ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Statistical Aspects of Wasserstein Distances

Victor M. Panaretos[1] and Yoav Zemel[2]

[1]Institute of Mathematics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; email: victor.panaretos@epfl.ch

[2]Institut für Mathematische Stochastik, Georg–August–Universität, 37077 Göttingen, Germany; email: yoav.zemel@mathematik.uni-goettingen.de

Annu. Rev. Stat. Appl. 2019. 6:405–31

First published as a Review in Advance on November 2, 2018

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-030718-104938

Keywords

deformation map, empirical optimal transport, Fréchet mean, goodness-of-fit, inference, Monge–Kantorovich problem, optimal coupling, probability metric, transportation of measure, warping, registration, Wasserstein space

Abstract

Wasserstein distances are metrics on probability distributions inspired by the problem of optimal mass transportation. Roughly speaking, they measure the minimal effort required to reconfigure the probability mass of one distribution in order to recover the other distribution. They are ubiquitous in mathematics, with a long history that has seen them catalyze core developments in analysis, optimization, and probability. Beyond their intrinsic mathematical richness, they possess attractive features that make them a versatile tool for the statistician: They can be used to derive weak convergence and convergence of moments, and can be easily bounded; they are well-adapted to quantify a natural notion of perturbation of a probability distribution; and they seamlessly incorporate the geometry of the domain of the distributions in question, thus being useful for contrasting complex objects. Consequently, they frequently appear in the development of statistical theory and inferential methodology, and they have recently become an object of inference in themselves. In this review, we provide a snapshot of the main concepts involved in Wasserstein distances and optimal transportation, and a succinct overview of some of their many statistical aspects.

1. INTRODUCTION

Wasserstein distances are metrics between probability distributions that are inspired by the problem of optimal transportation. These distances (and the optimal transport problem) are ubiquitous in mathematics, most notably in fluid mechanics, partial differential equations, optimization, and, of course, probability theory and statistics. In addition to their theoretical importance, they have provided a successful framework for the comparison of (at times complex) objects in fields of application such as image retrieval (Rubner et al. 2000), computer vision (Ni et al. 2009), pharmaceutical statistics (Munk & Czado 1998), genomics (Bolstad et al. 2003, Evans & Matsen 2012), economics (Gini 1914) and finance (Rachev et al. 2011), to name but a few. Indeed, while their origins lie with Monge's (primarily mathematical) inquiry into how to optimally transport a pile of earth of a given volume into a pit of equal volume but potentially different shape, Kantorovich's modern reformulation, which catalyzed the development of this rich theory, was inspired by the concrete problem of optimal resource allocation. Unsurprisingly, there is a vast literature on Wasserstein distances and optimal transportation, originally rooted primarily in analysis and probability, but later branching out to quantitative fields well beyond. In statistics, Wasserstein distances play a prominent role in theory and methodology, and more recently have become an object of inference in themselves. In his thousand-page book, Villani (2008, p. 2) writes that reviewing the optimal transport literature is a "dauntingly difficult task," and if one focuses more narrowly on statistical aspects of Wasserstein distances, it is still impossible to carry out a comprehensive review in the order of twenty-five pages. We thus restrict ourselves to a high level overview of some salient aspects and main concepts, admittedly influenced by our own perspective and interests, and apologize for the inevitable omissions.

1.1. Overview

Wasserstein distances appear in statistics in several ways. We delineate three broad categories of statistical use of these distances, according to which we will structure our review:

1. Wasserstein distances and the associated notion of an optimal coupling are often exploited as a versatile tool in asymptotic theory due to the topological structure they induce and their relatively easy majorization, and Section 2 reviews some of their appealing features in that context.
2. In other cases, Wasserstein distances are employed as a methodological tool in order to carry out statistical inference, primarily involving structural models and goodness-of-fit testing. Section 3 describes key methods and results in this vein.
3. Finally, a recent trend in functional data analysis is to consider the space of probability measures equipped with a Wasserstein distance as a sample/parameter space itself, a direction that is taken up in Section 4.

In contexts such as 2 and 3, it is often important to carry out explicit computations related to the Wasserstein distance, and Section 5 gives a brief overview of such numerical aspects. First, we review the basic definitions and relevant notions that we require throughout the article.

1.2. Basic Notions

The p-Wasserstein distance[1] between probability measures μ and ν on $\mathbb{R}^d$ is defined as

$$W_p(\mu,\nu) = \inf_{\substack{X\sim\mu\\ Y\sim\nu}} (\mathbb{E}\|X-Y\|^p)^{1/p}, \qquad p \geq 1, \tag{1.}$$

[1] This is also known as Mallows' distance, the earth mover's distance, the (Monge–)Kantorovich(–Rubinstein) distance, or the Fréchet distance (when $p = 2$). The term "Wasserstein distance" became popular, mainly in Western literature, following Dobrushin (1970), who studied some of its topological properties and referred to an earlier work by Wasserstein. Villani (2008, p. 118) and Bobkov & Ledoux (2018, p. 4) provide more details.

where the infimum is taken over all pairs of d-dimensional random vectors X and Y marginally distributed as μ and ν, respectively (an obviously nonempty set, since one can always construct independent random variables with prescribed marginals). For convenience, we use both notations $W_p(X, Y)$ and $W_p(\mu, \nu)$ interchangeably whenever $X \sim \mu$ and $Y \sim \nu$. The distance is finite provided the pth moments exist, $\mathbb{E}\|X\|^p + \mathbb{E}\|Y\|^p < \infty$, and this will be tacitly assumed in what follows. The definition generalizes to laws defined on much more general spaces: If $(\mathcal{X}, \rho)$ is any complete and separable metric space, W_p can be defined in the same way, with $\|X - Y\|$ replaced by the metric $\rho(X, Y)$. In particular, this setup incorporates laws on infinite-dimensional function spaces such as $L^2[0, 1]$. For simplicity, we restrict to the setting where $\mathcal{X}$ is a normed vector space, employing the notation $(\mathcal{X}, \|\cdot\|)$ henceforth.

The optimization problem defining the distance is typically referred to in the literature as optimal transport(ation) or the Monge–Kantorovich problem. When X and Y take values on the real line, their joint distribution is characterized by specifying their marginal distributions and a copula (Sklar 1959). Since the marginals here are fixed to be the laws of X and Y, the problem is to find a copula that couples X and Y together as tightly as possible in an L_p-sense, on average; if $p = 2$ then that copula is the one that maximizes the correlation (or covariance) between X and Y, i.e., the copula inducing maximal linear dependence.

The Wasserstein distances W_p are proper distances in that they are nonnegative, are symmetric in X and Y, and satisfy the triangle inequality. A compactness argument shows that the infimum in their definition is indeed attained (if $\mathcal{X}$ is complete). When endowed with the distance W_p, the space of measures with finite pth moments—the Wasserstein space $\mathcal{W}_p(\mathcal{X})$—is complete and separable if $\mathcal{X}$ is so. Although many other metrics can be defined on the space of probability measures (Rachev 1991, Gibbs & Su 2002), Wasserstein distances exhibit some particularly attractive features:

- They incorporate the geometry of the ground space $\mathcal{X}$: If X and Y are degenerate at points $x, y \in \mathcal{X}$, then $W_p(X, Y)$ is equal to the distance between x and y in $\mathcal{X}$. This property hints at why Wasserstein distances are successful in imaging problems and why they can capture the human perception of whether images are similar or not (see Section 4).
- Convergence of X_n to X in Wasserstein distance is equivalent to convergence in distribution, supplemented with $\mathbb{E}\|X_n\|^p \to \mathbb{E}\|X\|^p$. This makes W_p convenient for proving central limit theorem–type results (see Section 2).
- Since they are defined as the solution of minimization problems, they are quite easy to bound from above: Any joint distribution with the correct marginals provides an upper bound for the Wasserstein distance (see Section 2). Moreover, they enjoy some differentiability, allowing for application of the delta method (see Section 3).

In addition to the probabilistic definition (Equation 1), one can consider the analytic definition, which helps dissect the structure of the Monge–Kantorovich optimization problem:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}. \qquad 2.$$

Here, $\Gamma(\mu, \nu)$ is the set of probability measures γ on $\mathcal{X} \times \mathcal{X}$ satisfying $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$ for all Borel subsets $A, B \subseteq \mathcal{X}$. Elements $\gamma \in \Gamma(\mu, \nu)$ are called couplings of μ and ν, i.e., joint distributions on $\mathcal{X} \times \mathcal{X}$ with prescribed marginals μ and ν on each axis, which hopefully elucidates the equivalence to the probabilistic definition given by Equation 1. The analytical definition (Equation 2) has a simple intuitive interpretation in the discrete case: Given a $\gamma \in \Gamma(\mu, \nu)$ and any pair of locations (x, y), the value of $\gamma(x, y)$ tells us what proportion of μ's mass at x ought to be transferred to y in order to reconfigure μ into ν. Quantifying the

effort of moving a unit of mass from x to y by $\|x-y\|^p$ yields the interpretation of $W_p(\mu,\nu)$ as the minimal effort required to reconfigure μ's mass distribution into that of ν.

The analytical definition given by Equation 2 underlines that the feasible set Γ is convex and that the objective function is (up to the power $1/p$) linear in γ. Optimal γs can thus be expected to be extremal, that is, relatively sparse. Examples of such sparse couplings are deterministic ones, i.e., couplings supported on the graph of some deterministic function $T:\mathcal{X}\to\mathcal{X}$, rather than on $\mathcal{X}\times\mathcal{X}$, so that they can be realized as

$$\gamma(A\times B)=\mu(A\cap T^{-1}(B)).$$

Such a coupling reassigns all of μ's mass at a given location to a unique destination. When the vector (X,Y) is distributed according to such a γ, its two coordinates are completely dependent: $Y=T(X)$ for the deterministic function $T:\mathcal{X}\to\mathcal{X}$. Such T is called an optimal transport map and must satisfy $\nu(B)=\mu(T^{-1}(B))$ for all $B\subseteq\mathcal{X}$ if γ is to be in Γ, i.e., T pushes μ forward to ν (denoted by $T\#\mu=\nu$). **Figure 1** illustrates these definitions.

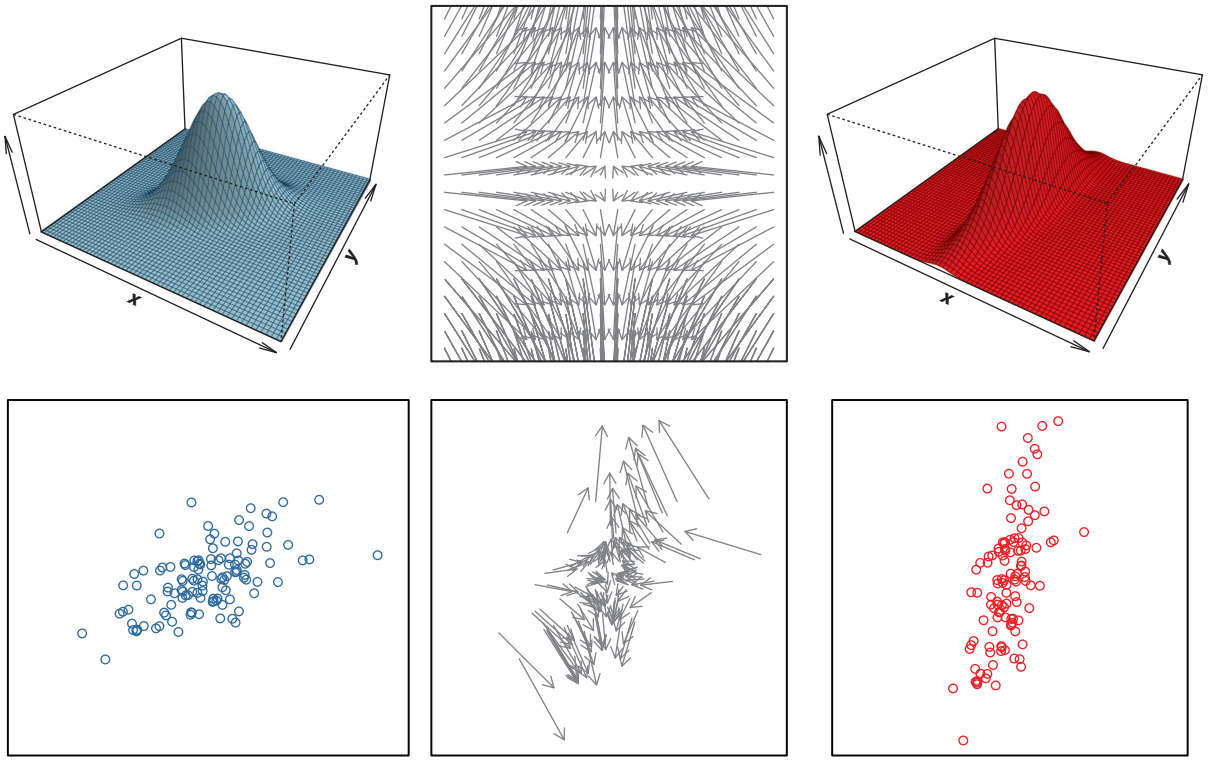

Figure 1

Illustration of the analytic and probabilistic definitions of the p-Wasserstein distance. The top row of plots shows the densities of two Gaussian probability measures μ (*left*, *blue*) and ν (*right*, *red*), and the optimal deterministic map T (*middle*, *gray*) that deforms μ into ν, i.e., $T\#\mu=\nu$. The map is plotted in the form of the vector field $T(x)-x$, where each arrow indicates the source and destination of the mass being transported. Reversing the direction of the arrows would produce the inverse map, optimally deforming the measure ν to obtain μ. The bottom row features two independent random samples $X_1,\dots,X_N \overset{\text{i.i.d.}}{\sim} \mu$ (*left*, *blue*) and $Y_1,\dots,Y_N \overset{\text{i.i.d.}}{\sim} \nu$ (*right*, *red*), for $N=120$. The sample $\{X_i\}_{i=1}^N$ was constructed by sampling μ directly. The sample $\{Y_i\}_{i=1}^N$ was constructed by applying the optimal map T to the sample $\{X_i\}_{i=1}^N$, i.e., $Y_i=T(X_i)$. The plot in the middle illustrates how the sample $\{X_i\}_{i=1}^N$ is rearranged in order to produce the sample $\{Y_i\}_{i=1}^N$, by plotting the vectors $T(X_i)-X_i$. The optimality of T can be understood in terms of minimizing the average squared length of these arrows. In all plots, the x and y axes range from -3 to 3. Abbreviation: i.i.d., independent and identically distributed.

