Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 138
DOI: 10.11159/icsta23.138

Lipschitz Variational Approximation of Total Variation Distance

Rui Ding*
*Department of Applied Mathematics and Statistics, Stony Brook University
Stony Brook, NY, USA, 11794
rui.ding.1@stonybrook.edu

Abstract – In this work, we build on the family of Integral Probability Metrics and design a new distance metric between probability distributions that belong to this family. The new metric is termed Lipschitz Variational Total Variation Distance and is a relaxation of the integral probability metric representation of the well-known total variation distance. We propose simple procedures to estimate this distance metric and demonstrate its convergence. Based on the Lipschitz smoothness of the proposed metric family, the proposed metrics, hence its empirical estimate, can provide meaningful and tight lower bounds for the total variation distance between two probability distributions. Finally, we extend our results to general measures and provide an application of the proposed estimators to bounding the Neyman-Pearson region.

Keywords: total variation distance; statistical distance; integral probability metric; statistical estimation

1. INTRODUCTION

Given samples from two unknown probability measures, P and Q, it is often of interest to estimate the distance (or divergence) between them. Two well-known families of such distances are the integral probability metrics (IPM) and the f-divergences. Both contain many well-known probability distances and divergences, among which the total variation distance (TVD) is the only intersection of the two families. Since it is well known from previous works (see for example, [1], [2]) that TVD cannot be estimated by simple procedures that work well on other instances from either family of probability distances (divergences), a variational approach is needed to bound TVD by some other distances and estimate the chosen bounding distance metric instead. To this end, we consider the family of IPMs and design a new metric in this family that closely approximates TVD but by relaxing the constraint on the family of functions, we obtain simple and provably convergent estimators based on empirical samples. There is a large literature on variational approximations and estimations for general f-divergences (see for example, [3], [4]) and inequalities between f-divergences and integral probability metrics (see for example, [5], [6]), although we discuss the IPM point of view only in this work.

We begin by considering the family of IPMs, which for two probability distributions P, Q defined on the metric space $(\mathcal{S}, \rho)$, can be defined as,

$$\gamma_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{S}} f\, dP - \int_{\mathcal{S}} f\, dQ \right| \tag{1}$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on $\mathcal{S}$, see [2], [7], [8]. Some classic distance metrics can be recovered under various choices of the class of functions, in particular:

(a) When $\mathcal{F}_W = \{f : ||f||_L \le 1\}$, $\gamma_{\mathcal{F}_W}(P,Q) = W_1(P,Q)$ is the Wasserstein-1 distance, which is also known by its dual as the Kantorovich metric. Here $||f||_L$ is the Lipschitz norm defined as,

$$||f||_L = \sup\{\tfrac{f(x)-f(y)}{\rho(x,y)} : x, y \in \mathcal{S}, x \neq y\} \tag{2}$$

(b) When $\mathcal{F}_\beta = \{f : ||f||_{BL} \le 1\}$, $\gamma_{\mathcal{F}_\beta}(P,Q) = \beta(P,Q)$ is the Dudley metric, or the dual-bounded Lipschitz distance, where $||f||_{BL} = ||f||_\infty + ||f||_L$ with the maximal norm defined as,

$$||f||_\infty = \sup\{|f(x)| : x \in \mathcal{S}\} \tag{3}$$

Dudley metric [9] metrizes weak convergence.

(c) When $\mathcal{F}_{TVD} = \{f : ||f||_\infty \le 1\}$, $\gamma_{\mathcal{F}_{TVD}}(P,Q) = TVD(P,Q)$ is the total variation distance. The optimal f^* always attain boundary values of $\{1, -1\}$ and the TVD can be equivalently written as,

$$TVD(P, Q) = \int_{x \in \mathcal{S}} |p(x) - q(x)| \, dx \tag{4}$$

where p, q are probability density functions of P, Q respectively.

Next, we develop a new distance metric within the IPM family, called Lipschitz Variational TVD (LV-TVD), and discuss its properties. We then discuss estimators for the proposed distance metric.

