Distances Between Probability Distributions of Different Dimensions

Yuhang Cai and Lek-Heng Lim, *Senior Member, IEEE*

***Abstract*— Comparing probability distributions is an indispensable and ubiquitous task in machine learning and statistics. The most common way to compare a pair of Borel probability measures is to compute a metric between them, and by far the most widely used notions of metric are the Wasserstein metric and the total variation metric. The next most common way is to compute a divergence between them, and in this case almost every known divergences such as those of Kullback–Leibler, Jensen–Shannon, Rényi, and many more, are special cases of the f-divergence. Nevertheless these metrics and divergences may only be computed, in fact, are only defined, when the pair of probability measures are on spaces of the same dimension. How would one quantify, say, a KL-divergence between the uniform distribution on the interval $[-1, 1]$ and a Gaussian distribution on $\mathbb{R}^3$? We show that these common notions of metrics and divergences give rise to natural distances between Borel probability measures defined on spaces of different dimensions, e.g., one on $\mathbb{R}^m$ and another on $\mathbb{R}^n$ where m, n are distinct, so as to give a meaningful answer to the previous question.**

***Index Terms*— Probability densities, probability measures, Wasserstein distance, total variation distance, KL-divergence, Rényi divergence.**

I. Introduction

MEASURING a distance, whether in the sense of a metric or a divergence, between two probability distributions is a fundamental endeavor in machine learning and statistics. We encounter it in clustering [1], density estimation [2], generative adversarial networks [3], image recognition [4], minimax lower bounds [5], and just about any field that undertakes a statistical approach towards data. It is well-known that the space of Borel probability measures on a measurable space $\Omega \subseteq \mathbb{R}^n$ may be equipped with many different metrics and divergences, each good for its own purpose, but two of the most common families are the p-Wasserstein metric

$$\mathrm{W}_p(\mu, \nu) \coloneqq \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} \|x - y\|_2^p \, d\gamma(x, y) \right]^{1/p}$$

and the f-divergence

$$\mathrm{D}_f(\mu \| \nu) \coloneqq \int_\Omega f\Big(\frac{d\mu}{d\nu}\Big) d\nu.$$

For $p = 1$ and 2, the p-Wasserstein metric gives the Kantorovich metric (also called earth mover's metric) and Lévy–Fréchet metric respectively. Likewise, for various choices of f, we obtain as special cases the Kullback–Liebler, Jensen–Shannon, Rényi, Jeffreys, Chernoff, Pearson chi-squared, Hellinger squared, exponential, and alpha–beta divergences, as well as the total variation metric (see Table I). Nevertheless, a p-Wasserstein metric cannot be expressed as an f-divergence.

All these distances are only defined when μ and ν are probability measures on a common measurable space $\Omega \subseteq \mathbb{R}^n$. This article provides an answer to the question:

> How can one define a distance between μ, a probability measure on $\Omega_1 \subseteq \mathbb{R}^m$, and ν, a probability measure on $\Omega_2 \subseteq \mathbb{R}^n$, where $m \neq n$?

We will show that this problem has a natural solution that works for any of the aforementioned metrics and divergences in a way that is consistent with recent extensions of distances to inequidimensional covariance matrices [6] and subspaces [7]. Although we will draw from the same high-level ideas in [6], [7], we require substantially different techniques in order to work with probability measures.

Given a p-Wasserstein metric or an f-divergence, which is defined between two probability measures of the same dimension, we show that it naturally defines *two* different distances for probability measures μ and ν on spaces of different dimensions — we call these the *embedding distance* and *projection distance* respectively. Both these distances are completely natural and are each befitting candidates for the distance we seek; the trouble is that there is not one but two of them, both equally reasonable. The punchline, as we shall prove, is that the two distances are always equal, giving us a unique distance defined on inequidimensional probability measures. We will state this result more precisely after introducing a few notations.

To the best of our knowledge — and we have one of our referees to thank for filling us in on this — the only alternative for defining a distance between probability measures of different dimensions is the *Gromov–Wasserstein distance* proposed in [8]. As will be evident from our description below, we adopt a 'bottom-up' approach that begins from first principles and requires nothing aside from the most basic definitions. On the other hand, the approach in [8] is a 'top-down' one by adapting

Manuscript received November 11, 2020; revised November 23, 2021; accepted January 26, 2022. Date of publication February 2, 2022; date of current version May 20, 2022. This work was supported in part by DARPA under Grant HR00112190040, in part by NSF IIS under Grant 1546413, in part by NSF DMS under Grant 1854831, and in part by the Eckhardt Faculty Fund. *(Corresponding author: Lek-Heng Lim.)*

Yuhang Cai is with the Department of Statistics, The University of Chicago, Chicago, IL 60637 USA (e-mail: yuhangc@uchicago.edu).

Lek-Heng Lim is with the Computational and Applied Mathematics Initiative, The University of Chicago, Chicago, IL 60637 USA (e-mail: lekheng@uchicago.edu).

Communicated by O. Johnson, Associate Editor for Probability and Statistics.

Digital Object Identifier 10.1109/TIT.2022.3148923

0018-9448 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

the vastly more general and powerful Gromov–Hausdorff distance to a special case. Our construction works with a wide variety of common metrics and divergences mentioned in first paragraph. Although the work in [8] is restricted to the 2-Wasserstein metric, it is conceivable that the framework therein would apply more generally to other metrics as well; however, it is not obvious how the framework might apply to divergences given that the Gromov–Hausdorff approach requires a metric. In the one case that allows a comparison, namely, applying the two different constructions to the 2-Wasserstein metric to obtain distances on probability measures of different dimensions, they lead to different results. We are of the opinion that both approaches are useful although our simplistic approach is more likely to yield distances that have closed-form expressions or are readily computable, as we will see in Section VI; the Gromov–Wasserstein distance tends to be NP-hard [9] and closed-form expression are rare and not easy to obtain [10].

A. Main Result

Let $\mathrm{M}(\Omega)$ denote the set of all Borel probability measures on $\Omega \subseteq \mathbb{R}^n$ and let $\mathrm{M}^p(\Omega) \subseteq \mathrm{M}(\Omega)$ denote those with finite pth moments, $p \in \mathbb{N}$. For any $m, n \in \mathbb{N}$, $m \leq n$, we write

$$\mathrm{O}(m,n) \coloneqq \{V \in \mathbb{R}^{m\times n} : VV^{\mathsf{T}} = I_m\},$$

i.e., the Stiefel manifold of $m \times n$ matrices with orthonormal rows. We write $\mathrm{O}(n) \coloneqq \mathrm{O}(n,n)$ for the orthogonal group. For any $V \in \mathrm{O}(m,n)$ and $b \in \mathbb{R}^m$, let

$$\varphi_{V,b} : \mathbb{R}^n \to \mathbb{R}^m, \qquad \varphi_{V,b}(x) = Vx + b;$$

and for any $\mu \in \mathrm{M}(\mathbb{R}^n)$, let $\varphi_{V,b}(\mu) \coloneqq \mu \circ \varphi_{V,b}^{-1}$ be the pushforward measure. For simplicity, we write $\varphi_V \coloneqq \varphi_{V,0}$ when $b = 0$. More generally, for any measurable map $\varphi : \mathbb{R}^n \to \mathbb{R}^m$, we let $\varphi(\mu) \coloneqq \mu \circ \varphi$ denote the pushforward measure.

For any $m, n \in \mathbb{N}$, there is no loss of generality in assuming that $m \leq n$ for the remainder of our article. Our goal is to define a distance $d(\mu,\nu)$ for measures $\mu \in \mathrm{M}(\Omega_1)$ and $\nu \in \mathrm{M}(\Omega_2)$ where $\Omega_1 \subseteq \mathbb{R}^m$ and $\Omega_2 \subseteq \mathbb{R}^n$, and where by 'distance' we include both metrics and divergences. Again, there is no loss of generality in assuming that

$$\Omega_1 = \mathbb{R}^m, \qquad \Omega_2 = \mathbb{R}^n \tag{1}$$

since we may simply restrict our attention to measures supported on smaller subsets. Henceforth, we will assume (1). We call μ and ν an m- and n-dimensional measure respectively.

We begin by defining the projection and embedding of measures. These are measure theoretic analogues of the Schubert varieties in [7] and we choose notations similar to [7].

Definition 1: Let $m, n \in \mathbb{N}$, $m \leq n$. For any $\mu \in \mathrm{M}(\mathbb{R}^m)$ and $\nu \in \mathrm{M}(\mathbb{R}^n)$, the *embeddings* of μ into $\mathbb{R}^n$ are the set of n-dimensional measures

$$\Phi^+(\mu,n) \coloneqq \{\alpha \in \mathrm{M}(\mathbb{R}^n) : \varphi_{V,b}(\alpha) = \mu \text{ for some } V \in \mathrm{O}(m,n),\ b \in \mathbb{R}^m\};$$

and the *projections* of ν onto $\mathbb{R}^m$ are the set of m-dimensional measures

$$\Phi^-(\nu,m) \coloneqq \{\beta \in \mathrm{M}(\mathbb{R}^m) : \varphi_{V,b}(\nu) = \beta \text{ for some } V \in \mathrm{O}(m,n),\ b \in \mathbb{R}^m\}.$$

Let d be any notion of distance on $\mathrm{M}(\mathbb{R}^n)$ for any $n \in \mathbb{N}$. Define the *projection distance*

$$d^-(\mu,\nu) \coloneqq \inf_{\beta\in\Phi^-(\nu,m)} d(\mu,\beta)$$

and the *embedding distance*

$$d^+(\mu,\nu) \coloneqq \inf_{\alpha\in\Phi^+(\mu,n)} d(\alpha,\nu).$$

Both $d^-(\mu,\nu)$ and $d^+(\mu,\nu)$ are natural ways of defining d on probability measures μ and ν of different dimensions. The trouble is that they are just as natural and there is no reason to favor one or the other. Our main result, which resolves this dilemma, may be stated as follows.

Theorem 1: Let $m, n \in \mathbb{N}$, $m \leq n$. Let d be a p-Wasserstein metric or an f-divergence. Then

$$d^-(\mu,\nu) = d^+(\mu,\nu). \tag{2}$$

The common value in (2), denoted $\widehat{d}(\mu,\nu)$, defines a distance between μ and ν and serves as our answer to the question on page 4020. We will prove Theorem 1 for p-Wasserstein metric (Theorem 4) and for f-divergence (Theorem 5). Jensen–Shannon divergence (Theorem 6) and total variation metric (Theorem 7) require separate treatments since the definition of p-Wasserstein metric requires that μ and ν have finite pth moments and the definition of f-divergence requires that μ and ν have densities, assumptions that we do not need for Jensen–Shannon divergence and total variation metric. While the proofs of Theorems 4, 5, 6, and 7 follow a similar broad outline, the subtle details are different and depend on the specific distance involved.

An important departure from the results in [6], [7] is that in general $\widehat{d}(\mu,\nu) \neq d(\mu,\nu)$ when $m = n$ although the Gromov–Wasserstein distance [8] mentioned earlier also lacks this property. To see this, we state a more general corollary.

Corollary 1: Let $m, n \in \mathbb{N}$, $m \leq n$. Let d be a p-Wasserstein metric, a Jensen–Shannon divergence, a total variation metric, or an f-divergence. Then $\widehat{d}(\mu,\nu) = d^-(\mu,\nu) = d^+(\mu,\nu) = 0$ if and only if $\varphi_{V,b}(\nu) = \mu$ for some $V \in \mathrm{O}(m,n)$ and $b \in \mathbb{R}^m$.

Corollary 1 gives a necessary and sufficient condition for $\widehat{d}(\mu,\nu)$ to be zero, saying that this happens if only if the two measures μ and ν are rotated and translated copies of each other, modulo embedding in a higher-dimensional ambient space when $m \neq n$. For any d that is not rotationally invariant, we will generally have $\widehat{d}(\mu,\nu) \neq d(\mu,\nu)$ when $m = n$.

The discussion in the previous paragraph notwithstanding, the distance $\widehat{d}$ has several nice features. Firstly, it preserves certain well-known relations satisfied by the original distance d. For example, we know that for $p \leq q$, the p- and q-Wasserstein metrics satisfy $\mathrm{W}_p(\mu,\nu) \leq \mathrm{W}_q(\mu,\nu)$ for measures μ, ν of the same dimension; we will see in Corollary 2 that

$$\widehat{\mathrm{W}}_p(\mu,\nu) \leq \widehat{\mathrm{W}}_q(\mu,\nu)$$

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

for measures μ, ν of *different* dimensions. Secondly, as our construction applies consistently across a wide variety of distances, both metrics and divergences, relations between different types of distances can also be preserved. For example, the total variation metric and KL-divergence satisfy Pinker's inequality $d_{\mathrm{TV}}(\mu,\nu)^2 \leq 1/2\, \mathrm{D}_{\mathrm{KL}}(\mu\|\nu)$ for measures μ, ν of the same dimension; we will see in Corollary 3 that

$$\widehat{d}_{\mathrm{TV}}(\mu,\nu)^2 \leq \frac{1}{2}\widehat{\mathrm{D}}_{\mathrm{KL}}(\mu\|\nu)$$

for measures μ, ν of *different* dimensions. As another example, the Hellinger squared divergence and total variation metric satisfy $\mathrm{D}_{\mathrm{H}}(\mu,\nu)^2 \leq 2\ d_{\mathrm{TV}}(\mu,\nu) \leq \sqrt{2}\,\mathrm{D}_{\mathrm{H}}(\mu,\nu)$ for measures μ, ν of the same dimension; we will see in Corollary 4 that

$$\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu)^2 \leq 2\widehat{d}_{\mathrm{TV}}(\mu,\nu) \leq \sqrt{2}\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu)$$

for measures μ, ν of *different* dimensions.