As it turns out, under sufficient regularity, such deterministic couplings are optimal. When $\mathcal{X} = \mathbb{R}^d$ is finite-dimensional and μ is absolutely continuous with respect to Lebesgue measure, the infimum (if finite) is attained (uniquely if $p > 1$) by such a deterministic coupling. In this case we denote the map T inducing the coupling by $\mathbf{t}_X^Y$ or $\mathbf{t}_\mu^\nu$. In the next paragraph we briefly sketch the arguments leading to this result. As the problem is analytical in nature, characterizing the solutions requires some tools from mathematical analysis. We have attempted to avoid technicalities to the extent possible, but with optimal transport, the devil is in the details, as the problem is qualitatively different depending on whether the random variables are discrete or continuous. The less mathematically inclined reader can skip to the paragraph containing Equation 3, simply retaining the loose statement that in the quadratic case $p = 2$, optimal maps are characterized as gradients of convex functions. Our presentation mainly follows Villani (2003); more references are given at the end of this section.

1.2.1. Uniqueness and characterization. Like any convex optimization problem, the Monge–Kantorovich problem admits a Lagrangian dual problem, consideration of which leads to a characterization of optimal maps. The dual problem can be seen as

$$\sup_{\phi,\psi} \left\{ \mathbb{E}\phi(X) + \mathbb{E}\psi(Y) \right\}, \qquad \text{subject to} \quad \phi(x) + \psi(y) \le \|x - y\|^p$$

for integrable ϕ and ψ. The inequality $\mathbb{E}\phi(X) + \mathbb{E}\psi(Y) \le \mathbb{E}\|X - Y\|^p$ implies weak duality, in that the above supremum is no larger than the infimum in Definition 1. But under mild conditions there is strong duality, and there exist a pair (ϕ, ψ) and a joint coupling γ such that $\mathbb{E}\phi(X) + \mathbb{E}\psi(Y) = \mathbb{E}_\gamma \|X - Y\|^p$. Furthermore, a version of complementary slackness holds between the two optimal solutions in such a way that one provides much information on the other. This is best demonstrated in the quadratic case $p = 2$, by virtue of the factorization $\|x-y\|^2 = \|x\|^2+\|y\|^2-2\langle x, y\rangle$. Algebraic manipulations then allow the dual to be recast as

$$\inf_{\varphi,\Psi} \left\{ \mathbb{E}\varphi(X) + \mathbb{E}\Psi(Y) \right\}, \qquad \text{subject to} \quad \varphi(x) + \Psi(y) \ge \langle x, y \rangle .$$

A simple yet consequential observation is that for a given φ, the best candidate for Ψ is the Legendre transform of φ,

$$\varphi^*(y) = \sup_{x \in \mathcal{X}} \{\langle x, y\rangle - \varphi(x)\},$$

the smallest function satisfying $\varphi^*(y) + \varphi(x) \ge \langle x, y\rangle$. Iterating this idea amounts to replacing φ by $\varphi^{**} = (\varphi^*)^*$, which is larger than φ yet still obeys the constraint $\varphi^{**}(x) + \varphi^*(y) \ge \langle x, y\rangle$. The choice $\Psi = \varphi^*$ makes the dual unconstrained, and φ is optimal if and only if $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$ with probability one with respect to X and Y. Going back to the primal problem, we see that once an optimal φ is found, a joint distribution will be optimal if and only if it assigns unit probability to the event $\varphi(X) + \varphi^*(Y) = \langle X, Y\rangle$. Furthermore, φ itself may be assumed to be the Legendre transform of φ^*, namely $\varphi = \varphi^{**}$.

At this stage one can invoke the rich theory of convex analysis. Legendre transforms are always convex, and the equality $\varphi(x) + \varphi^*(y) = \langle x, y\rangle$ holds if and only if y is a subgradient of φ at x. If φ has a unique subgradient y at x, then $y = \nabla\varphi(x)$ is the gradient of φ and is determined uniquely. The regularity of convex functions implies that this is the case for all x up to a set of Lebesgue measure 0. Thus, if X has a density, then the optimal map T is characterized as the unique gradient of a convex function that pushes X forward to Y. On the other hand, if X is discrete, then it might be concentrated on the small set where φ is not differentiable, in which case the optimal coupling will not be induced from a map.

Similar arguments apply for other values of $p > 1$. For a given ϕ, the best candidate for ψ is the c-transform[2] of ϕ,

$$\phi^c(y) = \inf_{x \in \mathcal{X}} \{\|x - y\|^p - \phi(x)\},$$

which again leads to an unconstrained dual problem $\sup_\phi \mathbb{E}\phi(X) + \phi^c(Y)$. A function ϕ is optimal if and only if $\phi(x) + \phi^c(y) = \|x - y\|^p$ with probability one, and ϕ itself can be assumed to be a c-transform. In analogy with the quadratic case, the equality $\phi(x) + \phi^c(y) = \|x - y\|^p$ entails a relation between y and the gradient of ϕ, and c-transforms enjoy differentiability properties similar to those of convex functions.

In summary, when X has a density, optimal maps $\mathbf{t}_X^Y$ are precisely functions of the form

$$\mathbf{t}_X^Y(x) = \begin{cases} \nabla\varphi(x) \text{ for some convex } \varphi, & p = 2, \\ x - \|\nabla\phi(x)\|^{1/(p-1)-1}\nabla\phi(x) \text{ for some } c\text{-transform } \phi, & p \neq 2. \end{cases} \quad 3.$$

This formula for general p is also valid if $p = 2$, with $\phi(x) = \|x\|^2/2 - \varphi(x)$. Importantly, this uniqueness and characterization result holds for two classes of spaces $\mathcal{X}$ extending $\mathbb{R}^d$: Riemannian manifolds and separable Hilbert spaces.

1.2.2. Regularity. The convex gradient characterization gives rise to a rich regularity theory in the quadratic case. When both μ and ν have densities f and g, the convex potential φ solves the Monge–Ampère equation,

$$\det\nabla^2\varphi(x) = \frac{f(x)}{g(\nabla\varphi(x))}.$$

The regularity theory of Monge–Ampère equations allows one to deduce the smoothness of the optimal map $T = \nabla\varphi$. Roughly speaking, if X and Y have convex supports and positive, bounded densities with derivatives up to order $k \geq 0$, then the optimal map $\mathbf{t}_\mu^\nu$ has continuous derivatives up to order $k + 1$.

1.2.3. Explicit solutions. Apart from the characterization of optimal maps T as gradients of convex functions (when $p = 2$) or c-transforms, typically neither T nor the Wasserstein distance W_p admit closed-form expressions. There are two special yet important cases with explicit formulae. When $d = 1$, denoting F_X and $F_X^{-1}(q) = \inf\{x : F_X(x) \geq q\}$, $q \in (0, 1)$, the distribution and quantile functions of X, we have

$$W_p(X, Y) = \|F_X^{-1} - F_Y^{-1}\|_p = \left(\int_0^1 |F_X^{-1}(\alpha) - F_Y^{-1}(\alpha)|^p \, d\alpha\right)^{1/p}, \qquad \mathbf{t}_X^Y = F_Y^{-1} \circ F_X,$$

where $\mathbf{t}_X^Y$ is optimal if X is a continuous random variable. This allows the quantile function F_Y^{-1} of any random variable Y to be interpreted as the optimal map from a uniform random variable to Y (also see Section 6 for an interesting interpretation/extension of this fact). In the special case $p = 1$, there is an alternative, often more convenient, formula:

$$W_1(X, Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| \, dt.$$

[2]Here, the cost of transferring a unit of mass from x to y is $c(x, y) = \|x - y\|^p$, but these ideas are valid for more general cost functions c, hence the name.

The function $\mathbf{t}_X^Y = F_Y^{-1} \circ F_X$ is still optimal but might not be unique. One can also bound W_p in terms of the distribution functions:

$$W_p^p(X, Y) \leq p 2^{p-1} \int_{\mathbb{R}} |t|^{p-1} |F_X(t) - F_Y(t)| \, \mathrm{d}t.$$

The other case where closed-form formulae are available is when X and Y are Gaussian. If $X \sim N(m_1, \Sigma_1)$ and $Y \sim N(m_2, \Sigma_2)$, then

$$\begin{aligned} W_2^2(X, Y) &= \|m_1 - m_2\|^2 + \mathrm{tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2}], \\ \mathbf{t}_X^Y(x) &= m_2 + \Sigma_1^{-1/2} [\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}]^{1/2} \Sigma_1^{-1/2} (x - m_1), \end{aligned} \tag{4.}$$

where $\mathbf{t}_X^Y$ is defined if Σ_1 is injective (more generally, if its kernel is included in that of Σ_2). These formulae are valid in infinite dimensions too, in which case $\mathbf{t}_X^Y$ may be unbounded, and only defined on an affine subspace of $\mathcal{X}$. Furthermore, this result holds in location-scale families that are not necessarily Gaussian.

1.3. Bibliographic Notes

In addition to the survey of Rachev (1985), there are a number of books dedicated to optimal transport: Rachev & Rüschendorf (1998a,b), Villani (2003, 2008), Ambrosio & Gigli (2013), Santambrogio (2015), and the forthcoming Panaretos & Zemel (2019), leaning to the statistical side of the subject. The reader interested in the extensive bibliography may consult in particular the first, second and fourth of these references. For space considerations, we only give a very brief historical overview and a summary list of references.

The origin of the optimal transport problem is the monograph by Monge (1781), in which he posed the question for the particular case $\mathcal{X} = \mathbb{R}^3$ and $p = 1$; Appell (1886) also provides an early reference. The probabilistic formulation of Kantorovich (1942) was a major breakthrough and one of the catalysts that led Kantorovich to develop linear programming, for which he was awarded the Nobel prize in 1975 (jointly with T.C. Koopman, who independently arrived at similar results after Kantorovich).

Duality results have a rich history dating back at least to Kantorovich & Rubinstein (1958). Very general results (for all Borel cost functions) in this context can be found in Beiglböck & Schachermayer (2011). Kellerer (1984) explores duality in a multimarginal formulation involving more than two measures (see also Section 4).

The one-dimensional case is intrinsically related to the Fréchet–Höffding bounds (Höffding 1940, Fréchet 1951). Readers are directed to Bass (1955) and Dall'Aglio (1956) for early references and to Cuesta-Albertos et al. (1993) for detailed discussion when $p = 2$. The bound for W_p in terms of distribution functions is due to Ebralidze (1971) and can be found in generalized form in Bobkov & Ledoux (2018, section 7.4). There are analogous results for measures on spaces with simple structure; see Delon et al. (2010) for the unit circle and Kloeckner (2015) for ultrametric spaces.

For the Gaussian case, readers are directed to Olkin & Pukelsheim (1982) or Givens & Shortt (1984) in finite dimensions, and Gelbrich (1990) and Cuesta-Albertos et al. (1996) for an infinite-dimensional extension.

The convex gradient characterization in the quadratic case was discovered independently by a number of authors: Knott & Smith (1984), Cuesta-Albertos & Matrán (1989), Rüschendorf & Rachev (1990), and Brenier (1991). For other values of the exponent p (and more general cost functions), see Gangbo & McCann (1996). The Riemannian version is due to McCann (2001), and Ambrosio et al. (2008, section 6.2.2) treat the infinite-dimensional case.

The regularity result was discovered by Caffarelli (1992); Figalli (2017) provides an accessible exposition. There are other (e.g., Sobolev) types of regularity results, as explained by Villani (2008, pp. 332–36) or Santambrogio (2015, section 1.7.6).

