}$ is the *total variation metric* when $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ while it is the *Kolmogorov distance* when $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}^d\}$. The Kolmogorov distance is used in proving the classical central limit theorem in $\mathbb{R}^d$, and also appears as the Kolmogorov-Smirnov statistic in hypothesis testing [39].

(d) *Kernel distance:* $\gamma_\mathcal{F}$ is called the *kernel distance* or *maximum mean discrepancy* [6, 20, 21, 46] when $\mathcal{F} = \{f : \|f\|_\mathcal{H} \leq 1\}$, where $\mathcal{H}$ represents a reproducing kernel Hilbert space (RKHS) [2] with k as its reproducing kernel (r.k.) — we write the space $(\mathcal{H}, k)$.[1] The kernel distance is used in statistical applications including homogeneity testing [20, 21], independence testing [22], testing for conditional independence [17] and mixture density estimation [42].

As described above, an important application of distance estimates between $\mathbb{P}$ and $\mathbb{Q}$ (based on random i.i.d. samples drawn from them) is in homogeneity testing and independence testing, where the estimate of the distance can be used as a test statistic (additional applications include classification of probability measures using empirically computed distances). While the kernel distance and the total variation distance have been successfully applied in testing, most other IPMs have not, due to the absence of good estimates for continuous random variables, especially in the multivariate case. Indeed, for testing, it is crucial that the statistic have a consistent estimator exhibiting fast convergence behavior and low computational complexity (i.e., the estimator must be easy to compute). In Section 2.1, we provide empirical estimates of the above mentioned IPMs, in particular the Kantorovich metric (and therefore the L_1-Wasserstein distance), Dudley metric, and kernel distance, based on finite samples drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$. The empirical estimators of the Kantorovich distance and Dudley metric are obtained by solving linear programs, while that of the kernel distance is computed in closed form, thereby demonstrating that the kernel distance is simpler to compute than the remaining IPMs.

We show in Section 2.2 that the empirical estimators derived in Section 2.1 exhibit a nice connection to the problem of binary classification. In Section 2.2, we first show that $\gamma_\mathcal{F}(\mathbb{P}, \mathbb{Q})$ (*resp.* its empirical estimator) is the negative of the optimal risk associated with a binary classifier that separates the class conditional distributions, $\mathbb{P}$ and $\mathbb{Q}$ (*resp.* $\mathbb{P}_m$ and $\mathbb{Q}_n$—see the last paragraph of

[1] A function $k : S \times S \to \mathbb{R},\ (x, y) \mapsto k(x, y)$ is a *reproducing kernel* of the Hilbert space $\mathcal{H}$ if and only if the following hold: *(i)* $\forall\, y \in S,\ k(\cdot, y) \in \mathcal{H}$ and *(ii)* $\forall\, y \in S,\ \forall\, f \in \mathcal{H},\ \langle f, k(\cdot, y)\rangle_\mathcal{H} = f(y)$. When a reproducing kernel exist, $\mathcal{H}$ is called a reproducing kernel Hilbert space, and is defined as the completion of the span of k, $\mathcal{H} = \overline{\text{span}\{k(\cdot, y)|y \in S\}}$. It can be shown that a real-valued k is a reproducing kernel if and only if it is symmetric and positive definite [6].

this section for the notation), where the classification rule is restricted to $\mathcal{F}$. In other words, the Kantorovich metric, Dudley metric and the kernel distance (and their empirical estimators) can be understood as the negative of the optimal risk associated with a classifier for which the classification rule is restricted to $\{f : \|f\|_L \leq 1\}$, $\{f : \|f\|_{BL} \leq 1\}$ and $\{f : \|f\|_{\mathcal{H}} \leq 1\}$, respectively. We then show that the empirical estimators of the Kantorovich metric, Dudley metric and kernel distance are related to the *margins* of the Lipschitz classifier [51], bounded Lipschitz classifier, and support vector machine, respectively. The significance of this result is that the smoothness of the classifier is inversely related to the empirical estimator of the IPM between class conditionals $\mathbb{P}$ and $\mathbb{Q}$. Although this is intuitively clear, our result provides a theoretical justification. Finally, we also establish the relation between the kernel distance and the Parzen window classifier [37, 38] (see *kernel classification rule* [14, Chapter 10]).

Next, in Section 3, we show that the empirical estimators derived in Section 2.1 are strongly consistent, and provide their rates of convergence using standard techniques from empirical process theory. Based on these results, it will be clear that the empirical estimator of the kernel distance exhibits a fast rate of convergence compared with that of other IPMs, and its rate of convergence is independent of the dimension d (for $S = \mathbb{R}^d$), unlike with other IPMs. Our experimental results in Section 4 confirm the convergence theory discussed in Section 3 and therefore demonstrate the practical viability of these estimators. Based on these convergence results, in Section 3, we also show how a homogeneity test can be constructed using the empirical estimator of IPM as a test statistic (see Remark 3.6(v)).

Since the total variation distance is also an IPM, we discuss its empirical estimator in Section 5, and show that it is not strongly consistent. Because of this, we provide new lower bounds for the total variation distance in terms of the Kantorovich metric, Dudley metric and the kernel distance, which can be consistently estimated. These bounds also translate as lower bounds on the Kullback-Leibler divergence through Pinsker's inequality [16].

We note that there exist other distance/divergence measures between probabilities besides those of the IPM family. One popular family of divergence measures are the *Ali-Silvey distances* [1], also called the *Csiszár's ϕ-divergences* [11], defined as

$$D_\phi(\mathbb{P}, \mathbb{Q}) := \int_S \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \text{ if } \mathbb{P} \ll \mathbb{Q},$$

where S 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-Leibler (KL) divergence ($\phi(t) = t \log t$), Hellinger distance ($\phi(t) = (\sqrt{t} - 1)^2$), χ^2-divergence ($\phi(t) = (t-1)^2$) and total variation distance ($\phi(t) = |t-1|$). The empirical estimation of ϕ-divergences, especially the KL-divergence, has recently been studied in depth [30, 32, 52]. We emphasize that ϕ-divergences and IPMs are fundamentally different, and intersect only at the total variation distance (our proof of this result is in Appendix A).

TABLE 1

Function space, $\mathcal{F}$	Unit ball	IPM, $\gamma_{\mathcal{F}}$
$\mathrm{Lip}(S,\rho) := \{f : S \to \mathbb{R} \,\vert\, \|f\|_L < \infty\}$	$\mathcal{F}_W := \{f : \|f\|_L \leq 1\}$	$W := \gamma_{\mathcal{F}_W}$
$BL(S,\rho) := \{f : S \to \mathbb{R} \,\vert\, \|f\|_{BL} < \infty\}$	$\mathcal{F}_\beta := \{f : \|f\|_{BL} \leq 1\}$	$\beta := \gamma_{\mathcal{F}_\beta}$
Bounded measurable functions	$\mathcal{F}_{TV} := \{f : \|f\|_\infty \leq 1\}$	$TV := \gamma_{\mathcal{F}_{TV}}$
$(\mathcal{H}, k)$	$\mathcal{F}_k := \{f : \|f\|_{\mathcal{H}} \leq 1\}$	$\gamma_k := \gamma_{\mathcal{F}_k}$

Before proceeding with our main presentation, we introduce the notation we will use throughout the paper. For a measurable function f and a probability measure $\mathbb{P}$, $\mathbb{P}f := \int f \, d\mathbb{P}$ denotes the expectation of $f(X)$ where X is distributed as $\mathbb{P}$. Given an i.i.d. sample $X_1, \ldots, X_n$ 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 $\frac{1}{n}\sum_{i=1}^n f(X_i)$. Table 1 defines the function spaces, unit balls in these function spaces, and the associated IPMs that we use throughout the paper.

2. Empirical estimators of integral probability metrics

Given $\{X_1^{(1)}, X_2^{(1)}, \ldots, X_m^{(1)}\}$ and $\{X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}\}$, which are i.i.d. samples drawn randomly from $\mathbb{P}$ and $\mathbb{Q}$, respectively, the empirical estimator of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ is given by

$$\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^N \widetilde{Y}_i f(X_i) \right|, \tag{2.1}$$

where $\mathbb{P}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i^{(1)}}$ and $\mathbb{Q}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i^{(2)}}$ represent the empirical distributions of $\mathbb{P}$ and $\mathbb{Q}$, respectively, $N = m + n$, $\widetilde{Y}_i = \frac{1}{m}$ when $X_i = X_i^{(1)}$ for $i = 1, \ldots, m$ and $\widetilde{Y}_{m+i} = -\frac{1}{n}$ when $X_{m+i} = X_i^{(2)}$ for $i = 1, \ldots, n$. The computation of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$ in (2.1) is not straightforward for arbitrary $\mathcal{F}$. To obtain meaningful results, in Section 2.1, we restrict ourselves to $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$ and compute (2.1). We show that the Kantorovich (and therefore L_1-Wasserstein) and Dudley metrics can be estimated by solving linear programs (see Theorems 2.1 and 2.3). By contrast, the empirical estimator for the kernel distance can be obtained in closed form (Theorem 2.4; proved in [20, 21]).

In Section 2.2, we present a novel interpretation of IPMs and their empirical estimators (especially of the Kantorovich metric, Dudley metric and kernel distance) by relating them to binary classification.

2.1. Empirical estimators of the Kantorovich metric (W), Dudley metric (β) and kernel distance (γ_k)

The following results present the empirical estimators of the Kantorovich metric W, Dudley metric β, and kernel distance γ_k.

Theorem 2.1 (Empirical estimator of the Kantorovich metric)**.** *For all* $\alpha \in [0,1]$*, the following function solves (2.1) for* $\mathcal{F} = \mathcal{F}_W$*:*

$$f_\alpha(x) := \alpha \min_{i=1,\ldots,N}(a_i^\star + \rho(x, X_i)) + (1-\alpha) \max_{i=1,\ldots,N}(a_i^\star - \rho(x, X_i)), \tag{2.2}$$

where

$$W(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star, \tag{2.3}$$

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

$$\max_{a_1,\ldots,a_N} \left\{ \sum_{i=1}^N \widetilde{Y}_i a_i \,:\, -\rho(X_i, X_j) \le a_i - a_j \le \rho(X_i, X_j), \forall\, i, j \right\}. \tag{2.4}$$

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

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

and hence

$$\begin{aligned} W(\mathbb{P}_m, \mathbb{Q}_n) &\le \sup \left\{ \sum_{i=1}^N \widetilde{Y}_i f(X_i) \,:\, \max_{X_i \ne X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)} \le 1 \right\} \\ &= \sup \left\{ \sum_{i=1}^N \widetilde{Y}_i f(X_i) \,:\, |f(X_i) - f(X_j)| \le \rho(X_i, X_j), \forall\, i, j \right\} \\ &= \sup \left\{ \sum_{i=1}^N \widetilde{Y}_i a_i \,:\, |a_i - a_j| \le \rho(X_i, X_j), \forall\, i, j \right\}, \end{aligned}$$

where we have set $a_i := f(X_i)$. 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 (2.4). Note that the objective in (2.4) is linear in $\{a_i\}_{i=1}^N$ with linear inequality constraints, and therefore by Theorem 32.1 in [36], 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 the Lipschitz extension theorem due to McShane and Whitney [27, 54], any g on $\{X_1, \ldots, X_N\}$ with $g(X_i) = a_i^\star$ and $\|g\|_L = 1$ can be extended to a function f_α (on S) as defined in (2.2), where $f_\alpha(X_i) = g(X_i)$, $\forall\, i$ and $\|f_\alpha\|_L = \|g\|_L$, which means f_α is a maximizer of (2.1) and $W(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star$. □

Remark 2.2. (a) The main result that is invoked in the proof of Theorem 2.1 is the extension of Lipschitz functions (defined on a subset of S) to S. Since such an extension is also possible for uniformly Hölder continuous functions, we obtain an empirical estimator of $\gamma_{\mathcal{F}}$ similar to (2.3) and (2.4)—with ρ in (2.4) replaced by ρ^θ—where $\mathcal{F} = \{\|f\|_\theta \le 1\}$ and

$$\|f\|_\theta := \sup \left\{ \frac{|f(x) - f(y)|}{\rho^\theta(x, y)} \,:\, x \ne y \text{ in } S \right\}, \, 0 < \theta \le 1.$$

(b) Applying a similar idea as in the proof of Theorem 2.1 to the empirical estimation of the Fortet-Mourier metric, it can be shown that $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) \leq \sum_{i=1}^N \widetilde{Y}_i a_i^\star$, where $\{a_i^\star\}_{i=1}^N$ solve the linear program in (2.4) with $\rho(X_i, X_j)$ replaced by $c(X_i, X_j)$. Since it is not clear whether an extension theorem similar to the one invoked in Theorem 2.1 (for Lipschitz functions) holds for $f \in \{g : \|g\|_c < \infty\}$, it is not clear whether $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star$ holds for any $\{X_i\}_{i=1}^N$.

Theorem 2.3 (Empirical estimator of the Dudley metric)**.** *For all $\alpha \in [0, 1]$, the following function solves (2.1) for $\mathcal{F} = \mathcal{F}_\beta$:*

$$g_\alpha(x) := \max\left(-\max_{i=1,\ldots,N} |a_i^\star|, \min\left(h_\alpha(x), \max_{i=1,\ldots,N} |a_i^\star|\right)\right), \tag{2.5}$$

where

$$h_\alpha(x) := \alpha \min_{i=1,\ldots,N} (a_i^\star + L^\star \rho(x, X_i)) + (1-\alpha) \max_{i=1,\ldots,N} (a_i^\star - L^\star \rho(x, X_i)), \tag{2.6}$$

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

$$\beta(\mathbb{P}_m, \mathbb{Q}_n) = \sum_{i=1}^N \widetilde{Y}_i a_i^\star, \tag{2.7}$$

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

$$\begin{aligned} \max_{a_1,\ldots,a_N,b,c} \ & \sum_{i=1}^N \widetilde{Y}_i a_i \\ s.t. \ & -b\,\rho(X_i, X_j) \leq a_i - a_j \leq b\,\rho(X_i, X_j), \ \forall\, i, j \\ & -c \leq a_i \leq c, \ \forall\, i, \ b + c \leq 1. \end{aligned} \tag{2.8}$$

Proof. The proof is similar to that of Theorem 2.1. Note that

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

which means

$$\begin{aligned} \beta(\mathbb{P}_m, \mathbb{Q}_n) &= \sup\left\{\sum_{i=1}^N \widetilde{Y}_i f(X_i) \ : \ \|f\|_{BL} \leq 1\right\} \\ &\leq \sup\left\{\sum_{i=1}^N \widetilde{Y}_i f(X_i) : \max_i |f(X_i)| + \max_{X_i \neq X_j} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)} \leq 1\right\}. \end{aligned}$$

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

$$\max_{a_1,\ldots,a_N} \left\{ \sum_{i=1}^N \widetilde{Y}_i a_i \ : \ \max_{X_i \neq X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} + \max_i |a_i| \leq 1 \right\}. \tag{2.9}$$

Introducing variables b and c such that $\max_{X_i \neq X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} \leq b$ and $\max_i |a_i| \leq c$ reduces the program in (2.9) to (2.8). In addition, it is easy to see that the optimum occurs at the boundary of the constraint set, i.e., $\max_{X_i \neq X_j} \frac{|a_i - a_j|}{\rho(X_i, X_j)} + \max_i |a_i| = 1$. Hence, by Proposition 11.2.3 of [15], g_α in (2.5) extends any g defined on $\{X_1, \ldots, X_N\}$ (with $g(X_i) = a_i^\star$ and $\|g\|_{BL} = 1$) to S, i.e., $g_\alpha(X_i) = g(X_i)$, $\forall\, i$ and $\|g_\alpha\|_{BL} = \|g\|_{BL}$. Note that h_α in (2.6) is the Lipschitz extension of f to S (by McShane-Whitney Lipschitz extension theorem). Therefore, g_α is a solution to (2.1) and (2.7) holds. □

Theorem 2.4 (Empirical estimator of the kernel distance [20, 21])**.** *Let k be a strictly positive definite kernel, i.e., for all $n \in \mathbb{N}$, $\{\alpha_i\}_{i=1}^n \subset \mathbb{R}\backslash\{0\}$ and all mutually distinct $\{\theta_i\}_{i=1}^n \subset S$, $\sum_{i,j=1}^n \alpha_i \alpha_j k(\theta_i, \theta_j) > 0$. Then, for $\mathcal{F} = \mathcal{F}_k$, the following function is the unique solution to (2.1):*

$$f = \frac{1}{\| \sum_{i=1}^N \widetilde{Y}_i k(\cdot, X_i)\|_{\mathcal{H}}} \sum_{i=1}^N \widetilde{Y}_i k(\cdot, X_i), \tag{2.10}$$

and

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \left\| \sum_{i=1}^N \widetilde{Y}_i k(\cdot, X_i) \right\|_{\mathcal{H}} = \sqrt{\sum_{i,j=1}^N \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)}. \tag{2.11}$$

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

$$\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = \sup \left\{ \left\langle f, \sum_{i=1}^N \widetilde{Y}_i k(\cdot, X_i) \right\rangle_{\mathcal{H}} \ : \ \|f\|_{\mathcal{H}} \leq 1 \right\},$$

where we have used the reproducing property of $\mathcal{H}$, i.e., $\forall\, f \in \mathcal{H}$, $\forall\, x \in S$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$. The result in (2.11) follows from the Cauchy-Schwartz inequality. Since k is strictly positive definite, $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) = 0$ if and only if $\mathbb{P}_m = \mathbb{Q}_n$, which therefore ensures that (2.10) is well-defined. □

It is clear from Theorems 2.1, 2.3 and 2.4 that the empirical estimator of the kernel distance is easy to implement (as it is available in closed form) compared with the empirical estimators of the Kantorovich and Dudley metrics, which involve solving linear programs. One important observation to be made about all these estimators is that they depend on $\{X_i\}_{i=1}^N$ through ρ or k, which means that, once $\{\rho(X_i, X_j)\}_{i,j=1}^N$ or $\{k(X_i, X_j)\}_{i,j=1}^N$ are known, the complexity of the corresponding estimators is independent of the dimension d when $S = \mathbb{R}^d$. Also note that while the maximizer of the kernel distance (see (2.10)) is unique, α in (2.2) and (2.5) signifies that the maximizers of the Kantorovich and Dudley metrics are not unique.