Another advantage of our construction is that for some common distributions, the distance $\widehat{d}$ obtained often has closed-form expression or is readily computable,[1] as we will see in Section VI. In particular, we will have an explicit answer for the rhetorical question in the abstract: What is the KL-divergence between the uniform distribution on $[-1, 1]$ and a Gaussian distribution on $\mathbb{R}^3$?

B. Background

For easy reference, we remind the reader of two results.

Theorem 2 (Hahn Decomposition): Let Ω be a measurable space and μ be a signed measure on the σ-algebra $\Sigma(\Omega)$. Then there exist P and $N \in \Sigma(\Omega)$ such that

(i) $P \cup N = \Omega$, $P \cap N = \varnothing$;
(ii) any $E \in \Sigma(\Omega)$ with $E \subseteq P$ has $\mu(E) \geq 0$;
(iii) any $E \in \Sigma(\Omega)$ with $E \subseteq N$ has $\mu(E) \leq 0$.

The Disintegration Theorem [11] rigorously defines the notion of a nontrivial "restriction" of a measure to a measure-zero subset of a measure space. It is famously used to establish the existence of conditional probability measures.

Theorem 3 (Disintegration Theorem): Let Ω_1 and Ω_2 be two Radon spaces. Let $\mu \in \mathrm{M}(\Omega_1)$ and $\varphi : \Omega_1 \to \Omega_2$ be a Borel measurable function. Set $\nu \in \mathrm{M}(\Omega_2)$ to be the pushforward measure $\nu = \mu \circ \varphi^{-1}$. Then there exists a ν-almost everywhere uniquely determined family of probability measures $\{\mu_y \in \mathrm{M}(\Omega_1) : y \in \Omega_2\}$ such that

(i) the function $\Omega_2 \to \mathrm{M}(\Omega_1)$, $y \mapsto \mu_y$ is Borel measurable, i.e., for any measurable $B \subseteq \Omega_1$, $y \to \mu_y(B)$ is a measurable function of y;
(ii) $\mu_y\big(\Omega_1 \setminus \varphi^{-1}(y)\big) = 0$;
(iii) for every Borel-measurable function $f : \Omega_1 \to [0, \infty]$,

$$\int_{\Omega_1} f(x)\, d\mu(x) = \int_{\Omega_2} \int_{\varphi^{-1}(y)} f(x)\, d\mu_y(x)\, d\nu(y).$$

In this article, we use the terms 'probability measure' and 'probability distribution' interchangeably since given a cumulative distribution function F and $A \in \Sigma(\Omega)$, $\mu(A) \coloneqq \int_{x\in A} dF(x)$ defines a probability measure.

[1]To the extent afforded by the original distance d — if d has no closed-form expression or is NP-hard, we would not expect $\widehat{d}$ to be any different.

II. Wasserstein Metric

We begin by properly defining the p-Wasserstein metric, filling in some details left out in Section I. Given two measures $\mu, \nu \in \mathrm{M}^p(\mathbb{R}^n)$ and any $p \in [1, \infty]$, the *p-Wasserstein metric*, also called the L^p-Wasserstein metric, between them is

$$\mathrm{W}_p(\mu,\nu) \coloneqq \left[\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^{2n}} \|x-y\|_2^p\, d\gamma(x,y)\right]^{1/p} \tag{3}$$

where, as usual, $p = \infty$ is interpreted in the limiting sense of essential supremum. Here

$$\Gamma(\mu,\nu) \coloneqq \big\{\gamma \in \mathrm{M}(\mathbb{R}^{2n}) : \pi_1^n(\gamma) = \nu,\ \pi_2^n(\gamma) = \mu\big\}$$

is the set of *couplings* between μ and ν, where $\pi_1^n : \mathbb{R}^{2n} \to \mathbb{R}^n$ is the projection onto the first n coordinates and $\pi_2^n : \mathbb{R}^{2n} \to \mathbb{R}^n$ the projection to the last n coordinates. The measure $\pi \in \Gamma(\mu,\nu)$ that attains the minimum in (3) is called the *optimal transport coupling*. For the purpose of this article, we use the standard Euclidean metric $d_{\mathrm{E}}(x,y) = \|x-y\|_2$ but this may be replaced by other metrics and $\mathbb{R}^n$ by other metric spaces; in which case (3) is just called the *Wasserstein metric* or *transportation distance*. The general definition is due to Villani [12] but the notion has a long history involving the works of Fréchet [13], Kantorovich [14], Lévy [15], Wasserstein [16], and many others. As we mentioned earlier, the 1-Wasserstein metric is often called the earth mover's metric or Kantorovich metric whereas the 2-Wasserstein metric is sometimes called the Lévy–Fréchet metric [13].

The Wasserstein metric is widely used in the imaging sciences for capturing geometric features [17]–[19], with a variety of applications including contrast equalization [20], texture synthesis [21], image matching [22], [23], image fusion [24], medical imaging [25], shape registration [26], image watermarking [27]. In economics, it is used to match job seekers with jobs, determine real estate prices, form matrimonial unions, among many other things [28]. Wasserstein metric and optimal transport coupling also show up unexpectedly in areas from astrophysics, where it is used to reconstruct initial conditions of the early universe [29]; to computer music, where it is used to automate music transcriptions [30]; to machine learning, where it is used for machine translation [31] and word embedding [32].

Unlike the f-divergence in Section III, a significant advantage afforded by the Wasserstein distance is that it is finite even when neither measure is absolutely continuous with respect to the other. Our goal is to use W_p to construct a new distance $\widehat{\mathrm{W}}_p$ so that $\widehat{\mathrm{W}}_p(\mu,\nu)$ would be well-defined for $\mu \in \mathrm{M}(\mathbb{R}^m)$ and $\nu \in \mathrm{M}(\mathbb{R}^n)$ where $m \neq n$. Note that any attempt to directly extend the definition in (3) to such a scenario would require that we make sense of $\|x-y\|_2$ for $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ — our approach would avoid this conundrum entirely. We begin by establishing a simple but crucial lemma.

Lemma 1: Let $m, n \in \mathbb{N}$, $m \leq n$, and $p \in [1, \infty]$. For any $\alpha, \nu \in \mathrm{M}^p(\mathbb{R}^n)$, any $V \in \mathrm{O}(m,n)$, and any $b \in \mathbb{R}^m$, we have

$$\mathrm{W}_p\big(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)\big) \leq \mathrm{W}_p(\alpha,\nu).$$

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

Proof: Let $\gamma \in \mathrm{M}(\mathbb{R}^{2n})$ be the optimal transport coupling for $\mathrm{W}_p(\alpha, v)$. Consider the measurable map

$$\overline{\varphi}_{V,b} : \mathbb{R}^{2n} \to \mathbb{R}^{2m}, \quad (x, y) \mapsto \big(\varphi_{V,b}(x), \varphi_{V,b}(y)\big),$$

and define $\gamma_+ = \overline{\varphi}_{V,b}(\gamma)$. As $\varphi_{V,b} \circ \pi_i^n = \pi_i^m \circ \overline{\varphi}_{V,b}$,

$$\pi_1^m(\gamma_+) = \varphi_{V,b}(\alpha), \quad \pi_2^m(\gamma_+) = \varphi_{V,b}(\beta),$$

and thus $\gamma_+(x, y) \in \Gamma\big(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)\big)$. Now

$$\begin{aligned}\mathrm{W}_p(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)) &\le \int_{x\in\mathbb{R}^m,\, y\in\mathbb{R}^m} \|x - y\|_2^p \, d\gamma_+(x, y) \\ &= \int_{z\in\mathbb{R}^n,\, w\in\mathbb{R}^n} \|\varphi_{V,b}(z) - \varphi_{V,b}(w)\|_2^p \, d\gamma(z, w) \\ &\le \int_{z\in\mathbb{R}^n,\, w\in\mathbb{R}^n} \|z - w\|_2^p \, d\gamma(z, w),\end{aligned}$$

and taking pth root gives the result. The last inequality follows from $\|\varphi_{V,b}(z) - \varphi_{V,b}(w)\|_2 \le \|z - w\|_2$ as $\varphi_{V,b}$ is an orthogonal projection plus a translation. □

Lemma 1 assures that $\Phi^-(\mu, m) \subseteq \mathrm{M}^p(\mathbb{R}^m)$ but in general we may not have $\Phi^+(\nu, n) \subseteq \mathrm{M}^p(\mathbb{R}^n)$. With this in mind, we introduce the set

$$\begin{aligned}\Phi_p^+(\mu, n) \coloneqq \{&\alpha \in \mathrm{M}^p(\mathbb{R}^n) : \varphi_{V,b}(\alpha) = \mu \\ &\text{for some } V \in \mathrm{O}(m, n),\ b \in \mathbb{R}^m\}\end{aligned}$$

for use in the next result, which shows that projection and embedding Wasserstein distances are always equal.

Theorem 4: Let $m, n \in \mathbb{N}$, $m \le n$, and $p \in [1, \infty]$. For $\mu \in \mathrm{M}^p(\mathbb{R}^m)$ and $\nu \in \mathrm{M}^p(\mathbb{R}^n)$, let

$$\begin{aligned}\mathrm{W}_p^-(\mu, \nu) &\coloneqq \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{W}_p(\mu, \beta), \\ \mathrm{W}_p^+(\mu, \nu) &\coloneqq \inf_{\alpha\in\Phi_p^+(\mu,n)} \mathrm{W}_p(\alpha, \nu).\end{aligned}$$

Then

$$\mathrm{W}_p^-(\mu, \nu) = \mathrm{W}_p^+(\mu, \nu). \tag{4}$$

Proof: It is easy to deduce that $\mathrm{W}_p^-(\mu, \nu) \le \mathrm{W}_p^+(\mu, \nu)$: For any $\alpha \in \Phi^+(\mu, n)$, there exists $V_\alpha \in \mathrm{O}(m, n)$ and $b_\alpha \in \mathbb{R}^m$ with $\varphi_{V_\alpha, b_\alpha}(\alpha) = \mu$. It follows from Lemma 1 that $\mathrm{W}_p(\alpha, \nu) \ge \mathrm{W}_p\big(\mu, \varphi_{V_\alpha, b_\alpha}(\nu)\big)$ and thus

$$\begin{aligned}\inf_{\alpha\in\Phi^+(\mu,n)} \mathrm{W}_p(\alpha, \nu) &\ge \inf_{\alpha\in\Phi^+(\mu,n)} \mathrm{W}_p\big(\mu, \varphi_{V_\alpha, b_\alpha}(\nu)\big) \\ &\ge \inf_{V\in\mathrm{O}(m,n),\, b\in\mathbb{R}^m} \mathrm{W}_p\big(\mu, \varphi_{V,b}(\nu)\big).\end{aligned}$$

The bulk of the work is to show that $\mathrm{W}_p^-(\mu, \nu) \ge \mathrm{W}_p^+(\mu, \nu)$. Let $\varepsilon > 0$ be arbitrary. Then there exists $\beta_* \in \Phi^-(\nu, m)$ with

$$\mathrm{W}_p(\mu, \beta_*) \le \mathrm{W}_p^-(\mu, \nu) + \varepsilon.$$

Let $V_* \in \mathrm{O}(m, n)$ and $b_* \in \mathbb{R}^m$ be such that $\varphi_{V_*, b_*}(\nu) = \beta_*$ and $W_* \in \mathrm{O}(n - m, n)$ be such that $\begin{bmatrix} V_* \\ W_* \end{bmatrix} \in \mathrm{O}(n)$. Then φ_{W_*} is the complementary projection of φ_{V_*, b_*}. Applying Theorem 3 to φ_{V_*, b_*}, we obtain a family of measures $\{\nu_y \in \mathrm{M}(\mathbb{R}^n) : y \in \mathbb{R}^m\}$ that satisfy

$$\int_{\mathbb{R}^n} f(x) \, d\nu(x) = \int_{\mathbb{R}^m} \int_{\varphi_{V_*,b_*}^{-1}(y)} f(x) \, d\nu_y(x) \, d\beta_*(y)$$

for any measurable function f.