2. OPTIMAL TRANSPORT AS A TECHNICAL TOOL

This section reviews some of the features of Wasserstein metrics that make them useful as a technical tool for deriving large sample theory results in statistics. To facilitate the presentation, we first state some simple facts that play a role in the development. Let X and Y be random vectors taking values in $\mathcal{X} = \mathbb{R}^d$; we maintain the notation $(\mathcal{X}, \|\cdot\|)$ to stress that the properties are valid in infinite dimensions as well.

- For any real number a, $W_p(aX, aY) = |a| W_p(X, Y)$.
- For any fixed vector $x \in \mathcal{X}$, $W_p(X + x, Y + x) = W_p(X, Y)$.
- If $\mathbb{E}(X) = \mathbb{E}(Y)$, then for any fixed $x \in X$, $W_2^2(X + x, Y) = \|x\|^2 + W_2^2(X, Y)$.
- For product measures and when $p = 2$, we have $W_2^2(\otimes_{i=1}^n \mu_i, \otimes_{i=1}^n \nu_i) = \sum_{i=1}^n W_2^2(\mu_i, \nu_i)$ in the analytic notation.

The proofs of the first three statements rely on the equivalence between the classes of the corresponding couplings. For example, $U = (X + x, Y + x)$ is a coupling of $X+x$ and $Y+y$ if and only if $U-(x, x)$ is a coupling of (X, Y). For the last property, observe that the map $x \mapsto [\mathbf{t}_{\mu_1}^{\nu_1}(x), \ldots, \mathbf{t}_{\mu_1}^{\nu_1}(x)]$ is a gradient of a convex function and pushes forward $\otimes\mu_i$ to $\otimes\nu_i$.

2.1. Deviations from Gaussianity

If $\{X_i\}_{i\geq 1}$ are independent and identically distributed random variables with mean zero and finite variance, then the central limit theorem asserts that the suitably rescaled averages $S_n = n^{1/2}\overline{X}_n$ converge in distribution to a normal random variable Z with the same variance. Since $\mathbb{E}S_n^2 = \mathbb{E}Z^2$, the convergence also holds in 2-Wasserstein distance. This property makes the 2-Wasserstein distance convenient for handling deviations from Gaussianity. The arguments generally involve the subadditivity of the Wasserstein distance with respect to convolutions, a property that can be established using the infimum-over-couplings definition of the Wasserstein distances. For example, assuming $\mathbb{E}X_i = 0$,

$$W_2^2\left(\sum_{i=1}^n a_i X_i, Z\right) \leq \sum_{i=1}^n a_i^2 W_2^2(X_i, Z), \qquad Z \sim N(0,1), \qquad \sum_{i=1}^n a_i^2 = 1. \tag{5.}$$

To see this, let $Z_i \sim N(0, 1)$ be independent and consider optimal couplings on $\mathbb{R}^2$ such that $\mathbb{E}|a_i X_i - a_i Z_i| = W_2^2(a_i X_i, a_i Z_i)$. Take the product π of all these couplings (a joint distribution on $\mathbb{R}^{2n}$). Then, under π, $\sum a_i Z_i$ is standard normal and

$$W_2^2\left(\sum_{i=1}^n a_i X_i, Z\right) \leq \mathbb{E}_\pi \left|\sum_{i=1}^n a_i X_i - \sum_{i=1}^n a_i Z_i\right|^2 = \sum_{i=1}^n \mathbb{E}\,|a_i X_i - a_i Z_i|^2 = \sum_{i=1}^n W_2^2(a_i X_i, a_i Z),$$

from which Equation 5 follows. Mallows (1972) used this property in order to derive necessary and sufficient conditions for a triangular array to be jointly asymptotically normal. Recall that $X_n = (X_{n1}, \ldots, X_{nd})$ converge in distribution to a standard multivariate $N(0, I_d)$ if and only if $a^t X_n \to Z$ for all $a \in \mathbb{R}^d$, $\|a\| = 1$. Now let X_{nj} $(j \leq n < \infty)$ be a triangular array. In analogy with a fixed dimension, we say that (X_{nj}) is jointly asymptotically normal if $a_n^t X_n \to Z$ for any sequence of vectors $a_n \in \mathbb{R}^n$, $\|a_n\| = 1$. This requires X_{nj} to converge to Z uniformly in j, i.e., $X_{nm_n} \to Z$ for any sequence of coordinates $m_n \leq n$. This condition is not sufficient, however.

Mallows (1972) observed that metrics inducing convergence in distribution are not subadditive, and this is remedied by the Wasserstein distance. If $\mathbb{E}X_{nj}^2 \to 1$ uniformly in j, in addition to the uniform convergence in distribution, then $W_2^2(X_{nj}, Z) \to 0$, and as a consequence of Equation 5, $W_2^2(a_n^t X_n, Z) \to 0$, and the array is jointly asymptotically normal. The length of the nth row of the array can be arbitrary, as long as it diverges to infinity with n.

When the X_is in Equation 5 have the same distribution as X and $a_i = 1/\sqrt{n}$, the inequality gives a bound that is uniform in n. Bickel & Freedman (1981) use this result in their study of the asymptotics of the bootstrap. For instance, denote by F_n the empirical distribution function corresponding to a sample $X_1, \ldots, X_n$ and the sample mean by $\mu_n = \overline{X}$. Let $X_1^*, \ldots, X_m^*$ be a bootstrapped sample from F_n with sample average μ_m^*. Then as $n, m \to \infty$, the conditional [upon (X_i)] distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ converges to $N(0, \mathrm{var}(X_1))$, which is the same asymptotic distribution of μ_n.

Another additive property, shown in a similar way to Equation 5, is

$$W_p\left(\sum_{i=1}^n U_i, \sum_{i=1}^n V_i\right) \le \sum_{i=1}^n W_p(U_i, V_i),$$

for independent (U_i) and (V_i). A particular case is that $W_p(X + Y, X) \le W_p(Y, 0) = [\mathbb{E}\|Y\|^p]^{1/p}$, and taking Y to be Gaussian with small variance allows one to approximate in W_p any probability law with a smooth surrogate law to arbitrary precision. In other words, smooth measures are dense in W_p, just as they are dense with respect to convergence in distribution. Discrete measures are also dense (see Section 3.3).

Actually, the subadditivity properties can be used in order to prove the central limit theorem. Tanaka (1973) does so by noticing that equality in Equation 5 holds only for Gaussian distributions. Johnson & Samworth (2005) obtain rates of convergence for the central limit theorem, and more generally, for convergence to stable laws. Berry–Esseen-type bounds for the Wasserstein distance can be found in Rio (2009). For random elements in Banach spaces, readers are directed to Rachev & Rüschendorf (1994).

2.2. Equilibrium, Concentration, and Poisson Approximations

A different class of settings where Wasserstein distances are used is in the study of convergence of Markov chains to their equilibrium distribution; this usage dates back to Dobrushin (1970). The idea is to show a sort of contraction property of the transition kernel with respect to the Wasserstein distance. Let P be the transition matrix. In studying convergence of the Kac random walk on the orthogonal group SO(n), Oliveira (2009) showed that

$$W_{D,2}(\mu P, \nu P) \le \xi W_{D,2}(\mu, \nu)$$

for some $\xi < 1$, where D is a distance between matrices, leading to exponential convergence to equilibrium. A result of similar spirit was derived by Eberle (2014) for the transition kernel of the Metropolis-adjusted Langevin algorithm, a Markov chain Monte Carlo method. The constant ξ above is related to the Wasserstein spectral gap of the transition kernel. Hairer et al. (2014) explored its behavior in infinite-dimensional state spaces, when taking finite-dimensional projections of P. They showed that for the preconditioned Crank–Nicolson algorithm, ξ remains stable, whereas for the random walk Metropolis algorithm, ξ may converge to one. Rudolf & Schweizer (2018) employ Wasserstein distances to bound the difference between the behavior of some nicely behaved Markov chain and a perturbed version thereof, obtained from a modification in the transition kernel.

Wasserstein distances also appear in concentration of measure, in the form of transportation inequalities (Ledoux 2005, chapter 6). A measure μ_0 satisfies such an inequality if, for any other measure ν,

$$W_1(\mu_0, \nu) \le C\sqrt{H(\mu_0, \nu)}, \qquad H(\mu, \nu) = \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\, \mathrm{d}\mu.$$

If this holds, and $\mu(A) \ge 1/2$, then

$$\mathbb{P}(X \notin A_r) \le e^{-r^2/C'}, \qquad A_r = \{x : \|x - A\| \le r\}.$$

Furthermore, the representation of W_1 as the supremum over Lipschitz functions (see the next subsection) yields concentration inequalities for $f(X) - \mathbb{E}f(X)$ with f Lipschitz.

In a different context, Barbour & Brown (1992) use Wasserstein metrics to quantify the error in approximating a point process Ξ by a Poisson point process P with the same mean measure λ. Suppose for simplicity that the sample space is $[0, 1]$, and for two (not necessarily probability) measures $\tilde{\mu}, \tilde{\nu}$ with total masses A and B, define the probabilities $\mu = \tilde{\mu}/A$, $\nu = \tilde{\nu}/B$ and $d(\tilde{\mu}, \tilde{\nu}) = W_1(\mu, \nu)$ if $A = B$ and 1 (the maximal value) if $A \ne B$. The processes Ξ and P can then be viewed as random elements in the metric space $\mathcal{X}$ of measures with the distance d, and their laws can be compared using the upper degree Wasserstein space W_1 on $(\mathcal{X}, d)$. Schuhmacher (2009) provides an extension where d is replaced by a Wasserstein distance of different order W_p.

2.3. Relation to Other Metrics

We conclude this section by reviewing some useful relations between W_p and other probability metrics. We first relate W_p to W_q by two simple results from Villani (2003, chapter 7), and then describe bounds (mostly borrowed from Gibbs & Su 2002) pertaining to W_1 and the Prokhorov, total variation, and bounded Lipschitz distances. For notational simplicity we state the bounds in the Euclidean setting, but they hold on any complete separable metric space $(\mathcal{X}, \rho)$. For random variables X and Y on $\mathcal{X}$, let Ω be the union of their ranges and set

$$D = \sup_{x,y\in\Omega} \|x - y\|, \qquad d_{\min} = \inf_{x\ne y\in\Omega} \|x - y\|.$$

In the analytic version $\Omega = \mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)$, where $X \sim \mu$, $Y \sim \nu$, and supp stands for support. If X and Y are bounded, then D is finite; if X and Y are (finitely) discrete, then $d_{\min} > 0$.

- If $p \le q$, then $W_p \le W_q$, by Jensen's inequality.
- A reverse version also holds, $W_q^q \le W_p^p D^{q-p}$.
- Duality arguments yield the particularly useful Kantorovich–Rubinstein (Kantorovich & Rubinstein 1958) representation for W_1 as

$$W_1(X, Y) = \sup_{\|f\|_{\mathrm{Lip}}\le 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|, \qquad \|f\|_{\mathrm{Lip}} = \sup_{x\ne y} \frac{|f(x) - f(y)|}{\|x - y\|},$$

  valid on any separable metric space (Dudley 2002, section 11.8).
- This shows that W_1 is larger than the bounded Lipschitz (BL) metric

$$W_1(X, Y) \ge \mathrm{BL}(X, Y) = \sup_{\|f\|_\infty + \|f\|_{\mathrm{Lip}}\le 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

  that metrizes convergence in distribution (Dudley 2002, theorem 11.3.3).
- Let P denote the Prokhorov distance. Then $P^2(X, Y) \le W_1(X, Y) \le (D + 1)P(X, Y)$.
- For the class of random variables supported on a fixed bounded subset $K \subseteq \mathcal{X}$, BL and W_1 are equivalent up to constant, and all metrics W_p are topologically equivalent.

- The Wasserstein distances W_p can be bounded by a version of total variation TV (Villani 2008, theorem 6.15). A weaker but more explicit bound for $p = 1$ is $W_1(X, Y) \leq D \times \mathrm{TV}(X, Y)$.
- For discrete random variables, there is an opposite bound $\mathrm{TV} \leq W_1/d_{\min}$.
- The total variation between convolutions with a sufficiently smooth measure is bounded above by W_1 (Mariucci & Reiß 2017, proposition 4).
- The Toscani (or Toscani–Fourier) distance is also bounded above by W_1 (Mariucci & Reiß 2017, proposition 2).

Beyond bounded random variables, W_p, W_q, BL, and TV induce different topologies, so that one cannot bound, for example, W_1 in terms of BL in the unbounded case. On a more theoretical note, we mention that the Kantorovich–Rubinstein formula yields an embedding of any Polish space $(\mathcal{X}, \rho)$ in the Banach space of finite signed measures on $\mathcal{X}$.