2. LIPSCHITZ VARIATIONAL TOTAL VARIATION DISTANCE

Consider the following function class $\mathcal{F}_{LVD}^{l} = \{f: ||f||_{\infty} \leq 1, ||f||_{L} \leq l\}$. Then the resulting IPM is called the Lipschitz variational Total Variation Distance (LV-TVD), which is,

$$\gamma_{LVD}^{l}(P, Q) = \sup_{f \in \mathcal{F}_{LVD}^{l}} \left| \int_{\mathcal{S}} f \, dP - \int_{\mathcal{S}} f \, dQ \right| \tag{5}$$

This represents a one-parameter family of IPMs where $l > 0$ is a Lipschitz smoothness parameter that controls the Lipschitzness of the chosen function class, which is also constrained to be maximally bounded by 1. We obtain some direct connections to well-known distances in the following theorem and lemmas.

Theorem 1. *For any probability distribution P, Q defined on metric space $(\mathcal{S}, \rho)$:*
(1) $\forall l > 0, \gamma_{LVD}^{l}(P, Q) \leq TVD(P, Q)$.
(2) $\forall l_1 \geq l_2 > 0, \gamma_{LVD}^{l_2}(P, Q) \leq \gamma_{LVD}^{l_1}(P, Q)$, and $\lim_{l \to \infty} \gamma_{LVD}^{l}(P, Q) = TVD(P, Q)$.
(3) When $l = 1$, $\beta(P, Q) \leq \gamma_{LVD}^{1}(P, Q) \leq W_1(P, Q)$.

Lemma 2. *γ_{LVD}^{l} is a metric on the space of probability distributions $\mathcal{P}(\mathcal{S}), \forall l > 0$.*

Lemma 3. *$\{\gamma_{LVD}^{l}\}_{l>0}$ metrizes weak convergence topology in the sense that $\forall l > 0$ and $\forall P, Q$ on $(\mathcal{S}, \rho)$,*

$$\min(1, l)\beta(P, Q) \leq \gamma_{LVD}^{l}(P, Q) \leq (1 + l)\beta(P, Q) \tag{6}$$

Lemma 4. *For any probability distributions P, Q defined on metric space $(\mathcal{S}, \rho)$ and $\forall c, l > 0$,*

$$\sup_{f \in \{f: ||f||_L \leq l, ||f||_{\infty} \leq c\}} \left| \int_{\mathcal{S}} f \, dP - \int_{\mathcal{S}} f \, dQ \right| = c \gamma_{LVD}^{l/c}(P, Q) \tag{7}$$

3. EMPIRICAL ESTIMATOR FOR LIPSCHITZ VARIATIONAL TVD

Similar to [2], we consider an empirical estimator for $\gamma_{LVD}^{l}(P, Q)$ which estimates $\gamma_{LVD}^{l}(P_m, Q_n)$ based on empirical distributions P_m, Q_n of P, Q where the empirical distributions are defined over finite samples $\{X_i^{(1)}\}_{i=1}^{m} \sim P$ and $\{X_j^{(2)}\}_{j=1}^{n} \sim Q$ with sample sizes m, n which may be potentially different. Let's make the following definitions similar to [2]. Let $\{X_i\}_{i=1}^{N}$ where $N = m + n$, $X_i = X_i^{(1)}, \forall i = 1, \dots, m$ and $X_i = X_i^{(2)}, \forall i = m + 1, \dots, N$. Define $\{\tilde{Y}_i\}_{i=1}^{N}$ such that $\tilde{Y}_i = \frac{1}{m}, \forall i = 1, \dots, m$ and $\tilde{Y}_i = -\frac{1}{n}, \forall i = m + 1, \dots, N$. The Lipschitz variational TVD for the empirical distributions P_m, Q_n is defined through the following maximization problem, for a chosen $l > 0$:

$$\gamma_{LVD}^{l}(P_m, Q_n) = \sup_{f \in \mathcal{F}_{LVD}^{l}} \left| \sum_{i=1}^{N} \tilde{Y}_i f(X_i) \right| = \sup_{f \in \mathcal{F}_{LVD}^{l}} \sum_{i=1}^{N} \tilde{Y}_i f(X_i) \tag{8}$$

where the second equality is true because the family $\mathcal{F}_{LVD}^{l}$ is closed under negation. The empirical LV-TVD can be obtained from the optimal objective value of the following linear programming (LP) problem,

$$\begin{aligned} \gamma_{LVD}^{l}(P_m, Q_n) \quad &= \max_{a_1, \dots, a_N} \sum_{i=1}^{N} \tilde{Y}_i a_i \\ s.t. \quad &-l\rho(X_i, X_j) \leq a_i - a_j \leq l\rho(X_i, X_j), \forall i, j = 1, \dots, N \\ &-1 \leq a_i \leq 1, \forall i = 1, \dots, N \end{aligned} \tag{9}$$

The two constraints in (9) translate the Lipschitz and maximal bounds on functions in the class $\mathcal{F}^{l}_{LVD}$. The LP in (9) can have many redundant constraints, hence we use a reduction of these constraints that makes the problem much faster to solve, especially in 1-D data setting where data points are ordered (details omitted). Let the optimal solution of (9) be denoted by $\{a_i^*\}_{i=1}^N$. By Lipschitz extension theorem and relevant results for Wasserstein-1 distance and Dudley metric (see [2], [9], [10]), the optimal function f^* in (9) is an extension of the optimal solution of the LP: $f^*(X_i) = a_i^*, \forall i = 1, \dots, N$. The resulting solution of the LP gives the LV-TVD distance of the empirical distributions as,

$$\gamma^{l}_{LVD}(P_m, Q_n) = \sum_{i=1}^{N} \tilde{Y}_i \, a_i^* \tag{10}$$

Theorem 5. *For any given $l > 0$ and samples from probability distributions P, Q, in the limit $m, n \to \infty$, the empirical LV-TVD distance estimator $\gamma^{l}_{LVD}(P_m, Q_n)$ converges almost surely to the true LV-TVD distance $\gamma^{l}_{LVD}(P, Q)$ in the sense that:* $\lim_{m,n\to\infty} \left|\gamma^{l}_{LVD}(P_m, Q_n) - \gamma^{l}_{LVD}(P, Q)\right| = 0$ *a.s.*

Based on the above theorem of convergence, we can get a consistent estimator of the desired distance metric $\gamma^{l}_{LVD}(P, Q)$ for any two probability distributions P, Q based on two sample estimates $\gamma^{l}_{LVD}(P_m, Q_n)$. The properties in Theorem 1 make it obvious that the proposed estimators $\gamma^{l}_{LVD}(P_m, Q_n)$ can serve as tight estimators of lower bounds for the TVD, given that sufficient samples are available and the parameter l is chosen to be large enough. [2] discussed why an empirical TVD estimator of the form in (8) is not consistent due to the function class of TVD not being Lipschitz continuous. The LV-TVD function class serves as a close relaxation of the TVD function class while being Lipschitz and bounded, hence guaranteeing consistency of the estimator in (8), (9). We point out that the estimator in (8), (9) is translation invariant, which is also a property of the LV-TVD distance metric γ^{l}_{LVD} itself.

4. NUMERICAL EXPERIMENTS

The convergence behavior of the proposed estimators is demonstrated via the following experiments, see Figure 1.

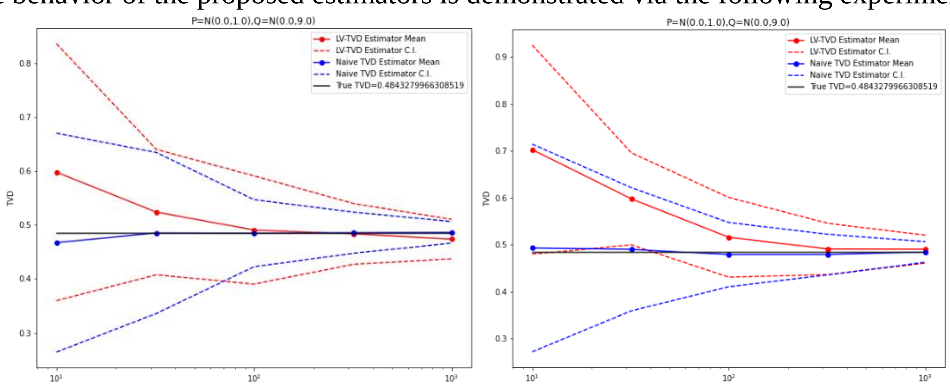

Fig. 1: Empirical LV-TVD Estimator for P = N(0, 1), Q = N(0, 9), (a): $l = 2$; (b): $l = 4$

In all our experiments in this section, the ground-truth TVD values (computed via numerical integration), a naive TVD estimator based on estimating the mean and variance of each data sample first and plugging these estimates into the numeric integration procedure for TVD distance between Gaussians, and the LV-TVD estimator values are divided by a factor of 2 to adhere to the conventional definition of TVD which can be shown to be bounded between 0 and 1. As the sample size $m = n = 10, 32, 100, 320, 1000$ increases, our estimator asymptotically converges to a value close to the true TVD value, which is an upper bound for $\gamma^{l}_{LVD}(P, Q)$. Each choice of sample size is run for 20 iterations so as to obtain both a mean and a confidence interval of the estimator results which demonstrates the estimator convergence. In particular, we observe that the estimator results for larger l are generally upper bounding those for smaller l for the same setting.

Besides pairs of Gaussians, another simpler example we considered is when $P = Exp(\lambda)$ is an exponential distribution and $Q = U(0,1)$ is a uniform distribution. Here $\lambda > 0$ is the rate parameter of an exponential distribution. $TVD(P, Q)$ for this case has an analytic solution. Specifically, $TVD(P, Q) = \int_x |p(x) - q(x)| \, dx$ can be explicitly computed as:

$$TVD(P,Q) = 2\left(1 - \frac{1}{\lambda} - \frac{\log\lambda}{\lambda} + e^{-\lambda}\right), \lambda \geq 1 \tag{11}$$

$$TVD(P,Q) = 2e^{-\lambda}, 0 < \lambda \leq 1 \tag{12}$$

For $0 < \lambda < 1$, we can exactly compute $\gamma_{LVD}^{l}(P,Q)$ through a maximization problem with respect to the location of the piecewise linear identifier function f. Take for example the case where $l = 2$, the only point of intersection of the density functions p(x) = $\lambda e^{-\lambda x}$,q(x) = $\mathbf{1}[0,1]$ is at $x = 1$. The optimal identifier function will be a piecewise linear function $f_z(x)$ with some $z \in [0,1]$ so that $f_z(x) = -1, -\infty < x \leq z$, $f_z(x) = -1 + 2(x - z), z < x < 1 + z$, and $f_z(x) = 1, 1 + z \leq x < \infty$. We can exactly integrate $\int_x f_z\,(x)\big(p(x) - q(x)\big)dx$ under the choice of identifier function parametrized by a single number $z \in [0,1]$ and the optimal z results in the LV-TVD value as the optimal objective of the IPM's maximization problem:

$$\gamma_{LVD}^{l=2}(P,Q) = \max_{z\in[0,1]} \{\tfrac{2}{\lambda}(1 - e^{-\lambda})e^{-\lambda z} - (z-1)^2\}, \forall 0 < \lambda \leq 1 \tag{13}$$

For example, when $\lambda = 0.5 < 1$ and $l = 2$, the optimal solution of the problem (13) is approximately $z^* \approx 0.726357$ and the optimal solution is approximately $\gamma_{LVD}^{l=2}(P,Q) \approx 1.01969$. The optimal identifier function is plotted in Figure 2(a). Notice that the TVD value for this pair of distributions can be explicitly computed via formula (12) as $TVD(P,Q) = 2e^{-0.5} \approx 1.21306$. The intrinsic gap between $\gamma_{LVD}^{l=2}$ and TVD for this pair of distributions is about 0.19337. For a fixed $l > 0$, this gap is instance dependent.

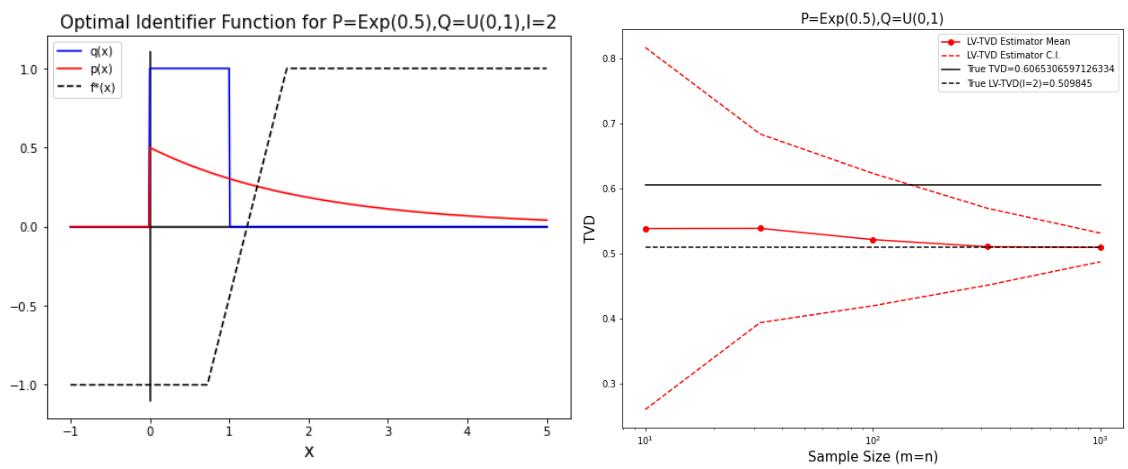

Fig. 2: Numerical Example for $P = Exp(0.5)$, $Q = U(0,1)$, $l = 2$, (a): Optimal Identifier Function; (b): Empirical LV-TVD Estimator

We performed the LV-TVD estimator with $l = 2$ for data samples from $P = Exp(0.5)$ and $Q = U(0,1)$, where sample size grows as $m = n = 10, 32, 100, 320, 1000$ and 20 iterations are performed for each setting, see Figure 2(b). In addition to the ground-truth TVD value, which is plotted in a solid black line, we add a dashed black line that indicates the ground-truth LV-TVD value analytically computed, both divided by 2 by convention. Clearly, the LV-TVD estimator is converging asymptotically to the ground-truth LV-TVD value $\gamma_{LVD}^{l=2}(P,Q)$, which provides a lower bound of $TVD(P,Q)$ for this choice of P, Q.

5. EXTENSION TO GENERAL MEASURES

Results in sections 2 and 3 can be readily extended to general measures P, Q which integrates to $0 < s, t < \infty$ respectively over the metric space $(\mathcal{S}, \rho)$. Equation (5) still defines a distance metric for these two measures, and we can prove similar results to Theorem 1 on the space of general measures P, Q. In particular, we remark the following relationship still holds:

$$\gamma_{LVD,s,t}^{l}(P,Q) = \sup_{f\in\mathcal{F}_{LVD}^{l}} \left|\int_{\mathcal{S}} f\,dP - \int_{\mathcal{S}} f\,dQ\right| \leq \sup_{f\in\mathcal{F}_{TVD}} \left|\int_{\mathcal{S}} f\,dP - \int_{\mathcal{S}} f\,dQ\right| = TVD_{s,t}(P,Q) \tag{14}$$

where we have defined the generalized TVD distance as $TVD_{s,t}(P,Q)$. Similar to section 2, we can show that the proposed distance metric is indeed a metric on general measures P, Q.