Let $\gamma \in \Gamma(\beta_*, \mu)$ be the optimal transport coupling attaining $\mathrm{W}_p(\beta_*, \mu)$. Then

$$\pi_1^m(\gamma) = \beta_*, \qquad \pi_2^m(\gamma) = \mu.$$

We define a new measure $\gamma_+ \in \mathrm{M}(\mathbb{R}^{2n})$ that will in turn give us a measure $\alpha_* \in \Phi_p^+(\mu, n)$ with $\mathrm{W}_p(\alpha_*, \nu) \le \mathrm{W}_p(\mu, \beta_*)$. Firstly, we will define an intermediate probability measure $\tilde{\gamma}$ in $\mathrm{M}(\mathbb{R}^{n+m})$. For any measurable set $S \subseteq \mathbb{R}^{n+m}$, we define

$$\tilde{\gamma}(S) \coloneqq \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{(x,z)\in S} \, d\nu_y(x) \, d\gamma(y, z)$$

where $\mathbb{I}$ denotes an indicator function. Consider the map

$$\rho : \mathbb{R}^{n+m} \to \mathbb{R}^{2n}, \quad (x, z) \mapsto \big(x, V_*^\intercal(z - b_*) + W_*^\intercal W_* x\big)$$

with $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$. Observe that this is an embedding of $\mathbb{R}^{n+m}$ into $\mathbb{R}^{2n}$. If we let $(x, y) = \rho\big((x, z)\big)$, then we find that $\varphi_{V_*,b_*}(y) = z$ and $\varphi_{W_*}(y) = \varphi_{W_*}(x)$. We define γ_+ to be the pushforward measure $\rho(\tilde{\gamma})$. Next we will prove that $\pi_1^n(\gamma_+) = \nu$. For any measurable set $S \subseteq \mathbb{R}^n$, we have

$$\begin{aligned}\pi_1^n(\gamma_+)(S) &= \gamma_+\big((\pi_1^n)^{-1}(S)\big) = \gamma_+(S \times \mathbb{R}^n) \\ &= \tilde{\gamma}\big(\rho^{-1}(S \times \mathbb{R}^n)\big) = \tilde{\gamma}(S \times \mathbb{R}^m) \\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{(x,z)\in S\times\mathbb{R}^m} \, d\nu_y(x) \, d\gamma(y, z) \\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{x\in S} \, d\nu_y(x) \, d\gamma(y, z) \\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{x\in S} \, d\nu_y(x) \, d\beta_*(y) = \nu(S).\end{aligned}$$

Note that the first four equalities follow from the definition of the pushforward measure. For the next-to-last equality, observe that the indicator function $\mathbb{I}_{x\in S}$ is only a function of y. Hence $\nu(x)$ is a marginal measure of γ_+. Let $\alpha_* \in \mathrm{M}(\mathbb{R}^n)$ be defined by $\alpha_* = \pi_2^n(\gamma_+)$. Then

$$\begin{aligned}&\int_{y\in\mathbb{R}^n} \|y\|_2^p \, d\alpha_*(y) = \int_{y\in\mathbb{R}^n} \big(\|y\|_2^2\big)^{p/2} \, d\alpha_*(y) \\ &= \int_{y\in\mathbb{R}^n} \big(\|\varphi_{V_*,b_*}(y) - b_*\|_2^2 + \|\varphi_{W_*}(y)\|_2^2\big)^{p/2} \, d\alpha_*(y) \\ &\le \max\{2^{\frac{p-2}{2}}, 1\} \int_{y\in\mathbb{R}^n} \|\varphi_{V_*,b_*}(y) - b_*\|_2^p + \|\varphi_{W_*}(y)\|_2^p \, d\alpha_*(y) \\ &= \max\{2^{\frac{p-2}{2}}, 1\} \bigg(\int_{y\in\mathbb{R}^m} \|y - b_*\|_2^p \, d\mu(y) \\ &\qquad + \int_{(x,y)\in\mathbb{R}^{2n}} \|\varphi_{W_*}(y)\|_2^p \, d\gamma_+(x, y) \bigg) \\ &\le \max\{2^{\frac{3p-4)}{2}}, 1\} \bigg(\int_{y\in\mathbb{R}^m} \|y\|_2^p \, d\mu(y) + \|b_*\|_2^p \\ &\qquad + \int_{x\in\mathbb{R}^n} \|\varphi_{W_*}(x)\|_2^p \, d\nu(x) \bigg) \\ &\le \max\{2^{\frac{3p-4}{2}}, 1\} \bigg(\int_{y\in\mathbb{R}^m} \|y\|_2^p \, d\mu(y) + \int_{x\in\mathbb{R}^n} \|x\|_2^p \, d\nu(x) \bigg).\end{aligned}$$

Some explanation is in order: In the first inequality we have used $(\sigma + \tau)^{p/2} \le \max\{2^{\frac{p-2}{2}}, 1\} \cdot (\sigma^{p/2} + \tau^{p/2})$; in the fourth equality we observe that the support of γ_+ is contained in the subspace $\{(x, y) : \varphi_{W_*}(x) = \varphi_{W_*}(y)\}$; in the fifth inequality

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

we have used $\|\sigma-\tau\|_2 \le \|\sigma\|_2+\|\tau\|_2$ and $(\sigma+\tau)^p \le 2^{p-1}(\sigma^p+\tau^p)$; and in the last inequality, $\|\varphi_{W_*}(x)\|_2 \le \|x\|_2$. Since the pth central moment of α_* is bounded by the pth central moments of μ and ν, we have $\alpha_* \in \mathrm{M}^p(\mathbb{R}^n)$. Finally,

$$\begin{aligned}\mathrm{W}_p^p(\alpha_*,\nu) &\le \int_{\mathbb{R}^{2n}} \|x-y\|_2^p \, d\gamma_+(x,y)\\ &= \int_{\mathbb{R}^{2n}} \|\varphi_{V_*,b_*}(x)-\varphi_{V_*,b_*}(y)\|_2^p \, d\gamma_+(x,y)\\ &= \int_{\mathbb{R}^{n+m}} \|\varphi_{V_*,b_*}(x)-z\|_2^p \, d\tilde{\gamma}(x,z)\\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \|\varphi_{V_*,b_*}(x)-z\|_2^p \, d\nu_y(x)\, d\gamma(y,z)\\ &= \int_{\mathbb{R}^{2m}} \|y-z\|_2^p \, d\gamma(y,z) = \mathrm{W}_p^p(\mu,\beta_*).\end{aligned}$$

Note that the first relation is an inequality as γ_+ may not be an optimal transport coupling between α_* and ν; the next equality follows from the support of γ_+ being contained in the subspace $\{(x,y) : \varphi_{W_*}(x) = \varphi_{W_*}(y)\}$; and the next-to-last equality comes from the definition of pushforward measure.

We next show that $\varphi_{V_*,b_*}(\alpha_*) = \mu$ under the projection φ_{V_*,b_*}, i.e., $\alpha_* \in \Phi_p^+(\mu,n)$. For a measurable $S \subseteq \mathbb{R}^m$,

$$\begin{aligned}\varphi_{V_*,b_*}(\alpha_*)(S) &= \alpha_*\big(\varphi_{V_*,b_*}^{-1}(S)\big)\\ &= \gamma_+\big(\mathbb{R}^n \times \varphi_{V_*,b_*}^{-1}(S)\big) = \tilde{\gamma}(\mathbb{R}^n \times S)\\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{(x,z)\in\mathbb{R}^n\times S} \, d\nu_y(x)\, d\gamma(y,z)\\ &= \int_{(y,z)\in\mathbb{R}^{2m}} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{z\in S} \, d\nu_y(x)\, d\gamma(y,z) = \mu(S),\end{aligned}$$

as required. Observe that the first three equalities are all consequences of the definition of a pushforward measure. Therefore, with Lemma 1, we have $\mathrm{W}_p(\alpha_*,\nu) = W_p(\mu,\beta_*)$. Hence

$$\begin{aligned}\mathrm{W}_p^+(\mu,\nu) &= \inf_{\alpha\in\Phi^+(\mu,n)} \mathrm{W}_p(\alpha,\nu) \le \mathrm{W}_p(\alpha_*,\nu)\\ &= \mathrm{W}_p(\mu,\beta_*) \le \mathrm{W}_p^-(\mu,\nu)+\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $\mathrm{W}_p^-(\mu,\nu) \ge \mathrm{W}_p^+(\mu,\nu)$. □

We denote the common value in (4) by $\widehat{\mathrm{W}}_p(\mu,\nu)$, and call it the *augmented p-Wasserstein distance* between $\mu \in \mathrm{M}_p(\mathbb{R}^m)$ and $\nu \in \mathrm{M}_p(\mathbb{R}^n)$. Note that this is a distance in the sense of a distance from a point to a set; it is not a metric since if we take $\mu,\nu \in \mathrm{M}^p(\mathbb{R}^m)$ with ν a nontrivial rotation of μ, we will have $\widehat{\mathrm{W}}_p(\mu,\nu) = 0$ even though $\mu \ne \nu$.

The augmented p-Wasserstein distance $\widehat{\mathrm{W}}_p$ preserves some properties of the p-Wasserstein metric W_p; an example is the following inequality, which is known to hold for W_p.

Corollary 2: Let $m,n \in \mathbb{N}$, $m \le n$. Let $p,q \in [1,\infty]$, $p \le q$. For any $\mu \in \mathrm{M}_q(\mathbb{R}^m)$ and $\nu \in \mathrm{M}_q(\mathbb{R}^n)$, we have

$$\widehat{\mathrm{W}}_p(\mu,\nu) \le \widehat{\mathrm{W}}_q(\mu,\nu).$$

Proof: This follows from $\widehat{\mathrm{W}}_p(\mu,\nu) = \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{W}_p(\mu,\beta) \le \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{W}_q(\mu,\beta) = \widehat{\mathrm{W}}_q(\mu,\nu)$. □

III. f-Divergence

The most useful notion of distance on probability densities is often not a metric. Divergences are in general asymmetric and do not satisfy the triangle inequality. The Kullback–Leibler divergence [33], [34] is probably the best known example, ubiquitous in information theory, machine learning, and statistics. It is used to characterize relative entropy in information systems [35], to measure randomness in continuous time series [36], to quantify information gain in comparison of statistical models of inference [37], among other things.

The KL-divergence is a special limiting case of a Rényi divergence [38], which is in turn a special case of a vast generalization called the f-divergence [39].

Definition 2: Let $\mu,\nu \in \mathrm{M}(\Omega)$ and μ be absolutely continuous with respect to ν. For any convex function $f : \mathbb{R} \to \mathbb{R}$ with $f(1) = 0$, the *f-divergence* of μ from ν is

$$\mathrm{D}_f(\mu\|\nu) = \int_\Omega f\Big(\frac{d\mu}{d\nu}\Big)\, d\nu = \int_\Omega f\big(g(x)\big)\, d\nu(x),$$

with g the Radon–Nikodym derivative $d\mu(x) = g(x)\, d\nu(x)$.

Aside from the Rényi divergence, the f-divergence includes just about every known divergences as special cases. These include the Pearson chi-squared [40], Hellinger squared [41], Chernoff [42], Jeffreys [43], alpha–beta [44], Jensen–Shannon [45], [46], and exponential [47] divergences, as well as the total variation metric. For easy reference, we provide a list in Table I. Note that taking limit as $\theta \to 1$ in the Rényi divergence gives us the KL-divergence.

These divergences are all useful in their own right. The Pearson chi-squared divergence is used in statistical test of categorical data to quantify the difference between two distributions [40]. The Hellinger squared divergence is used in dimension reduction for multivariate data [48]. The Jeffreys divergence is used in Markov random field for image classification [49]. The Chernoff divergence is used in image feature classification, indexing, and retrieval [50]. The Rényi divergence is used in quantum information theory as a measure of entanglement [51]. The alpha–beta divergence is used in geometrical analyses of parametric inference [52]. We will defer discussions of Jensen–Shannon divergence and total variation metric in Sections IV and V respectively.

By definition, an f-divergence $\mathrm{D}_f(\mu\|\nu)$ is only defined if μ is absolutely continuous with respect to ν. For convenience, in this section we will restrict our attention to probability measures with densities so that we do not have to keep track of which measure is absolutely continuous to which other measure. Let λ^n be the Lebesgue measure restricted to $\Omega \subseteq \mathbb{R}^n$. With respect to λ^n, we define

$$\begin{aligned}\mathrm{M}_d(\Omega) &\coloneqq \{\mu \in \mathrm{M}(\Omega) : \mu \text{ has density}\},\\ \mathrm{M}_{pd}(\Omega) &\coloneqq \{\mu \in \mathrm{M}_d(\Omega) : \mu \text{ has strictly positive density}\}.\end{aligned}$$

Note that $\mu \in \mathrm{M}_d(\Omega)$ iff it is absolutely continuous with respect to λ^n. The following lemma guarantees the existence of projection and embedding f-divergences, to be defined later.

Lemma 2: Let $m,n \in \mathbb{N}$, $m \le n$, and $\Phi^-(\nu,m)$ be as in Definition 1.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

TABLE I
IN THE FOLLOWING, $\zeta = (1-\theta)\mu + \theta\nu$, $\eta = (1-\theta)\nu + \theta\mu$, AND $\theta, \phi \in (0,1)$

	$f(t)$	$\mathrm{D}_f(\mu\|\nu)$
Kullback–Liebler	$t \log t$	$\int_\Omega \log\big(\frac{d\mu}{d\nu}\big) d\mu$
Exponential	$t \log^2 t$	$\int_\Omega \log^2\big(\frac{d\mu}{d\nu}\big) d\mu$
Pearson	$(t-1)^2$	$\int_\Omega \big(\frac{d\mu}{d\nu} - 1\big)^2 d\nu$
Hellinger	$(\sqrt{t}-1)^2$	$\int_\Omega \big[\big(\frac{d\mu}{d\nu}\big)^{1/2} - 1\big]^2 d\nu$
Jeffreys	$(t-1)\log t$	$\int_\Omega \big(\frac{d\mu}{d\nu} - 1\big) \log\big(\frac{d\mu}{d\nu}\big) d\nu$
Rényi	$\frac{t^\theta - t}{\theta(\theta-1)}$	$\frac{1}{\theta(\theta-1)} \int_\Omega \big[\big(\frac{d\mu}{d\nu}\big)^\theta - \frac{d\mu}{d\nu}\big] d\nu$
Chernoff	$\frac{4(1-t^{(1+\theta)/2})}{1-\theta^2}$	$\frac{4}{1-\theta^2} \int_\Omega \big[1 - \big(\frac{d\mu}{d\nu}\big)^{(1+\theta)/2}\big] d\nu$
alpha–beta	$\frac{2(1-t^{(1-\theta)/2})(1-t^{(1-\phi)/2})}{(1-\theta)(1-\phi)}$	$\frac{2}{(1-\theta)(1-\phi)} \int_\Omega \big[1 - \big(\frac{d\mu}{d\nu}\big)^{(1-\theta)/2}\big]\big[1 - \big(\frac{d\mu}{d\nu}\big)^{(1-\phi)/2}\big] d\nu$
Jensen–Shannon	$\frac{t}{2}\log\frac{t}{(1-\theta)t+\theta} + \frac{1}{2}\log\frac{1}{1-\theta+\theta t}$	$\frac{1}{2}\int_\Omega \log\big(\frac{d\mu}{d\zeta}\big) d\mu + \frac{1}{2}\int_\Omega \log\big(\frac{d\nu}{d\eta}\big) d\nu$
total variation	$\frac{\|t-1\|}{2}$	$\sup_{A\in\Sigma(\Omega)} \|\mu(A) - \nu(A)\|$

(i) If $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, then $\Phi^-(\nu, m) \subseteq \mathrm{M}_{pd}(\mathbb{R}^m)$.
(ii) If $\alpha \in \mathrm{M}_d(\mathbb{R}^n)$ and $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, then α is absolutely continuous with respect to ν.

Proof: Let $\beta \in \Phi^-(\nu, m)$ and let $V \in \mathrm{O}(m,n)$, $b \in \mathbb{R}^m$ be such that $\varphi_{V,b}(\nu) = \beta$. Let $d\nu(x) = t(x)\, d\lambda^n(x)$ and $W \in \mathrm{O}(n-m, n)$ be such that $\begin{bmatrix} V \\ W \end{bmatrix} \in \mathrm{O}(n)$. For any measurable $g : \mathbb{R}^m \to \mathbb{R}$,

$$\begin{aligned}\int_{y\in\mathbb{R}^m} g(y)\, d\beta(y) &= \int_{x\in\mathbb{R}^n} g(\varphi_{V,b}(x))\, d\nu(x) \\ &= \int_{x\in\mathbb{R}^n} g(\varphi_{V,b}(x)) t(x)\, d\lambda^n(x) \\ &= \int_{y\in\mathbb{R}^m} g(y) t'(y)\, d\lambda^m(y),\end{aligned}$$

where $t'(y) = \int_{\varphi_{V,b}^{-1}(y)} t(x)\, d\lambda^{n-m}\big(\varphi_W(x)\big)$. This is because $\varphi_{V,b}$ is an orthogonal projection plus a translation and for any measurable function f,

$$\begin{aligned}&\int_{x\in\mathbb{R}^n} f(x)\, d\lambda^n(x) \\ &\quad = \int_{y\in\mathbb{R}^m} \int_{\varphi_{V,b}^{-1}(y)} f(x)\, d\lambda^{n-m}(\varphi_W(x))\, d\lambda^m(y),\end{aligned}$$

where the existence of φ_W follows from Theorem 3. Hence $d\beta(y) = t'(y)\, d\lambda^m(y)$ and we have (i). For (ii), suppose $d\alpha(x) = t_\alpha(x)\, d\lambda^n(x)$ and $d\nu(x) = t_\nu(x)\, d\lambda^n(x)$ with $t_\nu(x) > 0$, then

$$d\alpha(x) = \frac{t_\alpha(x)}{t_\nu(x)}\, d\nu(x).$$

□

We deduce an f-divergence analogue of Lemma 1.

Lemma 3: Let $m, n \in \mathbb{N}$, $m \le n$, and $f : \mathbb{R} \to \mathbb{R}$ be convex with $f(1) = 0$. Let $\alpha \in \mathrm{M}_d(\mathbb{R}^n)$, $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, $V \in \mathrm{O}(m,n)$, and $b \in \mathbb{R}^m$. Then

$$\mathrm{D}_f(\alpha\|\nu) \ge \mathrm{D}_f\big(\varphi_{V,b}(\alpha)\|\varphi_{V,b}(\nu)\big).$$

Proof: Let $\mu = \varphi_{V,b}(\alpha)$ and $\beta = \varphi_{V,b}(\nu)$ with $d\alpha(x) = t(x)\, d\nu(x)$. By Theorem 3, for any measurable function g,

$$\int_{\mathbb{R}^n} g(x)\, d\alpha(x) = \int_{\mathbb{R}^m} \int_{\varphi_{V,b}^{-1}(y)} g(x)\, d\alpha_y(x)\, d\mu(y),$$
$$\int_{\mathbb{R}^n} g(x)\, d\nu(x) = \int_{\mathbb{R}^m} \int_{\varphi_{V,b}^{-1}(y)} g(x)\, d\nu_y(x)\, d\beta(y).$$

By Lemma 2, we have $d\mu(y) = t'(y)\, d\beta(y)$ with $t'(y) = \int_{\varphi_{V,b}^{-1}(y)} t(x)\, d\nu_y(x)$. By Definition 2 and Jensen inequality,

$$\begin{aligned}\mathrm{D}_f(\alpha\|v) &= \int_{x\in\mathbb{R}^n} f\big(t(x)\big)\, d\nu(x) \\ &= \int_{y\in\mathbb{R}^m} \left[\int_{\varphi_{V,b}^{-1}(y)} f\big(t(x)\big)\, d\nu_y(x)\right] d\beta(y) \\ &\ge \int_{y\in\mathbb{R}^m} f\left[\int_{\varphi_{V,b}^{-1}(y)} t(x)\, d\nu_y(x)\right] d\beta(y) \\ &= \int_{y\in\mathbb{R}^m} f\big(t'(y)\big)\, d\beta(y) = \mathrm{D}_f(\mu\|\beta).\end{aligned}$$

□

Lemma 2 assures that $\Phi^-(\nu, m) \subseteq \mathrm{M}_d(\mathbb{R}^m)$ but in general, it will not be true that $\Phi^+(\mu, n) \subseteq \mathrm{M}_d(\mathbb{R}^n)$. As such we introduce the following subset:

$$\begin{aligned}\Phi_d^+(\mu, n) \coloneqq \{&\alpha \in \mathrm{M}_d(\mathbb{R}^n) : \varphi_{V,b}(\alpha) = \mu \\ &\text{for some } V \in \mathrm{O}(m,n),\ b \in \mathbb{R}^m\}\end{aligned}$$

and with this, we establish Theorem 1 for f-divergence.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

Theorem 5: Let $m, n \in \mathbb{N}$ and $m \le n$. For $\mu \in \mathrm{M}(\mathbb{R}^m)$ and $\nu \in \mathrm{M}(\mathbb{R}^n)$, let

$$\mathrm{D}_f^-(\mu\|\nu) \coloneqq \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{D}_f(\mu\|\beta),$$
$$\mathrm{D}_f^+(\mu\|\nu) \coloneqq \inf_{\alpha\in\Phi_d^+(\mu,n)} \mathrm{D}_f(\alpha\|\nu).$$

Then

$$\mathrm{D}_f^-(\mu\|\nu) = \mathrm{D}_f^+(\mu\|\nu). \tag{5}$$

Proof: Again, $\mathrm{D}_f^-(\mu\|\nu) \le \mathrm{D}_f^+(\mu\|\nu)$ is easy: For any $\alpha \in \Phi_d^+(\mu,n)$, there exist $V_\alpha \in \mathrm{O}(m,n)$ and $b_\alpha \in \mathbb{R}^m$ with $\varphi_{V_\alpha,b_\alpha}(\alpha) = \mu$. It follows from Lemma 3 that $\mathrm{D}_f(\alpha\|\nu) \ge \mathrm{D}_f\big(\mu\|\varphi_{V_\alpha,b_\alpha}(\nu)\big)$ and thus

$$\begin{aligned}\inf_{\alpha\in\Phi_d^+(\mu,n)} \mathrm{D}_f(\alpha\|\nu) &\ge \inf_{\alpha\in\Phi_d^+(\mu,n)} \mathrm{D}_f\big(\mu\|\varphi_{V_\alpha,b_\alpha}(\nu)\big)\\ &\ge \inf_{V\in\mathrm{O}(m,n),\ b\in\mathbb{R}^m} \mathrm{D}_f\big(\mu\|\varphi_{V,b}(\nu)\big).\end{aligned}$$

It remains to show $\mathrm{D}_f^-(\mu\|\nu) \ge \mathrm{D}_f^+(\mu\|\nu)$. By the definition of $\mathrm{D}_f^-(\mu\|\nu)$, for any $\varepsilon > 0$, there exists $\beta_* \in \Phi^-(\nu,m)$ with

$$\mathrm{D}_f^-(\mu\|\nu) \le \mathrm{D}_f(\mu\|\beta_*) \le \mathrm{D}_f^-(\mu\|\nu) + \varepsilon.$$

Let $V_* \in \mathrm{O}(m,n)$ and $b_* \in \mathbb{R}^m$ be such that $\varphi_{V_*,b_*}(\nu) = \beta_*$ and $W_* \in \mathrm{O}(n-m,n)$ be such that $\begin{bmatrix} V_* \\ W_* \end{bmatrix} \in \mathrm{O}(n)$. Applying Theorem 3 to φ_{V_*,b_*}, we obtain a family of measures $\{\nu_y \in \mathrm{M}(\mathbb{R}^n) : y \in \mathbb{R}^m\}$ such that for any measurable function f,

$$\int_{\mathbb{R}^n} f(x)\,d\nu(x) = \int_{\mathbb{R}^m} \int_{\varphi_{V_*,b_*}^{-1}(y)} f(x)\,d\nu_y(x)\,d\beta_*(y).$$

Define $\alpha_* \in \mathrm{M}(\mathbb{R}^n)$ by

$$\alpha_*(S) = \int_{\mathbb{R}^n} \mathbb{I}_{x\in S}\,d\alpha_*(x) = \int_{\mathbb{R}^m} \int_{\varphi_{V_*,b_*}^{-1}(y)} \mathbb{I}_{x\in S}\,d\nu_y(x)\,d\mu(y)$$

for any measurable set $S \subseteq \mathbb{R}^n$. Since $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, we may identify $\{\nu_y \in \mathrm{M}(\mathbb{R}^n) : y \in \mathbb{R}^m\}$ as a subset of $\mathrm{M}_d(\mathbb{R}^{n-m})$. Let $d\nu_y(x) = s_y(x)\,d\lambda^{n-m}(\varphi_{W_*}(x))$ and $d\mu(y) = g(y)\,d\lambda^m(y)$. Then

$$\begin{aligned}d\alpha_*(x) &= g\big(\varphi_{V_*,b_*}(x)\big)\,s_{\varphi_{V_*,b_*}(x)}(x)\\ &\qquad d\lambda^{n-m}(\varphi_{W_*}(x))d\lambda^m(\varphi_{V_*,b_*}(x))\\ &= g\big(\varphi_{V_*,b_*}(x)\big)\,s_{\varphi_{V_*,b_*}(x)}(x)\,d\lambda^n(x).\end{aligned}$$

Hence we deduce that $\alpha_* \in \mathrm{M}_d(\mathbb{R}^n)$. We may also check that $\varphi_{V_*,b_*}(\alpha_*) = \mu$. Let $d\mu(y) = t(y)\,d\beta_*(y)$. Then

$$d\alpha_*(x) = t\big(\varphi_{V_*,b_*}(x)\big)\,d\nu(x).$$

Finally, by Definition 2, we have

$$\begin{aligned}\mathrm{D}_f^+(\mu\|\nu) &\le \mathrm{D}_f(\alpha_*\|\nu)\\ &= \int_{\mathbb{R}^n} f\big(t(\varphi_{V_*,b_*}(x))\big)\,d\nu(x)\\ &= \int_{\mathbb{R}^m} f\big(t(y)\big)\,d\beta_*(y)\\ &= \mathrm{D}_f(\mu\|\beta_*) \le \mathrm{D}_f^-(\mu\|\nu) + \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we have $\mathrm{D}_f^+(\mu\|\nu) \le \mathrm{D}_f^-(\mu\|\nu)$. □

As in the case of Wasserstein distance, we denote the common value in (5) by $\widehat{\mathrm{D}}_f(\mu\|\nu)$ and call it the *augmented* f-divergence, and likewise for all specific f-divergences.

Surprisingly, certain relations between these distances remain true with our extension to probability densities of different dimensions. For example, Pinker's inequality [53] between the total variation metric d_{TV} and KL-divergence D_{KL} holds for the augmented total variation distance $\widehat{d}_{\mathrm{TV}}$ and augmented KL-divergence $\widehat{\mathrm{D}}_{\mathrm{KL}}$; another standard relation between the Hellinger squared divergence D_{H} and total variation metric is preserved for their augmented counterparts too.

Corollary 3 (Pinker's Inequality for Probability Measures of Different Dimensions): Let $m, n \in \mathbb{N}$, $m \le n$. For any $\mu \in \mathrm{M}_d(\mathbb{R}^m)$ and $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, we have

$$\widehat{d}_{\mathrm{TV}}(\mu,\nu) \le \sqrt{\frac{1}{2}\widehat{\mathrm{D}}_{\mathrm{KL}}(\mu\|\nu)}.$$

Proof: This follows from $\widehat{\mathrm{D}}_{\mathrm{KL}}(\mu\|\nu) = \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{D}_{\mathrm{KL}}(\mu\|\beta) \ge \inf_{\beta\in\Phi^-(\nu,m)} 2d_{\mathrm{TV}}(\mu\|\beta)^2 = 2\widehat{d}_{\mathrm{TV}}(\alpha\|\nu)^2$. □

Corollary 4: Let $m, n \in \mathbb{N}$, $m \le n$. For any $\mu \in \mathrm{M}_d(\mathbb{R}^m)$ and $\nu \in \mathrm{M}_{pd}(\mathbb{R}^n)$, we have

$$\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu)^2 \le 2\widehat{d}_{\mathrm{TV}}(\mu,\nu) \le \sqrt{2}\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu).$$

Proof: Clearly the two inequalities hold for D_{H} and d_{TV} when $m = n$. The inequidimensional version then follows from

$$\begin{aligned}\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu)^2 &= \inf_{\beta\in\Phi^-(\nu,m)} \mathrm{D}_{\mathrm{H}}(\mu,\beta)^2\\ &\le \inf_{\beta\in\Phi^-(\nu,m)} 2d_{\mathrm{TV}}(\mu,\beta) = 2\widehat{d}_{\mathrm{TV}}(\mu,\nu)\\ &\le \inf_{\beta\in\Phi^-(\nu,m)} \sqrt{2}\,\mathrm{D}_{\mathrm{H}}(\mu,\beta) = \sqrt{2}\widehat{\mathrm{D}}_{\mathrm{H}}(\mu,\nu).\end{aligned}$$

□

IV. Jensen–Shannon Divergence

Let $\mu, \nu \in \mathrm{M}(\mathbb{R}^n)$ and $\theta \in (0,1)$. The *Jensen–Shannon divergence* is defined by

$$\mathrm{D}_{\mathrm{JS}}(\mu,\nu) \coloneqq \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(\mu\|\zeta) + \frac{1}{2}\,\mathrm{D}_{\mathrm{KL}}(\nu\|\eta), \tag{6}$$

where $\zeta \coloneqq (1-\theta)\mu + \theta\nu$ and $\eta \coloneqq (1-\theta)\nu + \theta\mu$. What we call Jensen–Shannon divergence here is slightly more general [46] than the usual definition [45],[2] which corresponds to the case when $\theta = 1/2$. When $\theta = 1$, we get the Jeffreys divergence in Table I as another special case. We have written $\mathrm{D}_{\mathrm{JS}}(\mu,\nu)$ instead of the usual $\mathrm{D}_{\mathrm{JS}}\,f(\mu\|\nu)$ for f-divergence to remind the reader that D_{JS} is symmetric in its arguments; in fact, $\mathrm{D}_{\mathrm{JS}}(\mu,\nu)^{1/2}$ defines a metric on $\mathrm{M}(\mathbb{R}^n)$.

The Jensen–Shannon divergence is often viewed as the symmetrization of the Kullback–Liebler divergence but this perspective hides an important distinction, namely, the JS-divergence may be defined on probability measures without densities: Observe that μ, ν are automatically absolutely continuous with respect to ζ and η. As such the definition in (6) is valid for any $\mu, \nu \in \mathrm{M}(\mathbb{R}^n)$ and we do not need to work over $\mathrm{M}_d(\mathbb{R}^n)$ like in Section III.

[2]Neither Jensen nor Shannon is a coauthor of [45]. The name comes from an application of Jensen inequality to Shannon entropy as a convex function to establish nonnegativity of the divergence.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

The JS-divergence is used in applications to compare genome [54], [55] and protein surfaces [56] in bioinformatics; to quantify information flow in social and biological systems [57], [58], and to detect anomalies in fire experiments [59]. It was notably used to establish the main theorem in the landmark paper on Generative Adversarial Nets [60].

Lemma 4: Let $m, n \in \mathbb{N}$, $m \leq n$, and $\theta \in (0,1)$. For any $\alpha, \nu \in \mathrm{M}(\mathbb{R}^n)$, $V \in \mathrm{O}(m,n)$, and $b \in \mathbb{R}^m$,

$$\mathrm{D}_{\mathrm{JS}}(\alpha, \nu) \geq \mathrm{D}_{\mathrm{JS}}\big(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)\big).$$

Proof: The proof is similar to that of Lemma 3. We only need to check that $\varphi_{V,b}(\zeta) = (1-\theta)\varphi_{V,b}(\alpha) + \theta\varphi_{V,b}(\nu)$, $\varphi_{V,b}(\eta) = \theta\varphi_{V,b}(\alpha) + (1-\theta)\varphi_{V,b}(\nu)$, where $\zeta = (1-\theta)\mu + \theta\nu$, $\eta = (1-\theta)\nu + \theta\mu$. □

We now prove Theorem 1 for Jensen–Shannon divergence.

Theorem 6: Let $m, n \in \mathbb{N}$ and $m \leq n$. For $\mu \in \mathrm{M}(\mathbb{R}^m)$ and $\nu \in \mathrm{M}(\mathbb{R}^n)$, let

$$\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) \coloneqq \inf_{\beta \in \Phi^-(\nu, m)} \mathrm{D}_{\mathrm{JS}}(\mu, \beta),$$
$$\mathrm{D}^+_{\mathrm{JS}}(\mu, \nu) \coloneqq \inf_{\alpha \in \Phi^+(\mu, n)} \mathrm{D}_{\mathrm{JS}}(\alpha, \nu).$$

Then

$$\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) = \mathrm{D}^+_{\mathrm{JS}}(\mu, \nu). \tag{7}$$

Proof: For any $\alpha \in \Phi^+(\mu, n)$, there exist $V_\alpha \in \mathrm{O}(m,n)$ and $b_\alpha \in \mathbb{R}^m$ with $\varphi_{V_\alpha, b_\alpha}(\alpha) = \mu$. It follows from Lemma 4 that $\mathrm{D}_{\mathrm{JS}}(\alpha, \nu) \geq \mathrm{D}_{\mathrm{JS}}\big(\mu, \varphi_{V_\alpha, b_\alpha}(\nu)\big)$ and thus

$$\begin{aligned}\inf_{\alpha \in \Phi^+(\mu,n)} \mathrm{D}_{\mathrm{JS}}(\alpha, \nu) &\geq \inf_{\alpha \in \Phi^+(\mu,n)} \mathrm{D}_{\mathrm{JS}}\big(\mu, \varphi_{V_\alpha, b_\alpha}(\nu)\big) \\ &\geq \inf_{V \in \mathrm{O}(m,n),\ b \in \mathbb{R}^m} \mathrm{D}_{\mathrm{JS}}\big(\mu, \varphi_{V,b}(\nu)\big)\end{aligned}$$

and thus $\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) \leq \mathrm{D}^+_{\mathrm{JS}}(\mu, \nu)$.

We next show that $\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) \geq \mathrm{D}^+_{\mathrm{JS}}(\mu, \nu)$. By the definition of $\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu)$, for each $\alpha \in \Phi^+(\mu, n)$ and any $\varepsilon > 0$, there exists $\beta_* \in \Phi^-(\nu, m)$ such that

$$\mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) \leq \mathrm{D}_{\mathrm{JS}}(\mu, \beta_*) \leq \mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) + \varepsilon.$$

Let $V_* \in \mathrm{O}(m,n)$ and $b_* \in \mathbb{R}^m$ be such that $\varphi_{V_*, b_*}(\nu) = \beta_*$. Applying Theorem 3 to φ_{V_*, b_*}, we obtain $\{\nu_y \in \mathrm{M}(\mathbb{R}^n) : y \in \mathbb{R}^m\}$ that satisfies

$$\int_{\mathbb{R}^n} f(x)\, d\nu(x) = \int_{\mathbb{R}^m} \int_{\varphi^{-1}_{V_*, b_*}(y)} f(x)\, d\nu_y(x)\, d\beta_*(y)$$

for any measurable function f. Let $\alpha_* \in \mathrm{M}(\mathbb{R}^n)$ be such that

$$\alpha_*(S) = \int_{\mathbb{R}^n} \mathbb{I}_{x \in S}\, d\alpha_*(x) = \int_{\mathbb{R}^m} \int_{\varphi^{-1}_{V_*, b_*}(y)} \mathbb{I}_{x \in S}\, d\nu_y(x)\, d\mu(y)$$

for any measurable set $S \subseteq \mathbb{R}^n$. Then $\varphi_{V_*, b_*}(\alpha_*) = \mu$ and so $\alpha_* \in \Phi^+(\mu, n)$. Consider the weighted measures

$$\zeta^* \coloneqq (1-\theta)\mu + \theta\beta_*, \qquad \eta^* \coloneqq \theta\mu + (1-\theta)\beta_*,$$
$$\xi_1 \coloneqq (1-\theta)\alpha_* + \theta\nu, \qquad \xi_2 \coloneqq \theta\alpha_* + (1-\theta)\nu.$$

Since μ is absolutely continuous with respect to ζ^* and β_* to η^*, we let $d\mu = g_1\, d\zeta^*$ and $d\beta_* = g_2\, d\eta^*$. Then we have

$$\varphi_{V_*, b_*}(\xi_1) = \zeta^*, \qquad d\alpha_*(x) = g_1\big(\varphi_{V_*, b_*}(x)\big)\, d\xi_1(x),$$
$$\varphi_{V_*, b_*}(\xi_2) = \eta^*, \qquad d\nu(x) = g_2\big(\varphi_{V_*, b_*}(x)\big)\, d\xi_2(x).$$

By the definition of $\mathrm{D}^+_{\mathrm{JS}}$,

$$\begin{aligned}\mathrm{D}^+_{\mathrm{JS}}(\mu, \nu) &\leq \mathrm{D}_{\mathrm{JS}}(\alpha_*, \nu) = \frac{1}{2}\big[\mathrm{D}_{\mathrm{KL}}(\alpha_* \| \xi_1) + \mathrm{D}_{\mathrm{KL}}(\nu \| \xi_2)\big] \\ &= \frac{1}{2}\int_{\mathbb{R}^n} \log\big[g_1\big(\varphi_{V_*, b_*}(x)\big)\big] g_1\big(\varphi_{V_*, b_*}(x)\big)\, d\xi_1(x) \\ &\quad + \log\big[g_2\big(\varphi_{V_*, b_*}(x)\big)\big] g_2\big(\varphi_{V_*, b_*}(x)\big)\, d\xi_2(x) \\ &= \frac{1}{2}\int_{\mathbb{R}^m} \log\big(g_1(y)\big) g_1(y)\, d\zeta^*(y) \\ &\quad + \log\big(g_2(y)\big) g_2(y)\, d\eta^*(y) \\ &= \frac{1}{2}\big[\mathrm{D}_{\mathrm{KL}}(\mu \| \zeta^*) + \mathrm{D}_{\mathrm{KL}}(\beta_* \| \eta^*)\big] \\ &= \mathrm{D}_{\mathrm{JS}}(\mu, \beta_*) \leq \mathrm{D}^-_{\mathrm{JS}}(\mu, \nu) + \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $\mathrm{D}^+_{\mathrm{JS}}(\mu, \nu) \leq \mathrm{D}^-_{\mathrm{JS}}(\mu, \nu)$. □

V. Total Variation Distance

The total variation metric is quite possibly the most classical notion of distance between probability measures. It is used in Markov models [61], [62], stochastic processes [63], Monte Carlo algorithms [64], geometric approximation [65], image restoration [66], among other areas.

The definition is straightforward: The *total variation metric* between $\mu, \nu \in \mathrm{M}(\mathbb{R}^n)$ is simply

$$d_{\mathrm{TV}}(\mu, \nu) \coloneqq \sup_{A \in \Sigma(\mathbb{R}^n)} |\mu(A) - \nu(A)|.$$

As we saw in Section III, when the probability measures have densities, the total variation metric is a special case of the f-divergence with $f(t) = |t-1|/2$. While it is not a special case of the Wasserstein metric in Section II, it is related in that

$$d_{\mathrm{TV}}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R}^{2n}} \mathbb{I}_{x \neq y}(x, y)\, d\gamma(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of couplings as in the definition of Wasserstein metric and $\mathbb{I}_{x \neq y}$ is an indicator function, i.e., takes value 1 when $x \neq y$ and 0 when $x = y$.

For any $\mu, \nu \in \mathrm{M}(\mathbb{R}^n)$, it follows from Hahn Decomposition, i.e., Theorem 2, that there exists $S \in \Sigma(\mathbb{R}^n)$ with

$$d_{\mathrm{TV}}(\mu, \nu) = \mu(S) - \nu(S). \tag{8}$$

So for any measurable $B \subseteq S$, $C \subseteq S^{\complement} \coloneqq \mathbb{R}^n \setminus S$,

$$\mu(B) - \nu(B) \geq 0, \quad \mu(C) - \nu(C) \leq 0; \tag{9}$$

or, equivalently, for all measurable function $f(x) \geq 0$,

$$\int_S f(x)\, d\big(\mu(x) - \nu(x)\big) \geq 0,$$
$$\int_{S^{\complement}} f(x)\, d\big(\mu(x) - \nu(x)\big) \leq 0. \tag{10}$$

Lemma 5: Let $\alpha, \nu \in \mathrm{M}(\mathbb{R}^n)$. Then for any $V \in \mathrm{O}(m,n)$ and $b \in \mathbb{R}^m$,

$$d_{\mathrm{TV}}(\alpha, \nu) \geq d_{\mathrm{TV}}\big(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)\big).$$

Proof: By (8) and (9), $d_{\mathrm{TV}}\big(\varphi_{V,b}(\alpha), \varphi_{V,b}(\nu)\big) = \varphi_{V,b}(\alpha)(S) - \varphi_{V,b}(\nu)(S) = \alpha\big(\varphi^{-1}_{V,b}(S)\big) - \nu\big(\varphi^{-1}_{V,b}(S)\big) \leq d_{\mathrm{TV}}(\alpha, \nu)$. □

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

Theorem 7: Let $m, n \in \mathbb{N}$ and $m \leq n$. For $\mu \in \mathrm{M}(\mathbb{R}^m)$ and $\nu \in \mathrm{M}(\mathbb{R}^n)$, let

$$d^-_{\mathrm{TV}}(\mu,\nu) \coloneqq \inf_{\beta\in\Phi^-(\nu,m)} d_{\mathrm{TV}}(\mu,\beta),$$
$$d^+_{\mathrm{TV}}(\mu,\nu) \coloneqq \inf_{\alpha\in\Phi^+(\mu,n)} d_{\mathrm{TV}}(\alpha,\nu).$$

Then

$$d^+_{\mathrm{TV}}(\mu,\nu) = d^-_{\mathrm{TV}}(\mu,\nu). \tag{11}$$

Proof: For any $\alpha \in \Phi^+(\mu,n)$, there exist $V_\alpha \in \mathrm{O}(m,n)$ and $b_\alpha \in \mathbb{R}^m$ with $\varphi_{V_\alpha,b_\alpha}(\alpha) = \mu$. So by Lemma 5, $d_{\mathrm{TV}}(\alpha,\nu) \geq d_{\mathrm{TV}}\big(\mu,\varphi_{V_\alpha,b_\alpha}(\nu)\big)$ and we get $d^+_{\mathrm{TV}}(\mu,\nu) \geq d^-_{\mathrm{TV}}(\mu,\nu)$ from

$$\begin{aligned}\inf_{\alpha\in\Phi^+(\mu,n)} d_{\mathrm{TV}}(\alpha,\nu) &\geq \inf_{\alpha\in\Phi^+(\mu,n)} d_{\mathrm{TV}}\big(\mu,\varphi_{V_\alpha,b_\alpha}(\nu)\big)\\ &\geq \inf_{V\in\mathrm{O}(m,n),b\in\mathbb{R}^m} d_{\mathrm{TV}}\big(\mu,\varphi_{V,b}(\nu)\big).\end{aligned}$$

We next show that $d^-_{\mathrm{TV}}(\mu,\nu) \geq d^+_{\mathrm{TV}}(\mu,\nu)$. By the definition of $d^-_{\mathrm{TV}}(\mu,\nu)$, for any $\varepsilon > 0$, there exists $\beta_* \in \Phi^-(\nu,m)$ with

$$d^-_{\mathrm{TV}}(\mu,\nu) \leq d_{\mathrm{TV}}(\mu,\beta_*) \leq d^-_{\mathrm{TV}}(\mu,\nu) + \varepsilon.$$

Let $V_* \in \mathrm{O}(m,n)$ and $b_* \in \mathbb{R}^m$ be such that $\varphi_{V_*,b_*}(\nu) = \beta_*$ and $S \in \Sigma(\mathbb{R}^m)$ be such that $d_{\mathrm{TV}}(\mu,\beta_*) = \mu(S) - \beta_*(S)$. Applying Theorem 3 to φ_{V_*,b_*}, we obtain $\{\nu_y \in \mathrm{M}(\mathbb{R}^n) : y \in \mathbb{R}^m\}$ that satisfies

$$\int_{\mathbb{R}^n} f(x)\,d\nu(x) = \int_{\mathbb{R}^m}\int_{\varphi^{-1}_{V_*,b_*}(y)} f(x)\,d\nu_y(x)\,d\beta_*(y)$$

for any measurable function f. Let $\alpha_* \in \mathrm{M}(\mathbb{R}^n)$ be such that

$$\alpha_*(S) = \int_{\mathbb{R}^n} \mathbb{I}_{x\in S}\,d\alpha_*(x) = \int_{\mathbb{R}^m}\int_{\varphi^{-1}_{V_*,b_*}(y)} \mathbb{I}_{x\in S}\,d\nu_y(x)\,d\mu(y)$$

for any measurable set $S \subseteq \mathbb{R}^n$. We can check α_* is indeed a probability measure and $\varphi_{V_*,b_*}(\alpha_*) = \mu$. Partition $\mathbb{R}^n$ into $\varphi^{-1}_{V_*,b_*}(S)$ and $\varphi^{-1}_{V_*,b_*}(S^{\complement})$. We claim that for any measurable $B \subseteq \varphi^{-1}_{V_*,b_*}(S)$ and $C \subseteq \varphi^{-1}_{V_*,b_*}(S^{\complement})$,

$$\alpha_*(B) - \nu(B) \geq 0, \quad \alpha_*(C) - \nu(C) \leq 0.$$

Let $g = \mathbb{I}_{x\in B}$. Then

$$\begin{aligned}\alpha_*(B) - \nu(B) &= \int_{\mathbb{R}^n} g(x)\,d\big(\alpha_*(x) - \nu(x)\big)\\ &= \int_{\mathbb{R}^m}\int_{\varphi^{-1}_{V_*,b_*}(y)} g(x)\,d\nu_y(x)\,d\big(\mu(y) - \beta_*(y)\big)\\ &= \int_S\int_{\varphi^{-1}_{V_*,b_*}(y)} g(x)\,d\nu_y(x)\,d\big(\mu(y) - \beta_*(y)\big)\\ &= \int_S h(y)\,d\big(\mu(y) - \beta_*(y)\big),\end{aligned}$$

where $h(y) = \int_{\varphi^{-1}_{V_*,b_*}(y)} g(x)d\nu_y(x) \geq 0$. By (9), we deduce that $\alpha_*(B) - \nu(B) \geq 0$. Likewise, $\alpha_*(C) - \nu(C) \leq 0$. Let $T = \varphi^{-1}_{V_*,b_*}(S)$. Then for any measurable $A \subseteq \mathbb{R}^n$,

$$\begin{aligned}\alpha_*(A) - \nu(A) &= \alpha_*(A\cap T) - \nu(A\cap T)\\ &\quad + \alpha_*(A\cap T^{\complement}) - \nu(A\cap T^{\complement})\\ &\leq \alpha_*(A\cap T) - \nu(A\cap T) \leq \alpha_*(T) - \nu(T).\end{aligned}$$

Hence we obtain

$$\begin{aligned}d^+_{\mathrm{TV}}(\mu,\nu) &\leq d_{\mathrm{TV}}(\alpha_*,\nu) = \alpha_*(T) - \nu(T)\\ &= \mu(S) - \beta_*(S) = d_{\mathrm{TV}}(\mu,\beta_*) \leq d^-_{\mathrm{TV}}(\mu,\nu) + \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, $d^+_{\mathrm{TV}}(\mu,\nu) \leq d^-_{\mathrm{TV}}(\mu,\nu)$. □

Theorem 7 is stronger than what we may deduce from Theorem 5 as measures are not required to have densities.

VI. Examples

Theorems 4, 5, 6, and 7 show that to compute any of the distances therein between probability measures of different dimensions, we may either compute the projection distance d^- or the embedding distance d^+. We will present five examples, three continuous and two discrete. In this section, we denote our probability measures by ρ_1, ρ_2 instead of μ, ν to avoid any clash with the standard notation for mean.

In the following, we will write $\mathcal{N}_n(\mu,\Sigma)$ for the n-dimensional normal measure with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$. For $\rho_1 = \mathcal{N}_n(\mu_1,\Sigma_1)$ and $\rho_2 = \mathcal{N}_n(\mu_2,\Sigma_2) \in \mathrm{M}(\mathbb{R}^n)$, recall that the 2-Wasserstein metric and the KL-divergence between them are given by

$$\mathrm{W}_2^2(\rho_1,\rho_2) = \|\mu_1-\mu_2\|_2^2 + \operatorname{tr}\big(\Sigma_1+\Sigma_2-2\Sigma_2^{\frac12}\Sigma_1\Sigma_2^{\frac12}\big)^{\frac12},$$
$$\mathrm{D}_{\mathrm{KL}}(\rho_1\|\rho_2) = \frac12\bigg[\operatorname{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^{\intercal}\Sigma_2^{-1}(\mu_2-\mu_1) - n + \log\bigg(\frac{\det\Sigma_2}{\det\Sigma_1}\bigg)\bigg]$$

respectively. The former may be found in [67] while the latter is a routine calculation.

We adopt the standard convention that a vector in $\mathbb{R}^m$ will always be assumed to be a column vector, i.e., $\mathbb{R}^m \equiv \mathbb{R}^{m\times 1}$. A matrix $X \in \mathbb{R}^{m\times n}$, when denoted $X = [x_1,\dots,x_n]$ implicitly means that $x_1,\dots,x_n \in \mathbb{R}^m$ are its column vectors, and when denoted $X = [y_1^{\intercal},\dots,y_m^{\intercal}]^{\intercal}$ implicitly means that $y_1,\dots,y_m \in \mathbb{R}^n$ are its row vectors. The notation $\operatorname{diag}(\lambda_1,\dots,\lambda_n)$ means an $n\times n$ diagonal matrix with diagonal entries $\lambda_1,\dots,\lambda_n \in \mathbb{R}$.

Example 1 (2-Wasserstein Distance Between One- and n-Dimensional Gaussians): Let $\rho_1 = \mathcal{N}_1(\mu_1,\sigma^2) \in \mathrm{M}(\mathbb{R})$ be a one-dimensional Gaussian measure and $\rho_2 = \mathcal{N}_n(\mu_2,\Sigma) \in \mathrm{M}(\mathbb{R}^n)$ be an n-dimensional Gaussian measure, $n \in \mathbb{N}$ arbitrary. We seek the 2-Wasserstein distance $\widehat{W}_2(\rho_1,\rho_2)$ between them. By Theorem 4, we have the option of computing either $\mathrm{W}_2^-(\rho_1,\rho_2)$ or $\mathrm{W}_2^+(\rho_1,\rho_2)$ but the choice is obvious, given that the former is considerably simpler:

$$\begin{aligned}\mathrm{W}_2^-(\rho_1,\rho_2)^2 &= \min_{\|x\|_2=1,\ y\in\mathbb{R}} \|\mu_1 - x^{\intercal}\mu_2 - y\|_2^2\\ &\qquad + \operatorname{tr}(\sigma^2 + x^{\intercal}\Sigma x - 2\sigma\sqrt{x^{\intercal}\Sigma x})\\ &= \min_{\|x\|_2=1} (\sigma - \sqrt{x^{\intercal}\Sigma x})^2.\end{aligned}$$

Let λ_1 and λ_n be the largest and smallest eigenvalues of Σ. Then $\lambda_n \leq x^{\intercal}\Sigma x \leq \lambda_1$ and thus we must have

$$\widehat{W}_2(\rho_1,\rho_2) = \begin{cases}\sqrt{\lambda_n} - \sigma & \text{if } \sigma < \sqrt{\lambda_n},\\ 0 & \text{if } \sqrt{\lambda_n} \leq \sigma \leq \sqrt{\lambda_1},\\ \sigma - \sqrt{\lambda_1} & \text{if } \sigma > \sqrt{\lambda_1}.\end{cases}$$

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

Example 2 (KL-Divergence Between One- and n-Dimensional Gaussians): Let $\rho_1 = \mathcal{N}_1(\mu_1, \sigma^2) \in \mathrm{M}(\mathbb{R})$ and $\rho_2 = \mathcal{N}_n(\mu_2, \Sigma) \in \mathrm{M}(\mathbb{R}^n)$ be as in Example 1. By Theorem 5, we may compute either $\mathrm{D}^-_{\mathrm{KL}}(\rho_1\|\rho_2)$ or $\mathrm{D}^+_{\mathrm{KL}}(\rho_1\|\rho_2)$ and again the simpler option is

$$\begin{aligned}\mathrm{D}^-_{\mathrm{KL}}(\rho_1\|\rho_2) &= \min_{\|x\|_2=1, y\in\mathbb{R}} \frac{1}{2}\Big[\frac{\sigma^2}{x^{\mathsf{T}}\Sigma x} + \frac{(\mu_1 - x^{\mathsf{T}}\mu_2 - y)^2}{x^{\mathsf{T}}\Sigma x} \\ &\qquad - 1 + \log\Big(\frac{x^{\mathsf{T}}\Sigma x}{\sigma^2}\Big)\Big] \\ &= \min_{\|x\|_2=1} \frac{1}{2}\Big[\frac{\sigma^2}{x^{\mathsf{T}}\Sigma x} - 1 + \log\Big(\frac{x^{\mathsf{T}}\Sigma x}{\sigma^2}\Big)\Big].\end{aligned}$$

Again $\lambda_n \le x^{\mathsf{T}}\Sigma x \le \lambda_1$ where λ_1 and λ_n are the largest and smallest eigenvalues of Σ. Since $f(\lambda) = \sigma^2/\lambda + \log(\lambda/\sigma^2)$ has $f'(\lambda) = (\lambda - \sigma^2)/\lambda^2$, we obtain

$$\widehat{\mathrm{D}}_{\mathrm{KL}}(\rho_1\|\rho_2) = \begin{cases} \frac{1}{2}\Big[\frac{\sigma^2}{\lambda_n} - 1 + \log\Big(\frac{\lambda_n}{\sigma^2}\Big)\Big] & \text{if } \sigma < \sqrt{\lambda_n}, \\ 0 & \text{if } \sqrt{\lambda_n} \le \sigma \le \sqrt{\lambda_1}, \\ \frac{1}{2}\Big[\frac{\sigma^2}{\lambda_1} - 1 + \log\Big(\frac{\lambda_1}{\sigma^2}\Big)\Big] & \text{if } \sigma > \sqrt{\lambda_1}. \end{cases}$$

Example 3 (KL-Divergence Between Uniform Measure on m-Dimensional Ball and n-Dimensional Gaussian): Let $\mathbb{B}^m = \{x \in \mathbb{R}^m : \|x\|_2 \le 1\}$ be the unit 2-norm ball in $\mathbb{R}^m$ and let $\rho_1 = \mathcal{U}(\mathbb{B}^m)$ be the uniform probability measure on $\mathbb{B}^m$. Let $\rho_2 = \mathcal{N}_n(\mu_2, \Sigma)$ be an n-dimensional Gaussian measure with mean $\mu_2 \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ symmetric positive definite. By Theorem 5,

$$\begin{aligned}\widehat{\mathrm{D}}_{\mathrm{KL}}(\rho_1\|\rho_2) &= \mathrm{D}^-_{\mathrm{KL}}(\rho_1\|\rho_2) \\ &= \inf_{V\in\mathrm{O}(m,n),\, b\in\mathbb{R}^m} \mathrm{D}_{\mathrm{KL}}\big(\rho_1\|\varphi_{V,b}(\rho_2)\big).\end{aligned}$$

Note that $\varphi_{V,b}(\rho_2) = \mathcal{N}_m(V\mu_2 + b, V\Sigma V^{\mathsf{T}})$ is an m-dimensional Gaussian. Let $\lambda_1 \ge \cdots \ge \lambda_n > 0$ be the eigenvalues of Σ and $\sigma_1 \ge \cdots \ge \sigma_m > 0$ be the eigenvalues of $V\Sigma V^{\mathsf{T}}$. Then

$$\begin{aligned}&\mathrm{D}_{\mathrm{KL}}\big(\rho_1\|\varphi_{V,b}(\rho_2)\big) \\ &\quad= \frac{1}{2}\Big[\sum_{i=1}^m \log(\sigma_i) + \frac{1}{(m+2)}\sum_{i=1}^m \frac{1}{\sigma_i}\Big] \\ &\qquad + \log\Gamma\Big(\frac{m}{2}+1\Big) + \frac{m\log 2}{2} \\ &\qquad + \min_{b\in\mathbb{R}^m} (V\mu_2 + b)^{\mathsf{T}} V\Sigma V^{\mathsf{T}}(V\mu_2 + b) \\ &\quad= \frac{1}{2}\Big[\sum_{i=1}^m \log(\sigma_i) + \frac{1}{(m+2)}\sum_{i=1}^m \frac{1}{\sigma_i}\Big] \\ &\qquad + \log\Gamma\Big(\frac{m}{2}+1\Big) + \frac{m\log 2}{2}, \end{aligned} \tag{12}$$

where the minimum is attained at $b = -V\mu_2$ and Γ is the Gamma function. Let $g(\sigma) \coloneqq \log(\sigma)/2 + 1/[2(m+2)\sigma]$, which has global minimum at $\sigma = 1/(m+2)$. For any $\alpha \ge \beta \ge 0$,

$$g_m(\alpha,\beta) \coloneqq \begin{cases} g(\beta) & \text{if } \beta > \frac{1}{m+2}, \\ g\big(\frac{1}{m+2}\big) & \text{if } \beta \le \frac{1}{m+2} \le \alpha, \\ g(\alpha) & \text{if } \alpha < \frac{1}{m+2}. \end{cases} \tag{13}$$

Thus when $m = 1$, we have

$$\begin{aligned}\widehat{\mathrm{D}}_{\mathrm{KL}}(\rho_1\|\rho_2) &= g_1(\lambda_1,\lambda_n) + \frac{1}{2}\log\frac{\pi}{2} \\ &= \begin{cases} \frac{1}{2}\log\frac{\pi}{2} + \frac{1}{6\lambda_n} + \frac{1}{2}\log\lambda_n & \text{if } \lambda_n > \frac{1}{3}, \\ \frac{1}{2}\log\frac{\pi}{6} + \frac{1}{2} & \text{if } \lambda_n \le \frac{1}{3} \le \lambda_1, \\ \frac{1}{2}\log\frac{\pi}{2} + \frac{1}{6\lambda_1} + \frac{1}{2}\log\lambda_1 & \text{if } \lambda_1 < \frac{1}{3}. \end{cases}\end{aligned}$$

Note that setting $n = 3$ answers the question we posed in the abstract: What is the KL-divergence between the uniform distribution $\rho_1 = \mathcal{U}([-1,1])$ and the Gaussian distribution $\rho_2 = \mathcal{N}_3(\mu_2, \Sigma)$ in $\mathbb{R}^3$.

More generally, suppose $m < n/2$. For any $\sigma_1 \ge \cdots \ge \sigma_m \ge 0$ with

$$\lambda_{n-m+i} \le \sigma_i \le \lambda_i \qquad i = 1,\dots,m,$$

we construct $V \in \mathrm{O}(m,n)$ with $V\Sigma V^{\mathsf{T}} = \mathrm{diag}(\sigma_1,\dots,\sigma_m)$. Let $\Sigma = Q\Lambda Q^{\mathsf{T}}$ be an eigenvalue decomposition with $Q = [q_1,\dots,q_n] \in \mathrm{O}(n)$. For each $i = 1,\dots,m$, let

$$\begin{aligned} v_i(\theta_i) &\coloneqq q_i \sin\theta_i + q_{n-m+i}\cos\theta_i \in \mathbb{R}^n, \\ \sigma_i(\theta_i) &\coloneqq \lambda_i \sin^2\theta_i + \lambda_{n-m+i}\cos^2\theta_i \in \mathbb{R}_+. \end{aligned}$$

Then $V = [v_1(\theta_1)^{\mathsf{T}},\dots,v_m(\theta_m)^{\mathsf{T}}]^{\mathsf{T}} \in \mathrm{O}(m,n)$ and $V\Sigma V^{\mathsf{T}} = \mathrm{diag}\big(\sigma_1(\theta_1),\dots,\sigma_m(\theta_m)\big)$. Choosing θ_i so that $\sigma_i(\theta_i) = \sigma_i$, $i = 1,\dots,m$, gives us the required result.

With this observation, it follows that when $m < n/2$, the minimum in (12) is attained when $\sigma_i = \lambda_i$, $i = 1,\dots,m$, and we obtain the closed-form expression

$$\widehat{\mathrm{D}}_{\mathrm{KL}}(\rho_1\|\rho_2) = \sum_{i=1}^m g_m(\lambda_i,\lambda_{n-m+i}) + \log\Gamma\Big(\frac{m}{2}+1\Big) + \frac{m\log 2}{2},$$

where g_m is as defined in (13).

Example 4 (2-Wasserstein Distance Between Dirac Measure on $\mathbb{R}^m$ and Discrete Measure on $\mathbb{R}^n$): Let $y \in \mathbb{R}^m$ and $\rho_1 \in \mathrm{M}(\mathbb{R}^m)$ be the Dirac measure with $\rho_1(y) = 1$, i.e., all mass centered at y. Let $x_1,\dots,x_k \in \mathbb{R}^n$ be distinct points, $p_1,\dots,p_k \ge 0$, $p_1 + \cdots + p_k = 0$, and let $\rho_2 \in \mathrm{M}(\mathbb{R}^n)$ be the discrete measure of point masses with $\rho_2(x_i) = p_i$, $i = 1,\dots,k$. We seek the 2-Wasserstein distance $\widehat{W}_2(\rho_1,\rho_2)$ and by Theorem 4, this is given by $\mathrm{W}^-_2(\rho_1,\rho_2)$. We will show that it has a closed-form solution. Suppose $m \le n$, then

$$\begin{aligned}\mathrm{W}^-_2(\rho_1,\rho_2)^2 &= \inf_{V\in\mathrm{O}(m,n),\, b\in\mathbb{R}^m} \sum_{i=1}^k p_i\|Vx_i + b + y\|_2^2 \\ &= \inf_{V\in\mathrm{O}(m,n)} \sum_{i=1}^k p_i \Big\|Vx_i - \sum_{i=1}^k p_i V x_i\Big\|_2^2 \\ &= \inf_{V\in\mathrm{O}(m,n)} \mathrm{tr}(VXV^{\mathsf{T}}),\end{aligned}$$

noting that the second infimum is attained by $b = -y - \sum_{i=1}^k p_i V x_i$ and defining X in the last infimum to be

$$X \coloneqq \sum_{i=1}^k p_i\Big(x_i - \sum_{i=1}^k p_i x_i\Big)\Big(x_i - \sum_{i=1}^k p_i x_i\Big)^{\mathsf{T}} \in \mathbb{R}^{n\times n}.$$

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

Let the eigenvalue decomposition of the symmetric positive semidefinite matrix X be $X = Q\Lambda Q^{\mathrm{T}}$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$. Then

$$\inf_{V \in \mathrm{O}(m,n)} \operatorname{tr}(VXV^{\mathrm{T}}) = \sum_{i=0}^{m-1} \lambda_{n-i}$$

and is attained when $V \in \mathrm{O}(m,n)$ has row vectors given by the last m columns of $Q \in \mathrm{O}(n)$.

Example 5 (2-Wasserstein Distance Between Discrete Measures on $\mathbb{R}^m$ and $\mathbb{R}^n$): More generally, we may seek the 2-Wasserstein distance between discrete probability measures $\rho_1 \in \mathrm{M}(\mathbb{R}^m)$ and $\rho_2 \in \mathrm{M}(\mathbb{R}^n)$. Let ρ_1 be supported on $x_1, \dots, x_k \in \mathbb{R}^m$ with values $\rho_1(x_i) = p_i$, $i = 1, \dots, k$; and $y_1, \dots, y_l \in \mathbb{R}^n$ with values $\rho_2(x_i) = q_i$, $i = 1, \dots, l$. The optimization problem for $\mathrm{W}_2^-(\rho_1, \rho_2)$ becomes

$$\inf_{V \in \mathrm{O}(m,n), b \in \mathbb{R}^m,\ \pi \in \Gamma(\rho_1, \rho_2)} \sum_{i=1}^{k} \sum_{j=1}^{l} \pi_{ij} \|Vx_i + b - y_j\|_2^2, \quad (14)$$

where

$$\Gamma(\rho_1, \rho_2) = \Bigg\{ \pi \in \mathbb{R}_+^{k \times l} : \sum_{j=1}^{l} \pi_{ij} = p_i,\ i = 1, \dots, k; \\ \sum_{i=1}^{k} \pi_{ij} = q_j,\ j = 1, \dots, l \Bigg\}.$$

While the solution to (14) may no longer be determined in closed-form, it is a polynomial optimization problem and can be solved using the Lasserre sum-of-squares technique as a sequence of semidefinite programs [68].

VII. Conclusion

We proposed a simple, natural framework for taking any p-Wasserstein metric or f-divergence, and constructing a corresponding distance for probability distributions on m- and n-dimensional measure spaces where $m \neq n$. The new distances preserve some well-known properties satisfied by the original distances. We saw from several examples that the new distances may be either determined in closed-form or near closed-form, or computed using Stiefel manifold optimization or sums-of-squares polynomial optimization. In future work, we hope to apply our framework to other distances like the Bhattacharyya distance [69], the Lévy–Prokhorov metric [15], [70], and the Łukaszyk–Karmowski metric [71].

Acknowledgment

The authors would like to thank the two anonymous referees for their exceptionally helpful comments and suggestions that vastly improved our article. Yuhang Cai would like to express his gratitude to Philippe Rigollet, Jonathan Niles-Weed, and Geoffrey Schiebinger for teaching him about optimal transport and for their hospitality when he visited MIT in 2018. Lek-Heng Lim would like to thank Louis H. Y. Chen for asking the question on p. 4020.

References

[1] A. Irpino and R. Verde, "Dynamic clustering of interval data using a Wasserstein-based distance," *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1648–1658, Aug. 2008.
[2] S. J. Sheather, "Density estimation," *Stat. Sci.*, vol. 19, pp. 588–597, Nov. 2004.
[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*.
[4] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: Fine-grained image generation through asymmetric training," 2017, *arXiv:1703.10155*.
[5] A. Guntuboyina, "Lower bounds for the minimax risk using F-divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, Jun. 2011.
[6] L.-H. Lim, R. Sepulchre, and K. Ye, "Geometric distance between positive definite matrices of different dimensions," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5401–5405, Sep. 2019.
[7] K. Ye and L.-H. Lim, "Schubert varieties and distances between subspaces of different dimensions," *SIAM J. Matrix Anal. Appl.*, vol. 37, no. 3, pp. 1176–1197, Sep. 2016.
[8] F. Mémoli, "Gromov–Wasserstein distances and the metric approach to object matching," *Found. Comput. Math.*, vol. 11, no. 4, pp. 417–487, Aug. 2011.
[9] E. M. Loiola, N. M. M. D. Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido, "A survey for the quadratic assignment problem," *Eur. J. Oper. Res.*, vol. 176, no. 2, pp. 657–690, Jan. 2007.
[10] A. Salmona, J. Delon, and A. Desolneux, "Gromov-Wasserstein distances between Gaussian distributions," 2021, *arXiv:2104.07970*.
[11] J. K. Pachl, "Disintegration and compact measures," *Math. Scandinavica*, vol. 43, pp. 157–168, Jun. 1978.
[12] C. Villani, *Optimal Transport: Old New* (Grundlehren der Mathematischen Wissenschaften), vol. 338. Berlin, Germany: Springer-Verlag, 2009.
[13] M. Fréchet, "Sur la distance de deux lois de probabilitë," *CR Acad. Sci. Paris*, vol. 244, pp. 689–692, 1957.
[14] L. V. Kantorovich, "On a problem of Monge," *Zap. Nauchn. S.-Peterburg. Otdel. Math. Inst. Steklov.*, vol. 312, no. 11, pp. 15–16, 2004.
[15] P. Lévy, *Theorie de l'Addition des Variables Aleatoires*, vol. 1, E. Borel, Ed. Paris, France: Gauthier–Villars, 1937.
[16] L. N. Vasershtein, "Markov processes over denumerable products of spaces describing large system of automata," *Problems Inform. Transmiss.*, vol. 5, no. 3, pp. 47–52, 1969.
[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, Nov. 2000.
[18] J. Solomon *et al.*, "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Trans. Graph.*, vol. 34, no. 4, p. 66, 2015.
[19] R. Sandler and M. Lindenbaum, "Nonnegative matrix factorization with earth mover's distance metric for image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1590–1602, Aug. 2011.
[20] J. Delon, "Midway image equalization," *J. Math. Imag. Vis.*, vol. 21, no. 2, pp. 119–134, 2004.
[21] J. Gutierrez, J. Rabin, B. Galerne, and T. Hurtut, "Optimal patch assignment for statistically constrained texture synthesis," in *Scale Space Variational Methods Computer Vision*. Cham, Switzerland: Springer, 2017, pp. 172–183.
[22] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 254–269, 2013.
[23] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum, "An image morphing technique based on optimal mass preserving mapping," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1481–1495, Jun. 2007.
[24] N. Courty, R. Flamary, D. Tuia, and T. Corpetti, "Optimal transport for data fusion in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 3571–3574.
[25] W. Wang, J. A. Ozolek, D. Slepáev, A. B. Lee, C. Chen, and G. K. Rohde, "An optimal transportation approach for nuclear structure-based pathology," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 621–631, Mar. 2011.
[26] Y. Makihara and Y. Yagi, "Earth mover's morphing: Topology-free shape morphing using cluster-based emd flows," in *Computer Vision*. Berlin, Germany: Springer, 2010, pp. 202–215.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.

[27] B. Mathon, F. Cayre, P. Bas, and B. Macq, "Optimal transport for secure spread-spectrum watermarking of still images," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1694–1705, Apr. 2014.
[28] A. Galichon, *Optimal Transport Methods in Economics*. Princeton, NJ, USA: Princeton Univ. Press, 2016.
[29] U. Frisch, S. Matarrese, R. Mohayaee, and A. Sobolevski, "A reconstruction of the initial conditions of the universe by optimal mass transportation," *Nature*, vol. 417, pp. 260–262, 2002.
[30] R. Flamary, C. Févotte, N. Courty, and V. Emiya, "Optimal spectral transportation with application to music transcription," in *Proc. 30th Int. Conf. Adv. Neural Inform. Process. Sys.*, 2016, pp. 703–711.
[31] M. Zhang, Y. Liu, H. Luan, M. Sun, T. Izuha, and J. Hao, "Building earth mover's distance on bilingual word embeddings for machine translation," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2870–2876.
[32] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 2015, pp. 957–966.
[33] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[34] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Dover, 1997.
[35] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[36] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Phys. Rev. Lett.*, vol. 103, no. 21, Nov. 2009, Art. no. 214101.
[37] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Stat. Sci.*, vol. 10, no. 3, pp. 273–304, Aug. 1995.
[38] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Berkeley, CA, USA, 1961, pp. 547–561.
[39] I. Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, vol. 8, pp. 85–108, Oct. 1963.
[40] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.
[41] E. Hellinger, "Neue Begr" undung der Theorie quadratischer Formen von unendlichvielen Veränderlichen," *J. Reine Angew. Math.*, vol. 136, pp. 210–271, Oct. 1909.
[42] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.
[43] H. Jeffreys, *Theory of Probability*. Oxford, U.K.: Clarendon Press, 1961.
[44] S. Eguchi, "A differential geometric approach to statistical inference on the basis of contrast functionals," *Hiroshima Math. J.*, vol. 15, no. 2, pp. 341–391, Jan. 1985.
[45] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[46] F. Nielsen, "A family of statistical symmetric divergences based on Jensen's inequality," 2010, *arXiv:1009.4004*.
[47] O. Calin and C. Udrişte, *Geometric Model. Probab. Statistics*. Cham, Switzerland: Springer, 2014.
[48] C. R. Rao, "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance," *Qüestiió*, vol. 19, nos. 1–3, pp. 23–63, Oct. 1995.
[49] R. Nishii and S. Eguchi, "Image classification based on Markov random field models with Jeffreys divergence," *J. Multivariate Anal.*, vol. 97, no. 9, pp. 1997–2008, Oct. 2006.
[50] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," Commun. Signal Process. Lab., University of Michigan, Ann Arbor, MI, USA, Tech. Rep. CSPL-328, 2001.
[51] M. B. Hastings, I. González, A. B. Kallin, and R. G. Melko, "Measuring Renyi entanglement entropy in quantum Monte Carlo simulations," *Phys. Rev. Lett.*, vol. 104, no. 15, Apr. 2010, Art. no. 157201.
[52] S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, *Differential Geometry Statisitics Inference* (Institute of Mathematical Statistics Lecture Notes), vol. 10. Hayward, CA, USA: Institute of Mathematical Statistics, 1987.
[53] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless System*. Cambridge, U.K.: Cambridge University Press, 2011.
[54] S. Itzkovitz, E. Hodis, and E. Segal, "Overlapping codes within protein-coding sequences," *Genome Res.*, vol. 20, no. 11, pp. 1582–1589, Nov. 2010.
[55] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim, "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 8, pp. 2677–2682, Feb. 2009.
[56] Y. Ofran and B. Rost, "Analysing six types of protein–protein interfaces," *J. Mol. Biol.*, vol. 325, no. 2, pp. 377–387, 2003.
[57] S. Dedeo, R. Hawkins, S. Klingenstein, and T. Hitchcock, "Bootstrap methods for the empirical study of decision-making and information flows in social systems," *Entropy*, vol. 15, no. 12, pp. 2246–2276, Jun. 2013.
[58] S. Klingenstein, T. Hitchcock, and S. Dedeo, "The civilizing process in London's old bailey," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 26, pp. 9419–9424, Jul. 2014.
[59] F.-C. Mitroi-Symeonidis, I. Anghel, and N. Minculete, "Parametric Jensen-Shannon statistical complexity and its applications on full-scale compartment fire data," *Symmetry*, vol. 12, no. 1, p. 22, Dec. 2019.
[60] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Adv. Neural Inform. Process. Sys.*, 2014, pp. 2672–2680.
[61] T. Chen and S. Kiefer, "On the total variation distance of labelled Markov chains," in *Proc. EACSL Annu. Conf. Comput. Sci.*, Jul. 2014, pp. 1–10.
[62] J. Ding, E. Lubetzky, and Y. Peres, "Total variation cutoff in birth-and-death chains," *Probab. Theory Relat. Fields*, vol. 146, nos. 1–2, p. 61, 2010.
[63] I. Nourdin and G. Poly, "Convergence in total variation on Wiener chaos," *Stochastic Processes their Appl.*, vol. 123, no. 2, pp. 651–674, Feb. 2013.
[64] S. P. Brooks, P. Dellaportas, and G. O. Roberts, "An approach to diagnosing total variation convergence of MCMC algorithms," *J. Comput. Graph. Statist.*, vol. 6, no. 3, pp. 251–265, Sep. 1997.
[65] E. A. Peköz, A. Röllin, and N. Ross, "Total variation error bounds for geometric approximation," *Bernoulli*, vol. 19, no. 2, pp. 610–632, May 2013.
[66] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, 2005.
[67] I. Olkin and F. Pukelsheim, "The distance between two random vectors with given dispersion matrices," *Linear Algebra Appl.*, vol. 48, pp. 257–263, Oct. 1982.
[68] J. B. Lasserre, *An Introduction to Polynomial and Semi-algebraic Optimization* (Cambridge Texts in Applied Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2015.
[69] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, no. 1, pp. 99–109, 1943.
[70] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Teor. Veroyatnost. Primenen.*, vol. 1, pp. 177–238, Oct. 1956.
[71] S. Lukaszyk, "A new concept of probability metric and its applications in approximation of scattered data sets," *Comput. Mech.*, vol. 33, no. 4, pp. 299–304, Mar. 2004.

Authorized licensed use limited to: UNIV OF CHICAGO LIBRARY. Downloaded on May 21,2022 at 03:09:22 UTC from IEEE Xplore. Restrictions apply.