3. OPTIMAL TRANSPORT AS A TOOL FOR INFERENCE

As a measure of distance between probability laws, the Wasserstein distance can be used for carrying out of goodness-of-fit tests, and indeed, this has been its main use as a tool for statistical inference. In the simplest one-sample setup, we are given a sample $X_1, \ldots, X_n$ with unknown law μ and wish to test whether μ equals some known fixed law μ_0 (e.g., standard normal or uniform). The empirical measure μ_n associated with the sample $(X_1, \ldots, X_n)$ is the (random) discrete measure that assigns mass $1/n$ to each observation X_i. In this sense, the strong law of large numbers holds in Wasserstein space: With probability one, $W_p(\mu_n, \mu) \to 0$ as $n \to \infty$ if and only if $\mathbb{E}\|X\|^p < \infty$. It is consequently appealing to use $W_p(\mu_n, \mu_0)$ as a test statistic. In the two-sample setup, one independently observes a sample $Y_1, \ldots, Y_m \sim \nu$ with corresponding empirical measure ν_m, and $W_p(\mu_n, \nu_m)$ is a sensible test statistic for the null hypothesis $\mu = \nu$.

3.1. Univariate Measures

We identify measures μ on the real line ($\mathcal{X} = \mathbb{R}$), with their distribution function F; the empirical distribution function corresponding to μ_n is $F_n(t) = n^{-1} \sum_{i=1}^n \mathbf{1}\{X_i \leq t\}$. Thus, $X_i \sim F$, $Y_j \sim G$, and we slightly abuse notation by writing $W_p(F, G)$ for $W_p(\mu, \nu)$.

Munk & Czado (1998) derive the asymptotic distribution of $W_2(F_n, F_0)$ (and trimmed versions thereof). The main tool for the derivation is a Brownian bridge representation for the quantile process $q_n = \sqrt{n}(F_n^{-1} - F^{-1})$ that holds under suitable assumptions on F. There are four types of limiting results, depending on the combination null/alternative and one/two-sample. Roughly speaking, the limits are of order $\sqrt{n}$ and normal under the alternative, and of order n and not normal under the null. The two-sample asymptotics entail that m/n converges to a finite positive constant. In symbols,

$$\begin{aligned}
\sqrt{n}[W_2^2(F_n, F_0) - W_2^2(F, F_0)] &\to \text{normal} \quad (F \neq F_0), \\
nW_2^2(F_n, F_0) &\to \text{something} \quad (F = F_0), \\
\sqrt{\frac{mn}{m+n}}[W_2^2(F_n, G_m) - W_2^2(F, G)] &\to \text{normal} \quad (F \neq G), \text{ and} \\
\frac{mn}{m+n} W_2^2(F_n, G_m) &\to \text{something} \quad (F = G). \qquad 6.
\end{aligned}$$

Similar results were obtained independently in del Barrio et al. (2000), where one can also find a nice survey of other goodness-of-fit tests.

If one instead wants to test whether F belongs to a parametric family $\mathcal{F}$ of distributions, then the test statistic is the infimum of the Wasserstein distance between the empirical measure and members of $\mathcal{F}$. For example, in order to test the fit to some normal distribution, del Barrio et al. (1999a) find the asymptotic distribution of the test statistic

$$R_n = \frac{\inf_{\mu,\sigma^2} W_2^2[F_n, N(\mu,\sigma^2)]}{S_n^2}, \qquad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

an infinite sum of rescaled and centered χ^2 random variables (under the null hypothesis). Using a weighted version of the Wasserstein distance, de Wet (2002) constructs a test for location or scale families. Here, the null hypothesis is that $F = F_0(\cdot - \theta)$ or $F = F_0(\cdot/\theta)$ for some known distribution F_0 and F and unknown $\theta \in \mathbb{R}$ [or $(0,\infty)$]. In a more general setup, Freitag & Munk (2005) consider the case of a structural relationship between F and F_0 in the form

$$F^{-1}(t) = \phi_1(F_0^{-1}(\phi_2(t,\theta)),\theta),$$

for some (known) functions $\phi_1, \phi_2 : \mathbb{R}\times\Theta \to \mathbb{R}$ and parameters $\theta \in \Theta$. This setup includes the location-scale model when $\phi_2(t,\theta) = t$ and $\phi_1(t,\theta_1,\theta_2) = (t-\theta_1)/\theta_2$, and the Lehmann alternatives model when $\phi_2(t,\theta) = 1-(1-t)^\theta$ and $\phi_1(t,\theta) = t$. Motivated by population bioequivalence problems, Freitag et al. (2007) treat the dependent two-sample case, where one observes a sample $(X_i, Y_i)_{i=1}^n$ and wishes to compare the Wasserstein distance between the marginals.

Some of the required regularity is apparent from the following observation. The empirical process $\sqrt{n}(F_n - F)$ converges to $\mathbb{B}\circ F$, where $\mathbb{B}$ is a Brownian bridge on $[0,1]$, without assumptions on F (this result is known as Donsker's theorem). But the quantile process q_n involves inversion, and the limiting distribution is $\mathbb{B}(t)/F'(F^{-1}(t))$, which requires assumptions on F. Csörgő & Horváth (1993) provide a detailed study of the quantile process and asymptotics of functionals thereof. In the context of Wasserstein distance, del Barrio et al. (2005) study the limiting behavior of the norm $\|q_n\|_{2,w}^2 = \int_0^1 q_n^2(t)w(t)\,\mathrm{d}t$, for an integrable weight function w on $(0,1)$. The covariance function of the process $\mathbb{B}/F'\circ F^{-1}$ is

$$\eta(s,t) = \frac{\min(s,t) - st}{F'(F^{-1}(t))F'(F^{-1}(s))}, \qquad s,t \in (0,1),$$

and the limits are qualitatively different depending on whether the integrals $\int_0^1 \eta(t,t)w(t)\,\mathrm{d}t$ and/or $\int_0^1\int_0^1 \eta^2(t,s)w(t)w(s)\,\mathrm{d}t\,\mathrm{d}s$ are finite or not.

3.2. Multivariate Measures

Results in the multivariate setup are more scarce. One apparent reason for this is that the Wasserstein space of measures with multidimensional support is no longer embeddable in the function space $L_p(0,1)$ via quantile functions, and has positive curvature (see Section 4). As perhaps can be expected, multivariate distributional results for the empirical p-Wasserstein distance are chiefly available when it admits a closed form; that is, when $p = 2$ and we consider Gaussian distributions. Assume that $\mu = N(m_1, \Sigma_1)$. Given a sample $X_1,\dots,X_n$ from μ, let $\widehat{\mu}_n$ be the empirical Gaussian measure

$$\widehat{\mu}_n = N(\widehat{m}, \widehat{\Sigma}), \qquad \widehat{m} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^t.$$

The test statistic is now $W_2^2(\widehat{\mu}_n, \mu_0)$ for one sample and $W_2^2(\widehat{\mu}_n, \widehat{\nu}_m)$ for two samples, and the analog of the four cases in Equation 6 holds true. The underlying idea is to combine the classical central limit theorem for $\widehat{m}$ and $\widehat{\Sigma}$ with a delta method, and Rippl et al. (2016) establish the

necessary differentiability of the squared Wasserstein distance in the Gaussian setup in order to apply the delta method. Importantly, Gaussianity can be replaced with any location-scatter family of d-dimensional distribution functions

$$\{F(x) = F_0(m + \Sigma^{1/2}x) : m \in \mathbb{R}^d; \Sigma \in \mathbb{R}^{d\times d} \text{ positive definite}\},$$

where F_0 is an arbitrary distribution function with finite nonsingular covariance matrix.

For sufficiently smooth measures μ, ν (with moment conditions), del Barrio & Loubes (2018) find the normal limit of

$$\sqrt{n}[W_2^2(\mu_n, \nu) - \mathbb{E}W_2^2(\mu_n, \nu)].$$

They establish stability of the convex potential with respect to perturbations of the measures and invoke the Efron–Stein inequality. Again in analogy with Equation 6, the limiting distribution is degenerate at 0 if $\mu = \nu$. This central limit theorem does not, however, yield a limit for $W_2^2(\mu_n, \nu) - W_2^2(\mu, \nu)$, since the speed at which $\mathbb{E}W_2^2(\mu_n, \mu)$ decays to zero [and consequently that of $\mathbb{E}W_2^2(\mu_n, \nu) - W_2^2(\mu, \nu)$] depends on μ in a rather delicate way and can be arbitrarily slow (see Section 3.3).

When μ and ν are finitely supported measures, they can be identified with vectors r in the unit simplex, and the empirical vector r_n obeys a central limit theorem. Sommerfeld & Munk (2018) apply a delta method to obtain the limiting distributions of the Wasserstein distance. The latter is only directionally Hadamard differentiable, leading to a nonstandard delta method with nonlinear derivative. Correspondingly, the limiting distributions are not Gaussian, in general. In analogy with Equation 6, they show that $n^{1/2}(W_p(r_n, s) - W_p(r, s))$ has a distributional limit under the alternative, whereas under the null, the rate is $n^{1/(2p)}$ in agreement with results in Section 3.3. Sommerfeld & Munk (2018) highlight the implications of the nonstandard delta method for the bootstrap, whose consistency requires subsampling.

These results extend to countably supported measures, where one needs to impose an extra summability condition on r in order to ensure convergence of $\sqrt{n}(r_n - r)$ to the Gaussian limit $\mathbb{G}$ (Tameling et al. 2017). The limiting distributions have a more explicit form when the underlying space has the metric structure of a tree. Bigot et al. (2017a) establish similar limits for a regularized version (see Section 5) of the Wasserstein distance.

Wasserstein distances have recently been proposed by Bernton et al. (2017) for parameter inference in approximate Bayesian computation (also known as plug-and-play methods). The setup is that one observes data on $\mathcal{X}$ and wishes to estimate the underlying distribution μ belonging to a parametrized set of distributions $\{\mu_\theta\}_{\theta\in\mathbb{R}^N}$. However, the densities of these measures are too complicated to evaluate/optimize a likelihood. Instead one can only simulate from them and retain parameters that yield synthetic data resembling the observed data. A core issue here is how to contrast the true and simulated data, and Bernton et al. (2017) suggest using W_p to carry out such comparisons.

A Wasserstein metric has also been employed to compare persistence diagrams, a fundamental tool in topological data analysis (see Wasserman 2018 for a recent review) summarizing the persistent homology properties of a data set. Readers are directed to, for example, Mileyko et al. (2011), who introduce a version of the Wasserstein distance on the space of persistence diagrams, endowing it with a metric structure that allows statistical inference.

3.3. Bounds for the Expected Empirical Wasserstein Distance

As discussed in the previous subsections, the speed of convergence of the empirical measure μ_n to μ in Wasserstein distance W_p is important for statistical inference. This topic has a history dating

back to the seminal work of Dudley (1969) and a very rich literature. For space considerations, we will focus on the average value $\mathbb{E}W_p(\mu_n, \mu)$, but see the bibliographical notes (Section 3.3.1) for concentration inequalities and almost sure results. Upper bounds for the one-sample version are also valid for the two-sample version since $\mathbb{E}W_p(\mu_n, \nu_n) \leq 2\mathbb{E}W_p(\mu_n, \mu)$ when ν_n is another empirical measure. For brevity we write W_p for $W_p(\mu_n, \mu)$, and inequalities such as $\mathbb{E}W_p \geq Cn^{-1/2}$ hold for given p, some $C = C(\mu)$, and all n. We also tacitly assume that $\mu \in \mathcal{W}_p$, i.e., it has a finite pth moment, when writing W_p.

The behavior of $\mathbb{E}W_p(\mu_n, \mu)$ is qualitatively different depending on whether the underlying dimension $d > 2p$ or $d < 2p$. For discrete measures, $\mathbb{E}W_p$ is generally of the order $n^{-1/(2p)}$, independently of the dimension. In high dimensions this is better than absolutely continuous measures, for which the rate is $n^{-1/d}$, but when $d = 1$, some smooth measures attain the optimal rate $n^{-1/2}$, faster than $n^{-1/(2p)}$. We first note that it is quite easy to see that $W_p \to 0$ almost surely. However, even for $p = 1 = d$, the decay of $\mathbb{E}W_p$ can be arbitrarily slow (see Bobkov & Ledoux 2018, theorem 3.3).

Lower bounds are easier to obtain, and here are some examples:

- Fundamental $\sqrt{n}$ bound: If μ is nondegenerate, then $\mathbb{E}W_p \geq Cn^{-1/2}$.
- Separated support: If $\mu(A) > 0$, $\mu(B) > 0$, $\mu(A \cup B) = 1$ and $\text{dist}(A, B) = \inf_{x \in A, y \in B} \|x - y\| > 0$, then $\mathbb{E}W_p \geq C_p n^{-1/(2p)}$. Any finitely discrete nondegenerate measure satisfies this condition, as well as most countably discrete ones. This agrees with the rates of Sommerfeld & Munk (2018) above.
- Curse of dimensionality: If μ is absolutely continuous on $\mathbb{R}^d$, then $\mathbb{E}W_p \geq Cn^{-1/d}$. (This result is void of content when $d \leq 2$ in view of the $n^{-1/2}$ bound.) More generally, μ only needs to have an absolutely continuous part (e.g., a mixture of a Gaussian with a discrete measure), and the bound holds when μ_n is replaced with any measure supported on n points. Equivalently, it holds for the quantizer of μ, the n-point measure that is W_p-closest to μ.