To estimate the quantity $TVD_{s,t}(P,Q) = \sup_{f \in \mathcal{F}_{TVD}} \left| \int_{\mathcal{S}} f\, dP - \int_{\mathcal{S}} f\, dQ \right|$ we resolve to the same estimator of the form in Equation (8), where for two empirical measures P_m, Q_n that sums to s, t, we now define $\{\tilde{Z}_i\}_{i=1}^N$ such that $\tilde{Z}_i = \frac{s}{m}, \forall i = 1, \dots, m$ and $\tilde{Z}_i = -\frac{t}{n}, \forall i = m+1, \dots, N$. Define, $\forall l > 0$:

$$\gamma_{LVD,s,t}^{l}(P_m, Q_n) = \sup_{f \in \mathcal{F}_{LVD}^{l}} \left| \sum_{i=1}^{N} \tilde{Z}_i\, f(X_i) \right| = \sup_{f \in \mathcal{F}_{LVD}^{l}} \sum_{i=1}^{N} \tilde{Z}_i\, f(X_i) \tag{15}$$

where the second equality is true because the family $\mathcal{F}_{LVD}^{l}$ is closed under negation. The empirical LV-TVD can be obtained from the optimal objective value of a similar linear program as in (9), which we omit here. The empirical measures P_m, Q_n converge almost surely in the weak sense to P, Q, hence we arrive at the following theorem, which states the convergence of the proposed estimators.

Theorem 6. *For any given $l > 0$ and samples from general measures P, Q (that integrate to s, t respectively), in the limit $m, n \to \infty$, the empirical LV-TVD distance estimator $\gamma_{LVD,s,t}^{l}(P_m, Q_n)$ converges almost surely to $\gamma_{LVD,s,t}^{l}(P, Q)$ in the sense that:* $\lim_{m,n\to\infty} \left| \gamma_{LVD,s,t}^{l}(P_m, Q_n) - \gamma_{LVD,s,t}^{l}(P, Q) \right| = 0$ *a.s.*

Without loss of generality, let P, Q have densities sp, tq respectively for some constants $0 < s, t < \infty$, where p, q are probability densities that integrate to 1 over $x \in \mathcal{S}$. We arrive at the following inequality:

$$\gamma_{LVD,s,t}^{l}(P, Q) \le TVD_{s,t}(P, Q) = \sup_{f \in \mathcal{F}_{LVD}^{l}} \left| \int_{\mathcal{S}} f\, dP - \int_{\mathcal{S}} f\, dQ \right| = \int_{x \in \mathcal{S}} |sp(x) - tq(x)|\, dx \tag{16}$$

Hence the proposed estimators in (15) will provide an approximate lower bound of the quantity (which we termed generalized TVD) $TVD_{s,t}(P,Q) = \int_{x \in \mathcal{S}} |sp(x) - tq(x)|\, dx$ for any two probability densities p, q, given only samples from them that have large enough size. As the choice of parameter l increases and data size increases, the approximation tends to be more accurate. Similar to the case for standard TVD estimators, these generalized estimators are also translation invariant.

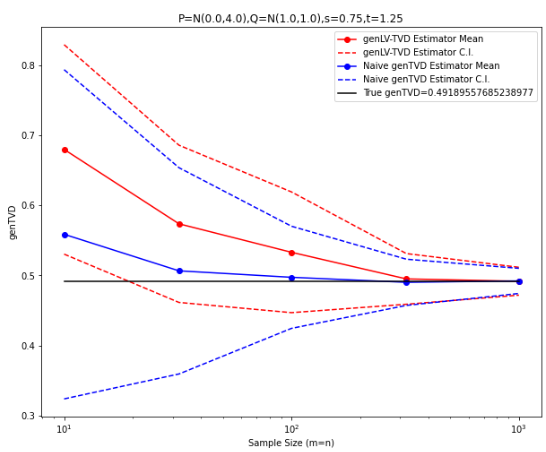

Fig. 3: Empirical LV-TVD Estimator for General Measures, $s = 0.75, t = 1.25, l = 3$

As a demonstration, consider P, Q which have densities sp, tq respectively where p, q are densities of two normal distributions $N(0,4), N(1,1)$. In Figure 3, we choose $s = 0.75, t = 1.25$ and show the convergence of the proposed estimator $\gamma_{LVD,s,t}^{l}(P_m, Q_n)$ for $l = 3$ (same setting and convention as before).

6. APPLICATION TO BOUNDING THE NEYMAN-PEARSON REGION

We propose to use the estimators in (15) to approximate the Neyman-Pearson region. For two probability distributions with densities p, q, and for $s, t > 0$ such that $s + t = 2$, the Neyman-Pearson region for type I ($\alpha(E)$) and type II ($\beta(E)$) errors (for distribution by q) satisfy (17) for an optimal choice of event $E^\star$ (see [1]):

$$s\alpha(E^\star) + t\beta(E^\star) = \frac{t+s}{2} - \frac{1}{2}\int_x |sp(x) - tq(x)|\, dx = \frac{t+s}{2} - \frac{1}{2} TVD_{s,t}(P, Q) \tag{17}$$

and can hence be bounded by the following inequality:

$$s\alpha(E^\star) + t\beta(E^\star) \lessapprox \frac{t+s}{2} - \frac{1}{2}\gamma^l_{LVD,s,t}(P, Q), \forall l > 0 \tag{18}$$

As a demonstration, we consider a case where p, q are densities of two normal distributions $N(0,4), N(1,1)$. In Figure 4 we replicate similar experiments as shown in Figure 3 for some chosen pairs of s, t values and plot the resulting generalized LV-TVD estimator mean value in red against the chosen s values, where $t = 2 - s$. (Here we consider sample size $m = n = 1000$ and Lipschitz parameter $l = 3$. For each of the 20 iterations, the same sampled data is used by all choices of (s, t) pairs, and the mean is taken over the 20 iterations for each choice of (s, t). Estimated values and the ground-truth generalized TVD values in the figure are divided by a factor of 2 according to convention.)

For the endpoints $s = 2, t = 0$ and $s = 0, t = 2$, the empirical estimator $\gamma^l_{LVD,s,t}(P_m, Q_n)$ by default produces the exact generalized TVD value of 2, a result that is independent of the choice of measures P, Q and the choice of l. The ground-truth generalized TVD values $TVD_{s,t}(P, Q) = \int_{x\in\mathcal{S}} |sp(x) - tq(x)|\, dx$ are computed via numerical integration for each pair of s, t values and plotted in blue together with the mean estimated values in Figure 4. Following (18), the estimated values can be used to approximate the true Neyman-Pearson region with a sufficient sample size. Notice that the results in Figure 4 demonstrate the estimated values are close enough to the true values of interest, and so is our approximation of $s\alpha(E^\star) + t\beta(E^\star)$ by $1 - \frac{1}{2}\gamma^l_{LVD,s,t}(P_m, Q_n)$, which is always bounded between 0 and 1, for $s + t = 2$ and $s, t > 0$. The Neyman-Pearson region's lower boundary can be directly computed from the estimated values by finding the convex hull above the supporting hyperplanes $s\alpha + t\beta = 1 - \frac{1}{2}\gamma^l_{LVD,s,t}(P_m, Q_n)$.