2.2. Interpretability of IPMs and their empirical estimators: Relation to binary classification

In this section, we provide a novel interpretation of IPMs and their empirical estimators by relating them to the problem of binary classification. We show in Proposition 2.5 that W, β and γ_k are related to the optimal risks associated with an appropriate binary classification problem, while in Proposition 2.6 we show their empirical estimators to be related to the *margins* (see footnote 2) of the Lipschitz classifier [51], bounded Lipschitz classifier, and support vector machine, respectively. The significance of the latter result is that the smoothness of these classifiers is inversely related to the distance between the empirical estimates of the class-conditional distributions, computed using W, β and γ_k, respectively. In addition, we also establish the relation between the kernel distance and the Parzen window classifier [37, 38] (also called the kernel classification rule [14, Chapter 10]).

Let us consider the binary classification problem with X being an S-valued random variable, Y being a $\{-1,1\}$-valued random variable and the product space, $S\times\{-1,1\}$, being endowed with a Borel probability measure μ. A discriminant function f is a real valued measurable function on S, whose sign is used to make a classification decision. Given a loss function, $L:\{-1,1\}\times\mathbb{R}\to\mathbb{R}$, the goal is to choose an $f\in\mathcal{F}$ that minimizes the risk associated with L, defined as,

$$\begin{aligned} R_{\mathcal{F}}^{L} &= \inf_{f\in\mathcal{F}} \int_{S\times\{-1,1\}} L(y,f(x))\,d\mu(x,y) \\ &= \inf_{f\in\mathcal{F}} \left\{ \varepsilon \int_{S} L(1,f(x))\,d\mathbb{P}(x) + (1-\varepsilon)\int_{S} L(-1,f(x))\,d\mathbb{Q}(x) \right\}, \end{aligned} \tag{2.12}$$

with the optimal L-risk being $R_{\mathcal{F}_\star}^{L}$, where $\mathcal{F}_\star$ is the set of all measurable functions on S, $\mathbb{P}(X) := \mu(X|Y=+1)$, $\mathbb{Q}(X) := \mu(X|Y=-1)$, $\varepsilon := \mu(S, Y=+1)$. Here, $\mathbb{P}$ and $\mathbb{Q}$ represent the class-conditional distributions and ε is the prior distribution of class $+1$. We now present the result that relates IPMs (between the class-conditional distributions) and the optimal L-risk of a binary classification problem.

Proposition 2.5 ($\gamma_{\mathcal{F}}$ and optimal L-risk)**.** *For $\alpha\in\mathbb{R}$ and $y\in\{-1,1\}$, define*

$$L(y,\alpha) = \frac{2y\alpha}{y(1-2\tau)-1}, \tag{2.13}$$

where $0<\tau<1$. Suppose $\mathcal{F}\subset\mathcal{F}_\star$ is such that $f\in\mathcal{F}\Rightarrow -f\in\mathcal{F}$. If $\varepsilon=\tau$, then $\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = -R_{\mathcal{F}}^{L}$.

Proof. Note that $L(1,\alpha) = -\alpha/\tau$ and $L(-1,\alpha) = \alpha/(1-\tau)$, which imply

$$\varepsilon\int_S L(1,f)\,d\mathbb{P} + (1-\varepsilon)\int_S L(-1,f)\,d\mathbb{Q} = \int_S f\,d\mathbb{Q} - \int_S f\,d\mathbb{P} = \mathbb{Q}f - \mathbb{P}f.$$

Therefore,

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

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

Proposition 2.5 shows that $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ (*resp.* $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$) is the negative of the optimal L-risk that is associated with a binary classifier that classifies the class-conditional distributions $\mathbb{P}$ and $\mathbb{Q}$ (*resp.* $\mathbb{P}_m$ and $\mathbb{Q}_n$) using the loss function L in (2.13), when the discriminant function is restricted to $\mathcal{F}$. Therefore, Proposition 2.5 provides a novel interpretation for the Kantorovich metric, Dudley metric and kernel distance (*resp.* their empirical estimators), which can be understood as the optimal L-risk associated with binary classifiers where the discriminant function f is restricted to $\mathcal{F}_W$, $\mathcal{F}_\beta$ and $\mathcal{F}_k$, respectively. We refer the reader to [31] for a similar result relating ϕ-divergences to the problem of binary classification.

While Proposition 2.5 is a general result relating any IPM to binary classification, the following result (in Proposition 2.6) relates specific IPMs such as the Kantorovich metric (W), Dudley metric (β) and kernel distance (γ_k) to certain well-known classification procedures such as the Lipschitz classifier (the classification rule belongs to $\mathrm{Lip}(S, \rho)$), bounded Lipschitz classifier (the classification rule belongs to $BL(S, \rho)$), and support vector machine (the classification rule belongs to $(\mathcal{H}, k)$), respectively. Before we present the result, we briefly introduce these classifiers.

Suppose $\{(X_i, Y_i)\}_{i=1}^N$ (with $X_i \in S$, $Y_i \in \{-1, 1\}$, $\forall\, i$) is a training sample drawn i.i.d. from μ and $m := |\{i : Y_i = 1\}|$. The Lipschitz classifier is defined as the solution, f_{lip} to the following program:

$$\inf \{\|f\|_L \,:\, f \in \mathrm{Lip}(S, \rho),\, Y_i f(X_i) \geq 1,\, i = 1, \ldots, N\}, \tag{2.14}$$

which is a large margin classifier with margin[2] $\frac{1}{\|f_{\mathrm{lip}}\|_L}$. The program in (2.14) computes a *smooth* function, f that classifies the training sample, $\{(X_i, Y_i)\}_{i=1}^N$ correctly (note that the constraints in (2.14) are such that $\mathrm{sign}(f(X_i)) = Y_i,\ i = 1, \ldots, N$, which means f classifies the training sample correctly, assuming it is separable). The smoothness is controlled by $\|f\|_L$ (the smaller the value of $\|f\|_L$, the smoother f and vice-versa). See [51] for a detailed study on the Lipschitz classifier. Replacing $\|f\|_L$ by $\|f\|_{BL}$ in (2.14) gives the bounded Lipschitz classifier, f_{BL} which is the solution to the following program:

$$\inf \{\|f\|_{BL} \,:\, f \in BL(S, \rho),\, Y_i f(X_i) \geq 1,\, i = 1, \ldots, N\}.$$

Replacing $\|f\|_L$ by $\|f\|_{\mathcal{H}}$ in (2.14), and taking the infimum over $f \in \mathcal{H}$, yields the hard-margin support vector machine, f_{svm} [10], i.e.,

$$f_{\mathrm{svm}} = \arg\inf \{\|f\|_{\mathcal{H}} \,:\, f \in \mathcal{H},\, Y_i f(X_i) \geq 1,\, i = 1, \ldots, N\}.$$

Proposition 2.6 (Empirical estimators and binary classification)**.** *The following hold:*

(a) $\frac{1}{\|f_{\mathrm{lip}}\|_L} \leq \frac{1}{2} W(\mathbb{P}_m, \mathbb{Q}_n)$.

[2]The margin is a technical term used to indicate how well the training sample can be separated. Large margin classifiers (i.e., smooth classifiers) are preferred as they generalize well to unseen samples (i.e., test samples). See [9, 37] for details.

(b) $\frac{1}{\|f_{\mathrm{BL}}\|_{BL}} \leq \frac{1}{2}\beta(\mathbb{P}_m, \mathbb{Q}_n)$.
(c) Suppose a bounded and measurable kernel, k satisfies

$$\int\int_S k(x,y)\, d\mu(x)\, d\mu(y) > 0, \tag{2.15}$$

for all non-zero finite signed Borel measures on a topological space, S. Then,

$$\frac{1}{\|f_{\mathrm{svm}}\|_{\mathcal{H}}} \leq \frac{1}{2}\gamma_k(\mathbb{P}_m, \mathbb{Q}_n).$$

Before we prove Proposition 2.6, let us discuss its significance. Proposition 2.6(a) shows that $\|f_{\mathrm{lip}}\|_L \geq \frac{2}{W(\mathbb{P}_m, \mathbb{Q}_n)}$, which means the smoothness of the classifier, f_{lip}, computed as $\|f_{\mathrm{lip}}\|_L$, is lower bounded by the inverse of the Kantorovich metric 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 separate the distributions $\mathbb{P}$ and $\mathbb{Q}$. A similar explanation holds for the bounded Lipschitz classifier and the support vector machine. Although it is intuitively clear that one would require a classifier that is "wiggly" (i.e., less smooth) to separate the class-conditional distributions that are not "well separated", the above result establishes this formally by defining the wiggliness of the classifier through its norm and the separation between class-conditionals through an IPM.

The condition on k in (2.15) is satisfied by a host of kernels that include the Gaussian kernel, $k(x,y) = \exp(-\sigma\|x-y\|_2^2)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$, the Laplacian kernel, $k(x,y) = \exp(-\sigma\|x-y\|_1)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$, etc. More generally, a large family of kernels that satisfy (2.15) can be obtained: if k is a bounded kernel on $\mathbb{R}^d$ such that $k(x,y) = \psi(x-y)$, where ψ is a continuous positive definite function, then (2.15) holds if and only if the support of the Fourier transform of ψ is $\mathbb{R}^d$ [44, Theorem 9].

To prove Proposition 2.6, we need the following lemma.

Lemma 2.7. *Let $\theta : V \to \mathbb{R}$ and $\psi : V \to \mathbb{R}$ be convex functions on a real vector space V. Suppose*

$$a = \sup\{\theta(x) : \psi(x) \leq b\}, \tag{2.16}$$

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

$$b = \inf\{\psi(x) : \theta(x) \geq a\}. \tag{2.17}$$

Proof. Note that $A := \{x : \psi(x) \leq b\}$ is a convex subset of V. Since θ is not constant on A, by Theorem 32.1 of [36], θ attains its supremum on the boundary of A. Therefore, any solution, x_* to (2.16) 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_* would not be a solution to (2.16). 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 $x^* \in A$ for some $x^* \in H$. From (2.16), this implies θ

attains its supremum relative to A at some point of the relative interior of A. By [36, Theorem 32.1], this implies θ is constant on A, leading to a contradiction. Therefore, $\inf\{\psi(x) : x \in H\} = b$ and the result in (2.17) follows. □

Proof of Proposition 2.6. Define $\mathbb{P}f := \int_S f\, d\mathbb{P}$. Note that $\|f\|_L$, $\|f\|_{BL}$ and $\|f\|_{\mathcal{H}}$ are convex functionals on the vector spaces $\mathrm{Lip}(S,\rho)$, $BL(S,\rho)$ and $U(S) := \{f : S \to \mathbb{R} \,|\, \|f\|_{\mathcal{H}} < \infty\}$, respectively. Similarly, $\mathbb{P}f - \mathbb{Q}f$ is a convex functional on $\mathrm{Lip}(S,\rho)$, $BL(S,\rho)$ and $U(S)$. Since $\mathbb{P} \neq \mathbb{Q}$, it is clear that $\mathbb{P}f - \mathbb{Q}f$ is not constant on $\mathcal{F}_W$ and $\mathcal{F}_\beta$. In fact this is also true for $\mathcal{F}_k$ if k satisfies the condition in (2.15). This is because if k satisfies (2.15), then γ_k is a metric on the space of probability measures [44, Theorem 7] and therefore for $\mathbb{P} \neq \mathbb{Q}$, $\mathbb{P}f - \mathbb{Q}f$ is not constant on $\mathcal{F}_k$. The results in (a)–(c) are then obtained by appropriately choosing ψ, θ, V and b in Lemma 2.7. Here, we only prove (a) as the proofs of (b) and (c) are similar to that of (a).

Since $W(\mathbb{P}_m, \mathbb{Q}_n) = \sup\{\sum_{j=1}^N \widetilde{Y}_j f(X_j) : \|f\|_L \le 1\}$, by Lemma 2.7, we have

$$1 = \inf\left\{ \|f\|_L : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge W(\mathbb{P}_m, \mathbb{Q}_n),\ f \in \mathrm{Lip}(S,\rho) \right\},$$

which can be written as

$$\frac{2}{W(\mathbb{P}_m, \mathbb{Q}_n)} = \inf\left\{ \|f\|_L : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge 2,\ f \in \mathrm{Lip}(S,\rho) \right\}.$$

Note that $\{f \in \mathrm{Lip}(S,\rho) : Y_j f(X_j) \ge 1, \forall\, j\} \subset \{f \in \mathrm{Lip}(S,\rho) : \sum_{j=1}^N \widetilde{Y}_j f(X_j) \ge 2\}$, and therefore

$$\frac{2}{W(\mathbb{P}_m, \mathbb{Q}_n)} \le \inf\left\{ \|f\|_L : Y_j f(X_j) \ge 1, \forall\, j,\ f \in \mathrm{Lip}(S,\rho) \right\},$$

proving (a). A similar analysis for β and γ_k yields (b) and (c). □

In the following, we present another interpretation for the kernel distance by relating it to the Parzen window classifier [37, 38] (also called the kernel classification rule [14]). Theorem 2.4 shows that f in (2.10) is the unique solution to (2.1) when $\mathcal{F} = \mathcal{F}_k$, which by Proposition 2.5 means that it is also the unique solution to $R_{\mathcal{F}}^L$ with empirical distribution. This implies f in (2.10) is the Bayes classifier with Bayes risk $-\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$, with the associated decision rule,

$$\mathrm{sign}(f(x)) = \begin{cases} 1, & \frac{1}{m}\sum_{Y_i=1} k(x, X_i) > \frac{1}{n}\sum_{Y_i=-1} k(x, X_i) \\ -1, & \frac{1}{m}\sum_{Y_i=1} k(x, X_i) \le \frac{1}{n}\sum_{Y_i=-1} k(x, X_i) \end{cases}, \tag{2.18}$$

which is exactly the Parzen window classifier[3].

[3]The classification rule in (2.18) 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, S need not be $\mathbb{R}^d$ and k need not be translation invariant. The rule in (2.18) can thus be seen as a generalization of the classical Parzen window rule. (ii) The kernel in (2.18) is positive definite unlike in the classical Parzen window rule where k need not be so.

3. Consistency and rate of convergence

In Section 2.1, we presented the empirical estimators of W, β and γ_k. For these estimators to be reliable, we need them to converge to their population values as $m, n \to \infty$. We would further like to have fast rates of convergence such that in practice, fewer samples are sufficient to obtain reliable estimates. We address these issues in this section. The strong consistency of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ is shown in Proposition 3.2, while their rates of convergence are analyzed in Corollary 3.5. Corollary 3.5 also proves the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ and analyzes its rate of convergence. We show that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ enjoys a fast rate of convergence compared to $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.

Before we start presenting the results, we 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 $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$, which follows from [49, Theorem 3.7].

Lemma 3.1. *Suppose the following conditions hold:*

(i) $\int_S F \, d\mathbb{P} < \infty$.
(ii) $\int_S F \, d\mathbb{Q} < \infty$.
(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. Note that $|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| \leq \sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{P} f| + \sup_{f \in \mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q} f|$. Therefore, by Theorem 3.7 of [49], $\sup_{f \in \mathcal{F}} |\mathbb{P}_m f - \mathbb{P} f| \xrightarrow{a.s.} 0$, $\sup_{f \in \mathcal{F}} |\mathbb{Q}_n f - \mathbb{Q} f| \xrightarrow{a.s.} 0$ and the result follows. □

The following corollary to Lemma 3.1 shows that $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ are strongly consistent.

Proposition 3.2 (Consistency of W and β)**.** *Let (S, 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. For any $f \in \mathcal{F}_W$,

$$f(x) \leq \sup_{x \in S} |f(x)| \leq \sup_{x,y} |f(x) - f(y)| \leq \|f\|_L \sup_{x,y} \rho(x, y) \leq \|f\|_L \operatorname{diam}(S) < \infty,$$

where $\operatorname{diam}(S)$ represents the diameter of S. Therefore, $\forall \, x \in S$, $F(x) \leq \operatorname{diam}(S) < \infty$, which satisfies (i) and (ii) in Lemma 3.1. Kolmogorov and Tihomirov [24]

have shown that

$$\mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty) \leq \mathcal{N}\left(\frac{\varepsilon}{4}, S, \rho\right) \log\left(2\left\lceil \frac{2\,\mathrm{diam}(S)}{\varepsilon}\right\rceil + 1\right). \tag{3.1}$$

Since $\mathcal{H}(\varepsilon, \mathcal{F}_W, L^1(\mathbb{P}_m)) \leq \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty)$, the conditions (iii) and (iv) in Lemma 3.1 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 $\mathcal{H}(\varepsilon, \mathcal{F}_\beta, \|\cdot\|_\infty) \leq \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty)$. Therefore, the result for β follows. □

Similar to Proposition 3.2, a strong consistency result for γ_k can be provided by estimating the entropy number of $\mathcal{F}_k$. See Cucker and Zhou [12, 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 (Theorem 3.3) 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 empirical estimators of W, β and γ_k. Using this result, we then prove the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$.

Theorem 3.3. *Define* $N := m + n$, $\left(\widetilde{Y}_1, \ldots, \widetilde{Y}_N\right) := \left(\frac{1}{m} \overset{m}{\ldots} \frac{1}{m}, -\frac{1}{n} \overset{n}{\ldots} -\frac{1}{n}\right)$ *and*

$$(X_1, \ldots, X_m, X_{m+1}, \ldots, X_N) := \left(X_1^{(1)}, \ldots, X_m^{(1)}, X_1^{(2)}, \ldots, X_n^{(2)}\right).$$

Let $\mathcal{F}$ *be the space of measurable functions such that* $\|f\|_\infty \leq \nu$, $\mathrm{Var}_{\mathbb{P}}(f) \leq \sigma_{\mathbb{P}}^2$ *and* $\mathrm{Var}_{\mathbb{Q}}(f) \leq \sigma_{\mathbb{Q}}^2$ *for all* $f \in \mathcal{F}$, *where* $\mathrm{Var}_{\mathbb{P}}(f) := \mathbb{P}f^2 - (\mathbb{P}f)^2$. *Then, with probability at least* $1 - 2e^{-\tau}$ *over the choice of* $\{X_i\}_{i=1}^N \sim \mathbb{P}^m \otimes \mathbb{Q}^n$ *and for all* $\alpha > 0$, $\delta \in (0, 1)$, *the following holds:*

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| \leq \frac{2(1+\alpha)}{1-\delta} R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) + \sqrt{\frac{2\tau(m+n)(\sigma_{\mathbb{P}}^2 \vee \sigma_{\mathbb{Q}}^2)}{mn}} + \frac{2\tau\nu(m+n)}{mn}\left(\frac{2}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{\delta(1-\delta)}\right), \tag{3.2}$$

where

$$R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^N \sigma_i \widetilde{Y}_i f(X_i)\right| \,\middle|\, \{X_i\}_{i=1}^N\right], \tag{3.3}$$

$\{\sigma_i\}_{i=1}^N$ *are independent Rademacher (symmetric* ± 1*-valued) random variables and* $a \vee b := \max(a, b)$.

Proof. We begin by noting that

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| \leq \sup_{f\in\mathcal{F}} |(\mathbb{P}_m - \mathbb{Q}_n)f - (\mathbb{P} - \mathbb{Q})f| =: g(X_1, \ldots, X_N).$$

The bound in (3.4) on g can be obtained from Proposition B.1 by using $\mu_i := \mathbb{P}$ for $i = 1, \ldots, m$ and $\mu_i := \mathbb{Q}$ for $i = m+1, \ldots, N$ so that $P := \mathbb{P}^m \otimes \mathbb{Q}^n$,

$z := (\omega_1, \ldots, \omega_N) = (X_1, \ldots, X_N)$, $\theta_i(f, \omega_i) = \frac{f(X_i) - \mathbb{P}f}{m}$ for $i = 1, \ldots, m$ and $\theta_i(f, \omega_i) = -\frac{f(X_i) - \mathbb{Q}f}{n}$ for $i = m+1, \ldots, N$. Note that $\int_{\Omega_i} \theta_i(f, \omega)\, d\mu_i(\omega) = 0$ for all i and $f \in \mathcal{F}$. Also note that $\|\theta_i(f, \cdot)\|_\infty \leq \frac{2\nu}{m} + \frac{2\nu}{n}$ for all $f \in \mathcal{F}$. In addition, for $i = 1, \ldots, m$, we have $\int_{\Omega_i} \theta_i^2(f, \omega)\, d\mu_i(\omega) = m^{-2}\mathbb{P}(f - \mathbb{P}f)^2 = m^{-2}\mathrm{Var}_{\mathbb{P}}(f) \leq \frac{\sigma_{\mathbb{P}}^2}{m^2}$ for all $f \in \mathcal{F}$. Similarly, for $i = m+1 \ldots, N$, we have $\int_{\Omega_i} \theta_i^2(f, \omega)\, d\mu_i(\omega) \leq \frac{\sigma_{\mathbb{Q}}^2}{n^2}$ for all $f \in \mathcal{F}$. By Proposition B.1, we then have that with probability at least $1 - e^{-\tau}$ and for all $\alpha > 0$,

$$
\begin{aligned}
g(X_1, \ldots, X_N) \leq\ & (1+\alpha)\mathbb{E}_P g + \sqrt{\frac{2\tau(m+n)(\sigma_{\mathbb{P}}^2 \vee \sigma_{\mathbb{Q}}^2)}{mn}} \\
& + \frac{2\tau\nu(m+n)}{mn}\left(\frac{1}{\alpha} + \frac{2}{3}\right) \qquad (3.4) \\
\overset{(a)}{\leq}\ & 2(1+\alpha)\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{N}\sigma_i \widetilde{Y}_i f(X_i)\right| + \sqrt{\frac{2\tau(m+n)(\sigma_{\mathbb{P}}^2 \vee \sigma_{\mathbb{Q}}^2)}{mn}} \\
& + \frac{2\tau\nu(m+n)}{mn}\left(\frac{1}{\alpha} + \frac{2}{3}\right) \\
=\ & 2(1+\alpha)\mathbb{E}_P R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) + \sqrt{\frac{2\tau(m+n)(\sigma_{\mathbb{P}}^2 \vee \sigma_{\mathbb{Q}}^2)}{mn}} \\
& + \frac{2\tau\nu(m+n)}{mn}\left(\frac{1}{\alpha} + \frac{2}{3}\right) \qquad (3.5)
\end{aligned}
$$

where (a) follows from bounding $\mathbb{E}\, g(X_1, \ldots, X_N)$ by using the idea of symmetrization (see [50]; for completeness, we prove the bound in Appendix B.2). By Proposition B.3, we have that with probability at least $1 - e^{-\tau}$ and for all $\delta \in (0, 1)$,

$$
\mathbb{E}_P R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) \leq \frac{R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right)}{1-\delta} + \frac{\tau\nu(m+n)}{mn\delta(1-\delta)}. \qquad (3.6)
$$

Combining (3.5) and (3.6), we have that with probability at least $1 - 2e^{-\tau}$, (3.2) holds. □

Theorem 3.3 holds for any $\mathcal{F}$ for which ν is finite (note that $\mathrm{Var}_{\mathbb{P}}(f) = \mathbb{P}f^2 - (\mathbb{P}f)^2 \leq \mathbb{P}f^2 \leq \nu^2 =: \sigma_{\mathbb{P}}^2$ and similarly $\mathrm{Var}_{\mathbb{Q}}(f) \leq \nu^2$). However, in order to comment about the consistency and rate of convergence of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$, we require an estimate of $R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right)$. Note that if $R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) \xrightarrow{\mathbb{P},\mathbb{Q}} 0$ as $m, n \to \infty$, then

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

therefore proving the consistency of $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$. In addition, if $R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) = O_{\mathbb{P},\mathbb{Q}}(r_{mn})$ where $r_{mn} \to 0$ as $m, n \to \infty$, then from (3.2),

$$|\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}\left(r_{mn} \vee \sqrt{\frac{m+n}{mn}}\right),$$

which provides a rate of convergence for $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$.

In Corollary 3.5 (below) to Theorem 3.3, we estimate R_{mn} in (3.3) for $\mathcal{F} = \mathcal{F}_W$, $\mathcal{F} = \mathcal{F}_\beta$ and $\mathcal{F} = \mathcal{F}_k$ to provide rates of convergence for $W(\mathbb{P}_m, \mathbb{Q}_n)$, $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ and $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$, respectively. Before we present and prove Corollary 3.5, we need the result in Proposition 3.4 (which is a slight modification of the Dudley entropy bound in [41, Lemma A.3], in turn based on the bound in [28]; also see [51, Theorem 16]) that bounds R_{mn} in terms of the entropy number of $\mathcal{F}$, which is then used in Corollary 3.5 to obtain bounds on R_{mn} for $\mathcal{F} = \mathcal{F}_W$ and $\mathcal{F} = \mathcal{F}_\beta$. While Proposition 3.4 could also be used to obtain a bound on R_{mn} for $\mathcal{F} = \mathcal{F}_k$, we instead use a direct and simple approach—a slight modification from [21] and [20, Appendix A.2]—, which does not require knowledge of the entropy number of $\mathcal{F}_k$.

Proposition 3.4. *Define* $\mathbb{T}_{mn} := \frac{m+n}{m}\mathbb{P}_m + \frac{m+n}{n}\mathbb{Q}_n$. *Then, for any* $\mathcal{F}$ *containing real valued functions on* S,

$$R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) \le \inf_{\alpha>0}\left\{4\alpha + 12\int_\alpha^{A_{\mathcal{F},\mathbb{T}_{mn}}} \sqrt{\frac{\mathcal{H}\left(\varepsilon, \mathcal{F}, L^2\left(\mathbb{T}_{mn}\right)\right)}{m+n}}\, d\varepsilon\right\}$$
$$\le \inf_{\alpha>0}\left\{4\alpha + 12\int_{\frac{\alpha\sqrt{mn}}{m+n}}^{\frac{\sqrt{mn}A_{\mathcal{F},\mathbb{T}_{mn}}}{m+n}} \sqrt{\frac{\mathcal{H}\left(\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right)}{mn/(m+n)}}\, d\varepsilon\right\}$$

where $A_{\mathcal{F},\mathbb{T}_{mn}} := \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mathbb{T}_{mn})}$. *Suppose* $\sup_{x\in S} F(x) \le \nu < \infty$. *Then,*

$$R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) \le \inf_{\alpha>0}\left\{4\alpha + 12\int_{\frac{\alpha\sqrt{mn}}{m+n}}^{\nu} \sqrt{\frac{\mathcal{H}\left(\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right)}{mn/(m+n)}}\, d\varepsilon\right\}. \tag{3.7}$$

Proof. See Appendix C. □

Corollary 3.5 (Rates of convergence for W, β and γ_k)**.** (i) *Let* S *be a bounded subset of* $(\mathbb{R}^d, \|\cdot\|_s)$ *for some* $1 \le s \le \infty$. *Then there exist finite constants* $\{C_j\}_{j=1}^4$ *(that depend only on* d *and* S*, and not on* m *and* n*) such that*

$$R_{mn}(\mathcal{F}_\beta, \{X_i\}_{i=1}^N) \le R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N)$$
$$\le \begin{cases} C_1\sqrt{\frac{m+n}{mn}} + C_2\sqrt{\frac{m+n}{mn}}\log(m+n), & d=1 \\ C_3\sqrt{\frac{m+n}{mn}} + C_4\frac{(m+n)^{2/3}}{\sqrt{mn}}, & d=2 \end{cases}. \tag{3.8}$$

For $d > 2$, *there exist finite constants* C_5, C_6 *and* N_0 *(that depend only on* d *and* S*, and not on* m *and* n*) such that for any* m, n *with* $(m\wedge n)^{d+1} > N_0(m\vee n)^d$,

$$R_{mn}(\mathcal{F}_\beta, \{X_i\}_{i=1}^N) \le R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) \le C_5\sqrt{\frac{m+n}{mn}} + C_6\frac{(m+n)^{d/(d+1)}}{\sqrt{mn}}. \tag{3.9}$$

Therefore, $|W(\mathbb{P}_m, \mathbb{Q}_n) - W(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_{mn})$ *and* $|\beta(\mathbb{P}_m, \mathbb{Q}_n) - \beta(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_{mn})$ *where*

$$r_{mn} = \begin{cases} \sqrt{\frac{m+n}{mn}} \log(m+n), & d = 1 \\ \frac{(m+n)^{d/(d+1)}}{\sqrt{mn}}, & d \geq 2 \end{cases}. \tag{3.10}$$

In addition, if S *is a bounded, convex subset of* $(\mathbb{R}^d, \|\cdot\|_s)$ *with non-empty interior, then there exist finite constants* $\{D_j\}_{j=1}^5$ *(that depend only on* d *and* S *and not on* m*) such that for* $(m \wedge n) > 9$*,*

$$\begin{aligned} R_{mn}(\mathcal{F}_\beta, \{X_i\}_{i=1}^N) &\leq R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) \\ &\leq \begin{cases} D_1 \sqrt{\frac{m+n}{mn}}, & d = 1 \\ D_2 \sqrt{\frac{m+n}{mn}} + D_3 \sqrt{\frac{m+n}{mn}} \log(m+n), & d = 2 \\ D_4 \sqrt{\frac{m+n}{mn}} + D_5 \frac{(m+n)^{(d-1)/d}}{\sqrt{mn}}, & d > 2 \end{cases}, \end{aligned} \tag{3.11}$$

and therefore, $|W(\mathbb{P}_m, \mathbb{Q}_n) - W(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_{mn})$ *and* $|\beta(\mathbb{P}_m, \mathbb{Q}_n) - \beta(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}(r_{mn})$ *where*

$$r_{mn} = \begin{cases} \sqrt{\frac{m+n}{mn}}, & d = 1 \\ \sqrt{\frac{m+n}{mn}} \log(m+n), & d = 2 \\ \frac{(m+n)^{(d-1)/d}}{\sqrt{mn}}, & d > 2 \end{cases}. \tag{3.12}$$

(ii) *Let* S *be a measurable space. Suppose* k *is measurable and* $\sup_{x \in S} \sqrt{k(x,x)} \leq \nu < \infty$*. Then,*

$$R_{mn}\left(\mathcal{F}_k; \{X_i\}_{i=1}^N\right) \leq \nu \sqrt{\frac{m+n}{mn}} \tag{3.13}$$

and therefore,

$$|\gamma_k(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_k(\mathbb{P}, \mathbb{Q})| = O_{\mathbb{P},\mathbb{Q}}\left(\sqrt{\frac{m+n}{mn}}\right). \tag{3.14}$$

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 empirical estimator of the kernel distance is strongly consistent.*

Proof. (i) Let $\mathcal{F} = \mathcal{F}_W$. Since S is a bounded subset of $\mathbb{R}^d$, it is totally bounded. Define $R := \operatorname{diam}(S)$. As shown in Proposition 3.2, we have $\sup_{f \in \mathcal{F}_W} F(x) \leq R$. Therefore, we obtain

$$\begin{aligned} \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty) &\leq \mathcal{N}\left(\frac{\varepsilon}{4}, S, \|\cdot\|_s\right) \log\left(2\left\lceil \frac{2R}{\varepsilon} \right\rceil + 1\right) \\ &\leq \eta \varepsilon^{-d} \left(4R\varepsilon^{-1} + 2\right), \end{aligned} \tag{3.15}$$

where we have used the facts that $\lceil x \rceil \leq x+1$, $\log(x) \leq x-1$ and there exists $\eta > 0$ (which depends on d and S) such that $\mathcal{N}(\varepsilon, S, \|\cdot\|_s) \leq \eta \varepsilon^{-d}$, $1 \leq s \leq \infty$.[4] Using (3.15) in (3.7), we have

$$R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) \leq \inf_{\alpha>0} \left\{ 4\alpha + 12\sqrt{2\eta} \int_{\frac{\alpha\sqrt{mn}}{m+n}}^{R} \frac{\frac{\sqrt{2R}}{\varepsilon^{(d+1)/2}} + \frac{1}{\varepsilon^{d/2}}}{\sqrt{mn/(m+n)}} d\varepsilon \right\}. \quad (3.16)$$

The bounds in (3.8) are simply obtained by bounding the right hand side of (3.16), which when used in (3.2) yields the rates in (3.10). See Appendix E for details.

Suppose S is convex. Then S is connected. It is easy to see that S is also centered, i.e., for all subsets $A \subset S$ with $\operatorname{diam}(A) \leq 2r$ there exists a point $x \in S$ such that $\|x-a\|_s \leq r$ for all $a \in A$. Since S is connected and centered, we have from [24] that

$$\begin{aligned} \mathcal{H}(\varepsilon, \mathcal{F}_W, \|\cdot\|_\infty) &\leq \mathcal{N}\left(\frac{\varepsilon}{2}, S, \|\cdot\|_s\right) \log 2 + \log\left(2\left\lceil \frac{2R}{\varepsilon} \right\rceil + 1\right) \\ &\leq \eta \varepsilon^{-d} \log 2 + 4R\varepsilon^{-1} + 2. \end{aligned} \quad (3.17)$$

Using (3.17) in (3.7), we have

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) \leq \inf_{\alpha>0} \left\{ \left(4 - \frac{12\sqrt{2}}{\sqrt{m+n}}\right)\alpha + 12 \int_{\frac{\alpha\sqrt{mn}}{m+n}}^{R} \frac{\frac{\sqrt{\eta \log 2}}{\varepsilon^{d/2}} + \frac{2\sqrt{R}}{\sqrt{\varepsilon}}}{\sqrt{mn/(m+n)}} d\varepsilon \right\} \\ + 12\sqrt{2}R\sqrt{\frac{m+n}{mn}}. \end{aligned} \quad (3.18)$$

The bound in (3.11) is obtained by bounding the right hand side of (3.18), which when used in (3.2) yields the rates in (3.12). See Appendix F for details.

Since $\mathcal{F}_\beta \subset \mathcal{F}_W$, we have $R_{mn}(\mathcal{F}_\beta; \{X_i\}_{i=1}^N) \leq R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N)$ and therefore, the result for $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ follows. The rates in (3.12) can also be directly obtained for β by using the entropy number of $\mathcal{F}_\beta$, i.e., $\mathcal{H}(\varepsilon, \mathcal{F}_\beta, \|\cdot\|_\infty) = O(\varepsilon^{-d})$ [50, Theorem 2.7.1] in (3.7).

(ii) The bounding technique on $R_{mn}(\mathcal{F}_k; \{X_i\}_{i=1}^N)$ is taken from [20], however we provide the proof in Appendix D for ease of reference. Omitting for simplicity the conditioning variables $\{X_i\}_{i=1}^N$ in the definition of $R_{mn}\left(\mathcal{F}_k; \{X_i\}_{i=1}^N\right)$, we have

$$\begin{aligned} R_{mn}\left(\mathcal{F}_k; \{X_i\}_{i=1}^N\right) &= \mathbb{E} \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i f(X_i) \right| = \mathbb{E} \left\| \sum_{i=1}^N \sigma_i \widetilde{Y}_i K(\cdot, X_i) \right\|_{\mathcal{H}} \quad (3.19) \\ &\leq \sqrt{\sum_{i=1}^N \widetilde{Y}_i^2 k(X_i, X_i)} + \sqrt{\mathbb{E} \sum_{i \neq j}^N \sigma_i \sigma_j \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)}. \end{aligned}$$

[4] Note that for any $x \in S \subset \mathbb{R}^d$, $\|x\|_\infty \leq \cdots \leq \|x\|_s \leq \cdots \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$. Therefore, $\forall\, s \geq 2$, $\mathcal{N}(\varepsilon, S, \|\cdot\|_s) \leq \mathcal{N}(\varepsilon, S, \|\cdot\|_2)$ and $\forall\, 1 \leq s \leq 2$, $\mathcal{N}(\varepsilon, S, \|\cdot\|_s) \leq \mathcal{N}(\varepsilon, S, \sqrt{d}\|\cdot\|_2) = \mathcal{N}(\varepsilon/\sqrt{d}, S, \|\cdot\|_2)$. Use $\mathcal{N}(\varepsilon, S, \|\cdot\|_2) \leq \gamma \varepsilon^{-d}$ [49, Lemma 2.5].

While the bound in (3.13) can be obtained from (3.19) by noting that $\mathcal{H}$ has *Rademacher type* 2 (see [5, p. 303] for more details), the Appendix D reasoning provides the constant explicitly. Substituting this bound in (3.2) yields (3.14). By the Borel-Cantelli lemma, the strong consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ follows. □

Remark 3.6. (i) Note that the rates of convergence of $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ are dependent on the dimension, d (for $S = \mathbb{R}^d$), which means that in large dimensions, more samples are needed to obtain useful estimates of $W(\mathbb{P}, \mathbb{Q})$ and $\beta(\mathbb{P}, \mathbb{Q})$. Also note that the rates are independent of the metric, $\|\cdot\|_s$, $1 \leq s \leq \infty$.

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

(iii) In the case of kernel distance, we have not made any assumptions on S except it being a measurable space. This means that for $S = \mathbb{R}^d$, the rate is independent of d (if ν in Corollary 3.5(ii) is independent of d), which is a very useful property. The boundedness condition is satisfied by many commonly used 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 [53] for more examples. As mentioned before, the estimates for $R_{mn}(\mathcal{F}_k; \{X_i\}_{i=1}^N)$ can be directly obtained by using the entropy numbers of $\mathcal{F}_k$. See Cucker and Zhou [12, Chapter 5] and Steinwart [45, Chapter 7] for the estimates of entropy numbers and $R_{mn}(\mathcal{F}_k; \{X_i\}_{i=1}^N)$ for various $\mathcal{H}$.

(iv) The rates obtained in (3.10) and (3.12) may not be optimal due to crude upper bounding techniques used to simplify the analysis. However, the idea is to demonstrate the dependence of these rates on d, in contrast to the case of kernel distance where the rates in (3.14) are independent of d.

(v) Combining Theorem 3.3 and Corollary 3.5, it is possible to construct a θ-level test for $H_0 : \mathbb{P} = \mathbb{Q}$ vs. $H_1 : \mathbb{P} \neq \mathbb{Q}$ as follows: For a fixed α, δ and θ, define

$$c_\theta := \frac{2(1+\alpha)}{1-\delta} R_{mn}(\mathcal{F}; \{X_i\}_{i=1}^N) + \sqrt{\frac{2\nu^2(m+n)\log\frac{2}{\theta}}{mn}} + \frac{2\nu(m+n)\log\frac{2}{\theta}}{mn}\left(\frac{2}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{\delta(1-\delta)}\right).$$

It is easy to see that under H_0, $\mathbb{P}^m \otimes \mathbb{Q}^n(\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) > c_\theta) \leq \theta$. Therefore, the test involves accepting H_0 when $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) \leq c_\theta$ and rejecting it if otherwise.

To summarize, in this section, we have shown that the empirical estimators of the Kantorovich metric, Dudley metric and kernel distance are strongly consistent and the empirical estimator of the kernel distance exhibits a fast rate of convergence compared with those of the Kantorovich and Dudley metrics. Therefore, based on the results in this section and Section 2.1, it is clear that the empirical estimator of the kernel distance has more favorable properties compared with the other empirical estimators under consideration, and hence is more suited for use in statistical inference applications.

4. Simulation results

So far, in Sections 2 and 3, we have presented the empirical estimators of W, β and γ_k and their convergence analysis. In this section, we demonstrate the performance of these estimators through simulations and, verify the dependence of the rate of convergence of the empirical estimators of W and β on d (when $S = \mathbb{R}^d$) as opposed to the dimension-independent rate of γ_k.

Given $\mathbb{P}$ and $\mathbb{Q}$, it is usually difficult to compute $W(\mathbb{P}, \mathbb{Q})$, $\beta(\mathbb{P}, \mathbb{Q})$ and $\gamma_k(\mathbb{P}, \mathbb{Q})$ in closed form. However, in order to test the performance of their empirical estimators, in the following, we consider some examples where $W(\mathbb{P}, \mathbb{Q})$, $\beta(\mathbb{P}, \mathbb{Q})$ and $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be computed exactly. Using these examples, we show that the proposed estimators of above IPMs can be used as good surrogates to their population versions, and therefore can be used in applications such as homogeneity testing.

4.1. Estimator of $W(\mathbb{P}, \mathbb{Q})$

For 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

$$W(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{d} W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}), \tag{4.1}$$

where

$$W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}) = \int_{\mathbb{R}} \left| F_{\mathbb{P}^{(i)}}(x) - F_{\mathbb{Q}^{(i)}}(x) \right| dx, \tag{4.2}$$

and $F_{\mathbb{P}^{(i)}}(x) = \mathbb{P}^{(i)}((-\infty, x])$ [48].[5] In the following, we consider two examples where W in (4.2) can be computed in closed form. We need S 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 3.5.

Example 1. Let $S = \times_{i=1}^d [a_i, s_i]$. 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

$$W(\mathbb{P}^{(i)}, \mathbb{Q}^{(i)}) = \frac{s_i + r_i - a_i - b_i}{2},$$

and $W(\mathbb{P}, \mathbb{Q})$ follows from (4.1).

[5] The explicit form for the L_1-Wasserstein distance in (1.2) is known for $(S, \rho(x, y)) = (\mathbb{R}, |x - y|)$ [47, 48], and 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 $F_{\mathbb{P}}^{-1}(u) = \inf\{x \in \mathbb{R} | F_{\mathbb{P}}(x) \ge u\}$, $0 < u < 1$. However, the exact computation (in closed form) of $W_1(\mathbb{P}, \mathbb{Q})$ is not straightforward for all $\mathbb{P}$ and $\mathbb{Q}$. Note that since $\mathbb{R}^d$ is separable, by the Kantorovich-Rubinstein theorem, $W(\mathbb{P}, \mathbb{Q}) = W_1(\mathbb{P}, \mathbb{Q})$, $\forall\, \mathbb{P}, \mathbb{Q}$.

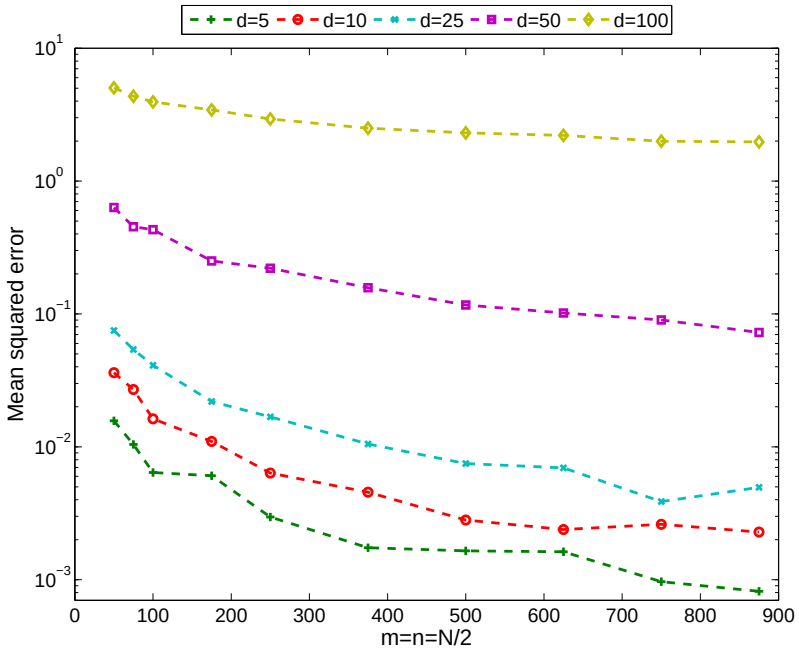

FIG 1. *Empirical mean squared error of the Kantorovich metric (W) between $\mathbb{P} = U[-\frac{1}{2}, \frac{1}{2}]^d$ and $\mathbb{Q} = U[0,1]^d$ with $\rho(x,y) = \|x-y\|_1$ for increasing sample size N and various d. Here, $U[l_1,l_2]^d$ represents a uniform distribution on $[l_1,l_2]^d$. The empirical mean squared error is computed by choosing $T = 100$. See Example 1 and footnote 6 for details.*

Figure 1 shows the behavior of $W(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error[6] for various d and various sample sizes, $m = n = \frac{N}{2}$. 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 $W(\mathbb{P}, \mathbb{Q}) = \frac{d}{2}$.

Example 2. Let $S = \times_{i=1}^{d}[0, c_i]$. Suppose $\mathbb{P}^{(i)}$, $\mathbb{Q}^{(i)}$ have densities

$$p_i(x) = \frac{d\mathbb{P}^{(i)}}{dx} = \frac{\lambda_i e^{-\lambda_i x}}{1 - e^{-\lambda_i c_i}}, \; q_i(x) = \frac{d\mathbb{Q}^{(i)}}{dx} = \frac{\mu_i e^{-\mu_i x}}{1 - e^{-\mu_i c_i}},$$

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

[6]Suppose $\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n)$ is an estimator of $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$. Then the mean squared error is given by $\mathbb{E}[\gamma_{\mathcal{F}}(\mathbb{P}_m, \mathbb{Q}_n) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})]^2$. Given T pairs of samples, $\{(\{X_i^{(1)}\}_{i=1}^m, \{X_i^{(2)}\}_{i=1}^n)_j\}_{j=1}^T$, the empirical mean squared error is computed as $\frac{1}{T}\sum_{j=1}^{T}[\gamma_{\mathcal{F}}(\mathbb{P}_m^j, \mathbb{Q}_n^j) - \gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})]^2$, where $\mathbb{P}_m^j$ and $\mathbb{Q}_n^j$ represent the empirical measures based on $(\{X_i^{(1)}\}_{i=1}^m, \{X_i^{(2)}\}_{i=1}^n)_j$.

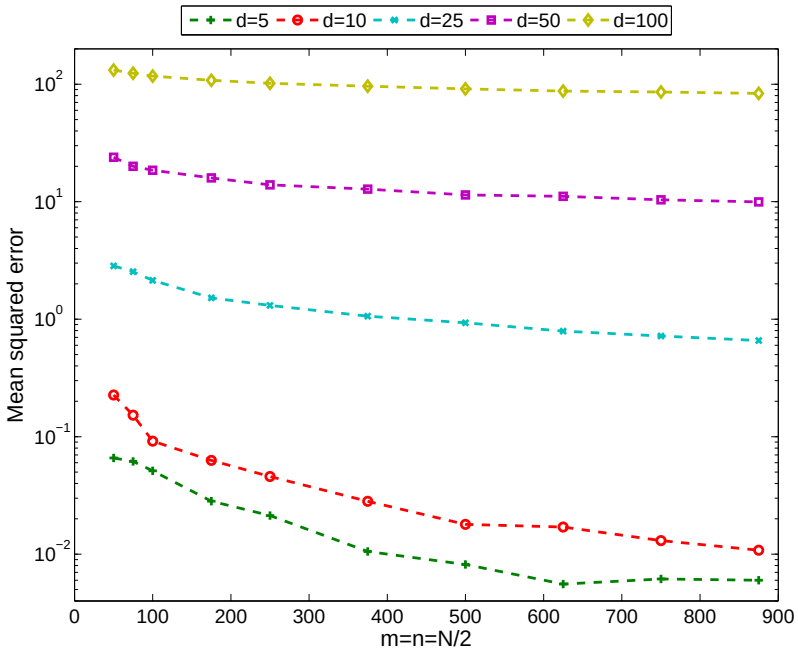

FIG 2. *Empirical mean squared error of the Kantorovich metric (W) between $\mathbb{P}$ and $\mathbb{Q}$, which are truncated exponential distributions on $\mathbb{R}_+^d$, with $\rho(x,y) = \|x-y\|_1$ for increasing sample size N and various d. The empirical mean squared error is computed by choosing $T = 100$. See Example 2 and footnote 6 for details.*

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 (4.1).

Figure 2 shows the behavior of $W(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error for various d and various sample sizes, $m = n = \frac{N}{2}$, where we chose $\lambda_i = 3$, $\mu_i = 1$ and $c_i = 5$ for all i.

The empirical estimates in Figures 1 and 2 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 (2.4). It is easy to see from these figures that $W(\mathbb{P}_m, \mathbb{Q}_n)$ 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. Note that instead of plotting the error bars around the bias of $W(\mathbb{P}_m, \mathbb{Q}_n)$, we plotted the empirical mean squared error so as not to crowd the plots. Figures 1 and 2 also show the effect of the dimensionality, d of the data on $W(\mathbb{P}_m, \mathbb{Q}_n)$, by showing that the rate of convergence of the estimator gets slower with increasing d (see the flattening of the curves at large d)—see Corollary 3.5.

4.2. Estimator of $\gamma_k(\mathbb{P}, \mathbb{Q})$

We now consider the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$. [21, 43] have shown that when k is measurable and bounded,

$$\begin{aligned}\gamma_k(\mathbb{P}, \mathbb{Q}) &= \left\| \int_S k(\cdot, x)\, d\mathbb{P}(x) - \int_S k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}} \\ &= \Big[\int \int_S k(x, y)\, d\mathbb{P}(x)\, d\mathbb{P}(y) + \int \int_S k(x, y)\, d\mathbb{Q}(x)\, d\mathbb{Q}(y) \\ &\qquad - 2 \int \int_S k(x, y)\, d\mathbb{P}(x)\, d\mathbb{Q}(y) \Big]^{\frac{1}{2}}, \end{aligned} \tag{4.3}$$

where the second equality follows from the reproducing property of the kernel (see footnote 1). Note that, although $\gamma_k(\mathbb{P}, \mathbb{Q})$ has a closed form in (4.3), 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 (4.3). In the following, we present four examples where we choose $\mathbb{P}$ and $\mathbb{Q}$ such that $\gamma_k(\mathbb{P}, \mathbb{Q})$ can be computed exactly. Also, note that for the consistency of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$, by Corollary 3.5, we just need the kernel, k to be measurable and bounded, and no assumptions on S are required.

Example 3. Let $S = \mathbb{R}^d$, $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}^{(i)}$ 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 (4.3) 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 (4.3) simply involve the convolution of Gaussian distributions.

Figure 3 shows the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error for various d and sample sizes $m = n = \frac{N}{2}$, where we chose $\tau = 1$, $\mu_i = 0$, $\lambda_i = 1$, $\sigma_i = \sqrt{2}$ and $\theta_i = \sqrt{2}$ for all i.

Example 4. Let $S = \mathbb{R}_+^d$, $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}^{(i)}$ and $\mathbb{Q} = \otimes_{i=1}^d \mathbb{Q}^{(i)}$. Suppose $\mathbb{P}^{(i)} = \mathit{Exp}(1/\lambda_i)$ and $\mathbb{Q}^{(i)} = \mathit{Exp}(1/\mu_i)$, 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 for $\lambda_i \neq \mu_i \neq \alpha$, $\forall\, i$, it is easy to verify that $\gamma_k(\mathbb{P}, \mathbb{Q})$ in (4.3) reduces to

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

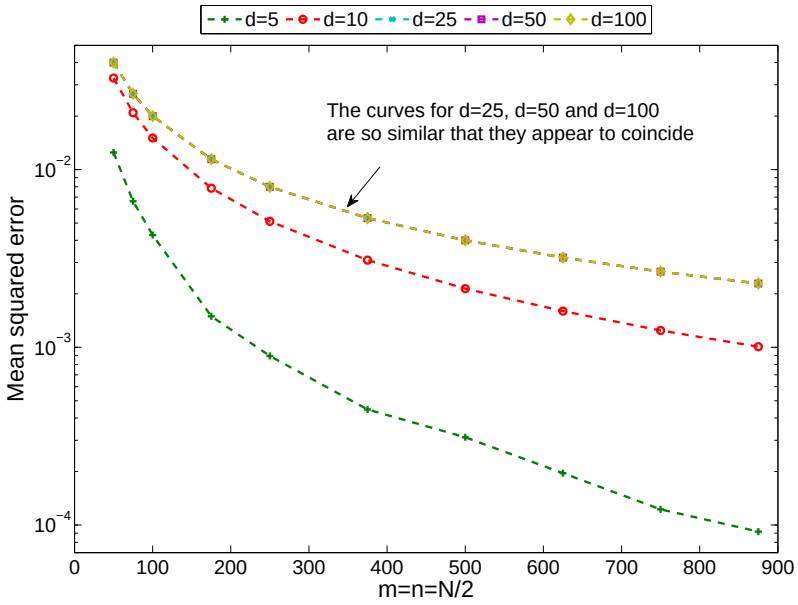

FIG 3. *Empirical mean squared error of the kernel distance (γ_k) between $\mathbb{P} = N(0, 2I_d)$ and $\mathbb{Q} = N(1, 2I_d)$ with $k(x, y) = \exp(-\frac{1}{2}\|x - y\|_2^2)$ for increasing sample size N and various d. Here, $N(\mu, \sigma^2 I_d)$ represents a normal distribution with mean vector $(\mu_1, \overset{d}{.}., \mu_d)$ and covariance matrix $\sigma^2 I_d$. I_d represents the $d \times d$ identity matrix. The empirical mean squared error is computed by choosing $T = 100$. See Example 3 and footnote 6 for details.*

Figure 4 shows the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error for various d and sample sizes $m = n = \frac{N}{2}$, where we chose $\alpha = 1$, $\lambda_i = 3$ and $\mu_i = 2$ for all i.

As in the case of $W(\mathbb{P}_m, \mathbb{Q}_n)$, 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 (2.11). It is easy to see from Figures 3 and 4 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. In addition, these figures also show that the dimensionality d does not greatly affect the rate of convergence of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$, as predicted by Corollary 3.5. In order to compare the performance of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ with $W(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of the dependence of the rate of convergence on d, in the following, we consider the estimation of $\gamma_k(\mathbb{P}, \mathbb{Q})$ for the distributions in Examples 1 and 2.

Example 5. Let $S = \times_{i=1}^d [a_i, s_i]$. 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 $-\infty < a_i \le r_i \le b_i \le s_i < \infty$. 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

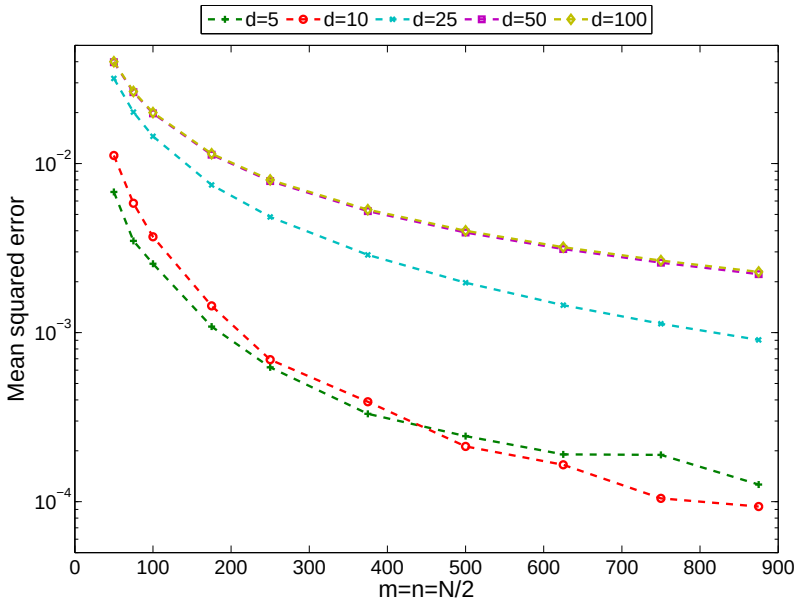

FIG 4. *Empirical mean squared error of the kernel distance (γ_k) between $\mathbb{P}$ and $\mathbb{Q}$, which are exponential distributions on $\mathbb{R}_+^d$ with $k(x,y) = \exp(-\|x-y\|_1)$ for increasing sample size N and various d. The empirical mean squared error is computed by choosing $T = 100$. See Example 4 and footnote 6 for details.*

$$\gamma_k^2(\mathbb{P},\mathbb{Q}) = \prod_{i=1}^{d} \frac{2(b_i - a_i - 1 + e^{a_i - b_i})}{(b_i - a_i)^2} + \prod_{i=1}^{d} \frac{2(s_i - r_i - 1 + e^{r_i - s_i})}{(s_i - r_i)^2} - 2\prod_{i=1}^{d} \frac{2(b_i - r_i) + e^{a_i - s_i} - e^{b_i - s_i} - e^{a_i - r_i} + e^{r_i - b_i}}{(b_i - a_i)(s_i - r_i)}.$$

Figure 5 shows the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error for various d and various sample sizes, $m = n = \frac{N}{2}$. As in Example 1, we chose $\alpha = 1$, $a_i = -\frac{1}{2}$, $b_i = \frac{1}{2}$, $r_i = 0$ and $s_i = 1$ for all $i = 1, \ldots, d$.

Example 6. Let $S = \times_{i=1}^{d}[0, c_i]$. Suppose $\mathbb{P}^{(i)}$, $\mathbb{Q}^{(i)}$ have densities

$$p_i(x) = \frac{d\mathbb{P}^{(i)}}{dx} = \frac{\lambda_i e^{-\lambda_i x}}{1 - e^{-\lambda_i c_i}}, \; q_i(x) = \frac{d\mathbb{Q}^{(i)}}{dx} = \frac{\mu_i e^{-\mu_i x}}{1 - e^{-\mu_i c_i}},$$

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. Let $k(x,y) = \exp(-\alpha\|x-y\|_1)$ for $\alpha > 0$. Then, it can be shown that for $\lambda_i \neq \alpha$ and $\mu_i \neq \alpha$ for all i, we have

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

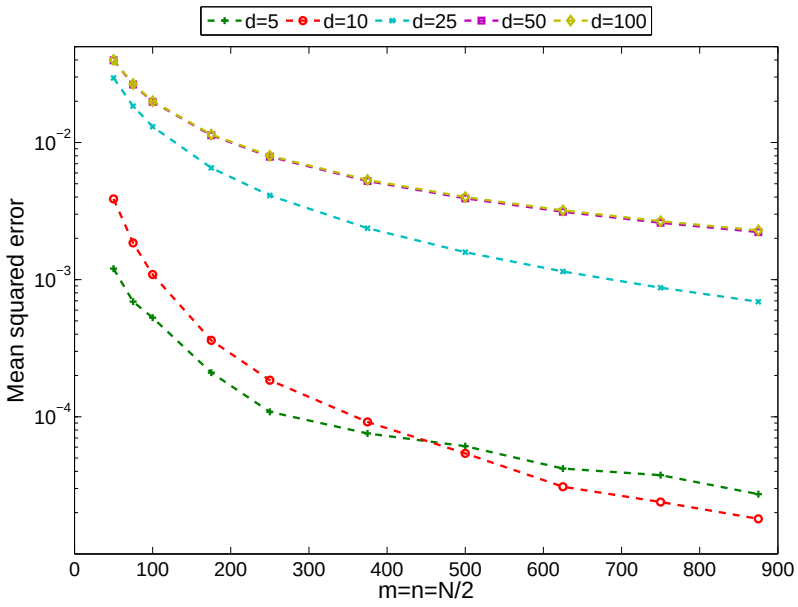

FIG 5. *Empirical mean squared error of the kernel distance* (γ_k) *between* $\mathbb{P} = U[-\frac{1}{2}, \frac{1}{2}]^d$ *and* $\mathbb{Q} = U[0,1]^d$ *with* $k(x,y) = \exp(-\|x-y\|_1)$ *for increasing sample size* N *and various* d. *Here* $U[l_1, l_2]^d$ *represents a uniform distribution on* $[l_1, l_2]^d$. *The empirical mean squared error is computed by choosing* $T = 100$. *See Example 5 and footnote 6 for details.*

where

$$\Theta(\lambda,\mu,c) := \frac{\alpha-\lambda-(\alpha+\mu)e^{-(\lambda+\mu)c}+(\lambda+\mu)e^{-(\alpha+\mu)c}}{(\lambda+\mu)(\alpha+\mu)(\alpha-\lambda)} + \frac{\alpha-\mu-(\alpha+\lambda)e^{-(\lambda+\mu)c}+(\lambda+\mu)e^{-(\alpha+\lambda)c}}{(\lambda+\mu)(\alpha+\lambda)(\alpha-\mu)}.$$

Figure 6 shows the behavior of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error for various d and various sample sizes, $m = n = \frac{N}{2}$, where we chose $\alpha = 2$, $\lambda_i = 3$, $\mu_i = 1$ and $c_i = 5$ for all i.

First note from Figures 5 and 6 that the quality of the estimate of $\gamma_k(\mathbb{P}, \mathbb{Q})$ improves with increasing sample size and that $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ estimates $\gamma_k(\mathbb{P}, \mathbb{Q})$ correctly. By comparing these figures with Figures 1 and 2, it can be seen that the rate of convergence of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is less strongly affected than $W(\mathbb{P}_m, \mathbb{Q}_n)$ by the dimensionality of data, again following Corollary 3.5.

4.3. Estimator of $\beta(\mathbb{P}, \mathbb{Q})$

In the case of W and γ_k, we have closed form expressions to start with—see (4.2) and (4.3)—which can be solved by numerical methods. The resulting values are then used as baselines to test the performance of the estimators of W

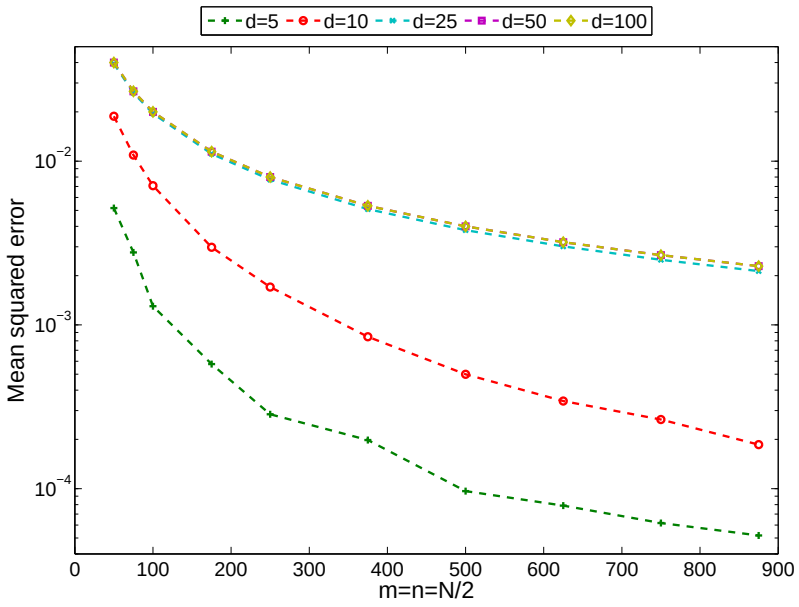

FIG 6. *Empirical mean squared error of the kernel distance (γ_k) between $\mathbb{P}$ and $\mathbb{Q}$, which are truncated exponential distributions on $\mathbb{R}^d_+$, with $k(x,y) = \exp(-\|x-y\|_1)$ for increasing sample size N and various d. The empirical mean squared error is computed by choosing $T = 100$. See Example 6 and footnote 6 for details.*

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 S, 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 $\sum_{i=1}^r \lambda_i = 1$, $\sum_{i=1}^s \mu_i = 1$, $\lambda_i \geq 0$, $\forall\, i$, $\mu_i \geq 0$, $\forall\, i$, and $X_i, Z_i \in S$. This is because, for this choice of $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}\beta(\mathbb{P},\mathbb{Q}) &= \sup\left\{\sum_{i=1}^r \lambda_i f(X_i) - \sum_{i=1}^s \mu_i f(Z_i) : \|f\|_{BL} \leq 1\right\} \\ &= \sup\left\{\sum_{i=1}^{r+s} \theta_i f(V_i) : \|f\|_{BL} \leq 1\right\}, \end{aligned} \tag{4.4}$$

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, (4.4) is of the form of (2.1) and so, by Theorem 2.3, $\beta(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{r+s} \theta_i a_i^\star$, where $\{a_i^\star\}$ solve the following linear program,

$$\begin{aligned} \max_{a_1,\ldots,a_{r+s},b,c} \quad & \sum_{i=1}^{r+s} \theta_i a_i \\ \text{s.t.} \quad & -b\,\rho(V_i,V_j) \leq a_i - a_j \leq b\,\rho(V_i,V_j),\ \forall\, i,j \\ & -c \leq a_i \leq c,\ \forall\, i \\ & b + c \leq 1. \end{aligned} \tag{4.5}$$

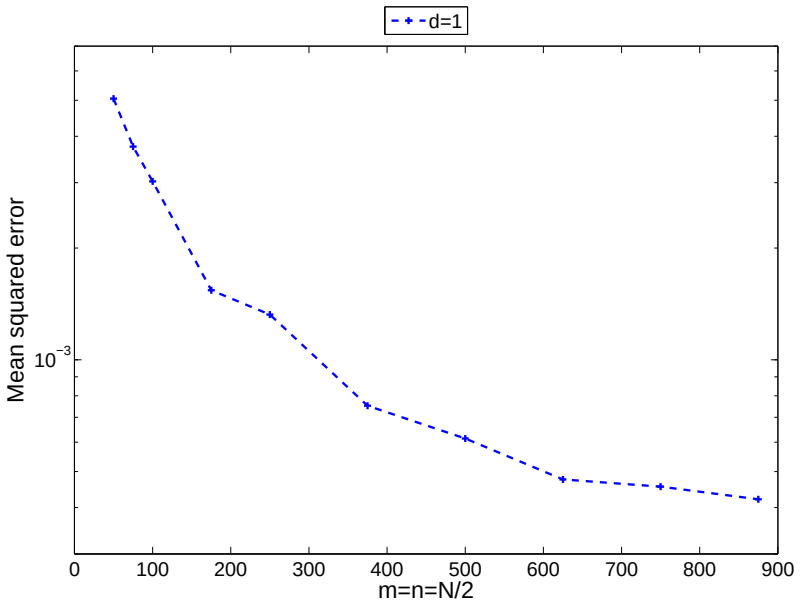

FIG 7. *Empirical mean squared error of the Dudley metric (β) between discrete distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}$ for increasing sample size N. The empirical mean squared error is computed by choosing $T = 100$. See Example 7 and footnote 6 for details.*

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 7. Let $S = \{0, 1, 2, 3, 4, 5\} \subset \mathbb{R}$, $\lambda = (\frac{1}{3}, \frac{1}{6}, \frac{1}{8}, \frac{1}{4}, \frac{1}{8})$, $\mu = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $X = (0, 1, 2, 3, 4)$ 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 (4.5) with $\rho(x, y) = |x - y|$, we get $\beta(\mathbb{P}, \mathbb{Q}) = 0.5278$.

Figure 7 shows the behavior of $\beta(\mathbb{P}_m, \mathbb{Q}_n)$ in terms of empirical mean squared error which is computed by drawing $T = 100$ sets of N i.i.d. samples (with $m = n = N/2$) from $\mathbb{P}$ and $\mathbb{Q}$ and solving the linear program in (2.8)—see footnote 6 for details. 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 $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.

5. Empirical estimation of total variation distance

In Sections 2–4, we derived and analyzed the empirical estimators of W, β and γ_k. Since the total variation distance,

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

is also an IPM, we consider in this section its empirical estimation and consistency analysis. Suppose S is a metric space. Let $TV(\mathbb{P}_m, \mathbb{Q}_n)$ be the empirical estimator of $TV(\mathbb{P}, \mathbb{Q})$. Using similar arguments as in Theorems 2.1 and 2.3, it can be shown that

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

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

$$\max_{a_1,\ldots,a_N} \left\{ \sum_{i=1}^{N} \widetilde{Y}_i a_i \,:\, -1 \le a_i \le 1, \forall\, i \right\}.$$

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

The issue in the empirical 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 the distributions are made. If certain reasonable assumptions are made on the distributions, however, then the total variation distance between such distributions can be estimated consistently in a strong sense.[7] On the other hand, instead of restricting the class of probability measures, 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 $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$. Examples of such choice of $\mathcal{F}$ 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 2.1 and 3. 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 [14, Theorem 12.4].

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

Proposition 5.1 (Lower bounds on TV)**.** (i) *Suppose (S, ρ) is a metric space. Then for all $\mathbb{P} \ne \mathbb{Q}$, we have*

$$TV(\mathbb{P}, \mathbb{Q}) \ge \frac{W(\mathbb{P}, \mathbb{Q})\beta(\mathbb{P}, \mathbb{Q})}{W(\mathbb{P}, \mathbb{Q}) - \beta(\mathbb{P}, \mathbb{Q})}. \tag{5.1}$$

[7] Suppose $S = \mathbb{R}^d$ 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 [13, Chapter 6] and references therein).

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

$$TV(\mathbb{P},\mathbb{Q}) \geq \frac{\gamma_k(\mathbb{P},\mathbb{Q})}{\sqrt{C}}. \tag{5.2}$$

Proof. (i) The proof is based on Lemma 2.7. Note that $\|f\|_L$, $\|f\|_{BL}$ and $\|f\|_\infty$ are convex functionals on the vector spaces $\mathrm{Lip}(S,\rho)$, $BL(S,\rho)$ and $U(S) := \{f : S \to \mathbb{R} \,|\, \|f\|_\infty < \infty\}$, respectively. Similarly, $\mathbb{P}f - \mathbb{Q}f$ is a convex functional on $\mathrm{Lip}(S,\rho)$, $BL(S,\rho)$ and $U(S)$. 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 2.7, the following sequence of inequalities is obtained:

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

which gives (5.1).

(ii) To prove (5.2), we use the coupling formulation for TV [25, p. 19] given by

$$TV(\mathbb{P},\mathbb{Q}) = 2 \inf_{\mu\in\mathcal{L}(\mathbb{P},\mathbb{Q})} \mu(X \neq Y), \tag{5.3}$$

where $\mathcal{L}(\mathbb{P},\mathbb{Q})$ is the set of all measures on $S\times S$ 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,

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

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

$$\gamma_k(\mathbb{P}, \mathbb{Q}) \leq \inf_{\lambda \in \mathcal{L}(\mathbb{P}, \mathbb{Q})} \int \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \, d\lambda(x, y). \tag{5.4}$$

Consider

$$\begin{aligned} \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} &\leq \mathbb{1}_{x \neq y} \|k(\cdot, x) - k(\cdot, y)\|_{\mathcal{H}} \\ &\leq \mathbb{1}_{x \neq y} \left[\|k(\cdot, x)\|_{\mathcal{H}} + \|k(\cdot, y)\|_{\mathcal{H}} \right] \\ &= \mathbb{1}_{x \neq y} \left[\sqrt{k(x, x)} + \sqrt{k(y, y)} \right] \leq 2\sqrt{C} \mathbb{1}_{x \neq y}. \end{aligned} \tag{5.5}$$

Using (5.5) in (5.4) yields (5.2) through (5.3). □

Remark 5.2. *(i)* A simple lower bound on TV can be obtained as $TV(\mathbb{P}, \mathbb{Q}) \geq \beta(\mathbb{P}, \mathbb{Q}), \forall \mathbb{P}, \mathbb{Q}$. However, it is easy to see that the bound in (5.1) 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 only if $\mathbb{P} = \mathbb{Q}$.

(ii) Theorem 4 in [18] shows that the total-variation and Kantorovich distances are related as $\frac{2W(\mathbb{P}, \mathbb{Q})}{diam(S)} \leq TV(\mathbb{P}, \mathbb{Q})$, which makes sense if S is bounded. By simple algebra, it is easy to check that this bound is weaker than the proposed bound in (5.1) if $W(\mathbb{P}, \mathbb{Q}) \leq (1 + diam(S)/2)\beta(\mathbb{P}, \mathbb{Q})$.

(iii) The bounds in (5.1) and (5.2) translate as lower bounds on the KL-divergence through Pinsker's inequality: $TV^2(\mathbb{P}, \mathbb{Q}) \leq 2 \, D_{t \log t}(\mathbb{P}, \mathbb{Q}), \forall \mathbb{P}, \mathbb{Q}$. See Fedotov *et al.* [16] 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 (5.1) and (5.2) also translate to lower bounds on other distance measures on probabilities. See [18] for a detailed discussion of the relation between various metrics.

6. Conclusion & discussion

In this work, we have studied the empirical estimation of integral probability metrics between two probability measures, based on finite samples drawn i.i.d. from each. We have provided empirical estimates, proved consistency, and obtained convergence rates for the empirical estimators of the Kantorovich metric, Dudley metric and kernel distance. We have shown that: (a) the empirical estimator of the kernel distance is easy to implement as it can be obtained in a closed form, unlike the Kantorovich and Dudley metrics, which require solving linear programs; (b) the empirical estimator of the kernel distance has a better rate of convergence than the empirical estimators of the other two metrics, though all these estimators are strongly consistent. Due to these favorable properties, the empirical estimator of the kernel distance might be more useful in statistical inference applications than the remaining two. We also provided a

novel interpretation of these empirical estimators by relating them to the binary classification problem.

There are several interesting questions yet to be explored in connection with this work:

(i) The minimax rate for estimating W, β and γ_k has not been established, nor is it known whether the empirical estimators achieve this rate.
(ii) Although the limiting distribution of $\gamma_k(\mathbb{P}_m, \mathbb{Q}_n)$ is known for the cases of $\mathbb{P} = \mathbb{Q}$ and $\mathbb{P} \neq \mathbb{Q}$ [21, Theorem 8]; [20, Theorem 12], it is not clear whether a limiting distribution exists for $W(\mathbb{P}_m, \mathbb{Q}_n)$ and $\beta(\mathbb{P}_m, \mathbb{Q}_n)$.
(iii) An empirical estimate of the Fortet-Mourier metric has yet to be obtained.

Appendix A: Relation between IPMs and ϕ-divergences

In this appendix, we discuss the relation between IPMs and ϕ-divergences, and show that IPMs are essentially different from ϕ-divergences.

Based on the definitions of IPM and ϕ-divergence, it is clear that $\{\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) : \mathcal{F}\}$ and $\{D_\phi(\mathbb{P}, \mathbb{Q}) : \phi\}$ represent classes of IPMs and ϕ-divergences (on $\mathbb{P}$ and $\mathbb{Q}$) indexed by $\mathcal{F}$ and ϕ, respectively. 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 on S. For $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_\lambda$, let $p = \frac{d\mathbb{P}}{d\lambda}$ and $q = \frac{d\mathbb{Q}}{d\lambda}$ be the Radon-Nikodym derivatives of $\mathbb{P}$ and $\mathbb{Q}$ with respect to λ. For $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_\lambda$ (so that $\mathbb{P} \ll \mathbb{Q}$), it is easy to check that the above two classes intersect at $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ and $\phi(t) = |t-1|$, i.e., $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q}) = \int_S |p - q| \, d\lambda$, which is the total-variation distance. A natural question to consider is for what conditions on $\mathcal{F}$ and ϕ is $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q})$ for all $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_\lambda$? This shows the degree of overlap between the class of IPMs and the class of ϕ-divergences. We answer this in the following theorem, where we show that the total-variation distance is the only "non-trivial"[8] IPM that is also a ϕ-divergence.

Theorem A.1 (Necessary and sufficient conditions)**.** *Suppose $\mathcal{F}_\star$ be the set of all real-valued measurable functions on S and Φ be the class of all convex functions $\phi : [0, \infty) \to (-\infty, \infty]$ continuous at 0 and finite on $(0, \infty)$. 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) $\mathcal{F} = \{f : \|f\|_\infty \leq \frac{\beta - \alpha}{2}\}, \phi(u) = \alpha(u-1)\mathbb{1}_{[0,1]}(u) + \beta(u-1)\mathbb{1}_{[1,\infty)}(u)$ *for some $\alpha < \beta < \infty$, i.e., $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q}) = \frac{\beta - \alpha}{2} \int_S |p - q| \, d\lambda$.*
(ii) $\mathcal{F} = \{f : f = c, c \in \mathbb{R}\}, \phi(u) = \alpha(u-1)\mathbb{1}_{[0,\infty)}(u)$, $\alpha \in \mathbb{R}$, *i.e., $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q}) = 0$.*

[8] Choosing $\mathcal{F}$ to be the set of all real-valued measurable functions on S and $\phi(t) = 0$ if $t = 1$ and $+\infty$ if $t \neq 1$ yields $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = D_\phi(\mathbb{P}, \mathbb{Q}) = 0$ if $\mathbb{P} = \mathbb{Q}$ and $+\infty$ if $\mathbb{P} \neq \mathbb{Q}$. It is easy to show that the converse also holds. For this choice of $\mathcal{F}$ and ϕ, the IPM is trivially a ϕ-divergence.

Proof. ($\Leftarrow$) Suppose (i) holds. Then for any $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_\lambda$, we have

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

where (a) follows from simple algebra after substituting ϕ in $D_\phi(\mathbb{P}, \mathbb{Q})$ (see [23]). This means $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$ and $D_\phi(\mathbb{P}, \mathbb{Q})$ are equal to the total variation distance between $\mathbb{P}$ and $\mathbb{Q}$.

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

($\Rightarrow$) 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[9] on $\mathscr{P}_\lambda$. Through a simple modification of Theorem 2 in [23], it can be shown that if D_ϕ is a pseudometric then $\phi(u) = \alpha(u-1)\mathbb{1}_{[0,1]}(u) + \beta(u-1)\mathbb{1}_{[1,\infty)}(u)$ for some $\beta \geq \alpha$, which means for $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_\lambda$, $D_\phi(\mathbb{P}, \mathbb{Q}) = \frac{\beta-\alpha}{2}\int_S |p-q|\,d\lambda$ if $\beta > \alpha$ and $D_\phi(\mathbb{P}, \mathbb{Q}) = 0$ if $\beta = \alpha$. Now, let us consider two cases.

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

Case 2: $\beta = \alpha$
$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f\in\mathcal{F}}|\mathbb{P}f - \mathbb{Q}f| = 0$ 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 S, i.e., $\mathcal{F} = \{f : f = c,\, c \in \mathbb{R}\}$. □

Note that in Theorem A.1, 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. 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.

Appendix B: Proof of Theorem 3.3: Supplementary results

In this section, we present supplementary results to prove Theorem 3.3.

B.1. Talagrand's inequality

The following is a general result, a special case of which is used to prove Theorem 3.3.

[9] Given a set S, a *metric* for S is a function $\rho : S\times S \to \mathbb{R}_+$ such that *(i)* $\forall\, x$, $\rho(x, x) = 0$, *(ii)* $\forall\, x, y$, $\rho(x, y) = \rho(y, x)$, *(iii)* $\forall\, x, y, z$, $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$, 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 (S, ρ), points in a pseudometric space need not be distinguishable: one may have $\rho(x, y) = 0$ for $x \neq y$.

Proposition B.1. *Let $B \geq 0$, $n \geq 1$, $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$ be probability spaces and $\theta_i : \mathcal{F} \times \Omega_i \to \mathbb{R}$ be bounded measurable functions, where $\mathcal{F}$ is the space of real-valued $\mathcal{A}_i$-measurable functions for all i. Suppose*

(a) $\int_{\Omega_i} \theta_i(f, \omega)\, d\mu_i(\omega) = 0$ *for all* i *and* $f \in \mathcal{F}$
(b) $\int_{\Omega_i} \theta_i^2(f, \omega)\, d\mu_i(\omega) \leq \sigma_i^2$ *for all* $f \in \mathcal{F}$
(c) $\|\theta_i(f, \cdot)\|_\infty \leq B$ *for all* i *and* $f \in \mathcal{F}$.

Define $Z := \times_{i=1}^n \Omega_i$ *and* $P := \otimes_{i=1}^n \mu_i$. *Furthermore, define* $g : Z \to \mathbb{R}$ *by*

$$g(z) := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \theta_i(f, \omega_i) \right|, \ z = (\omega_1, \ldots, \omega_n) \in Z.$$

Then, for all $\tau > 0$, *we have*

$$P\left(\left\{ z \in Z : g(z) \geq \mathbb{E}_P g + \sqrt{2\tau \left(\sum_{i=1}^n \sigma_i^2 + 2B\mathbb{E}_P g \right)} + \frac{2\tau B}{3} \right\} \right) \leq e^{-\tau}.$$

In addition, for all $\tau > 0$ *and* $\alpha > 0$,

$$P\left(\left\{ z \in Z : g(z) \geq (1+\alpha)\mathbb{E}_P g + \sqrt{2\tau \sum_{i=1}^n \sigma_i^2} + \tau B \left(\frac{2}{3} + \frac{1}{\alpha} \right) \right\} \right) \leq e^{-\tau}.$$

To prove Proposition B.1, we need the following result, which we quote from [45, Theorem A.9.14 followed by the simplification in p. 551–552]. Before we quote this result, we need some definitions. Define $Z := \Omega_1 \times \cdots \times \Omega_n$ and $Z_i' := \Omega_1 \times \cdots \times \Omega_{i-1} \times \Omega_{i+1} \times \cdots \times \Omega_n$ for $i = 1, \ldots, n$. Let $\pi_i' : Z \to Z_i'$ denote the projection of Z onto Z_i'. For fixed $i \in \{1, \ldots, n\}$ and $z := (\omega_1, \ldots, \omega_n) \in Z$, define $I_{i,z} : \Omega_i \to Z$ by $I_{i,z}(\omega) := (\omega_1, \ldots, \omega_{i-1}, \omega, \omega_{i+1}, \ldots, \omega_n)$, $\omega \in \Omega_i$.

Theorem B.2. *Let* $n \geq 1$ *and* $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1 \ldots, n$, *be probability spaces. Define* $P := \otimes_{i=1}^n \mu_i$. *Assume that there exist bounded measurable functions* $g : Z \to \mathbb{R}$, $g_i : Z_i' \to \mathbb{R}$ *and* $u_i : Z \to \mathbb{R}$ *such that*

(A) $u_i(z) \leq g(z) - g_i \circ \pi_i'(z) \leq 1$
(B) $\sum_{i=1}^n (g(z) - g_i \circ \pi_i'(z)) \leq g(z)$
(C) $\int_{\Omega_i} u_i \circ I_{i,z}(\omega)\, d\mu_i(\omega) \geq 0$
(D) $\int_{\Omega_i} |u_i \circ I_{i,z}(\omega)|^2\, d\mu_i(\omega) \leq \sigma_i^2$

for some constants $\sigma_i > 0$ *and all* $i = 1, \ldots, n$, $z \in Z$. *Then for all* $\tau > 0$, *we have*

$$P\left(\left\{ z \in Z : g(z) \geq \mathbb{E}_P g + \sqrt{2\tau \left(\sum_{i=1}^n \sigma_i^2 + 2\mathbb{E}_P g \right)} + \frac{2\tau}{3} \right\} \right) \leq e^{-\tau}.$$

Proof of Proposition B.1. Define $\eta_i(f,\omega_i) := B^{-1}\theta_i(f,\omega_i)$ and $h(z) := B^{-1}g(z)$. Let $u_i(z) = h(z) - h_i \circ \pi_i'(z)$ where

$$h_i(z_i') := \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) \right|.$$

It is easy to check that

$$u_i(z) = \sup_{f\in\mathcal{F}} \left| \sum_{j=1}^{n} \eta_j(f,\omega_j) \right| - \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) \right| \leq \sup_{f\in\mathcal{F}} |\eta_i(f,\omega_i)| \leq 1,$$

which implies h, h_i and u_i satisfy (A) in Theorem B.2.

For a fixed $\delta > 0$, let $f_\delta^* \in \mathcal{F}$ be such that

$$h(z) = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \eta_i(f,\omega_i) \right| \leq \left| \sum_{i=1}^{n} \eta_i(f_\delta^*,\omega_i) \right| + \delta,$$

which implies,

$$\begin{aligned}
(n-1)h(z) &\leq (n-1)\delta + \left| \sum_{i=1}^{n} \sum_{j\neq i} \eta_j(f_\delta^*,\omega_j) \right| \\
&\leq (n-1)\delta + \sum_{i=1}^{n} \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) \right| \\
&= (n-1)\delta + \sum_{i=1}^{n} h_i \circ \pi_i'(z).
\end{aligned}$$

Since δ is arbitrary, taking $\delta \to 0$, it is easy to see that h and h_i satisfy (B).

Consider

$$\begin{aligned}
\int_{\Omega_i} u_i \circ I_{i,z}(\omega)\, d\mu_i(\omega) &= \int_{\Omega_i} \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) + \eta_i(f,\omega) \right| d\mu_i(\omega) - h_i(z_i') \\
&\overset{(\star)}{\geq} \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) + \int_{\Omega_i} \eta_i(f,\omega)\, d\mu_i(\omega) \right| - h_i(z_i') \\
&\overset{(a)}{=} 0,
\end{aligned}$$

where we invoked Jensen's inequality in $(\star)$. The above ensures that u_i satisfies (C).

Consider

$$|u_i \circ I_{i,z}(\omega)|^2 = \left(\sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) + \eta_i(f,\omega) \right| - \sup_{f\in\mathcal{F}} \left| \sum_{j\neq i} \eta_j(f,\omega_j) \right| \right)^2$$
$$\leq \left| \sup_{f\in\mathcal{F}} |\eta_i(f,\omega)| \right|^2 = \sup_{f\in\mathcal{F}} \eta_i^2(f,\omega).$$

For a given $\delta > 0$, let $f_\delta \in \mathcal{F}$ be such that $\sup_{f\in\mathcal{F}} \eta_i^2(f,\omega) \leq \eta_i^2(f_\delta,\omega) + \delta$. Then we have,

$$\int_{\Omega_i} \sup_{f\in\mathcal{F}} \eta_i^2(f,\omega)\, d\mu_i(\omega) \leq \int_{\Omega_i} \eta_i^2(f_\delta,\omega)\, d\mu_i(\omega) + \delta \overset{(b)}{\leq} \frac{\sigma_i^2}{B^2} + \delta.$$

Since δ is arbitrary, taking $\delta \to 0$, we have that u_i satisfies (D) with σ_i^2 replaced by $\frac{\sigma_i^2}{B^2}$. Substituting for h proves the first inequality in Theorem B.1. The second inequality is obtained by applying $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ and $2\sqrt{uv} \leq \alpha u + \alpha^{-1} v$ for all $\alpha > 0$ to $\sqrt{2\tau\,(\sum_{i=1}^n \sigma_i^2 + 2B\mathbb{E}_P g)}$. □

B.2. Symmetrization inequality

The proof is a minor modification of the reasoning in [20, Appendix A.2, p. 757], the main difference being that we do not split the expectation into two separate Rademacher averages. We need to bound $h(X_1,\ldots,X_N) := \mathbb{E}\sup_{f\in\mathcal{F}} |(\mathbb{P}_m - \mathbb{Q}_n)f - (\mathbb{P}-\mathbb{Q})f|$. Let $\{\widetilde{X}_i^{(1)}\}_{i=1}^m$ and $\{\widetilde{X}_i^{(2)}\}_{i=1}^n$ be independent samples drawn from $\mathbb{P}$ and $\mathbb{Q}$ respectively. Define $\widetilde{\mathbb{P}}_m := \frac{1}{m}\sum_{i=1}^m \delta_{\widetilde{X}_i^{(1)}}$ and $\widetilde{\mathbb{Q}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{X}_i^{(2)}}$. Also define

$$(\widetilde{X}_1,\ldots,\widetilde{X}_m,\widetilde{X}_{m+1},\ldots,\widetilde{X}_N) := \left(\widetilde{X}_1^{(1)},\ldots,\widetilde{X}_m^{(1)},\widetilde{X}_1^{(2)},\ldots,\widetilde{X}_n^{(2)}\right).$$

Since $\mathbb{P}f = \mathbb{E}\,\widetilde{\mathbb{P}}_m f$ and $\mathbb{Q}f = \mathbb{E}\,\widetilde{\mathbb{Q}}_m f$, we have

$$\begin{aligned}
h(X_1,\ldots,X_N) &= \mathbb{E}\sup_{f\in\mathcal{F}} |(\mathbb{P}_m - \mathbb{Q}_n)f - \mathbb{E}\,(\widetilde{\mathbb{P}}_m - \widetilde{\mathbb{Q}}_n)f| \\
&\leq \mathbb{E}\sup_{f\in\mathcal{F}} |(\mathbb{P}_m - \mathbb{Q}_n)f - (\widetilde{\mathbb{P}}_m - \widetilde{\mathbb{Q}}_n)f| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^N \widetilde{Y}_i \left(f(X_i) - f(\widetilde{X}_i) \right) \right| \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i \left(f(X_i) - f(\widetilde{X}_i) \right) \right| \\
&\leq 2\,\mathbb{E}\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i f(X_i) \right|.
\end{aligned}$$

B.3. Concentration of $R_{mn}(\mathcal{F}; \{X_i\}_{i=1}^N)$

The following result shows that $R_{mn}(\mathcal{F}; \{X_i\}_{i=1}^N)$ is concentrated around its mean, which is a generalization of Lemma A.4 in [4].

Proposition B.3. *Let $\mathcal{F}$ and $R_{mn}(\mathcal{F}; \{X_i\}_{i=1}^N)$ be defined as in Theorem 3.3. Define $P := \mathbb{P}^m \otimes \mathbb{Q}^n$, $g(X_1, \ldots, X_N) := R_{mn}(\mathcal{F}; \{X_i\}_{i=1}^N)$. Then for all $\tau > 0$,*

$$P\left(\left\{(X_1, \ldots, X_N) \in S^N : g(X_1, \ldots, X_N) \leq \mathbb{E}_P g - \sqrt{\frac{2\tau\nu(m+n)\mathbb{E}_P g}{mn}}\right\}\right) \leq e^{-\tau}.$$

In addition, with probability at least $1 - e^{-\tau}$,

$$\mathbb{E}_P g \leq \frac{1}{1-\theta} g(X_1, \ldots, X_N) + \frac{\tau\nu(m+n)}{mn\theta(1-\theta)}$$

for all $\theta \in (0, 1)$.

To prove Proposition B.3, we need the following result, which we quote from [8, (7) and Theorem 6].

Theorem B.4. *Let $n \geq 1$ and $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1 \ldots, n$, be probability spaces. Let Z, Z_i', π_i' and P are defined as in Theorem B.2. Assume there exist bounded measurable functions $g : Z \to [0, \infty)$ and $g_i : Z_i' \to \mathbb{R}$ such that*

(A_1) $0 \leq g(z) - g_i \circ \pi_i'(z) \leq 1$
(B_1) $\sum_{i=1}^n (g(z) - g_i \circ \pi_i'(z)) \leq g(z)$

for all $i = 1, \ldots, n$, $z \in Z$. Then for all $\tau > 0$, we have

$$P\left(\left\{z \in Z : g(z) \leq \mathbb{E}_P g - \sqrt{2\tau\mathbb{E}_P g}\right\}\right) \leq e^{-\tau}.$$

Proof of Proposition B.3. Define $h(X_1, \ldots, X_N) := \frac{mn}{\nu(m+n)} g(X_1, \ldots, X_N)$ and $X^{\setminus i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$. It is clear that h is non-negative. Define

$$g_i(X^{\setminus i}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{j \neq i}^N \sigma_j \widetilde{Y}_j f(X_j)\right| \middle| X^{\setminus i}\right]$$

and $h_i(X^{\setminus i}) := \frac{mn}{\nu(m+n)} h(X_1, \ldots, X_N)$. Consider

$$\begin{aligned} g(X_1, \ldots, X_N) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{j=1}^N \sigma_j \widetilde{Y}_j f(X_j)\right| \middle| \{X_i, X^{\setminus i}\}\right] \\ &\geq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\sum_{j \neq i}^N \sigma_j \widetilde{Y}_j f(X_j) + \mathbb{E}\left[\sigma_i \widetilde{Y}_i f(X_i) \middle| X_i\right]\right| \middle| X^{\setminus i}\right] \\ &= g_i(X^{\setminus i}). \end{aligned}$$

Also consider

$$\begin{aligned} g(X_1,\ldots,X_N) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^N \sigma_j \widetilde{Y}_j f(X_j)\right| \Big| \{X_i, X^{\setminus i}\}\right] \\ &\le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j\neq i}^N \sigma_j \widetilde{Y}_j f(X_j)\right| \Big| X^{\setminus i}\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sigma_i \widetilde{Y}_i f(X_i)\right| \Big| X_i\right] \\ &\le g_i(X^{\setminus i}) + \frac{\nu(m+n)}{mn}, \end{aligned}$$

which implies $0 \le g(X_1,\ldots,X_n) - g_i(X^{\setminus i}) \le \frac{\nu(m+n)}{mn}$ for all i, and therefore h satisfies (A_1) in Theorem B.4. The proof that h satisfies (B_1) follows the technique in the proof of Proposition B.1 to show that h (in the proof of Proposition B.1) satisfies (B). For a fixed $\delta > 0$, let $f_\delta \in \mathcal{F}$ be such that

$$\begin{aligned} g(X_1,\ldots,X_N) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^N \sigma_j \widetilde{Y}_j f(X_j)\right| \Big| \{X_i, X^{\setminus i}\}\right] \\ &\le \mathbb{E}\left[\left|\sum_{j=1}^N \sigma_j \widetilde{Y}_j f_\delta(X_j)\right| \Big| \{X_i, X^{\setminus i}\}\right] + \delta, \end{aligned}$$

which implies

$$\begin{aligned} (N-1)g(X_1,\ldots,X_N) &\le (N-1)\delta + \mathbb{E}\left[\left|\sum_{i=1}^N \sum_{j\neq i} \sigma_j \widetilde{Y}_j f_\delta(X_j)\right| \Big| \{X_i, X^{\setminus i}\}\right] \\ &\le (N-1)\delta + \mathbb{E}\left[\sum_{i=1}^N \sup_{f\in\mathcal{F}}\left|\sum_{j\neq i} \sigma_j \widetilde{Y}_j f(X_j)\right| \Big| \{X_i, X^{\setminus i}\}\right] \\ &= (N-1)\delta + \sum_{i=1}^N \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{j\neq i} \sigma_j \widetilde{Y}_j f(X_j)\right| \Big| X^{\setminus i}\right] \\ &= (N-1)\delta + \sum_{i=1}^N g_i(X^{\setminus i}). \end{aligned}$$

Since δ is arbitrary, taking $\delta \to 0$, we have that h satisfies (B_1). The result therefore follows from Theorem B.4. □

Appendix C: Proof of Proposition 3.4

The following result is a simple modification of Massart's finite class lemma [26], which will be used to prove Proposition 3.4.

Lemma C.1. *Let $\mathcal{A}$ be some finite subset of $\mathbb{R}^l$ and $\{\sigma_i\}_{i=1}^l$ be independent Rademacher random variables. For any $a \in \mathbb{R}^l$, define $a_i := (a)_i$. Then for any $\nu \in \mathbb{R}^l$ satisfying $\sup_{a\in\mathcal{A}} \sqrt{\sum_{i=1}^l a_i^2 \nu_i^2} \le R$, we have*

$$\mathbb{E} \sup_{a\in\mathcal{A}} \sum_{i=1}^{l} \sigma_i \nu_i a_i \le \sqrt{2R^2 \log |\mathcal{A}|}.$$

Proof of Proposition 3.4. Let $\delta_0 := \sup_{f\in\mathcal{F}} \|f\|_{L^2(\mathbb{T}_{mn})}$ and for any $j \in \mathbb{N} \cup \{0\}$, let $\delta_j = 2^{-j}\delta_0$. For each j, let C_j be an $L^2(\mathbb{T}_{mn})$-cover at scale δ_j of $\mathcal{F}$. For each $f \in \mathcal{F}$, pick an $f_j \in C_j$ so that $\|f - f_j\|_{L^2(\mathbb{T}_{mn})} \le \delta_j$. For any M, f can be expressed by chaining as $f = f - f_M + \sum_{j=1}^M (f_j - f_{j-1})$, where $f_0 = 0$. For simplicity, we denote the conditional expectation in $R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right)$ by ignoring the conditioning variables $\{X_i\}_{i=1}^N$. Therefore, for any M, we have

$$\begin{aligned}
R_{mn}\left(\mathcal{F}; \{X_i\}_{i=1}^N\right) &= \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i \left(f(X_i) - f_M(X_i) + \sum_{j=1}^M f_j(X_i) - f_{j-1}(X_i) \right) \right| \\
&\le \mathbb{E} \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i \left(f(X_i) - f_M(X_i) \right) \right| \\
&\qquad + \sum_{j=1}^M \mathbb{E} \sup_{f_j - f_{j-1} \in \mathcal{F}_j} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i \left(f_j - f_{j-1} \right)(X_i) \right| \\
&\overset{(a)}{\le} \mathbb{E} \sqrt{\sum_{i=1}^N \sigma_i^2} \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^N \widetilde{Y}_i^2 \left(f(X_i) - f_M(X_i) \right)^2} \\
&\qquad + \sum_{j=1}^M \mathbb{E} \sup_{f_j - f_{j-1} \in \mathcal{F}_j} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i \left(f_j - f_{j-1} \right)(X_i) \right|,
\end{aligned}$$

where we have used Cauchy-Schwartz inequality in (a) and $\mathcal{F}_j$ is defined below (C.1). Note that

$$\begin{aligned}
&\sqrt{\sum_{i=1}^N \sigma_i^2} \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^N \widetilde{Y}_i^2 \left(f(X_i) - f_M(X_i) \right)^2} = \sup_{f\in\mathcal{F}} \sqrt{N \sum_{i=1}^N \widetilde{Y}_i^2 \left(f(X_i) - f_M(X_i) \right)^2} \\
&= \sup_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^m \frac{m+n}{m^2} \left(f(X_i^{(1)}) - f_M(X_i^{(1)}) \right)^2 + \sum_{i=1}^n \frac{m+n}{n^2} \left(f(X_i^{(2)}) - f_M(X_i^{(2)}) \right)^2} \\
&= \sup_{f\in\mathcal{F}} \|f - f_M\|_{L^2(\mathbb{T}_{mn})} \le \delta_M.
\end{aligned}$$

Also, it can be seen that

$$\|f_j - f_{j-1}\|_{L^2(\mathbb{T}_{mn})} \le \|f - f_j\|_{L^2(\mathbb{T}_{mn})} + \|f - f_{j-1}\|_{L^2(\mathbb{T}_{mn})} \le \delta_j + \delta_{j-1} = 3\delta_j.$$

Therefore,

$$R_{mn}\left(\mathcal{F};\{X_i\}_{i=1}^N\right) \leq \delta_M + \sum_{j=1}^{M} \mathbb{E} \sup_{f_j - f_{j-1} \in \mathcal{F}_j} \sum_{i=1}^{N} \sigma_i \widetilde{Y}_i \left(f_j(X_i) - f_{j-1}(X_i)\right), \quad \text{(C.1)}$$

where $\mathcal{F}_j := \{f_j - f_{j-1} : f_j \in C_j,\, f_{j-1} \in C_{j-1}, \|f_j - f_{j-1}\|_{L^2(\mathbb{T}_{mn})} \leq 3\delta_j\}$. Applying Lemma C.1 to $\mathcal{F}_j$ with $\nu_i = \frac{\sqrt{m+n}}{m}$ for $1 \leq i \leq m$ and $\nu_i = \frac{\sqrt{m+n}}{n}$ for $m+1 \leq i \leq N$, we obtain from (C.1) that

$$\begin{aligned} R_{mn}\left(\mathcal{F};\{X_i\}_{i=1}^N\right) &\leq \delta_M + \sum_{j=1}^{M} \sqrt{\frac{18\delta_j^2 \log\left(|C_j||C_{j-1}|\right)}{m+n}} \\ &\leq \delta_M + 6\sum_{j=1}^{M} \delta_j \sqrt{\frac{\log\left(|C_j|\right)}{m+n}} \\ &\leq \delta_M + 12\sum_{j=1}^{M} (\delta_j - \delta_{j+1}) \sqrt{\frac{\mathcal{H}\left(\delta_j, \mathcal{F}, L^2(\mathbb{T}_{mn})\right)}{m+n}} \\ &\leq \delta_M + 12\int_{\delta_{M+1}}^{\delta_0} \sqrt{\frac{\mathcal{H}\left(\varepsilon, \mathcal{F}, L^2(\mathbb{T}_{mn})\right)}{m+n}}\, d\varepsilon. \end{aligned} \quad \text{(C.2)}$$

For any $\alpha > 0$, pick $M = \sup\{j : \delta_j > 2\alpha\}$. This means, $\delta_{M+1} \leq 2\alpha$ and therefore $\delta_M = 2\delta_{M+1} \leq 4\alpha$. In addition, since $\delta_M > 2\alpha$, we have $\delta_{M+1} > \alpha$. Using these bounds in (C.2), we obtain

$$R_{mn}\left(\mathcal{F};\{X_i\}_{i=1}^N\right) \leq 4\alpha + 12\int_{\alpha}^{A_{\mathcal{F},\mathbb{T}_{mn}}} \sqrt{\frac{\mathcal{H}\left(\varepsilon, \mathcal{F}, L^2(\mathbb{T}_{mn})\right)}{m+n}}\, d\varepsilon. \quad \text{(C.3)}$$

Since α is arbitrary, taking infimum over α provides the result. Note that if $\|f\|_\infty \leq \varepsilon\frac{\sqrt{mn}}{m+n}$, then $\|f\|_{L^2(\mathbb{T}_{mn})} \leq \varepsilon$, which therefore implies

$$\mathcal{H}\left(\varepsilon, \mathcal{F}, L^2\left(\mathbb{T}_{mn}\right)\right) \leq \mathcal{H}\left(\frac{\sqrt{mn}}{m+n}\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right).$$

Hence,

$$R_{mn}\left(\mathcal{F};\{X_i\}_{i=1}^N\right) \leq \inf_{\alpha>0}\left\{4\alpha + 12\int_{\alpha}^{A_{\mathcal{F},\mathbb{T}_{mn}}} \sqrt{\frac{\mathcal{H}\left(\frac{\sqrt{mn}}{m+n}\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right)}{m+n}}\, d\varepsilon\right\}.$$

A simple change of variables provides the result in Proposition 3.4.

Suppose $\sup_{x\in S} F(x) \leq \nu < \infty$. Then it is easy to check that $A_{\mathcal{F},\mathbb{T}_{mn}} \leq \nu\frac{m+n}{\sqrt{mn}}$. Using this in the previous bound yields (3.7). □

Appendix D: Proof of Corollary 3.5(ii)

In the following, we bound $R_{mn}(\mathcal{F}_k; \{X_i\}_{i=1}^N)$ following [20, Theorem 8, Appendix A.2], which uses the proof technique in [3, Lemma 22]. For simplicity, we do not display the conditioning variables $\{X_i\}_{i=1}^N$ in the definition of $R_{mn}\left(\mathcal{F}_k; \{X_i\}_{i=1}^N\right)$. We have

$$\begin{aligned}
R_{mn}\left(\mathcal{F}_k; \{X_i\}_{i=1}^N\right) &= \mathbb{E} \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \sum_{i=1}^N \sigma_i \widetilde{Y}_i f(X_i) \right| \\
&\overset{(*)}{=} \mathbb{E} \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \left\langle \sum_{i=1}^N \sigma_i \widetilde{Y}_i k(\cdot, X_i), f \right\rangle_{\mathcal{H}} \right| \\
&= \mathbb{E} \left\| \sum_{i=1}^N \sigma_i \widetilde{Y}_i K(\cdot, X_i) \right\|_{\mathcal{H}} \\
&= \mathbb{E} \sqrt{\sum_{i,j=1}^N \sigma_i \sigma_j \widetilde{Y}_i \widetilde{Y}_j \langle k(\cdot, X_i), k(\cdot, X_j) \rangle_{\mathcal{H}}} \\
&\overset{(*)}{=} \mathbb{E} \sqrt{\sum_{i,j=1}^N \sigma_i \sigma_j \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)} \\
&= \mathbb{E} \sqrt{\sum_{i=1}^N \sigma_i^2 \widetilde{Y}_i^2 k(X_i, X_i) + \sum_{i \neq j}^N \sigma_i \sigma_j \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)} \\
&\leq \sqrt{\sum_{i=1}^N \widetilde{Y}_i^2 k(X_i, X_i)} + \sqrt{\mathbb{E} \sum_{i \neq j}^N \sigma_i \sigma_j \widetilde{Y}_i \widetilde{Y}_j k(X_i, X_j)}.
\end{aligned} \tag{D.1}$$

Since σ_i and σ_j are independent random variables with zero mean, the second term in the last line is zero, while the first term is bounded by $\nu\sqrt{\frac{m+n}{mn}}$, therefore yielding (3.13). Note that we have invoked the reproducing property (see footnote 1) of the kernel in $(*)$.

Appendix E: Proof of (3.8) and (3.9)

Define $\theta_{mn} := \sqrt{\frac{mn}{m+n}}$ and $\phi_{mn} := \frac{\sqrt{mn}}{m+n}$. We showed in (3.16) that for every $\alpha > 0$,

$$R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) \leq 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \int_{\alpha\phi_{mn}}^R \left(\frac{\sqrt{2R}}{\varepsilon^{(d+1)/2}} + \frac{1}{\varepsilon^{d/2}} \right) d\varepsilon.$$

We now consider three cases.

$d = 1$:

$$\begin{aligned} R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) &\leq 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \int_{\alpha\phi_{mn}}^{R} \left(\frac{\sqrt{2R}}{\varepsilon} + \frac{1}{\sqrt{\varepsilon}} \right) d\varepsilon \\ &= 4\alpha - \frac{24\sqrt{\eta R}}{\theta_{mn}} \log(\alpha\phi_{mn}) - \frac{24\sqrt{2\eta}}{\theta_{mn}} \sqrt{\alpha\phi_{mn}} + \frac{A_1}{\theta_{mn}} \\ &\leq 4\alpha - \frac{A_2}{\theta_{mn}} \log(\alpha\phi_{mn}) + \frac{A_1 - 24\sqrt{2\eta}}{\theta_{mn}}, \end{aligned}$$

where $A_1 := 24\sqrt{\eta R}\left(\log R + \sqrt{2}\right)$, $A_2 := 24\sqrt{\eta R} + 12\sqrt{2\eta}$ and therefore

$$\begin{aligned} R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) &\leq \inf_{\alpha>0} \left[4\alpha - \frac{A_2}{\theta_{mn}} \log(\alpha\phi_{mn}) \right] + \frac{A_1 - 24\sqrt{2\eta}}{\theta_{mn}} \\ &= \frac{C_1}{\theta_{mn}} + C_2 \frac{\log(m+n)}{\theta_{mn}}, \end{aligned}$$

where $C_1 := A_1 + A_2 - A_2 \log(A_2/4) - 24\sqrt{2\eta}$ and $C_2 := \frac{A_2}{2}$.

$d = 2$:

$$\begin{aligned} R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) &\leq 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \int_{\alpha\phi_{mn}}^{R} \left(\frac{\sqrt{2R}}{\varepsilon^{3/2}} + \frac{1}{\varepsilon} \right) d\varepsilon \\ &= 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \left(-2\sqrt{2} + \log R + \frac{2\sqrt{2R}}{\sqrt{\alpha\phi_{mn}}} + 2\log \frac{1}{\sqrt{\alpha\phi_{mn}}} \right) \\ &\leq 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \left(-2 - 2\sqrt{2} + \log R + \frac{2\sqrt{2R}+2}{\sqrt{\alpha\phi_{mn}}} \right) \end{aligned}$$

and therefore

$$\begin{aligned} R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) &\leq \inf_{\alpha>0} \left[4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \left(-2 - 2\sqrt{2} + \log R + \frac{2\sqrt{2R}+2}{\sqrt{\alpha\phi_{mn}}} \right) \right] \\ &= \frac{C_3}{\theta_{mn}} + C_4 \frac{(m+n)^{2/3}}{\sqrt{mn}}, \end{aligned}$$

where $C_3 := 12\sqrt{2\eta}(\log R - 2 - 2\sqrt{2})$, $A_1 := 12\sqrt{2\eta}(2\sqrt{2R} + 2)$ and $C_4 := (1 + 8^{1/3})A_1^{2/3}$.

$d > 2$:

$$\begin{aligned} R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) &\leq 4\alpha + \frac{12\sqrt{2\eta}}{\theta_{mn}} \int_{\alpha\phi_{mn}}^{R} \left(\frac{\sqrt{2R}}{\varepsilon^{(d+1)/2}} + \frac{1}{\varepsilon^{d/2}} \right) d\varepsilon \\ &= 4\alpha + \frac{24\sqrt{2\eta}}{\theta_{mn}} \left(\frac{\sqrt{2R}/(d-1)}{(\alpha\phi_{mn})^{\frac{d-1}{2}}} + \frac{1/(d-2)}{(\alpha\phi_{mn})^{\frac{d-2}{2}}} \right) + \frac{A_1}{\theta_{mn}}, \end{aligned}$$

where $A_1 := -24\sqrt{2\eta}R^{\frac{2-d}{2}}(\frac{\sqrt{2}}{d-1} + \frac{1}{d-2})$. Define $A_2 := \frac{48\sqrt{\eta R}}{d-1}$ and $A_3 := \frac{24\sqrt{2\eta}}{d-2}$. Then, we have

$$R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) \leq \begin{cases} 4\alpha + \left(\frac{A_2}{\theta_{mn}\phi_{mn}^{\frac{d-1}{2}}} + \frac{A_3}{\theta_{mn}\phi_{mn}^{\frac{d-2}{2}}}\right)\alpha^{-\frac{d-1}{2}} + \frac{A_1}{\theta_{mn}}, & \alpha \in (0,1] \\ 4\alpha + \left(\frac{A_2}{\theta_{mn}\phi_{mn}^{\frac{d-1}{2}}} + \frac{A_3}{\theta_{mn}\phi_{mn}^{\frac{d-2}{2}}}\right)\alpha^{-\frac{d-2}{2}} + \frac{A_1}{\theta_{mn}}, & \alpha \geq 1 \end{cases}$$

and therefore

$$R_{mn}(\mathcal{F}_W, \{X_i\}_{i=1}^N) \leq \min(A_4, A_5), \tag{E.1}$$

where

$$A_4 := \inf_{0<\alpha\leq 1}\left[4\alpha + \eta_{mn}\alpha^{-\frac{d-1}{2}} + \frac{A_1}{\theta_{mn}}\right]$$

and

$$A_5 := \inf_{\alpha\geq 1}\left[4\alpha + \eta_{mn}\alpha^{-\frac{d-2}{2}} + \frac{A_1}{\theta_{mn}}\right],$$

where $\eta_{mn} := \frac{A_2}{\theta_{mn}\phi_{mn}^{\frac{d-1}{2}}} + \frac{A_3}{\theta_{mn}\phi_{mn}^{\frac{d-2}{2}}}$. It is easy to check that

$$A_4 = 4\alpha^* + \eta_{mn}\left(\alpha^*\right)^{-\frac{d-1}{2}} + \frac{A_1}{\theta_{mn}}$$

and

$$A_5 = 4\alpha^{**} + \eta_{mn}\left(\alpha^{**}\right)^{-\frac{d-2}{2}} + \frac{A_1}{\theta_{mn}},$$

where $\alpha^* := 1 \wedge (\frac{(d-1)\eta_{mn}}{8})^{\frac{2}{d+1}}$ and $\alpha^{**} := 1 \vee (\frac{(d-2)\eta_{mn}}{8})^{\frac{2}{d}}$. Define $N_0 := \frac{(d-1)^2(A_2+A_3)^2 2^d}{64}$. Then for all m, n such that $(m\wedge n)^{d+1} > N_0(m\vee n)^d$, it is easy to check that $\eta_{mn} < \frac{8}{d-1}$, which implies $\alpha^* = (\frac{(d-1)\eta_{mn}}{8})^{\frac{2}{d+1}}$ and $\alpha^{**} = 1$. Therefore, for all m, n such that $(m\wedge n)^{d+1} > N_0(m\vee n)^d$,

$$A_4 = A_6\eta_{mn}^{\frac{2}{d+1}} + \frac{A_1}{\theta_{mn}}$$

and

$$A_5 = 4 + \eta_{mn} + \frac{A_1}{\theta_{mn}},$$

where $A_6 := 4(\frac{d-1}{8})^{\frac{2}{d+1}} + (\frac{8}{d-1})^{\frac{d-1}{d+1}}$. Using these results in (E.1), we have for all m, n such that $(m\wedge n)^{d+1} > N_0(m\vee n)^d$,

$$R_m(\mathcal{F}_W, \{X_i^{(1)}\}_{i=1}^m) \leq A_4 \leq \frac{A_7}{\theta_{mn}^{\frac{2}{d+1}}\phi_{mn}^{\frac{d-1}{d+1}}} + \frac{A_1}{\theta_{mn}} = A_7\frac{(m+n)^{\frac{d}{d+1}}}{\sqrt{mn}} + A_1\sqrt{\frac{m+n}{mn}},$$

where $A_7 := A_6(A_2+A_3)^{\frac{2}{d+1}}$.

Appendix F: Proof of (3.11)

Define $\theta_{mn} := \sqrt{\frac{mn}{m+n}}$, $\phi_{mn} := \frac{\sqrt{mn}}{m+n}$ and $\eta_{mn} := 4 - \frac{12\sqrt{2}\phi_{mn}}{\theta_{mn}}$. We showed in (3.18) that for any $\alpha > 0$,

$$R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) \leq \eta_{mn}\alpha + \frac{12}{\theta_{mn}} \int_{\alpha\phi_{mn}}^R \left(\frac{\sqrt{\eta \log 2}}{\varepsilon^{d/2}} + \frac{2\sqrt{R}}{\sqrt{\varepsilon}} \right) d\varepsilon + \frac{12\sqrt{2}R}{\theta_{mn}}.$$

As in Appendix E, we consider three cases:

$d = 1$: Define $A := \sqrt{\eta \log 2} + 2\sqrt{R}$.

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \eta_{mn}\alpha + \frac{12}{\theta_{mn}} \int_{\alpha\phi_{mn}}^R \frac{A}{\sqrt{\varepsilon}}\, d\varepsilon + \frac{12\sqrt{2}R}{\theta_{mn}} \\ &= \eta_{mn}\alpha - \frac{24A\sqrt{\phi_{mn}}}{\theta_{mn}}\sqrt{\alpha} + \frac{24A\sqrt{R} + 12\sqrt{2}R}{\theta_{mn}} \end{aligned}$$

and therefore

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \inf_{\alpha>0} \left[\eta_{mn}\alpha - \frac{24A\sqrt{\phi_{mn}}}{\theta_{mn}}\sqrt{\alpha} \right] + \frac{24A\sqrt{R} + 12\sqrt{2}R}{\theta_{mn}} \\ &= \frac{144A^2}{\theta_{mn}(12\sqrt{2} - 4\sqrt{m+n})} + \frac{24A\sqrt{R} + 12\sqrt{2}R}{\theta_{mn}}. \end{aligned}$$

Since $m \wedge n > 9$, we have $\frac{144A^2}{\theta_{mn}(12\sqrt{2}-4\sqrt{m+n})} < \frac{6\sqrt{2}A^2}{\theta_{mn}}$ and therefore,

$$R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) \leq \frac{24A\sqrt{R} + 12\sqrt{2}R + 6\sqrt{2}}{\theta_{mn}}.$$

$d = 2$: Define $B := 12\left(\sqrt{2}R + 4R + \log R\sqrt{\eta \log 2}\right)$ and $C := 12\sqrt{\eta \log 2} + 24\sqrt{R}$.

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \eta_{mn}\alpha + \frac{12}{\theta_{mn}} \int_{\alpha\phi_{mn}}^R \left(\frac{\sqrt{\eta \log 2}}{\varepsilon} + \frac{2\sqrt{R}}{\sqrt{\varepsilon}} \right) d\varepsilon + \frac{12\sqrt{2}R}{\theta_{mn}} \\ &= \eta_{mn}\alpha - \frac{24}{\theta_{mn}}\left(\sqrt{\eta \log 2} \log \sqrt{\alpha\phi_{mn}} + 2\sqrt{R}\sqrt{\alpha\phi_{mn}} \right) + \frac{B}{\theta_{mn}} \\ &\leq \eta_{mn}\alpha - \frac{C}{\theta_{mn}} \log \alpha\phi_{mn} + \frac{B - 48\sqrt{R}}{\theta_{mn}} \end{aligned}$$

and therefore,

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \inf_{\alpha>0} \left[\eta_{mn}\alpha - \frac{C}{\theta_{mn}} \log \alpha\phi_{mn} \right] + \frac{B - 48\sqrt{R}}{\theta_{mn}} \\ &\leq \frac{C}{\theta_{mn}} \log(m+n) + \frac{B - 48\sqrt{R} + C - C\log(C/4)}{\theta_{mn}}. \end{aligned}$$

$d > 2$: Define $A_1 := \frac{24\sqrt{\eta \log 2}}{d-2}$ and $A_2 := 48R + 12\sqrt{2}R - \frac{24\sqrt{\eta \log 2}R^{\frac{2-d}{2}}}{d-2}$.

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \eta_{mn}\alpha + \frac{12}{\theta_{mn}} \int_{\alpha\phi_{mn}}^{R} \left(\frac{\sqrt{\eta \log 2}}{\varepsilon^{d/2}} + \frac{2\sqrt{R}}{\sqrt{\varepsilon}} \right) d\varepsilon + \frac{12\sqrt{2}R}{\theta_{mn}} \\ &\leq \eta_{mn}\alpha + \frac{12}{\theta_{mn}} \int_{\alpha\phi_{mn}}^{R} \frac{\sqrt{\eta \log 2}}{\varepsilon^{d/2}}\, d\varepsilon + \frac{24\sqrt{R}}{\theta_{mn}} \int_0^R \frac{1}{\sqrt{\varepsilon}}\, d\varepsilon \\ &\quad + \frac{12\sqrt{2}R}{\theta_{mn}} \\ &= \eta_{mn}\alpha + \frac{A_1}{\theta_{mn}\phi_{mn}^{\frac{d-2}{2}}} \left(\frac{1}{\alpha} \right)^{\frac{d-2}{2}} + \frac{A_2}{\theta_{mn}}, \end{aligned}$$

and therefore

$$\begin{aligned} R_{mn}(\mathcal{F}_W; \{X_i\}_{i=1}^N) &\leq \inf_{\alpha > 0} \left[\eta_{mn}\alpha + \frac{A_1}{\theta_{mn}\phi_{mn}^{\frac{d-2}{2}}} \left(\frac{1}{\alpha} \right)^{\frac{d-2}{2}} \right] + \frac{A_2}{\theta_{mn}} \\ &= A_3 \frac{(m+n)^{\frac{d-1}{d}}}{\sqrt{mn}} + A_2 \sqrt{\frac{m+n}{mn}}, \end{aligned}$$

where $C := (\frac{(d-2)A_1}{2})^{2/d}$ and $A_3 := 2^{d-2}(C + A_1 C^{\frac{2-d}{2}})$.

Acknowledgments

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. B. K. S. also thanks the Gatsby Foundation for their generous support. A. G. was supported by grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747, and ARO MURI W911NF0810242. The authors thank Robert Williamson and Mark Reid for helpful conversations on the relation between ϕ-divergences and IPMs.

References

[1] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological)* **28**, 131–142. MR0196777

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337–404. MR0051437

[3] Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3**, 463–482. MR1984026

[4] Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *Annals of Statistics* **33**, 4, 1497–1537. MR2166554

[5] Beauzamy, B. (1985). *Introduction to Banach spaces and their Geometry*. North-Holland, The Netherlands. MR0889253

[6] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, London, UK. MR2239907

[7] Bickel, P. J. (1969). A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics* **40**, 1, 1–23. MR0256519

[8] Boucheron, S., Lugosi, G., and Massart, P. (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms* **16**, 3, 277–292. MR1749290

[9] Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation* **11**, 7, 1493–1517.

[10] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

[11] Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica* **2**, 299–318. MR0219345

[12] Cucker, F. and Zhou, D.-X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, UK. MR2354721

[13] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York. MR0780746

[14] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York. MR1383093

[15] Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK. MR1932358

[16] Fedotov, A. A., Harremoës, P., and Topsøe, F. (2003). Refinements of Pinsker's inequality. *IEEE Trans. Information Theory* **49**, 6, 1491–1498. MR1984937

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

[18] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70**, 3, 419–435.

[19] Gray, R. M., Neuhoff, D. L., and Shields, P. C. (1975). A generalization of Ornstein's $\overline{d}$ distance with applications to information theory. *Annals of Probability* **3**, 315–328. MR0368127

[20] Gretton, A., Borgwardt, K., Rasch, M., Schoelkopf, B., and Smola, A. (2012). A kernel two-sample test. *JMLR* **13**, 723–773.

[21] Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2007). 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, 513–520.

[22] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). 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, 585–592.

[23] Khosravifard, M., Fooladivanda, D., and Gulliver, T. A. (2007). Confliction of the convexity and metric properties in f-divergences. *IEICE Trans. Fundamentals* **E90-A**, 9, 1848–1853.

[24] Kolmogorov, A. N. and Tihomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in functional space. *American Mathematical Society Translations* **2**, 17, 277–364. MR0124720

[25] Lindvall, T. (1992). *Lectures on the Coupling Method*. John Wiley & Sons, New York. MR1180522

[26] Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.* **9**, 6, 245–303. MR1813803

[27] McShane, E. J. (1934). Extension of range of functions. *Bulletin of the American Mathematical Society* **40**, 837–842. MR1562984

[28] Mendelson, S. (2002). Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory* **48**, 1, 251–263. MR1872178

[29] Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability* **29**, 429–443. MR1450938

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

[31] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2009). On surrogate loss functions and f-divergences. *Annals of Statistics* **37**, 2, 876–904. MR2502654

[32] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* **56**, 11, 5847–5861. MR2808937

[33] Rachev, S. T. (1984). On a class of minimum functionals in a space of probability measures. *Theory of Probability and its Applications* **29**, 41–48. MR0739499

[34] Rachev, S. T. (1985). The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications* **29**, 647–676.

[35] Rachev, S. T. and Rüschendorf, L. (1998). *Mass transportation problems. Vol. I Theory, Vol. II Applications*. Probability and its Applications. Springer-Verlag, Berlin. MR1619170

[36] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ. MR0274683

[37] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA. MR1949972

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

[39] Shorack, G. R. (2000). *Probability for Statisticians*. Springer-Verlag, New York. MR1762415

[40] Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow University Mathematics Bulletin* **2**, 3–26. University of Moscow.

[41] Srebro, N., Sridharan, K., and Tewari, A. (2010). Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. 2199–2207.

[42] Sriperumbudur, B. (2011). Mixture density estimation via Hilbert space embedding of measures. In *Proceedings of International Symposium on Information Theory*. 1027–1030.

[43] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., and Schölkopf, B. (2008). Injective Hilbert space embeddings of probability measures. In *Proc. of the 21st Annual Conference on Learning Theory*, R. Servedio and T. Zhang, Eds. 111–122.

[44] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* **11**, 1517–1561. MR2645460

[45] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer. MR2450103

[46] Suquet, C. (2009). Reproducing kernel Hilbert spaces and random measures. In *Proc. of the 5th International ISAAC Congress, Catania, Italy, 25-30 July 2005*, H. G. W. Begehr and F. Nicolosi, Eds. World Scientific, 143–152.

[47] Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publishers, Boston.

[48] Vallander, S. S. (1973). Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.* **18**, 784–786.

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

[50] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York. MR1385671

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

[52] Wang, Q., Kulkarni, S. R., and Verdú, S. (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Information Theory* **51**, 9, 3064–3074. MR2239136

[53] Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK. MR2131724

[54] Whitney, H. (1934). Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society* **36**, 63–89. MR1501735

[55] Zolotarev, V. M. (1983). Probability metrics. *Theory of Probability and its Applications* **28**, 278–302. MR0700210