We briefly comment on how these bounds are obtained. The $\sqrt{n}$ bound is a corollary of the central limit theorem on $f(X)$, where $X \sim \mu$ and f is a suitable Lipschitz function. If μ has separated support and $k \sim B(n, \mu(A))$ is the number of points in μ_n falling in A, then a mass of $|k/n - \mu(A)|$ must travel at least $\text{dist}(A, B) > 0$ units of distance, yielding a lower bound on the Wasserstein distance. One then invokes the central limit theorem for k. For the curse of dimensionality, note that the number of balls of radius ϵ needed to cover the support of μ is proportional to ϵ^{-d}. If we take $\epsilon = Kn^{-1/d}$ with an appropriate $K > 0$, then n balls of radius ϵ centered at the points of the empirical measure miss mass τ from μ, and this mass has to travel at least ϵ, so $W_p^p \geq C'\tau n^{-p/d}$.

The last lower bound was derived by counting the number of balls needed in order to cover μ, which turns out to be a determining quantity for the upper bounds, too. To account for unbounded supports, we need to allow covering only a (large) fraction of the mass. Let

$$N(\mu, \epsilon, \tau) = \text{minimal number of } \epsilon\text{-balls whose union has } \mu \text{ measure at least } 1 - \tau.$$

These covering numbers increase as ϵ and τ approach zero and are finite for all $\epsilon, \tau > 0$. To put the next upper bound in context, we remark that any compactly supported μ on $\mathbb{R}^d$ satisfies $N(\mu, \epsilon, 0) \leq K\epsilon^{-d}$.

- If for some $d > 2p$, $N(\mu, \epsilon, \epsilon^{dp/(d-2p)}) \leq \epsilon^{-d}$, then $\mathbb{E}W_p \leq C_p n^{-1/d}$.

This covering number condition is verified if μ has finite moment of order large enough (Dudley 1969, proposition 3.4).

The exact formulae on the real line lead to a characterization of the measures attaining the optimal $n^{-1/2}$ rate:

- If $\mu \in \mathcal{W}_p(\mathbb{R})$ has compact support, then $\mathbb{E}W_1 \leq Cn^{-1/2}$, and consequently $\mathbb{E}W_p \leq C_p n^{-1/(2p)}$.
- A necessary and sufficient condition for $\mathbb{E}W_1 \leq Cn^{-1/2}$ is that

$$J_1(\mu) = J_1(F) = \int_{\mathbb{R}} \sqrt{F(t)(1-F(t))}\,\mathrm{d}t < \infty.$$

- The same holds for $\mathbb{E}W_p$, with the integrand in J_1 replaced by $[F(t)(1-F(t))]^{p/2}/[f(t)]^{p-1}$, where f is the density of the absolutely continuous part of μ.

Using the representation of W_1 as the integral of $|F_n - F|$, one sees that $J_1 < \infty$ suffices for the $n^{-1/2}$ rate, since the integrand has variance $n^{-1}F(t)(1-F(t))$. The condition $J_1 < \infty$ is essentially a moment condition, as it implies $\mathbb{E}X^2 < \infty$ and is a consequence of $\mathbb{E}X^{2+\delta}$ for some $\delta > 0$. But for $p > 1$, $J_p < \infty$ entails some smoothness of μ. In particular, the above lower bounds show that μ must be supported on a (possibly unbounded) interval, and the J_p condition means that the density should not vanish too quickly in the interior of the support.

3.3.1. Bibliographic notes. The lower bounds were adapted from Dudley (1969), Fournier & Guillin (2015), and Weed & Bach (2018).

The upper bound with the coverings dates back to Dudley (1969), who showed it for $p = 1$ and with the bounded Lipschitz metric. The version given here can be found in Weed & Bach (2018) and extends Boissard & Le Gouic (2014). We emphasize that their results are not restricted to Euclidean spaces. For Gaussian measures in a Banach space, Boissard & Le Gouic (2014) relate $\mathbb{E}W_2$ to small ball probabilities. Weed & Bach (2018) also show that absolutely continuous measures that are almost low dimensional enjoy better rates for moderate values of n, until eventually giving in to the curse of dimensionality.

In the limiting case $d = 2p$, there is an additional logarithmic term. For $p = 1$ the sufficiency of this term was noted by Dudley (1969, p. 44), and the necessity follows from a classical result of Ajtai et al. (1984) for μ uniform on $[0,1]^2$. For $p > 1$ and $d = 2p$, readers are directed to, for example, Fournier & Guillin (2015).

That absolutely continuous measures are the ones exhibiting slow convergence rates in high dimensions was already observed by Dobrić & Yukich (1995) in an almost sure sense: $n^{1/d}W_p$ has a positive limit if and only if μ has an absolutely continuous part. There are results for more general cost functions than powers of Euclidean distance; see Talagrand (1994) for μ uniform on $[0,1]^d$ and Barthe & Bordenave (2013) for a careful study of the two-sample version $W_p(\mu_n, \nu_n)$. Fournier & Guillin (2015) also deal with the Euclidean case, with some emphasis on deviation bounds and the limit cases $d = 2p$.

del Barrio et al. (1999b) showed that $J_1 < \infty$ is necessary and sufficient for the empirical process $\sqrt{n}(F_n - F)$ to converge in distribution to $\mathbb{B} \circ F$, with $\mathbb{B}$ Brownian bridge. A thorough treatment of the univariate case, including but not restricted to the J_p condition, can be found in Bobkov & Ledoux (2018), using an order statistic representation for the Wasserstein distance. One may also consult Mason (2016) for the alternative approach of weighted Brownian bridge approximations.

The topic is one of intense study, and the references here are far from exhaustive. We also mention some extensions for dependent data: Dédé (2009), Cuny (2017), and Dedecker & Merlevède (2017).

4. OPTIMAL TRANSPORT AS THE OBJECT OF INFERENCE

The previous section described applications of Wasserstein distances for carrying out statistical tasks such as goodness-of-fit testing. The topic of this section is a more recent trend, where one

views the Wasserstein space as a sample space for statistical inference. In this setup, one observes a sample $\mu_1, \ldots, \mu_n$ from a random measure Λ taking values in Wasserstein space $\mathcal{W}_p$ of measures with finite pth moment, and seeks to infer some quantity pertaining to the law of Λ using the observed data, typically in a nonparametric fashion. Such questions can be seen as part of next-generation functional data analysis, borrowing the terminology of Wang et al. (2016, section 6).

4.1. Fréchet Means of Random Measures

Perhaps the most basic question here, as anywhere, is estimating a mean. Clearly we could estimate the mean of Λ by the average $n^{-1}(\mu_1 + \cdots + \mu_n)$, which is also a probability measure. While this may often be a good estimator, in certain modern applications, such as imaging, it exhibits some unsatisfactory properties. As a simple example, consider two Dirac measures at distinct points $x \neq y$. Their average is the blurred measure putting mass $1/2$ at x and y. In contrast, as we see below, the Wasserstein distance leads to an average that is a Dirac measure at the midpoint $(x + y)/2$.

We focus on the special case $p = 2$, which is the most elegant and provides the canonical setup in deformation models (see Section 4.2). One way of giving a meaning to the notion of expectation in general metric space is to consider the Fréchet mean (better known in analysis as barycenter), named after Fréchet (1948) and defined as the minimizer of the Fréchet functional

$$F(\mu) = \mathbb{E}W_2^2(\Lambda, \mu) = \int_{\mathcal{W}_2} W_2^2(\lambda, \mu)\, \mathrm{d}\mathbb{P}(\lambda), \qquad \mu \in \mathcal{W}_2,$$

where $\mathbb{P}$ is the law of Λ. We shall refer to such a minimizer as the population (Fréchet) mean to distinguish it from the empirical version, where $\mathbb{E}W_2^2(\Lambda, \mu)$ is replaced with $\sum W_2^2(\mu_i, \mu)$.

Existence, uniqueness, computation, laws of large numbers, and central limit theorems for Fréchet means with respect to general metrics have been studied extensively under the umbrella of non-Euclidean statistics (e.g., Huckemann et al. 2010, Kendall & Le 2011). Even existence and uniqueness are nontrivial questions for many metrics and depend subtly on the induced geometry. It turns out that $\mathcal{W}_2$ induces a geometry that is very close to Riemannian (see Section 4.4). Despite posing challenges in that it is infinite-dimensional, has unbounded curvature, and presents an abundance of singularities, its geometry exhibits many favorable (indeed, quite unusual for nonlinear spaces) properties owing to the structure of the optimal transport problem.

By means of convex analysis, Agueh & Carlier (2011) deduce existence, uniqueness, and a characterization of empirical Fréchet means in $\mathcal{W}_2(\mathbb{R}^d)$ in what has become a seminal paper. Existence always holds, whereas the mean is unique provided that one of the measures μ_i is absolutely continuous. The results extend to the population version (Pass 2013): The condition is that with positive probability, Λ is absolutely continuous (assuming that the Fréchet functional F is finite). A notable exception is, again, when $d = 1$, in which case Fréchet means are unique with the sole restriction that F is finite.

A law of large numbers in Wasserstein space was proved by Le Gouic & Loubes (2017) in a very general setting (for arbitrary $p > 1$, and for spaces more general than $\mathbb{R}^d$). Since $\mathcal{W}_2(\mathbb{R}^d)$ is itself a complete and separable metric space, one can view $\mathbb{P}$, the law of Λ, as an element in the second level Wasserstein space $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$. Le Gouic & Loubes show that if $\mathbb{P}_n$ is a sequence of laws converging to $\mathbb{P}$ in the second level Wasserstein space, then the Fréchet means of $\mathbb{P}_n$ converge to that of $\mathbb{P}$ (if unique) in the first level $\mathcal{W}_2(\mathbb{R}^d)$. This setup covers the case where $\mathbb{P}_n$ is the empirical measure [in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$] corresponding to a sample from Λ. Álvarez-Esteban et al. (2018) provide an extension to trimmed Fréchet means.

4.2. Fréchet Means and Generative Models

From a statistical perspective, the choice of a metric and the consideration of the corresponding Fréchet mean often implicitly assume a certain underlying data-generating mechanism for the data. In the case of the Wasserstein metric, this mechanism is inextricably linked to warping or phase variation (Ramsay & Silverman 2005, Marron et al. 2015, Wang et al. 2016), where one wishes to infer the law of a process Y on (say) $[0, 1]$ but only has access to realizations of $\tilde{Y} = Y \circ T^{-1}$, where $T : [0, 1] \to [0, 1]$ is a random warp/deformation map. This setup is quite natural in physiological data such as growth curves or spike trains where each individual may have an intrinsic timescale, a sort of functional random effect. The problem would then be to correct for the effect of T that distorts time and recover the sample paths in the correct or objective timescale. Typically, it is natural to assume that T is an increasing homeomorphism, on the basis that time should always move forward, rather than backward, and, for identifiability reasons, that $\mathbb{E}T(t) = t$, $t \in [0, 1]$.

Now, when the functional datum Y is a random probability measure in $\mathcal{W}_2(\mathbb{R}^d)$ with intensity $\mathbb{E}[Y] = \lambda$, the warped version $\tilde{Y} = T\#Y$ is a random measure with conditional intensity $\Lambda = \mathbb{E}[\tilde{Y}|T] = T\#\lambda$. Assuming that T is increasing with $\mathbb{E}T$ equal to the identity then implies that λ is a Fréchet mean of Λ. More generally, if $\lambda \in \mathcal{W}_2(\mathbb{R}^d)$ and T is a random continuous function with mean identity that can be written as the gradient of a convex function on $\mathbb{R}^d$, then λ is a Fréchet mean of the random measure $\Lambda = T\#\lambda$. In other words, the Wasserstein geometry is canonical under the deformation model, and estimation of a Fréchet mean implicitly assumes a deformation model. The result in this form is due to Zemel & Panaretos (2018), but a parametric version is due to Bigot & Klein (2018). When λ is absolutely continuous and T is sufficiently injective, $\Lambda = T\#\lambda$ is absolutely continuous, and the Fréchet mean of Λ is unique and equals λ. In the particular case of Gaussian measures, the result even holds in infinite dimensions (Masarotto et al. 2018).