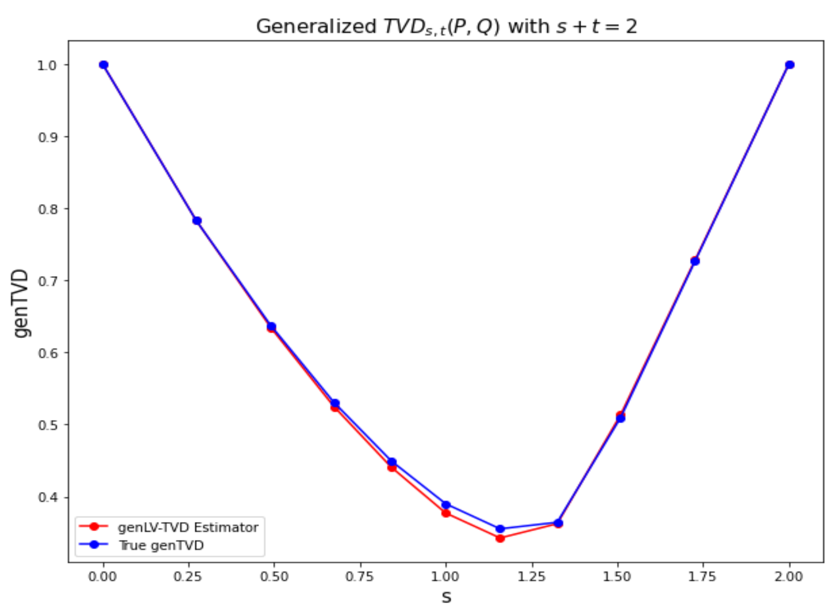

Fig. 4: Empirical LV-TVD Estimator ($l = 3$) for General Measures and Varying $s + t = 2$

7. CONCLUSIONS

In this paper, we proposed LV-TVD, a variational lower bound for total variation distance from the perspective of an integral probability metric family. We discussed properties of the proposed distance metrics as relating to well-known IPMs such as total variation distance, Wasserstein distance, and Dudley metric. We developed a consistent estimator for this proposed distance metric on the space of probability distributions and generalized it to the space of general measures. Our numerical results indicate that the proposed family of estimators provides good approximations to the true total variation distance between the underlying distributions of two given finite data samples. As an application, the extension of our estimation procedure to general measures can provide effective bounds to the Neyman-Pearson region. As an extension, we propose to consider in future works data-adaptive choices of the Lipschitz parameter l to enhance the performance of the proposed estimators and provide a tighter bound for TVD.

Our proposed estimators see applications in settings where we need to quantify the difference between two data samples, and it naturally fits the goal of quantifying distribution shifts between different time periods where the samples are taken. Hence it can be applied to many real-world problems where stability of the distribution over time needs to be quantified. For example, consider Explainability Index (EI) introduced in [11] which evaluates asset performance over a historic period balancing different categories of performance measures according to specified preferences, and can be applied to security selection. In the computation of EI, the distribution shift score is an important input risk component and can be computed by either approximating the empirical distributions with known densities or alternatively using our empirical estimators to directly estimate the total variation distances between data samples.

Acknowledgements
The author thanks Dr. Andrew Mullhaupt for valuable comments and suggestions.

References
[1] R. Ding and A. Mullhaupt, "Empirical Squared Hellinger Distance Estimator and Generalizations to a Family of α-Divergence Estimators," *Entropy*, 2023, vol. 25, no. 4, pp. 612-635.

[2] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, "On the Empirical Estimation of Integral Probability Metrics," *Electronic Journal of Statistics*, 2012, vol. 6, pp. 1550-1599.

[3] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization," *IEEE Transactions on Information Theory*, 2010, vol. 56, no. 11, pp. 5847–5861.

[4] A. Ruderman, M. Reid, D. Garcia-Garcia, and J. Petterson, "Tighter Variational Representations of f-Divergences via Restriction to Probability Measures," in *Proceedings of the 29th International Conference on Machine Learning*, 2012, vol. 1, pp. 1155–1162.

[5] R. Agrawal and T. Horel, "Optimal Bounds between f-Divergences and Integral Probability Metrics," *Journal of Machine Learning Research*, 2021, vol. 22, pp. 1-59.

[6] I. Sason and S. Verdú, "f-Divergence Inequalities," *IEEE Transactions on Information Theory*, 2016, vol. 62, no. 11, pp. 5973–6006.

[7] A. Muller, "Integral Probability Metrics and their Generating Classes of Functions," *Advances in Applied Probability*, 1997, vol. 29, pp. 429–443.

[8] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, "Non-parametric Estimation of Integral Probability Metrics," in *2010 IEEE International Symposium on Information Theory*, Austin, TX, USA, 2010, pp. 1428-1432.

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

[10] E. J. McShane, "Extension of Range of Functions," *Bulletin of the American Mathematical Society*, 1934, vol. 40, pp. 837–842.

[11] A. Hirsa, R. Ding., and S. Malhotra, "Explainability Index (EI): Unifying Framework of Performance Measures and Risk of Target (RoT): Variability from Target EI," available at *SSRN: https://ssrn.com/abstract=4335455*, 2023.