4.3. Fréchet Means and Multicouplings

The Fréchet mean problem is related to a multimarginal formulation of optimal transport considered by Gangbo & Święch (1998). Given $\mu_1, \ldots, \mu_n \in \mathcal{W}_2(\mathbb{R}^d)$, an optimal multicoupling is a joint distribution of a random vector $(X_1, \ldots, X_n)$ such that $X_i \sim \mu_i$ and

$$\frac{1}{2n^2}\mathbb{E}\sum_{1\le i<j\le n}\|X_i - X_j\|^2 = \frac{1}{2n}\mathbb{E}\sum_{i=1}^{n}\|X_i - \overline{X}\|^2$$

is minimized. Agueh & Carlier (2011) show that if $(X_1, \ldots, X_n)$ is an optimal multicoupling, then the law of $\overline{X} = n^{-1}\sum_i X_i$ is a Fréchet mean of $\{\mu_i\}_{i=1}^n$. Inspection of their argument shows that it can also give the only if direction. And, when at least one measure μ_i is regular, necessity and sufficiency combined can be used to construct the optimal multicoupling as $X_i = \mathbf{t}_\lambda^{\mu_i}(Z)$, where $Z \sim \lambda$ and λ is the Fréchet mean (see Pass 2013 and Zemel & Panaretos 2018 for more details). This illustrates how constructing the optimal multicoupling is inextricably linked to finding the Fréchet mean (for the latter, see Section 4.5). In fact, the argument of Agueh & Carlier (2011) extends to infinite-dimensional and even nonlinear space. Let $(\mathcal{X}, \rho)$ be a complete separable barycentric metric space: For any $x_1, \ldots, x_n \in \mathcal{X}$ there exists a unique Fréchet mean $\overline{x}$. Fréchet means of given measures $\mu_1, \ldots, \mu_n \in \mathcal{W}_2(\mathcal{X})$ are precisely the laws of $\overline{X}$, where $(X_1, \ldots, X_n)$ is an optimal multicoupling with respect to the cost $\mathbb{E}\sum_{i=1}^n \rho(X_i, \overline{X})^2$. This relation illustrates the idea that the Wasserstein space captures the geometry of the underlying space. As a particular special case, the Fréchet mean of Dirac measures is a Dirac measure at the Fréchet mean of the underlying points. Finally, we stress that the relation extends to any $p > 1$, where $\overline{x}^{(p)}$ minimizes $\sum \rho(x_i, x)^p$ and optimality is with respect to $\mathbb{E}\sum \rho(X_i, \overline{X}^{(p)})^p$. [Strictly speaking, these are not Fréchet means, as one minimizes $\sum W_p^p(\mu_i, \mu)$ instead of $\sum W_p^2(\mu_i, \mu)$.]

4.4. Geometry of Wasserstein Space

A typical step in estimating Fréchet means in non-Euclidean settings is approximation of the manifold by a linear space, the tangent space. In the Wasserstein case, the latter is a function space. Let λ be the Fréchet mean, and assume sufficient regularity that λ is unique and absolutely continuous. Then, convergence of a sample Fréchet mean $\widehat{\lambda}_n$ to λ can be quantified by that of the optimal map $\mathbf{t}_\lambda^{\widehat{\lambda}_n}$ to the identity map $\mathbf{i}$ because

$$W_2^2(\widehat{\lambda}_n, \lambda) = \int_{\mathbb{R}^d} \|\mathbf{t}_\lambda^{\widehat{\lambda}_n}(x) - x\|^2 \, \mathrm{d}\lambda(x) = \|\mathbf{t}_\lambda^{\widehat{\lambda}_n} - \mathbf{i}\|_{\mathcal{L}^2(\lambda)}^2.$$

Here $\mathcal{L}^2(\lambda)$ is the L^2-like space of measurable functions $\mathbf{r} : \mathbb{R}^d \to \mathbb{R}^d$ such that the real-valued function $x \mapsto \|\mathbf{r}(x)\|$ is in $L^2(\lambda)$, and whose $L^2(\lambda)$-norm defines $\|\mathbf{r}\|_{\mathcal{L}^2(\lambda)}$. Thus, we can linearize the Wasserstein space by identifying an arbitrary measure μ with the function $\mathbf{t}_\lambda^\mu - \mathbf{i}$ in the linear space $\mathcal{L}_2(\lambda)$; subtracting the identity centers this linear space at λ.

4.4.1. The tangent bundle. Ambrosio et al. (2008) consider absolutely continuous curves in Wasserstein space and show that optimal maps arise as minimal tangent vectors to such curves. With that in mind, they define the tangent space at λ as the span of such maps minus the identity

$$\mathrm{Tan}_\lambda = \overline{\{t(\mathbf{t}_\lambda^\mu - \mathbf{i}) : \mu \in \mathcal{W}_2; t \in \mathbb{R}\}}^{\mathcal{L}^2(\lambda)}.$$

By definition, each $\mathbf{t}_\lambda^\mu$ (and the identity) is in $\mathcal{L}^2(\lambda)$, so $\mathrm{Tan}_\lambda \subseteq \mathcal{L}^2(\lambda)$, from which it inherits the inner product. The definition can be adapted to a nonabsolutely continuous λ by restricting μ in the definition of Tan_λ to those μ for which $\mathbf{t}_\lambda^\mu$ exists (this optimal map might not be unique, and any possible choice of $\mathbf{t}_\lambda^\mu$ leads to a tangent vector). There is an alternative equivalent definition of the tangent space in terms of gradients of smooth functions (see Ambrosio et al. 2008, definition 8.4.1 and theorem 8.5.1). The alternative definition highlights that it is essentially the inner product that depends on λ, but not the elements of the tangent space.

The exponential map $\exp_\lambda : \mathrm{Tan}_\lambda \to \mathcal{W}_2$ at λ is the restriction of the transformation that sends $\mathbf{r} \in \mathcal{L}^2(\lambda)$ to $(\mathbf{r} + \mathbf{i})\#\lambda \in \mathcal{W}_2$. Specifically,

$$\exp_\lambda(t(\mathbf{t} - \mathbf{i})) = [t(\mathbf{t} - \mathbf{i}) + \mathbf{i}]\#\lambda = [t\mathbf{t} + (1-t)\mathbf{i}]\#\lambda \quad (t \in \mathbb{R}).$$

When λ is absolutely continuous, the log map $\log_\lambda : \mathcal{W}_2 \to \mathrm{Tan}_\lambda$ is

$$\log_\lambda(\mu) = \mathbf{t}_\lambda^\mu - \mathbf{i}$$

and is the right inverse of the exponential map (which is therefore surjective). Segments in the tangent space are retracted to the Wasserstein space under $\exp_\lambda$ to McCann's (1997) interpolant

$$\left[t\mathbf{t}_\lambda^\mu + (1-t)\mathbf{i}\right]\#\lambda,$$

and these are the unique (constant speed) geodesics in Wasserstein space (Santambrogio 2015, proposition 5.32). If λ is singular, then the log map is only defined on a subset of Wasserstein space. Gigli (2011) provides a description of the tangent bundle when the underlying space $\mathbb{R}^d$ is replaced by a Riemannian manifold.

4.4.2. Curvature and compatible measures. If $\mu, \nu, \rho \in \mathcal{W}_2$, then a coupling argument shows that

$$\|\log_\rho(\mu) - \log_\rho(\nu)\|_{\mathcal{L}^2(\rho)}^2 = \|\mathbf{t}_\rho^\mu - \mathbf{t}_\rho^\nu\|_{\mathcal{L}^2(\rho)}^2 = \int \|\mathbf{t}_\rho^\mu(x) - \mathbf{t}_\rho^\nu(x)\|^2 \, \mathrm{d}\rho(x) \geq W_2^2(\mu, \nu). \qquad 7.$$

In differential geometry terminology, this means that $\mathcal{W}_2$ has nonnegative sectional curvature. In the special case $d = 1$, there is equality, and the Wasserstein space is flat; the correspondence $\mu \iff \mathbf{t}_\rho^\mu - \mathbf{i}$ is an isometry, and $\mathcal{W}_2(\mathbb{R})$ can be viewed as a subset of the Hilbert space $L^2(\mu)$. Computation of Fréchet means is then particularly simple: If $\mu_1, \ldots, \mu_n$ are arbitrary measures in $\mathcal{W}_2(\mathbb{R})$ and ν is any absolutely continuous measure, then the Fréchet mean of (μ_i) is $[(1/n)\sum \mathbf{t}_\nu^{\mu_i}]\#\nu$; this extends to the population version. An important extension to $\mathbb{R}^d$ was obtained by Boissard et al. (2015). Equality will hold in Equation 7, provided some compatibility holds between the measures μ, ν, ρ. The composition $\mathbf{t}_\rho^\nu \circ \mathbf{t}_\mu^\rho$ pushes μ forward to ρ by definition, but might not do so optimally. We say that μ, ν, ρ are compatible if $\mathbf{t}_\rho^\nu \circ \mathbf{t}_\mu^\rho$ is optimal, i.e., equals $\mathbf{t}_\mu^\nu$. Boissard et al. (2015) show that if the collection $(\mu_1, \ldots, \mu_n, \nu)$ is compatible (in their terminology, the optimal maps are admissible) in this sense, then, again, the Fréchet mean is $[(1/n)\sum \mathbf{t}_\nu^{\mu_i}]\#\nu$. This setup covers the one-dimensional setup, but also multivariate measures with structure that mimics the one-dimensional case. For example, a collection of measures having the same d-dimensional copula (and potentially different marginals) is compatible, and so is a collection of measures having the same angular behavior but different marginal distributions for their norms.

4.4.3. Gaussian measures. Without such structural restrictions, the Wasserstein space is positively curved, and computation of the Fréchet mean of a sample is not straightforward. As an important example, if $\mu_i \sim N(0, \Sigma_i)$ are nonsingular Gaussian measures on $\mathbb{R}^d$, then the Fréchet mean is also Gaussian and its covariance is the unique nonsingular solution of the matrix equation

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\Sigma^{1/2}\Sigma_i\Sigma^{1/2})^{1/2}. \qquad 8.$$

The μ_is will be compatible if the covariances commute, in which case we have the explicit solution $\Sigma^{1/2} = n^{-1}(\Sigma_1^{1/2} + \cdots + \Sigma_n^{1/2})$, but otherwise there is no explicit expression for the Fréchet mean. The restriction of $\mathcal{W}_2(\mathbb{R}^d)$ to Gaussian measures leads to a stratified space whose geometry was studied carefully by Takatsu (2011), including expressions for the curvature. In particular, the curvature grows without bound as one approaches singular covariance matrices.

4.5. Fréchet Means via Steepest Descent

A common procedure for finding Fréchet means is differentiating the Fréchet functional F and moving in the negative direction of the gradient (Karcher 1977, Afsari et al. 2013). The gradient at x_0 typically takes the form

$$\nabla F(x) = \frac{1}{n}\sum_{i=1}^{n} -\log_x(x_i).$$

This formula also holds true in Wasserstein space, where the log map is as given in Section 4.4. Steepest descent can then be defined using the exponential map as

$$\rho_{j+1} = \exp_{\rho_j}(\nabla F(\rho_j)) = \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{t}_{\rho_j}^{\mu_i}\right]\#\rho_j.$$

The resulting iteration was independently arrived at in this steepest descent form by Zemel & Panaretos (2018) and in the form of a fixed point equation iteration by Álvarez-Esteban et al. (2016). It has the advantage of reducing the multitransport problem of finding the Fréchet mean to a succession of pairwise problems that are simpler in nature, in the same spirit as generalized Procrustes analysis (Dryden & Mardia 1998). This benefit is best illustrated in the Gaussian case, where the optimal maps have the explicit expression given in Equation 4. The algorithm converges

to the unique Fréchet mean in this Gaussian case and in general will reach at least a stationary point (where ∇F vanishes). There are local minima that are not global: Álvarez-Esteban et al. (2016) construct measures $\mu_1, \ldots, \mu_4, \mu$ in $\mathbb{R}^2$ such that the average of $\mathbf{t}_\mu^{\mu_i}$ is the identity, but μ is not the Fréchet mean. Their example shows that the problem cannot be solved by smoothness conditions on the measures. But smoothness and convexity of the supports yield an optimality criterion for local minima (Zemel & Panaretos 2018), essentially, that a sufficiently smooth local minimum is a global minimum.

4.6. Large Sample Statistical Theory in Wasserstein Space

The general consistency result of Le Gouic & Loubes (2017) is the important and necessary first step in providing a sound statistical theory for random measures in Wasserstein space. The next step would be to establish the rate of convergence and a central limit theorem. By exploiting the central limit theorem in Hilbert spaces, the one-dimensional case can be well understood, even under sampling noise: The empirical mean $\widehat{\lambda}_n$, viewed as the L^2 map, $\sqrt{n}(\mathbf{t}_\lambda^{\widehat{\lambda}_n} - \mathbf{i})$, converges in distribution to a zero-mean Gaussian process whose covariance structure is that of the random element $\mathbf{t}_\lambda^\Lambda$ (Panaretos & Zemel 2016). Bigot et al. (2018b) provide minimax-type results in this vein. Since the Wasserstein space on $\mathbb{R}^d$ stays embedded in a Hilbert under the compatible setup of Boissard et al. (2015), these results can certainly be extended to that setup. In fact, Boissard et al. (2015) use this embedding to carry out principal component analysis (PCA) in Wasserstein space. Bigot et al. (2017c) provide an alternative procedure, convex PCA.

The only central limit theorem–type result we know of beyond the compatible setup was found recently by Agueh & Carlier (2017). Suppose that Λ takes finitely many values: $\mathbb{P}(\Lambda = \lambda_k) = p_k$, $k = 1, \ldots, K$, and λ_k is Gaussian $N(0, \Sigma_k)$ with Σ_k nonsingular. Given a sample $\mu_1, \ldots, \mu_n$ from Λ, let $\widehat{p}_n(k)$ be the proportion of (μ_i)s that equal λ_k. Then $\sqrt{n}(\widehat{p}_n - p)$ has a Gaussian limit. Equation 8 extends to weighted Fréchet means and defines Σ in a sufficiently smooth way, so one can invoke the delta method to obtain a central limit theorem for $\sqrt{n}(\widehat{\Sigma}_n - \Sigma)$. Agueh & Carlier (2017) also cover the case where $K = 2$ and λ_i are arbitrary, though this setup falls under the umbrella of compatibility since any pair of measures is a compatible collection. Ongoing work by Kroshnin & Suvorikova (2018) focuses on extending the results of Agueh & Carlier (2017) to arbitrary random Gaussian/elliptical measures. Beyond this location-scatter setup, very recent results by Ahidar-Coutrix et al. (2018) suggest that the rate of convergence of the empirical Fréchet mean to its population counterpart can be slower than $n^{-1/2}$.

5. COMPUTATIONAL ASPECTS

Beyond the one-dimensional and Gaussian cases, explicit expressions for the Wasserstein distance and/or the optimal couplings are rare. When $\mu = (1/n)\sum_{i=1}^n \delta_{x_i}$ and $\nu = (1/m)\sum_{j=1}^m \delta_{y_j}$ are uniform discrete measures on n and m points, a coupling γ can be identified with an $n \times m$ matrix Γ, where Γ_{ij} represents the mass to be transferred from x_i to y_j. The cost function reduces to a cost matrix $c_{ij} = \|x_i - y_j\|^p$, and the total cost associated with it is $\sum_{ij} \Gamma_{ij} c_{ij}$. This double sum is to be minimized over Γ subject to the $m + n$ mass preservation constraints

$$\sum_{i=1}^n \Gamma_{ij} = 1/m \quad (j = 1, \ldots, m), \qquad \sum_{j=1}^m \Gamma_{ij} = 1/n \quad (i = 1, \ldots, n), \qquad \Gamma_{ij} \geq 0.$$

One can easily write the constraints in the weighted version of the problem. This optimization problem can be solved using standard linear programming techniques. In particular, there exists an optimal solution Γ with at most $n + m - 1$ nonzero entries. In the special case $n = m$ and

uniform measures, the extremal points of the constraints polytope are the permutation matrices, and these correspond precisely to deterministic couplings that have n (rather than $2n-1$) nonzero entries.

The specific structure of the constraints matrix allows the development of specialized algorithms: The Hungarian method of Kuhn (1955) and its variant by Munkres (1957) are classical examples, with alternatives such as network simplex, min flow–type algorithms, and others (see Luenberger & Ye 2008, chapter 6). The best algorithms have the prohibitive complexity $n^3 \log n$ in the worst-case scenario. Sommerfeld et al. (2018) propose sampling $s \ll n$ points from μ and ν and estimating $W_p(\mu, \nu)$ by the empirical distance $W_p(\mu_s, \nu_s)$. They provide bounds on the computational and statistical trade-off regulated by s.

The multimarginal problem can also be recast as a linear program whose solution yields the Fréchet mean (see Section 4.3). If we have n measures μ_i supported on m_i points ($i = 1, \dots, n$), then the number of variables in the problem is $\prod m_i$, and the number of equality constraints is $\sum m_i$, of which $n-1$ are redundant. Anderes et al. (2016) provide a detailed account of the problem, in which they show the peculiar property that the optimal maps $\mathbf{t}_{\overline{\mu}}^{\mu_i}$ exist, where $\overline{\mu}$ is a Fréchet mean. This is far from obvious, since besides the uniform discrete setup with an equal number of points, the optimal coupling between discrete measures is rarely induced from a map. There are alternative formulations with fewer variables and fewer constraints: exact ones (Borgwardt & Patterson 2018) as well as polynomial-time approximations (Borgwardt 2017).

One can certainly approximate $W_p(\mu, \nu)$ by $W_p(\mu_n, \nu_n)$ for some μ_n, ν_n supported on, say, n points. The approximated problem can be solved exactly, as it is a finite linear program. How to best approximate a measure by discrete measures amounts to quantization and is treated in detail by Graf & Luschgy (2007). Unfortunately, quantization is extremely difficult in practice, and even one-dimensional measures rarely admit explicit solutions; moreover, the computational cost of solving the n-to-n points case scales badly with n.

Another class of algorithm is continuous in nature. Recall from Section 4.4 that optimal maps $\mathbf{t}_\mu^\nu$ are equivalent to the unique geodesics in $\mathcal{W}_2$. Benamou & Brenier (2000) exploit this equivalence and develop a numerical scheme to approximate the entire geodesic. Although this dynamic formulation adds an extra time dimension to the problem, it can be recast as a convex problem, unlike the formulation with the optimal map as variable. Chartrand et al. (2009) carry out steepest descent in the dual variable φ in order to maximize the dual $\varphi \mapsto \int \varphi \, \mathrm{d}\mu + \int \varphi^* \, \mathrm{d}\nu$.

In an influential paper, Cuturi (2013) advocated adding an entropy penalty term $\kappa \sum \Gamma_{ij} \log \Gamma_{ij}$ to the objective function. This yields a strictly convex problem with complexity n^2, much smaller than the linear programming complexity $n^3 \log n$. This entropy term enforces Γ to be diffuse (strictly positive), in stark contrast with the unpenalized optimal coupling, but the regularized solution converges to the sparse one as $\kappa \searrow 0$. This idea is extended to the Fréchet mean problem in Cuturi & Doucet (2014), where the Fréchet mean is computed with respect to the penalized Wasserstein distance, and in Bigot et al. (2017b), where the penalization is imposed on the mean itself rather than the distance. Bigot et al. (2018a) suggest a data-driven choice of the regularization parameter according to the Goldenshluger–Lepski principle.

This field of research is very active, and there are tens of extensions and new algorithms. One can find a short survey in Tameling & Munk (2018), and we refer to Santambrogio (2015, chapter 6) and especially the book by Peyré & Cuturi (2018) for more details and references.

6. ON SOME RELATED DEVELOPMENTS

An interesting recent development that is, strictly speaking, not so much about Wasserstein distances as about measure transportation itself, considers how to generalize notions related to

quantiles to several dimensions. In one dimension, the quantile function F_Y^{-1} is the optimal map from a uniform variable U to Y. This observation can be used in order to define a multivariate quantile function of Y using the optimal transport map $\mathbf{t}_U^Y$ from some reference random variable U (e.g., uniform on the unit ball). Chernozhukov et al. (2017) describe the resulting form of the quantile contours and the induced notions of depth and ranks, and estimate them from data. Further work by Hallin (2017) considers extensions of the approach that do not require finite variance for Y (as is the case in one dimension). This measure-transportation approach also allows us to extend quantile regression to multivariate setups (Carlier et al. 2016).

Finally, due to space considerations, we have not attempted to describe the machine learning side of optimal transport, though there is a fast-growing literature for such tasks. Indicative examples include estimation of a low-dimensional measure in high-dimensional space (Canas & Rosasco 2012), regression in the space of histograms (Bonneel et al. 2016), dictionary learning (Rolet et al. 2016), Gaussian processes indexed by measures on $\mathbb{R}$ (Bachoc et al. 2017) or $\mathbb{R}^d$ (Bachoc et al. 2018), clustering in Wasserstein space (del Barrio et al. 2018), and unsupervised alignment of point clouds in high dimensions (Grave et al. 2018).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This was supported in part by a European Research Council Starting Grant Award to Victor M. Panaretos. Yoav Zemel is funded by Swiss National Science Foundation grant #178220. We thank a reviewer for comments on a preliminary version of the article.

LITERATURE CITED

Afsari B, Tron R, Vidal R. 2013. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM J. Control Optim.* 51:2230–60

Agueh M, Carlier G. 2011. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* 43:904–24

Agueh M, Carlier G. 2017. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *C. R. Math.* 355:812–18

Ahidar-Coutrix A, Le Gouic T, Paris Q. 2018. On the rate of convergence of empirical barycentres in metric spaces: curvature, convexity and extendible geodesics. arXiv:1806.02740 [math.ST]

Ajtai M, Komlós J, Tusnády G. 1984. On optimal matchings. *Combinatorica* 4:259–64

Álvarez-Esteban PC, del Barrio E, Cuesta-Albertos J, Matrán C. 2016. A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.* 441:744–62

Álvarez-Esteban PC, del Barrio E, Cuesta-Albertos JA, Matrán C. 2018. Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. *Bernoulli* 24:3147–79

Ambrosio L, Gigli N. 2013. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, ed. B Piccoli, M Rascle, pp. 1–155. Berlin: Springer

Ambrosio L, Gigli N, Savaré G. 2008. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Basel, Switz.: Birkhäser

Anderes E, Borgwardt S, Miller J. 2016. Discrete Wasserstein barycenters: optimal transport for discrete data. *Math. Methods Oper. Res.* 84:389–409

Appell P. 1886. *Mémoire sur les déblais et les remblais des systèmes continus ou discontinus*. Paris: Impr. Nat.

Bachoc F, Gamboa F, Loubes JM, Venet N. 2017. A Gaussian process regression model for distribution inputs. *IEEE Trans. Inf. Theory*. **https://dx.doi.org/10.1109/TIT.2017.2762322**

Bachoc F, Suvorikova A, Loubes JM, Spokoiny V. 2018. Gaussian process forecast with multidimensional distributional entries. arXiv:1805.00753 [stat.ME]
Barbour AD, Brown TC. 1992. Stein's method and point process approximation. *Stoch. Process. Appl.* 43:9–31
Barthe F, Bordenave C. 2013. Combinatorial optimization over two random point sets. In *Séminaire de Probabilités XLV*, ed. C Donati-Martin, A Lejay, A Rouault, pp. 483–535. Berlin: Springer
Bass J. 1955. Sur la compatibilité des fonctions de répartition. *C. R. Hebd. Séa. Acad. Sci.* 240:839–41
Beiglböck M, Schachermayer W. 2011. Duality for Borel measurable cost functions. *Trans. Am. Math. Soc.* 363:4203–24
Benamou JD, Brenier Y. 2000. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* 84:375–93
Bernton E, Jacob PE, Gerber M, Robert CP. 2017. Inference in generative models using the Wasserstein distance. arXiv:1701.05146 [stat.ME]
Bickel PJ, Freedman DA. 1981. Some asymptotic theory for the bootstrap. *Ann. Stat.* 9:1196–217
Bigot J, Cazelles E, Papadakis N. 2017a. Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications. arXiv:1711.08947 [math.ST]
Bigot J, Cazelles E, Papadakis N. 2017b. Penalized barycenters in the Wasserstein space. arXiv:1606.01025 [math.ST]
Bigot J, Cazelles E, Papadakis N. 2018a. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. arXiv:1804.08962 [stat.ME]
Bigot J, Gouet R, Klein T, López A. 2017c. Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Stat.* 53:1–26
Bigot J, Gouet R, Klein T, López A. 2018b. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electron. J. Stat.* 12:2253–89
Bigot J, Klein T. 2018. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM Probab. Stat.* 22:35–57
Bobkov S, Ledoux M. 2018. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Mem. Am. Math. Soc.* In press
Boissard E, Le Gouic T. 2014. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. H. Poincaré Probab. Stat.* 50:539–63
Boissard E, Le Gouic T, Loubes JM. 2015. Distribution's template estimate with Wasserstein metrics. *Bernoulli* 21:740–59
Bolstad BM, Irizarry RA, Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–93
Bonneel N, Peyré G, Cuturi M. 2016. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.* 35:1–10
Borgwardt S. 2017. Strongly polynomial 2-approximations of discrete Wasserstein barycenters. arXiv:1704.05491 [math.OC]
Borgwardt S, Patterson S. 2018. Improved linear programs for discrete barycenters. arXiv:1803.11313 [math.OC]
Brenier Y. 1991. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* 44:375–417
Caffarelli LA. 1992. The regularity of mappings with a convex potential. *J. Am. Math. Soc.* 5:99–104
Canas G, Rosasco L. 2012. Learning manifolds with K-means and K-flats. In *Advances in Neural Information Processing Systems 25*, ed. F Pereira, C Burges, L Bottou, K Weinberger, pp. 2465–73. Red Hook, NY: Curran
Carlier G, Chernozhukov V, Galichon A. 2016. Vector quantile regression: an optimal transport approach. *Ann. Stat.* 44:1165–92
Chartrand R, Wohlberg B, Vixie K, Bollt E. 2009. A gradient descent solution to the Monge–Kantorovich problem. *Appl. Math. Sci.* 3:1071–80
Chernozhukov V, Galichon A, Hallin M, Henry M. 2017. Monge–Kantorovich depth, quantiles, ranks and signs. *Ann. Stat.* 45:223–56
Csörgő M, Horváth L. 1993. *Weighted Approximations in Probability and Statistics*. New York: Wiley

Cuesta-Albertos JA, Matrán C. 1989. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.* 17:1264–76

Cuesta-Albertos JA, Matrán-Bea C, Tuero-Diaz A. 1996. On lower bounds for the L_2-Wasserstein metric in a Hilbert space. *J. Theor. Probab.* 9:263–83

Cuesta-Albertos JA, Rüschendorf L, Tuero-Diaz A. 1993. Optimal coupling of multivariate distributions and stochastic processes. *J. Multivar. Anal.* 46:335–61

Cuny C. 2017. Invariance principles under the Maxwell–Woodroofe condition in Banach spaces. *Ann. Probab.* 43:1578–611

Cuturi M. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, ed. CJC Burges, L Bottou, M Welling, Z Ghahramani, K Weinberger, pp. 2292–300. Red Hook, NY: Curran

Cuturi M, Doucet A. 2014. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning*, ed. EP Xing, T Jebara, pp. 685–93. Brookline, MA: Microtome

Dall'Aglio G. 1956. Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Ann. Scuola Norm. Sup. Pisa Classe Sci.* 10:35–74

de Wet T. 2002. Goodness-of-fit tests for location and scale families based on a weighted L_2-Wasserstein distance measure. *Test* 11:89–107

Dédé S. 2009. An empirical central limit theorem in L_1 for stationary sequences. *Stoch. Process. Appl.* 119:3494–515

Dedecker J, Merlevède F. 2017. Behavior of the Wasserstein distance between the empirical and the marginal distributions of stationary α-dependent sequences. *Bernoulli* 23:2083–127

del Barrio E, Cuesta-Albertos JA, Matrán C. 2000. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test* 9:1–96

del Barrio E, Cuesta-Albertos JA, Matrán C, Mayo-Íscar A. 2018. Robust clustering tools based on optimal transportation. *Stat. Comput.* **https://doi.org/10.1007/s11222-018-9800-z**

del Barrio E, Cuesta-Albertos JA, Matrán C, Rodríguez-Rodríguez JM. 1999a. Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Stat.* 27:1230–39

del Barrio E, Giné E, Matrán C. 1999b. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* 27:1009–71

del Barrio E, Giné E, Utzet F. 2005. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* 11:131–89

del Barrio E, Loubes JM. 2018. Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.* In press

Delon J, Salomon J, Sobolevski A. 2010. Fast transport optimization for Monge costs on the circle. *SIAM J. Appl. Math.* 70:2239–58

Dobrić V, Yukich JE. 1995. Asymptotics for transportation cost in high dimensions. *J. Theor. Probab.* 8:97–118

Dobrushin RL. 1970. Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.* 15:458–86

Dryden IL, Mardia KV. 1998. *Statistical Shape Analysis*. New York: Wiley

Dudley RM. 1969. The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Stat.* 40:40–50

Dudley RM. 2002. *Real Analysis and Probability*. Cambridge, UK: Cambridge Univ. Press

Eberle A. 2014. Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann. Probab.* 24:337–77

Ebralidze SS. 1971. Inequalities for the probabilities of large deviations in the multidimensional case. *Theory Probab. Appl.* 16:733–37

Evans SN, Matsen FA. 2012. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. B* 74:569–92

Figalli A. 2017. *The Monge–Ampère Equation and Its Applications*. Zürich, Switz.: Eur. Math. Soc.

Fournier N, Guillin A. 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* 162:707–38

Fréchet M. 1948. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré Probab. Stat.* 10:215–310

Fréchet M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon* 14:53–77
Freitag G, Czado C, Munk A. 2007. A nonparametric test for similarity of marginals—with applications to the assessment of population bioequivalence. *J. Stat. Plan. Inference* 137:697–711
Freitag G, Munk A. 2005. On Hadamard differentiability in *k*-sample semiparametric models—with applications to the assessment of structural relationships. *J. Multivar. Anal.* 94:123–58
Gangbo W, McCann RJ. 1996. The geometry of optimal transportation. *Acta Math.* 177:113–61
Gangbo W, Święch A. 1998. Optimal maps for the multidimensional Monge–Kantorovich problem. *Comm. Pure Appl. Math.* 51:23–45
Gelbrich M. 1990. On a formula for the L_2-Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachr.* 147:185–203
Gibbs AL, Su FE. 2002. On choosing and bounding probability metrics. *Int. Stat. Rev.* 70:419–35
Gigli N. 2011. On the inverse implication of Brenier–McCann theorems and the structure of $(P_2(M), W_2)$. *Meth. Appl. Anal.* 18:127–58
Gini C. 1914. Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazione statistiche. *Atti Reale Inst. Veneto Sci. Lett. Arti* 74:185–213
Givens CR, Shortt RM. 1984. A class of Wasserstein metrics for probability distributions. *Mich. Math. J.* 31:231–40
Graf S, Luschgy H. 2007. *Foundations of Quantization for Probability Distributions*. Berlin: Springer
Grave E, Joulin A, Berthet Q. 2018. Unsupervised alignment of embeddings with Wasserstein Procrustes. arXiv:1805.11222 [cs.LG]
Hairer M, Stuart AM, Vollmer SJ. 2014. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* 24:2455–90
Hallin M. 2017. *On distribution and quantile functions, ranks and signs in* $\mathbb{R}^d$. ECARES Work. Pap. 2017-34, Univ. Libre Brux. **https://ideas.repec.org/p/eca/wpaper/2013-258262.html**
Höffding W. 1940. Masstabinvariante Korrelationstheorie. *Schr. Math. Inst. Angew. Math. Univ. Berlin* 5:181–233
Huckemann S, Hotz T, Munk A. 2010. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Stat. Sin.* 20:1–58
Johnson O, Samworth R. 2005. Central limit theorem and convergence to stable laws in Mallows distance. *Bernoulli* 11:829–45
Kantorovich LV. 1942. On the translocation of masses. *Dokl. Acad. Nauk. SSSR* 37:227–29
Kantorovich LV, Rubinstein GS. 1958. On a space of completely additive functions. *Vestnik Leningr. Univ.* 13:52–59
Karcher H. 1977. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* 30:509–41
Kellerer HG. 1984. Duality theorems for marginal problems. *Z. Wahrscheinlichkeitstheorie Verwandte Gebiete* 67:399–432
Kendall WS, Le H. 2011. Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Braz. J. Probab. Stat.* 25:323–52
Kloeckner BR. 2015. A geometric study of Wasserstein spaces: ultrametrics. *Mathematika* 61:162–78
Knott M, Smith CS. 1984. On the optimal mapping of distributions. *J. Optim. Theory Appl.* 43:39–49
Kroshnin A, Suvorikova A. 2018. *Central limit theorem for Wasserstein barycenters of Gaussian measures*. Presented at the 4th Conference of the International Society for Nonparametric Statistics, Salerno, Italy, June 11–15
Kuhn HW. 1955. The Hungarian method for the assignment problem. *Naval Res. Log.* 2:83–97
Le Gouic T, Loubes JM. 2017. Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields* 168:901–17
Ledoux M. 2005. *The Concentration of Measure Phenomenon*. Providence, RI: Am. Math. Soc.
Luenberger DG, Ye Y. 2008. *Linear and Nonlinear Programming*. New York: Springer
Mallows C. 1972. A note on asymptotic joint normality. *Ann. Math. Stat.* 43:508–15
Mariucci E, Reiß M. 2017. Wasserstein and total variation distance between marginals of Lévy processes. arXiv:1710.02715 [math.PR]
Marron JS, Ramsay JO, Sangalli LM, Srivastava A. 2015. Functional data analysis of amplitude and phase variation. *Stat. Sci.* 30:468–84

Masarotto V, Panaretos VM, Zemel Y. 2018. Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A*. **https://doi.org/10.1007/s13171-018-0130-1**
Mason DM. 2016. A weighted approximation approach to the study of the empirical Wasserstein distance. In *High Dimensional Probability VII*, ed. C Houdré, DM Mason, P Reynaud-Bouret, J Rosiński, pp. 137–54. Basel, Switz.: Birkhäuser
McCann RJ. 1997. A convexity principle for interacting gases. *Adv. Math.* 128:153–79
McCann RJ. 2001. Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.* 11:589–608
Mileyko Y, Mukherjee S, Harer J. 2011. Probability measures on the space of persistence diagrams. *Inverse Probl.* 27:124007
Monge G. 1781. Mémoire sur la théorie des déblais et des remblais. In *Histoire de l'Académie Royale des Sciences de Paris*, pp. 666–704. Paris: Impr. R.
Munk A, Czado C. 1998. Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. B* 60:223–41
Munkres J. 1957. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.* 5:32–38
Ni K, Bresson X, Chan T, Esedoglu S. 2009. Local histogram based segmentation using the Wasserstein distance. *Int. J. Comput. Vis.* 84:97–111
Oliveira RI. 2009. On the convergence to equilibrium of Kac's random walk on matrices. *Ann. Appl. Probab.* 19:1200–31
Olkin I, Pukelsheim F. 1982. The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* 48:257–63
Panaretos VM, Zemel Y. 2016. Amplitude and phase variation of point processes. *Ann. Stat.* 44:771–812
Panaretos VM, Zemel Y. 2019. *An Invitation to Statistics in Wasserstein Space*. Berlin: Springer. In press
Pass B. 2013. Optimal transportation with infinitely many marginals. *J. Funct. Anal.* 264:947–63
Peyré G, Cuturi M. 2018. *Computational Optimal Transport*. arXiv:1803.00567 [stat.ML]
Rachev ST. 1985. The Monge–Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.* 29:647–76
Rachev ST. 1991. *Probability Metrics and the Stability of Stochastic Models*. New York: Wiley
Rachev ST, Rüschendorf L. 1994. On the rate of convergence in the CLT with respect to the Kantorovich metric. In *Probability in Banach Spaces 9*, ed. J Hoffmann-Jørgensen, J Kuelbs, MB Marcus, pp. 193–207. New York: Springer
Rachev ST, Rüschendorf L. 1998a. *Mass Transportation Problems*. Vol. I: *Theory*. New York: Springer
Rachev ST, Rüschendorf L. 1998b. *Mass Transportation Problems*. Vol. II: *Applications*. New York: Springer
Rachev ST, Stoyanov SV, Fabozzi FJ. 2011. *A Probability Metrics Approach to Financial Risk Measures*. New York: Wiley
Ramsay JO, Silverman BW. 2005. *Functional Data Analysis*. New York: Springer
Rio E. 2009. Upper bounds for minimal distances in the central limit theorem. *Ann. Inst. H. Poincaré Probab. Stat.* 45:802–17
Rippl T, Munk A, Sturm A. 2016. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivar. Anal.* 151:90–109
Rolet A, Cuturi M, Peyré G. 2016. Fast dictionary learning with a smoothed Wasserstein loss. *PMLR* 51:630–38
Rubner Y, Tomasi C, Guibas LJ. 2000. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* 40:99–121
Rudolf D, Schweizer N. 2018. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* 24:2610–39
Rüschendorf L, Rachev ST. 1990. A characterization of random variables with minimum L^2-distance. *J. Multivar. Anal.* 32:48–54
Santambrogio F. 2015. *Optimal Transport for Applied Mathematicians*. Basel, Switz.: Birkhäuser
Schuhmacher D. 2009. Stein's method and Poisson process approximation for a class of Wasserstein metrics. *Bernoulli* 15:550–68
Sklar M. 1959. Fonctions de répartition en n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* 8:229–31
Sommerfeld M, Munk A. 2018. Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. B* 80:219–38

Sommerfeld M, Schrieber J, Munk A. 2018. Optimal transport: fast probabilistic approximation with exact solvers. arXiv:1802.05570 [stat.CO]
Takatsu A. 2011. Wasserstein geometry of Gaussian measures. *Osaka J. Math.* 48:1005–26
Talagrand M. 1994. The transportation cost from the uniform measure to the empirical measure in dimension ≥3. *Ann. Probab.* 22:919–59
Tameling C, Munk A. 2018. Computational strategies for statistical inference based on empirical optimal transport. In *2018 IEEE Data Science Workshop (DSW)*, pp. 175–79. New York: IEEE
Tameling C, Sommerfeld M, Munk A. 2017. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. arXiv:1707.00973 [math.PR]
Tanaka H. 1973. An inequality for a functional of probability distributions and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Wahrscheinlichkeitstheorie Verwandte Gebiete* 27:47–52
Villani C. 2003. *Topics in Optimal Transportation*. Providence, RI: Am. Math. Soc.
Villani C. 2008. *Optimal Transport: Old and New*. Berlin: Springer
Wang JL, Chiou JM, Müller HG. 2016. Functional data analysis. *Annu. Rev. Stat. Appl.* 3:257–95
Wasserman L. 2018. Topological data analysis. *Annu. Rev. Stat. Appl.* 5:501–32
Weed J, Bach F. 2018. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*. In press
Zemel Y, Panaretos VM. 2018. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*. In press

Annual Review of Statistics and its Application

Volume 6, 2019

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics