Bernoulli **25**(2), 2019, 932–976
https://doi.org/10.3150/17-BEJ1009

Fréchet means and Procrustes analysis in Wasserstein space

YOAV ZEMEL[*] and VICTOR M. PANARETOS[**]

Institut de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.
E-mail: [*]*yoav.zemel@epfl.ch;* [**]*victor.panaretos@epfl.ch*

We consider two statistical problems at the intersection of functional and non-Euclidean data analysis: the determination of a Fréchet mean in the Wasserstein space of multivariate distributions; and the optimal registration of deformed random measures and point processes. We elucidate how the two problems are linked, each being in a sense dual to the other. We first study the finite sample version of the problem in the continuum. Exploiting the tangent bundle structure of Wasserstein space, we deduce the Fréchet mean via gradient descent. We show that this is equivalent to a Procrustes analysis for the registration maps, thus only requiring successive solutions to pairwise optimal coupling problems. We then study the population version of the problem, focussing on inference and stability: in practice, the data are i.i.d. realisations from a law on Wasserstein space, and indeed their observation is discrete, where one observes a proxy finite sample or point process. We construct regularised nonparametric estimators, and prove their consistency for the population mean, and uniform consistency for the population Procrustes registration maps.

Keywords: functional data analysis; manifold statistics; Monge–Kantorovich problem; multimarginal transportation; optimal transportation; phase variation; point process; random measure; registration; shape theory; warping

1. Introduction

Functional data analysis (e.g., Hsing and Eubank [40]) and non-Euclidean statistics (e.g., Patrangenaru and Ellingson [60]) represent modern areas of statistical research, whose key challenges arise from the intrinsic complexity of the data and the peculiarities of their ambient space. In the first case, the data are random elements in a separable Hilbert space of functions (typically $L^2[0, 1]$), and resulting challenges are linked to infinite dimensionality (e.g., ill-posed studentisation, Munk *et al.* [55], and discrete measurements of continuum random objects, Zhang and Wang [73]). In the second case, the data are seen as random elements of a finite-dimensional Riemannian manifold (often a shape space), and resulting challenges are linked to the non-linear structure of the space (e.g., existence/uniqueness of Fréchet means, Le [48] and Kendall [45], and analysis of manifold variation, Huckemann, Munk and Hotz [41]).

At the intersection of these two domains, with manifestations in neurophysiology, imaging, and environmetrics, one finds data objects that are best modelled as *distributions* over $\mathbb{R}^d$, that is, *random measures* (Stoyan, Kendall and Mecke [24], Kallenberg [43]). Such random measures carry the infinite dimensional traits of functional data, but at the same time are characterised by intrinsic non-linearities due to their positivity and integrability constraints, requiring a non-Euclidean point of view. Indeed, despite their functional nature, their dominating variational

1350-7265 © 2019 ISI/BS

feature is not due to additive amplitude fluctuations (as can be seen in the Karhunen–Loève expansion of functional data), but rather to *random deformation* of a structural mean (as in Freitag and Munk [33]) or template (as in *morphometrics*, Bookstein [20]). Still, being infinite dimensional, their observation is typically done discretely, for example, noisily over a grid (e.g., Amit *et al.* [8], Allassonnière *et al.* [4]) or via random sampling (e.g., Panaretos and Zemel [58]), requiring tools and techniques from nonparametric statistics, as used in functional data analysis.

In this setting, the typical statistical objective is to estimate the underlying template that gives rise to the data by random deformation. This can often be modelled as a Fréchet mean with respect to some metric structure; dual to this problem is the recovery the deformation maps themselves, in order to *register* the individual realisations in a common coordinate system, given by *registration maps*. These problems are interwoven in *shape theory*, where the template and registration maps are the two ingredients of Procrustes analysis (Gower [37]; Dryden and Mardia [29]) and non-Euclidean PCA (Huckemann, Munk and Hotz [41]; Huckemann and Ziezold [42]). Obviously, the methods and algorithms for estimating a mean and carrying out a registration/Procrustes analysis are inextricably linked with the geometry of the sample space, which can be a matter of modelling choice or of first principles.

In this paper, we choose to study the problem of *Fréchet averaging* and *Procrustes registration* when the data are viewed as elements of the L^2-Wasserstein space of multivariate measures on $\mathbb{R}^d$. We choose this setting since it has a long history in assessing compatibility and fit of distributions related via deformations (Munk and Czado [54]; Freitag and Munk [33]), and as it can be seen to be a natural analogue of using L^2, in the case of measures[1] (Panaretos and Zemel [58]; Bigot and Klein [14]). We work at both a sample level and a population level, as well as both at the level of continuum and discrete observation: our object of study is the determination of the Fréchet mean and registration maps at the level of a sample, as well as at their estimation when the observed measures are discretely observed realisations from a population of random measures. When $d = 1$, the problem is well understood, owing to the flat geometry of Wasserstein space (Panaretos and Zemel [58]). When $d > 1$, however, the Wasserstein space has non-negative curvature, and one encounters the classical difficulties of non-Euclidean statistics, augmented by the infinite dimensionality and discrete measurement of the problem (see Anderes *et al.* [9], Sommerfeld and Munk [67] and Tameling *et al.* [69] for challenges involved in the discrete setting).

In more detail, our contributions are:

(A) *At the sample level*: we illustrate how knowledge of the Fréchet mean (template) gives an explicit solution to the optimal registration/multicoupling problem (Section 3.1, Proposition 2). We study the tangent space geometry, using it to determine the gradient of the Fréchet functional (Section 3.2.2, Theorem 1), and characterise Karcher means via its zeroes (Corollary 1, Section 3.2.3). We give criteria for determining when a Karcher mean (local optimum) is a Fréchet mean (global optimum; Theorem 2). We construct a gradient descent algorithm (Algorithm 1), and find its optimal stepsize (Lemma 2) illustrating the algorithm structurally equivalent to a Procrustes algorithm (Section 3.3), reducing the determination of the mean to the successive solution of pairwise optimal transport problems. We prove that the gradient iterate converges to a

[1] In the sense that the Wasserstein space is topologically homeomorphic to a convex subset of $L^2([0,1]^d)$; when $d = 1$, this homeomorphism is an isometry, whereas for $d > 1$, it is a local isometry.

Karcher mean in the Wasserstein metric (Section 3.3.2, Theorem 3); and that the induced transportation maps converge uniformly to the Procrustes maps (required for optimal mutlicoupling; Theorem 4, Section 3.3.3). The latter is particularly involved and requires techniques from the geometry of monotone operators on $\mathbb{R}^d$. As a noteworthy corollary, we deduce convergence of the multicouplings (Corollary 3).

(B) *At the population level*: we consider a population level model linking Fréchet means and optimal registration and give conditions for model identifiability (Section 4.1, Theorem 5); We then tackle the problem of point estimation of the population mean and registration maps in a functional data analysis setup, where instead of observing an i.i.d. sample $\{\mu^1, \dots, \mu^N\}$ from the population, we observe samples or point processes with these measures as distributions/intensities. In this setting, we construct regularised nonparametric estimators of the Fréchet means and Procrustes maps, and prove that they are consistent in Wasserstein distance and uniform norm, respectively (Theorems 6 and 7).

Before presenting our main results, we first provide a short introduction to Wasserstein space in Section 2. Section 5 gathers the main proofs, for the sake of tidiness, and Section 6 presents several interesting special examples as an illustration. An online Supplement [72] provides further technical details omitted from the main paper, including some important measurability issues.

In reviewing an earlier version of our paper, a referee brought to our attention independent parallel work by Álvarez-Esteban *et al.*, that has since been published in [6]. Their work overlaps with part of ours in (A) above (Sections 3.3.1 and 3.3.2). In particular, they too arrive at a (structurally) same algorithm (Algorithm 1). Their motivation, construction, and convergence proof differ substantially from ours (theirs is a fixed point iteration heuristically motivated by the Gaussian case, while their proof uses almost sure representations). Indeed, our geometrical framework and proof techniques is what allows us to study the problem of optimal registration (Procrustes analysis), requiring a careful study of the stochastic convergence of monotone operators on $\mathbb{R}^d$ (Section 5.5).

2. Optimal transportation and Wasserstein space

The reason the Wasserstein space arises as the natural space to capture deformation-based variation of random measures lies in its deep connection with the problem of *optimal transportation of measure*. This consists in solving the *Monge problem* (Villani [70]): given a pair of measures (μ, ν), find a mapping $\mathbf{t}_\mu^\nu : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that $\mathbf{t}_\mu^\nu \# \mu = \nu$, and

$$\int_{\mathbb{R}^d} \|\mathbf{t}_\mu^\nu(x) - x\|^2 \, \mathrm{d}\mu(x) \le \int_{\mathbb{R}^d} \|\mathbf{q}(x) - x\|^2 \, \mathrm{d}\mu(x),$$

for any other $\mathbf{q}$ such that $\mathbf{q}\#\mu = \nu$. Here, "#" denotes the push-forward operation, where $[\mathbf{t}\#\mu](A) = \mu(\mathbf{t}^{-1}(A))$ for all Borel sets A of $\mathbb{R}^d$. The map $\mathbf{t}_\mu^\nu$ is called an optimal transport plan, and a solution to this problem yields an optimal deformation of μ into ν with respect to the *transport cost* given by squared Euclidean distance.

An optimal transport map may fail to exist, and instead, one may need to solve the relaxed Monge problem, known as the *Kantorovich* problem (Villani [70]). Here instead of seeking a map $\mathbf{t}_\mu^\nu \# \mu = \nu$, one seeks a distribution ξ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, minimising the functional

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\xi(x, y)$$

over all measures ξ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. In probabilistic terms, ξ yields a coupling of random variables $X \sim \mu$ and $Y \sim \nu$ that minimises the quantity

$$\mathbb{E}\|X - Y\|^2,$$

over all possible couplings of X and Y. It can be shown that when the measure μ is regular (absolutely continuous with respect to Lebesgue measure), the Kantorovich problem reduces to the Monge problem, and the optimal coupling ξ is supported on the graph of the function. That is, the optimal coupling exists, is unique, and can be realised by a proper transport map $\mathbf{t}_\mu^\nu$.

One may consider the space $\mathcal{P}_2(\mathbb{R}^2)$ of all probability measures μ on $\mathbb{R}^d$ with finite variance (that is, $\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < \infty$) as a metric space, endowed with the L^2-Wasserstein distance

$$d(\mu, \nu) = \inf_{\xi \in \Gamma(\mu,\nu)} \sqrt{\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\xi(x, y)},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. The induced metric space is colloquially called *Wasserstein space* and will form the geometrical context for our study of *deformation-based* variation of random measures. This space has been used extensively in statistics, as it metrises the topology of weak convergence, and convergence with respect to the metric yields both convergence in law, as well as convergence of the first two moments (for instance, in applications to the bootstrap, e.g., Bickel and Freedman [12], and to goodness-of-fit, e.g., Rippl, Munk and Sturm [62]).

The appropriateness of this distance when modeling deformations of measures becomes clear based on our previous remark concerning regularity: one can imagine an initial regular template μ, that is *deformed* according to maps $\mathbf{q}_i$ to yield new measures $\mu^i = (\mathbf{q}_i)\#\mu$. It is then natural to quantify the distance of the template to its perturbations by means of the minimal transportation (or deformation) cost

$$d\bigl(\mu, \mu^i\bigr) = \sqrt{\int_{\mathbb{R}^d} \bigl\|\mathbf{t}_\mu^{\mu^i}(x) - x\bigr\|^2 \, \mathrm{d}\mu(x)}.$$

That the distance can be expressed via a proper map, is due to the assumed regularity of μ. Note that the maps $\mathbf{q}_i$ themselves will, in general, not be identifiable (many Borel maps can push μ forward to μ^i). But they can be assumed to be exactly optimal, that is, $\mathbf{q}_i = \mathbf{t}_\mu^{\mu^i}$ as a matter of parsimony, and in any case without loss of generality, leading to identifiability. These maps will also solve the registration problem: a map of the form $\mathbf{t}_\mu^{\mu^i} - \mathbf{i}$, with $\mathbf{i}$ the identity mapping, shows how the coordinate system of μ should be deformed to be registered to the coordinate system of μ^i.

This raises the question of how to *characterise* the optimal transportation maps. For instance, in the one-dimensional case, if μ and ν are probability measures on $\mathbb{R}$, and μ is diffuse we may write

$$\mathbf{t}_{\mu}^{\nu} = G_{\nu}^{-1} \circ G_{\mu}, \tag{2.1}$$

where $G_{\mu}(t) = \int_{-\infty}^{t} \mathrm{d}\mu(x)$, $G_{\nu}(t) = \int_{-\infty}^{t} \mathrm{d}\nu(x)$ are their distribution functions and G_{ν}^{-1} is the quantile function of ν. This characterises optimal maps in one dimension as non-decreasing functions. More generally, when one has measures on $\mathbb{R}^d$, the class of optimal maps can be seen to be that of *monotone maps* (see Section 5.5), defined as fields $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ that are obtained as gradients of convex functions $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbf{t} = \nabla\varphi.$$

This is known as Brenier's characterisation (Villani [70], Theorem 2.12). With these basic definitions in place, we are now ready to consider the problem of finding a Fréchet mean of a collection of measures – the latter viewed as the common template measure that was deformed to give rise to these measures.

3. Sample setting

3.1. Fréchet means and optimal registration

The notion of a Fréchet mean (Fréchet [31]) generalises that of the mean in a normed vector space to a general metric space. Though it has primarily been studied on Riemannian manifolds, the generality of its definition allows it to be used very broadly: it replaces the usual "sum of squares", with a "sum of squared distances", the *Fréchet functional*. A closely related notion is that of a *Karcher mean* (Karcher [44]; Le [50]), a term that describes stationary points of the sum of squares functional, when the latter is differentiable. See Kendall [45], and Kendall and Le [46] for an overview and a detailed review, respectively. In the context of Wasserstein space, a Fréchet mean of a collection of measures $\{\mu^1, \dots, \mu^N\}$, is a minimiser of the Fréchet functional

$$F(\gamma) := \frac{1}{2N} \sum_{i=1}^{N} d^2\big(\mu^i, \gamma\big) \tag{3.1}$$

over elements γ in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, and a Karcher mean is a stationary point of F. The functional will be finite for any $\gamma \in \mathcal{P}_2(\mathbb{R}^d)$, provided that it is so for some γ_0. Population versions, assuming $\mathcal{P}_2(\mathbb{R}^d)$ is endowed with a probability measure, can also be defined, replacing summation by expectation with respect to that law. Interestingly, Fréchet himself [32] considered the Wasserstein metric between probability measures on $\mathbb{R}$, and some refer to this as the *Fréchet distance* (e.g., Dowson and Landau [28]). In general, existence and uniqueness of a sample Fréchet mean can be subtle, but Agueh and Carlier [2] have shown that it *will uniquely* *exist* in the Wasserstein space, provided that some regularity is asserted.[2] Here and in the following, we call a measure *regular* if it is absolutely continuous with respect to Lebesgue measure (this condition can be slightly weakened [2]).

Proposition 1 (Agueh and Carlier [2]). *Let* $\{\mu^1, \dots, \mu^N\}$ *be a collection in the Wasserstein space of measures* $\mathcal{P}_2(\mathbb{R}^d)$. *If at least one of the measures is regular with bounded density, then their Fréchet mean exists, is unique, and is regular.*

We will now show that, once the Fréchet mean $\bar{\mu}$ of $\{\mu^1, \dots, \mu^N\}$ has been determined, it may be used to optimally multi-couple the measures $\{\mu^1, \dots, \mu^n\}$ in $\mathbb{R}^{d\times N}$, in terms of pairwise mean square distances, thus providing a solution to the *multidimensional Monge–Kantorovich problem* considered by Gangbo and Święch [35]. That is, $\bar{\mu}$ can be used to construct a random vector whose marginals are as concentrated as possible in terms of pairwise mean-square distance, subject to the constraint of having laws $\{\mu^1, \dots, \mu^N\}$.

Our first result combines results of [2] and [35] to illustrate precisely how (also see Pass [59], Theorem 4.2.2, for an analogous result when considering continuous flows of measures).

Proposition 2 (Optimal multicoupling via Fréchet means). *Let* $\{\mu^1, \dots, \mu^N\}$ *be regular probability measures in* $\mathcal{P}_2(\mathbb{R}^d)$, *one with bounded density, and let* $\bar{\mu}$ *be their (unique) Fréchet mean with respect to the Wasserstein metric. Let* $Z \sim \bar{\mu}$ *and define*

$$\boldsymbol{X} = (X_1, \dots, X_N), \qquad X_i = \mathbf{t}_{\bar{\mu}}^{\mu^i}(Z), \qquad i = 1 \dots, N,$$

where $\mathbf{t}_{\bar{\mu}}^{\mu^i}$ *is the optimal transport plan pushing* $\bar{\mu}$ *forward to* μ^i. *Then* $X_i \sim \mu^i$ *for* $i = 1, \dots, N$ *and furthermore,*

$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathbb{E}\|X_i - X_j\|^2 \le \sum_{i=1}^{N} \sum_{j=i+1}^{N} \mathbb{E}\|Y_i - Y_j\|^2$$

for any other $\boldsymbol{Y} = (Y_1, \dots, Y_N)$ *such that* $Y_i \sim \mu^i$, $i = 1, \dots, N$.

In the language of shape theory, the Fréchet mean $\bar{\mu}$ may be used as a *template* to *jointly register* the collection of measures, just as Euclidean configurations can be registered to their Procrustes mean by a Procrustes analysis (Goodall [36]). Only in this case, instead of the similarity group of shape theory, registration is *deformation based*, by means of the collection of maps $\{\mathbf{t}_{\bar{\mu}}^{\mu^i}\}_{i=1}^N$, where $\mathbf{t}_{\bar{\mu}}^{\mu^i}$ is the optimal transport map

$$\mathbf{t}_{\bar{\mu}}^{\mu^i} \# \bar{\mu} = \mu^i.$$

By analogy to shape theory, we shall refer to these as *Procrustes maps*. These yield a common coordinate system (corresponding to $\bar{\mu}$) where one can best compare samples from each measure,

[2] For a population version, one needs to tackle measurability and identifiability issues, see Section 4.1.

similarly to “quantile renormalisation” in one dimension, for example, Bolstad *et al.* [17], Gallon *et al.* [34]. The Procrustes maps can also be used in order to produce a Principal Component Analysis, capturing the main modes of deformation-based variation (Bigot *et al.* [13], Panaretos and Zemel [58]; Huckemann, Munk and Hotz [41], Wang *et al.* [71]).

3.2. Wasserstein geometry and the gradient of the Fréchet functional

In this section, we determine the conditions for the Fréchet derivative of the Fréchet functional (3.1) to be well defined, and determine its functional form. Furthermore, we characterise Karcher means and give criteria for their optimality, opening the way for the determination of the Fréchet mean. The key to our analysis will be to exploit the tangent bundle over the Wasserstein space of regular measures.

3.2.1. *The tangent bundle*

Let $\mathcal{P}_2(\mathbb{R}^d)$ be the Wasserstein space of probability measures μ on $\mathbb{R}^d$ such that $\int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)$ is finite, as defined in Section 2. An absolutely continuous measure on $\mathbb{R}^d$ will be called *regular*. When $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ is regular and $\mu^1 \in \mathcal{P}_2(\mathbb{R}^d)$, the transportation map $\mathbf{t}_{\mu^0}^{\mu^1}$ uniquely exists, in which case there is a unique geodesic curve between μ^0 and μ^1. Using again the notation $\mathbf{i}$ for the identity map, this geodesic is given by

$$\mu_t = \bigl[\mathbf{i} + t\bigl(\mathbf{t}_{\mu^0}^{\mu^1} - \mathbf{i}\bigr)\bigr]\#\mu^0, \qquad t \in [0, 1].$$

This curve is known as McCann’s interpolation (McCann [52], Villani [70]). The tangent space at an arbitrary $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is then (Ambrosio *et al.* [7], Definition 8.4.1, p. 189)

$$\mathrm{Tan}_\mu = \mathrm{Tan}_\mu \mathcal{P}_2\bigl(\mathbb{R}^d\bigr) = \overline{\bigl\{\nabla\varphi : \varphi \in C_c^\infty\bigl(\mathbb{R}^d\bigr)\bigr\}}^{L^2(\mu)},$$

where $C_c^\infty(\mathbb{R}^d)$ denotes infinitely differentiable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ with compact support, and the closure operation is taken with respect to the space $L^2(\mu)$. Note the interesting fact that the closure operation is the only aspect of the tangent space that directly involves the measure μ. An equivalent definition, which is more useful to us, is given by Ambrosio *et al.* [7], Definition 8.5.1, p. 195:

$$\mathrm{Tan}_\mu = \overline{\bigl\{\lambda(\mathbf{r} - \mathbf{i}) : \mathbf{r} \text{ optimal between } \mu \text{ and } \mathbf{r}\#\mu;\ \lambda > 0\bigr\}}^{L^2(\mu)},$$

that is, we take the collection of $\mathbf{r}$’s that are optimal maps from μ to $\mathbf{r}\#\mu$; i.e. the gradients of convex functions. This is a linear space (not just a cone) by the first definition, even though it is not obvious from the second. The definitions are equivalent by Theorem 8.5.1 of Ambrosio *et al.* [7], p. 195. As was mentioned above, when $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ is regular, every measure $\mu^1 \in \mathcal{P}_2(\mathbb{R}^d)$ admits a unique optimal map $\mathbf{t}_{\mu^0}^{\mu^1}$ that pushes μ^0 forward to μ^1. Thus, the exponential map

$$\exp_{\mu^0}(\mathbf{r} - \mathbf{i}) = \mathbf{r}\#\mu^0$$

is surjective, and its inverse, the log map

$$\log_{\mu^0}(\mu^1) = \mathbf{t}_{\mu^0}^{\mu^1} - \mathbf{i},$$

is well-defined throughout $\mathcal{P}_2(\mathbb{R}^d)$. In particular, the geodesic $[\mathbf{i} + t(\mathbf{t}_{\mu^0}^{\mu^1} - \mathbf{i})]\#\mu^0$ is mapped bijectively to the line segment $t(\mathbf{t}_{\mu^0}^{\mu^1} - \mathbf{i}) \in \mathrm{Tan}_{\mu^0}$ through the log map.

3.2.2. *Gradient of the Fréchet functional*

We will now exploit the tangent bundle structure described in the previous section in order to determine the gradient of the empirical Fréchet functional. Fix $\mu^0 \in \mathcal{P}_2(\mathbb{R}^d)$ and consider the function

$$F_0 : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}, \qquad F_0(\mu) = \frac{1}{2} d^2(\mu, \mu^0).$$

When μ is regular, we have that ([7], Corollary 10.2.7, p. 239), for any μ^0

$$\lim_{\nu \to \mu} \frac{F_0(\nu) - F_0(\mu) + \int_{\mathbb{R}^d} \langle \mathbf{t}_\mu^{\mu^0}(x) - x, \mathbf{t}_\mu^\nu(x) - x \rangle \, \mathrm{d}\mu(x)}{d(\nu, \mu)} = 0,$$

where the convergence $\nu \to \mu$ is with respect to the Wasserstein distance. The integral above can be seen as the inner product

$$\langle \mathbf{t}_\mu^{\mu^0} - \mathbf{i}, \mathbf{t}_\mu^\nu - \mathbf{i} \rangle$$

in the space $L^2(\mu)$ that includes as a (closed) subspace the tangent space Tan_μ. In terms of this inner product and the log map, we can write

$$F_0(\nu) - F_0(\mu) = -\langle \log_\mu(\mu^0), \log_\mu(\nu) \rangle + o(d(\nu, \mu)), \qquad \nu \to \mu,$$

so that F_0 is Fréchet-differentiable at μ with derivative

$$F_0'(\mu) = -\log_\mu(\mu^0) = -(\mathbf{t}_\mu^{\mu^0} - \mathbf{i}) \in \mathrm{Tan}_\mu .$$

We have proven:

Theorem 1 (Gradient of the Fréchet functional). *Fix a collection of measures* $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$. *When* γ *is regular, the Fréchet functional*

$$F(\gamma) = \frac{1}{2N} \sum_{i=1}^N d^2(\gamma, \mu^i), \qquad \gamma \in \mathcal{P}_2(\mathbb{R}^d) \tag{3.2}$$

is Fréchet-differentiable, and its gradient satisfies

$$F'(\gamma) = -\frac{1}{N} \sum_{i=1}^N \log_\gamma(\mu^i) = -\frac{1}{N} \sum_{i=1}^N (\mathbf{t}_\gamma^{\mu^i} - \mathbf{i}). \tag{3.3}$$

3.2.3. *Karcher and Fréchet means*

We can now characterise Karcher means, and also show that the empirical Fréchet mean must be sought amongst them, by an immediate corollary to Theorem 1:

Corollary 1. *Let $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$ be regular measures, one of which with bounded density. A measure μ is a Karcher mean of $\{\mu^i\}$ if and only if*

$$\frac{1}{N}\sum_{i=1}^{N}\big(\mathbf{t}_\mu^{\mu^i} - \mathbf{i}\big) = 0, \qquad \mu\text{-almost everywhere.}$$

Furthermore, the Fréchet mean of $\{\mu^i\}$ is itself a Karcher mean.

In fact, the corollary suggests that a Karcher mean is "almost" a Fréchet mean: Agueh and Carlier [2] show by convex optimisation methods that if $\sum_{i=1}^N (\mathbf{t}_\mu^{\mu^i} - \mathbf{i}) = 0$ *everywhere* on $\mathbb{R}^d$ (rather than just μ-almost everywhere), then μ is in fact the unique Fréchet mean. Thus one hopes that this "gap of measure zero" can be bridged: that a sufficiently regular Karcher mean should in fact be a Fréchet mean. We now show that this is indeed the case; if $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$ are smooth measures with convex support, then a smooth Karcher mean of same support *must be the unique Fréchet mean*:

Theorem 2 (Optimality criterion for Karcher means). *Let μ^i for $i = 1, \dots, N$ be probability measures on an open convex $X \subseteq \mathbb{R}^d$ whose densities g^i are bounded and strictly positive on X and let μ be a regular Karcher mean of $\{\mu^i\}$ with density f. Then μ is the unique Fréchet mean of $\{\mu^i\}$, provided one of the following holds:*

1. *$X = \mathbb{R}^d$, f is bounded and strictly positive, and the densities $f, g^1, \dots, g^N$ are of class C^1;*
2. *X is bounded, $\mu(X) = 1$, f is bounded, and the densities $f, g^1, \dots, g^N$ are bounded from below on X.*

Remark 1. In the first condition, the C^1 assumption can be weakened to Hölder continuity of the densities for some exponent $\alpha \in (0, 1]$.

Remark 2. We conjecture that a stronger result should be valid: specifically, if $\mu^1, \dots, \mu^N$ satisfy the conditions of Theorem 2, then we conjecture the Fréchet functional F to in fact have a unique Karcher mean, coinciding with the Fréchet mean.

3.3. Gradient descent and Procrustes analysis

3.3.1. *Elements of the algorithm*

Let $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$ be regular and let $\gamma_j \in \mathcal{P}_2(\mathbb{R}^d)$ be a regular measure, representing our current estimate of the Fréchet mean of $\mu^1, \dots, \mu^N$ at step j. Following the discussion above, it makes sense to introduce a step size $\tau_j > 0$, and to carry out a steepest descent in the space of measures (e.g. Molchanov and Zuyev [53]), following the negative of the gradient:

$$\gamma_{j+1} = \exp_{\gamma_j}\big(-\tau_j F'(\gamma_j)\big) = \left[\mathbf{i} + \tau_j \frac{1}{N}\sum_{i=1}^{N} \log_{\gamma}\big(\mu^i\big)\right]\#\gamma_j = \left[\mathbf{i} + \tau_j \frac{1}{N}\sum_{i=1}^{N} \big(\mathbf{t}_{\gamma_j}^{\mu^i} - \mathbf{i}\big)\right]\#\gamma_j.$$

In order to guarantee that the descent is well-defined, we must make sure that the gradient itself will remain well-defined as we iterate over j. In view of Theorem 1, this requires showing that γ_{j+1} remains regular whenever γ_j is regular. This is indeed the case, at least if the step size is contained in [0, 1]:

Lemma 1 (Regularity of the iterates). *If γ_0 is regular and $\tau_0 \in [0, 1]$, then so is γ_1.*

Lemma 1 suggests that the step size must be restricted to [0, 1]. The next result suggests that the objective function essentially tells us that the *optimal* step size, achieving the maximal reduction of the objective function (thus corresponding to an approximate line search), is exactly equal to 1:

Lemma 2 (Optimal stepsize). *If $\gamma_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is regular then*

$$F(\gamma_1) - F(\gamma_0) \le -\big\|F'(\gamma_0)\big\|^2\left[\tau - \frac{\tau^2}{2}\right]$$

and the bound on the right-hand side of the last display is minimised when $\tau = 1$.

In light of the results in Lemmas 1 and 2, one needs only concentrate on the case $\tau_j = 1$. This has an interesting ramification: when $\tau = 1$, the gradient descent iteration is structurally equivalent to a Procrustes analysis. Specifically, the gradient descent algorithm proceeds by iterating the two steps of a Procrustes analysis (Gower [37]; Dryden and Mardia [29], p. 90):

(1) *Registration*: Each of the measures $\{\mu^1, \dots, \mu^N\}$ is *registered* to the current template γ_j, via the optimal transportation (registration) maps $\mathbf{t}_{\gamma_j}^{\mu^i}$. In geometrical terms, the measures $\{\mu^1, \dots, \mu^N\}$ are lifted to the tangent space at γ_j (via the log map), and their linear representation on the tangent space is expressed in local coordinates which coincide with the maps $\mathbf{t}_{\gamma_j}^{\mu^i} - \mathbf{i} = \log_{\gamma_j}(\mu^i)$. These can be seen as a common coordinate system for $\{\mu^1, \dots, \mu^N\}$, that is, a registration.

(2) *Averaging*: The registered measures are *averaged coordinate-wise*, using the common coordinates system by the registration step (1). In geometrical terms, the linear representation of $\{\mu^1, \dots, \mu^N\}$ afforded by their local coordinates $\mathbf{t}_{\gamma_j}^{\mu^i} - \mathbf{i} = \log_{\gamma_j}(\mu^i)$ is averaged linearly. The linear average is then retracted back onto the manifold via the exponential map to yield the estimate at the $(j + 1)$-step.

That the gradient descent reduces to Procrustes analysis is not simply of aesthetic value. It is of the essence, as it shows that the algorithm relies entirely on solving a succession of *pairwise op-*

Algorithm 1 Gradient Descent via Procrustes Analysis

```
(A) Set a tolerance threshold $\varepsilon > 0$.
(B) For $j = 0$, let $\gamma_j$ be an arbitrary regular measure.
(C) For $i = 1, \dots, N$ solve the (pairwise) Monge problem and find the optimal transport map
    $\mathbf{t}_{\gamma_j}^{\mu^i}$ from $\gamma_j$ to $\mu^i$.
(D) Define the map $T_j = N^{-1} \sum_{i=1}^{N} \mathbf{t}_{\gamma_j}^{\mu^i}$.
(E) Set $\gamma_{j+1} = T_j \# \gamma_j$, that is, push-forward $\gamma_j$ via $T_j$ to obtain $\gamma_{j+1}$.
(F) If $\|F'(\gamma_{j+1})\| < \varepsilon$, stop, and output $\gamma_{j+1}$ as the approximation of $\bar{\mu}$ and $\mathbf{t}_{\gamma_{j+1}}^{\mu^i}$ as the
    approximation of $\mathbf{t}_{\bar{\mu}}^{\mu^i}$, $i = 1, \dots, N$. Otherwise, return to step (C).
```

timal transportation problems, thus reducing the determination of the Fréchet mean to the classical Monge problem of optimal transportation (e.g., Benamou and Brenier [10], Haber *et al.* [39], Chartrand *et al.* [23]). After all, this is precisely the point of a Procrustes algorithm: exploiting the (easier) problem of pairwise registration to solve the (harder problem) of multi-registration. We note that, further to requiring the ability to solve the pairwise optimal transportation problem, and the regularity conditions on the measures, the algorithm does not require additional structural assumptions/workarounds to reduce the problem to the one-dimensional case (as in, for example the "admissibility" approach of Boissard *et al.* [16]). An additional practical advantage is that Procrustes algorithms are easily parallelisable, since one can distribute the solution of the pairwise transport problems at each step j. Any regular measure can serve as an initial point for the algorithm, for instance one of the μ^i. We should mention at this point that, if one is content with obtaining an *approximate* or *regularised* Fréchet mean, then there are several numerical strategies available, and there is a rapidly growing literature for the efficient computation of such schemes – we briefly summarise some such approaches in the concluding remarks section (Section 7).

The gradient/Procrustes iteration is presented succinctly as Algorithm 1.

3.3.2. *Convergence of the algorithm*

In order to tackle the issue of convergence, we will use an approach that is specific to the nature of optimal transportation. The reason is that Hessian type arguments that are used to prove similar convergence results for gradient descent on Riemmanian manifolds (Afsari *et al.* [1]) or Procrustes algorithms (Le [49], Groisser [38]) do not apply here, since the Fréchet functional may very well fail to be twice differentiable. Still, this specific geometry of Wasserstein space affords some advantages; for instance, we will place no restriction on the starting point for the iteration, except that it be regular.

Theorem 3 (Limit points are Karcher means). *Let $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$ be absolutely continuous probability measures, one of which with bounded density. Then, the sequence generated by Algorithm 1 stays in a compact set of the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$, and any limit point of the sequence is a Karcher mean of $(\mu^1, \dots, \mu^N)$.*

In view of Corollary 1, this immediately implies the following.

Corollary 2 (Wasserstein convergence of gradient descent). *Under the conditions of Theorem 3, if F has a unique stationary point, then the sequence $\{\gamma_j\}$ generated by Algorithm 1 converges to the Fréchet mean of $\{\mu^1, \dots, \mu^N\}$ in the Wasserstein metric,*

$$d(\gamma_j, \bar{\mu}) \overset{j\to\infty}{\longrightarrow} 0.$$

Of course, combining Theorem 3 with Theorem 2 shows that the conclusion of Corollary 2 holds when the appropriate assumptions on $\{\mu^i\}$ and the Karcher mean μ are satisfied. The proof of Theorem 3 is elaborate, and is constructed via a series of intermediate results in a separate section (Section 5.3.1) in the interest of tidiness. The main challenge is that the standard condition used for convergence of gradient descent algorithms, that gradients be Lipschitz, fails to hold in this setup. Indeed, F is not differentiable on discrete measures, and these constitute a dense subset of the Wasserstein space.

3.3.3. *Uniform convergence of Procrustes maps and multicoupling*

We conclude our analysis of the algorithm by turning to the Procrustes maps $\mathbf{t}_{\mu^i}^{\bar{\mu}}$, which optimally couple each sample observation μ^i to their Fréchet mean $\bar{\mu}$. These are the key objects required for the solution of the multicoupling problem (as established in Proposition 2), and one would use the limit of $\mathbf{t}_{\mu^i}^{\gamma_j}$ in j as their approximation. However, the fact that $d(\gamma_j, \bar{\mu}) \to 0$ does not immediately imply the convergence of $\mathbf{t}_{\mu^i}^{\gamma_j}$ to $\mathbf{t}_{\mu^i}^{\bar{\mu}}$: the Wasserstein convergence only means that certain integrals of the warp maps converge. Still, convergence of the warp maps *does* hold, indeed uniformly so on compacta, $\bar{\mu}$-almost everywhere:

Theorem 4 (Uniform convergence of Procrustes maps). *Under the conditions of Corollary 2, there exist sets $A, B^1, \dots, B^N \subseteq \mathbb{R}^d$ such that $\bar{\mu}(A) = 1 = \mu^1(B^1) = \cdots = \mu^N(B^N)$ and*

$$\sup_{\Omega_1} \| \mathbf{t}_{\gamma_j}^{\mu^i} - \mathbf{t}_{\bar{\mu}}^{\mu^i} \| \overset{j\to\infty}{\longrightarrow} 0, \qquad \sup_{\Omega_2} \| \mathbf{t}_{\mu^i}^{\gamma_j} - \mathbf{t}_{\mu^i}^{\bar{\mu}} \| \overset{j\to\infty}{\longrightarrow} 0, \qquad i = 1, \dots, N,$$

for any pair of compacta $\Omega_1 \subseteq A$, $\Omega_2 \subseteq B^i$, where the sequence $\mathbf{t}_{\mu^i}^{\gamma_j}$ and $\mathbf{t}_{\gamma_j}^{\mu^i} = (\mathbf{t}_{\mu^i}^{\gamma_j})^{-1}$ are the Procrustes maps generated by Algorithm 1. If in addition all the measures $\{\mu^1, \dots, \mu^N\}$ have the same support, then one can choose the sets so that $B^1 = \cdots = B^N$.

With both ingredients of the registration problem in hand, we deduce a solution to the latter:

Corollary 3 (Convergence of multicouplings). *Under the conditions of Corollary 2, the sequence of multicouplings*

$$\big(\mathbf{t}_{\gamma_j}^{\mu^1}, \dots, \mathbf{t}_{\gamma_j}^{\mu^n}\big)\#\gamma_j$$

of $\{\mu^1, \dots, \mu^N\}$ *converges* (*in Wasserstein distance on* $(\mathbb{R}^d)^N$) *to the optimal multicoupling* $(\mathbf{t}_{\bar{\mu}}^{\mu^1}, \dots, \mathbf{t}_{\bar{\mu}}^{\mu^n})\#\bar{\mu}$.

4. Population setting

In order to carry out *inference*, we must relate the sample collection of measures to a population, and show that the relevant quantities are identifiable parameters. Furthermore, in practice the sample measures will only be discretely observed, and this must be taken into account. We now formulate such a model, and study its nonparametric estimation from discrete observations.

4.1. Deformation models and discrete observation

Let λ be a regular probability measure with a strictly positive density on a convex compact $K \subset \mathbb{R}^d$ of positive Lebesgue measure,[3] and let $\{\Pi_1, \dots, \Pi_N\}$ be i.i.d. point processes with intensity measure λ,

$$\mathbb{E}\big[\Pi_i(A)\big] = \lambda(A),$$

for all Borel subsets $A \subseteq K$. Instead of observing the true processes $\{\Pi_1, \dots, \Pi_N\}$, we are able to observe *warped* versions

$$\widetilde{\Pi}_i := T_i \# \Pi_i, \qquad i = 1, \dots, N,$$

with conditional warped mean measures

$$\mathbb{E}[\widetilde{\Pi}_i | T_i] = \mathbb{E}[T_i \# \Pi_i | T_i] = \Lambda_i = T_i \# \lambda,$$

where the $\{T_i : \mathbb{R}^d \to \mathbb{R}^d\}$ are i.i.d. random homeomorphisms on K, satisfying the properties of

1. Unbiasedness: the Fréchet mean of $\Lambda_i = T_i\#\lambda$ is λ.
2. Regularity: T_i is a gradient of a convex function on K.

The conditional mean measures $\{\Lambda_i = T_i\#\lambda\}_{i=1}^N$ play the role of the unobservable sample of random measures generated from a population law constructed via random deformations of the template λ. The processes $\{\widetilde{\Pi}_i\}_{i=1}^N$ play the role of the discretely observed versions of the $\{\Lambda_i\}_{i=1}^N$. Conditions (1) and (2) state that the deformations $\{T_i\}$ are identifiable. They can also be motivated from first principles: (1) states that the maps do not deform the template λ on average (otherwise this "average deformation" would be by definition the template); and (2) states that among all possible deformations that could have mapped λ to Λ_i, we take the parsimonious choice of the optimal deformation. The importance and canonicity of these two assumptions has been discussed in depth in Panaretos and Zemel [58], Section 3.3, who study a one-dimensional

[3] In applied settings, the point processes will be observed on a bounded *observation window* K. For this reason as well as the sake of simplicity, we restrict our discussion to a given compact set (but remark that it could be extended to unbounded observation windows subject to further conditions).

version of the above problems (which is qualitatively very different, given the flat nature of 1d Wasserstein space, and the availability of explicit closed form expressions).

The connection of this deformation model to Fréchet means, via the optimal maps, is now given as follows (in a general setup, encompassing our model setup). Let $C_b(K, \mathbb{R}^d)$ be the space of continuous bounded functions $f : K \to \mathbb{R}^d$ endowed with the supremum norm $\|f\|_\infty = \sup_{x\in K} \|f(x)\|$.

Theorem 5 (Mean identity warp functions and Fréchet means). *Let $K \subset \mathbb{R}^d$ be a compact convex set of positive Lebesgue measure, and let $\lambda \in \mathcal{P}_2(K)$ be regular. Consider the random measure $\Lambda = T\#\lambda$, where $T : K \to K$ is a random deformation (viewed as a random element in $C_b(K, \mathbb{R}^d)$), almost surely injective, and satisfying*

1. *almost surely there exists a convex function ϕ such that $T = \nabla\phi$ on the interior of K;*
2. *$\mathbb{E}[T(x)] = x$ for all $x \in K$ (or on a dense subset of K);*
3. *almost surely T is differentiable with a nonsingular derivative for almost all $x \in K$.*

Then λ is the unique Fréchet mean of Λ, that is, the unique minimiser of the population Fréchet functional $\gamma \mapsto \mathbb{E}d^2(\Lambda, \gamma)$.

An important requirement for the statement and proof of Theorem 5 is that ϕ, ϕ^* and Λ are measurable as random elements in the appropriate spaces; this is not a priori obvious, but is established as part of the proof.

The statistical problem will now be to estimate the unknown structural mean measure λ, and the registration maps T_i non-parametrically, by smoothing the observed point processes $\{\widetilde{\Pi}_1, \dots, \widetilde{\Pi}_N\}$. Once λ and $\{T_i\}$ have been estimated, the processes $\{\widetilde{\Pi}_1, \dots, \widetilde{\Pi}_N\}$ can be registered by applying the inverses of the estimated maps T_i, allowing for further analysis of the point processes in a functional data context. Theorem 5 guarantees that the estimands considered are identifiable.

4.2. Regularised nonparametric estimation

In order to estimate the λ and the $\{\Lambda_i, T_i\}$, we will follow the steps below:

1. *Regularisation*: Estimate $\Lambda_i = T_i\#\lambda$ by a regular kernel estimator $\widehat{\Lambda}_i$ restricted on K,

$$\widehat{\Lambda}_i = \frac{1}{m}\sum_{j=1}^{m} \frac{\delta\{x_j\} * \psi_\sigma}{[\delta\{x_j\} * \psi_\sigma](K)}\Big|_K, \tag{4.1}$$

   where $\psi : \mathbb{R}^d \to (0, \infty)$ is a unit-variance isotropic density function, $\psi_\sigma(x) = \sigma^{-d}\psi(x/\sigma)$ for $\sigma > 0$ (more generally, ψ could be non-isotropic, having a bandwidth matrix, but we focus on the isotropic case for simplicity), and $\widetilde{\Pi}_i$ is the sum of dirac masses $\sum_{j=1}^m \delta\{x_i\}$. If $\widetilde{\Pi}_i$ contains no points (that is, $m = 0$), define $\hat{\Lambda}_i$ to be the (normalised) Lebesgue measure on K.
2. *Fréchet Mean Estimation*: Estimate λ by the empirical Fréchet mean $\hat{\lambda}$ of $\widehat{\Lambda}_1, \dots, \widehat{\Lambda}_N$, using the Procrustes Algorithm 1.

3. *Procrustes Analysis*: Estimate T_i by the optimal transportation map of $\widehat{\lambda}$ onto $\widehat{\Lambda}_i$, as given by the final step in the iteration of Algorithm 1. Estimate the map T_i^{-1} by $\widehat{T_i^{-1}} = \widehat{T}_i^{-1}$.
4. *Registration*: Register the observed point processes to a common coordinate system by defining $\widehat{\Pi}_i = \widehat{T_i^{-1}} \# \widetilde{\Pi}_i$.

In the next section, we will prove that our estimates are consistent for their population version, as the number of observed processes, and the number of points per process diverge.

4.3. Asymptotic theory

To establish consistency, we will use the *dense* asymptotics regime of functional data analysis, adapted to the current setting. We will consider a setup where the number of observed point processes n diverges, and the (mean) number of points in each observed process, $\mathbb{E}[\widetilde{\Pi}_i(K)]$, diverge too. Here we use the index notation "n" rather than "N" to emphasize that the index is no longer held fixed. Specifically, let $(\Pi_1^{(n)}, \Pi_2^{(n)}, \dots, \Pi_n^{(n)})_{n=1}^{\infty}$ be a triangular array of row-independent and identically distributed point processes on K following the same infinitely divisible distribution and having mean measure $\tau_n \lambda$, where $\tau_n > 0$ are constants. Let $T_1, \dots, T_n$ be independent and identically distributed realisations of a random homeomorphism T of K satisfying the unbiasedness and regularity assumptions of Section 4.1. Let $\widetilde{\Pi}_i^{(n)} = T_i \# \Pi_i^{(n)}$ and set $\Lambda_i = T_i \# \lambda = \tau_n^{-1} \mathbb{E}[\widetilde{\Pi}_i^{(n)} | T_i]$. Suppose that $\widehat{\Lambda}_i$ is an estimator of Λ_i, constructed by kernel smoothing of $\Pi_i^{(n)}$ using a (possibly random) bandwidth $\sigma_i^{(n)}$, as described in the previous section. Correspondingly, let $\widetilde{\Pi}_i^{(N)} = T_i \# \Pi_i^{(n)}$ and set $\Lambda_i = T_i \# \lambda = \tau_n^{-1} \mathbb{E}[\widetilde{\Pi}_i^{(n)} | T_i]$.

Theorem 6 (Consistency of the regularised Fréchet mean). *If $\tau_n / \log n \to \infty$ and $\sigma_n = \max_i \sigma_i^{(n)} \xrightarrow{\mathrm{p}} 0$ then*

1. *For any i,*

$$d(\widehat{\Lambda}_i, \Lambda_i) \xrightarrow{\mathrm{p}} 0;$$

2. *The estimator $\widehat{\lambda}_n$ is strongly consistent*

$$d(\widehat{\lambda}_n, \lambda) \xrightarrow{\mathrm{as}} 0.$$

If the smoothing is carried out independently across trains, that is, $\sigma_i^{(n)}$ depends only on $\widetilde{\Pi}_i^{(n)}$, then the result still holds if merely $\tau_n \to \infty$.

If $\mathbb{E}[\Pi_1^{(1)}]^4 < \infty$, $\sum_n \tau_n^{-2} < \infty$ and $\sigma_n \xrightarrow{\mathrm{as}} 0$ then convergence almost surely holds.

Remark 3. There is no lower bound on σ_n, and it can vanish at any rate, provided it is strictly positive. In practice, however, if σ_n is very small, then the densities of $\widehat{\Lambda}_i$ will have very high peaks, and the constant C_μ in Proposition 4 (with $\mu^i = \widehat{\Lambda}_i$) will be large (essentially proportional to $1/\sigma_n$). The proof of Proposition 3 suggests that this may slow down the convergence of Algorithm 1.

Remark 4. It is worth remarking that Le Gouic and Loubes [51], Theorem 3, consider the stability of Fréchet means in a rather general setting; verification of their assumptions in our particular setting, however, is quite involved and in fact essentially amounts to directly proving Theorem 6.

Our next two results concern the (uniform) consistency of the Procrustes registration procedure. Though the results themselves parallel their one-dimensional counterparts (see Panaretos and Zemel [58]), their proofs are entirely different, and substantially more involved (because the geometry of monotone mappings in $\mathbb{R}^d$ is far more rich than the geometry of monotone maps on $\mathbb{R}$). In particular, we have the following theorem.

Theorem 7 (Consistency of Procrustes maps). *Under the same conditions of Theorem 6, for any i and any compact set $\Omega \subseteq \operatorname{int}(K)$,*

$$\sup_{x\in\Omega} \big\| \widehat{T}_i^{-1}(x) - T_i^{-1}(x) \big\| \xrightarrow{\text{p}} 0, \qquad \sup_{x\in\Omega} \big\| \widehat{T}_i(x) - T_i(x) \big\| \xrightarrow{\text{p}} 0.$$

The same remarks at the end of the statement of Theorem 6 apply here as well.

Corollary 4 (Consistency of Procrustes Registration). *Under the same conditions of Theorem 6, the registration procedure is consistent: for any i*

$$d\left(\frac{\widehat{\Pi}_i}{\widehat{\Pi}_i(K)}, \frac{\Pi_i}{\Pi_i(K)} \right) \xrightarrow{\text{p}} 0, \qquad n \to \infty,$$

provided one of the following conditions holds:

1. *Every point of the boundary of K is exposed, that is, for any $y \in \partial K$ there exists $\alpha \in \mathbb{R}^d$ such that*

$$\langle y, \alpha \rangle > \langle y', \alpha \rangle, \qquad y' \in K \setminus \{y\}.$$

2. *The warp map T_i is* strictly monotone

$$\langle T_i(x') - T_i(x), x' - x \rangle > 0, \qquad x, x' \in \operatorname{int}(K), x \neq x'.$$

The first condition is satisfied by any ellipsoid in $\mathbb{R}^d$ and more generally if the boundary of K can be written as $\partial K = \{x : \varphi_K(x) = 0\}$, for a strictly convex function φ_K. Indeed, if α creates a supporting hyperplane to K at y and $\langle \alpha, y \rangle = \langle \alpha, y' \rangle$ for $y \neq y'$, then as φ_K is strictly convex on the line segment $[y, y']$, it is impossible that $y' \in K$ without the hyperplane intersecting the interior of K. Although this condition excludes some interesting cases, perhaps most prominently polyhedral sets such as $K = [0, 1]^d$, such sets can be approximated by convex sets that do satisfy it (Krantz [47], Proposition 1.12).

As for the second condition, in general it will hold almost surely. Indeed, as $T_i \# \lambda = \Lambda_i$ and both measures are absolutely continuous, there exists a λ-null set $\mathcal{N}$ such that T_i is strictly monotone outside $\mathcal{N}$ [7], Proposition 6.2.12. By assumption λ has a strictly positive density on K, so that λ-null subsets of K are precisely the Lebesgue null subsets of K. In that sense, this condition is not overly restrictive, and will most likely be satisfied under additional regularity assumptions on the warp maps T_i and, possibly, K.

5. Proofs of formal statements

Our proofs will require us to establish some analytical results that are intrinsic to the optimal transportation problem. These are essential for the proofs, especially of our main results, and some are nontrivial. For tidiness, we will state and prove these results separately at the end of this section (Section 5.5), developing our main results first, and referring to the analytical background when necessary.

5.1. Proofs of statements in Section 3.1

Proof of Proposition 2. The optimisation problem

$$\min_{Y_i\sim\mu^i}\mathbb{E}\sum_{i=1}^N\sum_{j=i+1}^N\|Y_i-Y_j\|^2=\min_{\xi\in\Gamma(\mu^1,\dots,\mu^N)}\int_{\mathbb{R}^{Nd}}\sum_{i=1}^N\sum_{j=i+1}^N\|t_i-t_j\|^2\,\mathrm{d}\xi(t_1,\dots,t_N)$$

is equivalent to minimising

$$G(\xi)=\frac{1}{2N}\int_{\mathbb{R}^{Nd}}\sum_{i=1}^N\left\|t_i-\frac{1}{N}\sum_{j=1}^N t_j\right\|^2\mathrm{d}\xi(t_1,\dots,t_N),\qquad \xi\in\Gamma\bigl(\mu^1,\dots,\mu^N\bigr),$$

and Agueh and Carlier [2], Proposition 4.2, show that $\min_\mu F(\mu)=\min_\xi G(\xi)$.

Since $\bar{\mu}$ is regular ([2], Proposition 5.1), $\boldsymbol{X}$ is well defined and has joint distribution

$$\xi'=h\#\bar{\mu},\qquad h:\mathbb{R}^d\to\mathbb{R}^{Nd},\qquad h=\bigl(\mathbf{t}_{\bar{\mu}}^{\mu^1},\dots,\mathbf{t}_{\bar{\mu}}^{\mu^N}\bigr).$$

Since the coordinates of h have mean identity (see [2] or Corollary 1, Equation (3.9)),

$$G\bigl(\xi'\bigr)=\frac{1}{2N}\int_{\mathbb{R}^d}\sum_{i=1}^N\bigl\|\mathbf{t}_{\bar{\mu}}^{\mu^i}-\mathbf{i}\bigr\|^2\,\mathrm{d}\bar{\mu}=\frac{1}{2N}\sum_{i=1}^N d^2\bigl(\bar{\mu},\mu^i\bigr)=F(\bar{\mu})=\inf_\mu F(\mu).$$

Thus ξ' is optimal. □

5.2. Proofs of statements in Section 3.2

Proof of Corollary 1. The characterisation of Karcher means is immediate from Theorem 1. The fact that the Fréchet mean μ satisfies $\sum_{i=1}^N(\mathbf{t}_\mu^{\mu^i}-\mathbf{i})=0$ μ-almost everywhere follows by a result of Agueh and Carlier [2]. For an alternative proof using the tangent bundle, see the supplementary material [72]. □

Proof of Theorem 2. The result exploits Caffarelli's regularity theory for Monge–Ampère equations. In the first case, by Theorem 4.14(iii) in Villani [70] there exist C^1 (in fact, $C^{2,\alpha}$) convex potentials φ_i on $\mathbb{R}^d$ with $\mathbf{t}_\mu^{\mu^i}=\nabla\varphi_i$, so that $\mathbf{t}_\mu^{\mu^i}(x)$ is a singleton for all $x\in\mathbb{R}^d$. The set

$\{x \in \mathbb{R}^d : \sum \mathbf{t}_\mu^{\mu^i}(x)/N \neq x\}$ is μ-negligible (and hence Lebesgue-negligible) and open by continuity. It is therefore empty, so $F'(\mu) = 0$ everywhere, and μ is the Fréchet mean (see the discussion after Corollary 1).

In the second case, by the main theorem in Caffarelli [21], p. 99, and the same argument, we have $\sum \mathbf{t}_\mu^{\mu^i}(x)/N = x$ for all $x \in X$. Since X is convex, there must exist a constant C such that $\sum \varphi_i(x) = C + N\|x\|^2/2$ for all $x \in X$. Hence, Equation (3.9) in [2] holds with $\mathbb{R}^d$ replaced by X. Repeating the proof of Proposition 3.8 in [2], we see that μ minimises F on $\mathcal{P}_2(X)$, the set of measures supported on X. (All the integrals that appear in the proof can be taken on X, where we know the inequality holds.) Again by convexity of X, the minimiser of F must be[4] in $\mathcal{P}_2(X)$ (see the existence proof at the beginning of the proof of Theorem 5 in the supplementary material [72]). □

5.3. Proofs of statements in Section 3.3

Proof of Lemma 1. By [7], Proposition 6.2.12, there exists a γ_0-null set A_i such that on $\mathbb{R}^d \setminus A_i$, $\mathbf{t}_{\gamma_0}^{\mu^i}$ is differentiable, $\nabla \mathbf{t}_{\gamma_0}^{\mu^i} > 0$ (positive definite), and $\mathbf{t}_{\gamma_0}^{\mu^i}$ is strictly monotone

$$\left\langle \mathbf{t}_{\gamma_0}^{\mu^i}(x) - \mathbf{t}_{\gamma_0}^{\mu^i}(x'), x - x' \right\rangle > 0, \qquad x, x' \notin A_i, x \neq x'.$$

Since $\mathbf{t}_{\gamma_0}^{\gamma_1} = (1-\tau)\mathbf{i} + \tau N^{-1} \sum_{i=1}^N \mathbf{t}_{\gamma_0}^{\mu^i}$, it stays strictly monotone (hence injective) and $\nabla \mathbf{t}_{\gamma_0}^{\gamma_1} > 0$ outside $A = \bigcup A_i$, which is a γ_0-null set.

Let h_0 denote the density of γ_0 and set $\Sigma = \mathbb{R}^d \setminus A$. Then $\mathbf{t}_{\gamma_0}^{\gamma_1}|_\Sigma$ is injective and $\{h_0 > 0\} \setminus \Sigma$ is Lebesgue negligible because

$$0 = \gamma_0(A) = \gamma_0(\mathbb{R}^d \setminus \Sigma) = \int_{\mathbb{R}^d \setminus \Sigma} h_0(x)\,\mathrm{d}x = \int_{\{h_0 > 0\} \setminus \Sigma} h_0(x)\,\mathrm{d}x,$$

and the integrand is strictly positive. Since $|\det \nabla \mathbf{t}_{\gamma_0}^{\mu^i}| > 0$ on Σ we obtain that $\gamma_1 = \mathbf{t}_{\gamma_0}^{\mu^i} \# \gamma_0$ is absolutely continuous by [7], Lemma 5.5.3. □

Proof of Lemma 2. Let $S_i = \mathbf{t}_{\gamma_0}^{\mu^i}$ be the optimal map from γ_0 to μ^i, and set $W_i = S_i - \mathbf{i}$. Then

$$2NF(\gamma_0) = \sum_{i=1}^N d^2(\gamma_0, \mu^i) = \sum_{i=1}^N \int_{\mathbb{R}^d} \|S_i - \mathbf{i}\|^2 \,\mathrm{d}\gamma_0 = \sum_{i=1}^N \langle W_i, W_i \rangle = \sum_{i=1}^N \|W_i\|^2, \tag{5.1}$$

with the inner product being in $L^2(\gamma_0)$. By definition

$$\gamma_1 = \left[(1-\tau)\mathbf{i} + \frac{\tau}{N} \sum_{j=1}^N S_j\right] \# \gamma_0 = \left[(1-\tau)S_i^{-1} + \frac{\tau}{N} \sum_{j=1}^N S_j \circ S_i^{-1}\right] \# \mu^i.$$

[4] We know that the minimiser must be in $\mathcal{P}_2(\overline{X})$, but minimising on $\mathcal{P}_2(X)$ suffices by continuity of F.

This is a map that pushes forward μ^i to γ_1 (not necessarily optimally). Hence,

$$d^2(\gamma_1, \mu^i) \le \int_{\mathbb{R}^d} \left\| \left[(1-\tau) S_i^{-1} + \frac{\tau}{N} \sum_{j=1}^N S_j \circ S_i^{-1} \right] - \mathbf{i} \right\|_{\mathbb{R}^d}^2 \mathrm{d}\mu^i.$$

Now $\mu^i = S_i \# \gamma_0$, which means that $\int f \, \mathrm{d}\mu^i = \int (f \circ S_i) \, \mathrm{d}\gamma_0$ for any measurable f. This change of variables gives

$$d^2(\gamma_1, \mu^i) \le \int_{\mathbb{R}^d} \left\| \left[(1-\tau)\mathbf{i} + \frac{\tau}{N} \sum_{j=1}^N S_j \right] - S_i \right\|_{\mathbb{R}^d}^2 \mathrm{d}\gamma_0 = \left\| -W_i + \frac{\tau}{N} \sum_{j=1}^N W_j \right\|_{L^2(\gamma_0)}^2 .$$

The norm is always in $L^2(\gamma_0)$, regardless of i. Developing the squares, summing over $i = 1, \dots, N$ and using (5.1) gives

$$\begin{aligned} 2NF(\gamma_1) &\le \sum_{i=1}^N \|W_i\|^2 - 2\frac{\tau}{N} \sum_{i,j=1}^N \langle W_i, W_j \rangle + \frac{\tau^2}{N^2} \sum_{i,j,k=1}^N \langle W_j, W_k \rangle \\ &= 2NF(\gamma_0) - 2N\tau \left\| \sum_{i=1}^N \frac{1}{N} W_i \right\|^2 + N\tau^2 \left\| \sum_{i=1}^N \frac{1}{N} W_i \right\|^2, \end{aligned}$$

and recalling that $W_i = S_i - \mathbf{i}$ yields

$$F(\gamma_1) - F(\gamma_0) \le \frac{\tau^2 - 2\tau}{2} \left\| \frac{1}{N} \sum_{i=1}^N W_i \right\|^2 = -\|F'(\gamma_0)\|^2 \left[\tau - \frac{\tau^2}{2} \right].$$

Since $\tau - \tau^2/2$ is clearly maximised at $\tau = 1$, the proof is complete. □

5.3.1. *Proof of Theorem* 3

We will prove the theorem by establishing the following facts:

1. The sequence $\|F'(\gamma_j)\|$ converge to zero as $j \to \infty$.
2. The sequence $\{\gamma_j\}$ is stays in a compact subset of $\mathcal{P}_2(\mathbb{R}^d)$.
3. The mapping $\gamma \mapsto \|F'(\gamma)\|^2$ is continuous.

The first two are relatively straightforward, and are proven in the form of the following two lemmas.

Lemma 3. *The objective value of the Fréchet functional decreases at each step of Algorithm* 1, *and* $\|F'(\gamma_j)\|$ *vanishes as* $j \to \infty$.

Proof. The first statement is clear from Lemma 2, from which it also follows that

$$\frac{1}{2}\sum_{j=0}^{k}\left\|F'(\gamma_j)\right\|^2 \le \sum_{j=0}^{k} F(\gamma_j) - F(\gamma_{j+1}) = F(\gamma_0) - F(\gamma_{k+1}) \le F(\gamma_0).$$

Consequently, the series at the left-hand side converges whence $\|F'(\gamma_j)\|^2 \to 0$. □

Lemma 4. *The sequence generated by Algorithm 1 stays in a compact subset of the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$.*

Proof. For any $\varepsilon > 0$ there exists a compact convex set K_ε such that $\mu^i(K_\varepsilon) > 1 - \varepsilon/N$ for $i = 1, \dots, N$. Let $A^i = (\mathbf{t}_{\gamma_j}^{\mu^i})^{-1}(K_\varepsilon)$, $A = \bigcap_{i=1}^N A^i$. Then $\gamma_j(A^i) > 1 - \varepsilon/N$, so that $\gamma_j(A) > 1 - \varepsilon$. Since K_ε is convex, $T_j(x) \in K_\varepsilon$ for any $x \in A$, so that

$$\gamma_{j+1}(K_\varepsilon) = \gamma_j\bigl(T_j^{-1}(K_\varepsilon)\bigr) \ge \gamma_j(A) > 1 - \varepsilon, \qquad j = 0, 1, \dots.$$

We shall now show that any weakly convergent subsequence of $\{\gamma_j\}$ is in fact convergent in the Wasserstein space. By Theorem 7.12 in Villani [70], it suffices to show that

$$\lim_{R\to\infty} \sup_{j\in\mathbb{N}} \int_{\{x:\|x\|>R\}} \|x\|^2 \, \mathrm{d}\gamma_j(x) = 0. \tag{5.2}$$

For simplicity, we shall show this under the stronger assumption that the measures $\mu^1, \dots, \mu^N$ have a finite third moment

$$\int_{\mathbb{R}^d} \|x\|^3 \, \mathrm{d}\mu^i(x) \le M(3), \qquad i = 1, \dots, N. \tag{5.3}$$

In Section 2 of the supplementary material [72], we show that (5.2) holds even if (5.3) does not.

For any $j \ge 1$, it holds that

$$\begin{aligned}\int_{\mathbb{R}^d} \|x\|^3 \, \mathrm{d}\gamma_j(x) &= \int_{\mathbb{R}^d} \left\| \frac{1}{N}\sum_{i=1}^N \mathbf{t}_{\gamma_{j-1}}^{\mu^i}(x)\right\|^3 \mathrm{d}\gamma_{j-1}(x) \le \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}^d} \left\|\mathbf{t}_{\gamma_{j-1}}^{\mu^i}(x)\right\|^3 \mathrm{d}\gamma_{j-1}(x) \\ &= \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}^d} \|x\|^3 \, \mathrm{d}\mu^i(x) \le M(3).\end{aligned}$$

This implies that for any $R > 0$ and any $j > 0$,

$$\int_{\{x:\|x\|>R\}} \|x\|^2 \, \mathrm{d}\gamma_j(x) \le \frac{1}{R}\int_{\{x:\|x\|>R\}} \|x\|^3 \, \mathrm{d}\gamma_j(x) \le \frac{1}{R} M(3),$$

and (5.2) follows. □

The third statement (continuity of the gradient) is much more subtle to establish. We will prove it in two steps: first we establish a proposition, giving sufficient conditions for the third statement to hold true. Then, we will verify that the conditions of the proposition are satisfied in the setting of Theorem 3.3, in the form of a lemma and a corollary. We start with the proposition.

Proposition 3 (Continuity of F'). *Let $\mu^1, \dots, \mu^N \in \mathcal{P}_2(\mathbb{R}^d)$ be given regular measures, and consider a sequence γ_n of regular measures that converges in $\mathcal{P}_2(\mathbb{R}^d)$ to a regular measure γ. If the densities of γ_n are uniformly bounded, then $\|F'(\gamma_n)\|^2 \to \|F'(\gamma)\|^2$.*

Proof. The regularity of γ_n and γ implies that F is indeed differentiable there, and so it needs to be shown that

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{t}_{\gamma_n}^{\mu^i} - \mathbf{i} \right\|_{L^2(\gamma_n)}^2 \longrightarrow \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{t}_{\gamma}^{\mu^i} - \mathbf{i} \right\|_{L^2(\gamma)}^2, \qquad n \to \infty.$$

Denote the integrands by g_n and g respectively. At a given $x \in \mathbb{R}^d$, $g_n(x)$ can be undefined, either because some $\mathbf{t}_{\gamma_n}^{\mu^i}(x)$ is empty, or because they can be multivalued. Redefine $g_n(x)$ at such points by setting it to 0 in the former case and choosing an arbitrary representative otherwise. Since the set of these ambiguity points is a γ_n-null set (because γ_n is absolutely continuous), this modification does not affect the value of the integral $\int g_n \, \mathrm{d}\gamma_n$. Apply the same procedure to g. Then g_n and g are finite and nonnegative throughout $\mathbb{R}^d$. Absolute continuity of γ, Remark 2.3 in [3] and Proposition 5 imply together that the set of points where g is not continuous is a γ-null set.

Next, we approximate g_n and g by bounded functions as follows. Since γ_n converge in the Wasserstein space, they satisfy (5.2) by [70], Theorem 7.12. It is easy to see that this implies the uniform absolute continuity

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall j \ge 1 \; \forall A \subseteq \mathbb{R}^d \text{ Borel}: \qquad \gamma_j(A) \le \delta \quad \Longrightarrow \quad \int_A \|x\|^2 \, \mathrm{d}\gamma_j(x) < \varepsilon. \tag{5.4}$$

The δ's can be chosen in such a way that (5.4) holds true for the finite collection $\{\mu^1, \dots, \mu^N\}$ as well. Fix $\varepsilon > 0$, set $\delta = \delta_\varepsilon$ as in (5.4), and let $A_n = \{x : g_n(x) \ge 4R\}$, where $R = R_\varepsilon \ge 1$ is such that (using (5.2))

$$\forall i \; \forall n: \qquad \int_{\{\|x\|^2 > R\}} \|x\|^2 \, \mathrm{d}\gamma_n(x) + \int_{\{\|x\|^2 > R\}} \|x\|^2 \, \mathrm{d}\mu^i(x) < \frac{\delta}{2N}.$$

The bound

$$g_n(x) \le 2\|x\|^2 + \frac{2}{N} \sum_{i=1}^{N} \left\| \mathbf{t}_{\gamma_n}^{\mu^i}(x) \right\|^2,$$

implies that

$$A_n \subseteq \bigl\{x : \|x\|^2 > R\bigr\} \cup \bigcup_{i=1}^{N} \bigl\{x : \bigl\|\mathbf{t}_{\gamma_n}^{\mu^i}(x)\bigr\|^2 > R\bigr\}.$$

To deal with the sets in the union observe that (since $\mathbf{t}_{\gamma_n}^{\mu^i}$ is γ_n-almost surely injective),

$$\gamma_n\bigl(\bigl\{x : \bigl\|\mathbf{t}_{\gamma_n}^{\mu^i}(x)\bigr\|^2 > R\bigr\}\bigr) = \mu^i\bigl(\bigl\{x : \|x\|^2 > R\bigr\}\bigr) < \frac{\delta}{2N},$$

so that $\gamma_n(A_n) < \delta$. We use this in conjunction with (5.4) to bound

$$\begin{aligned}\int_{A_n} g_n(x)\,\mathrm{d}\gamma_n(x) &\le 2\int_{A_n} \|x\|^2\,\mathrm{d}\gamma_n(x) + \frac{2}{N}\sum_{i=1}^{N}\int_{A_n}\bigl\|\mathbf{t}_{\gamma_n}^{\mu^i}(x)\bigr\|^2\,\mathrm{d}\gamma_n(x)\\ &\le 2\varepsilon + \frac{2}{N}\sum_{i=1}^{N}\int_{\mathbf{t}_{\gamma_n}^{\mu^i}(A_n)} \|x\|^2\,\mathrm{d}\mu^i(x) \le 4\varepsilon,\end{aligned}$$

where we have used the measure-preservation property $\mu^i(\mathbf{t}_{\gamma_n}^{\mu^i}(A_n)) = \gamma_n(A_n) < \delta$.

Define the truncation $g_{n,R}(x) = \min(g_n(x), 4R)$. Then $0 \le g_n - g_{n,R} \le g_n \mathbf{1}\{g_n > 4R\}$, so

$$\int \bigl[g_n(x) - g_{n,R}(x)\bigr]\,\mathrm{d}\gamma_n(x) \le \int_{A_n} g_n(x)\,\mathrm{d}\gamma_n(x) \le 4\varepsilon, \qquad n = 1, 2, \ldots.$$

The analogous truncated function g_R satisfies

$$0 \le g_R(x) \le 4R \qquad \forall x \in \mathbb{R}^d \quad \text{and} \quad \{x : g_R \text{ is continuous}\} \text{ is of } \gamma\text{-full measure.} \tag{5.5}$$

Let $E = \operatorname{supp}(\gamma)$. Proposition 6 (Section 5.5) implies pointwise convergence of $\mathbf{t}_{\gamma_n}^{\mu^i}(x)$ to $\mathbf{t}_{\gamma}^{\mu^i}(x)$ for any $i = 1, \ldots, N$ and any $x \in E \setminus \mathcal{N}$, where $\mathcal{N} = \bigcup_{i=1}^{N} \mathcal{N}^i$ and

$$\mathcal{N}^i = \bigl(E \setminus E^{\mathrm{den}}\bigr) \cup \bigl\{x : \mathbf{t}_{\gamma}^{\mu^i}(x) \text{ contains more than one element}\bigr\}.$$

Thus, g_n and g are univalued functions defined throughout $\mathbb{R}^d$, and $g_n \to g$ pointwise on $x \in E \setminus \mathcal{N}$ (for whatever choice of representatives selected to define g_n); consequently, $g_{n,R} \to g_R$ on $E \setminus \mathcal{N}$.

In order to restrict the integrands to a bounded set we invoke the tightness of the sequence (γ_n) and introduce a compact set K_ε such that $\gamma_n(\mathbb{R}^d \setminus K_\varepsilon) < \varepsilon/R$ for all n. Clearly, $g_{n,R} \to g_R$ on $E' = K_\varepsilon \cap E \setminus \mathcal{N}$, and by Egorov's theorem (valid as $\mathrm{Leb}(E') \le \mathrm{Leb}(K_\varepsilon) < \infty$), there exists a Borel set $\Omega = \Omega_\varepsilon \subseteq E'$ on which the convergence is uniform, and $\mathrm{Leb}(E' \setminus \Omega) < \varepsilon/R$. Let us write

$$\int g_{n,R}\,\mathrm{d}\gamma_n - \int g_R\,\mathrm{d}\gamma = \int g_R\,\mathrm{d}(\gamma_n - \gamma) + \int_{\Omega}(g_{n,R} - g_R)\,\mathrm{d}\gamma_n + \int_{\mathbb{R}^d\setminus\Omega}(g_{n,R} - g_R)\,\mathrm{d}\gamma_n,$$

and bound each of the three integrals at the right-hand side as $n \to \infty$.

The first integral vanishes as $n \to \infty$, by (5.5) and the Portmanteau lemma (Lemma 9, Section 5.5). For a given Ω, the second integral vanishes as $n \to \infty$, since $g_{n,R}$ converge to g_R uniformly. The third integral is bounded by $8R\gamma_n(\mathbb{R}^d \setminus \Omega)$. The latter set is a subset of $\mathcal{N} \cup (E' \setminus \Omega) \cup (\mathbb{R}^d \setminus E) \cup (\mathbb{R}^d \setminus K_\varepsilon)$, where the first set is Lebesgue-negligible and the second has Lebesgue measure smaller than ε/R. The hypothesis of the densities of γ_n implies that $\gamma_n(A) \le C\,\mathrm{Leb}(A)$ for any Borel set $A \subseteq \mathbb{R}^d$ and any $n \in \mathbb{N}$; it follows from this and $\gamma_n(\mathbb{R}^d \setminus K_\varepsilon) < \varepsilon/R$ that

$$\left|\int_{\mathbb{R}^d\setminus\Omega} (g_{n,R} - g_R)\,\mathrm{d}\gamma_n\right| \le 8R\bigl(C\varepsilon/R + \gamma_n\bigl(\mathbb{R}^d \setminus E\bigr) + \varepsilon/R\bigr) = 8\bigl(R\gamma_n\bigl(\mathbb{R}^d \setminus E\bigr) + C\varepsilon + \varepsilon\bigr).$$

Write the open set $E_1 = \mathbb{R}^d \setminus E$ as a countable union of closed sets A_k with $\mathrm{Leb}(E_1 \setminus A_k) < 1/k$, and conclude that

$$\limsup_{n\to\infty} \gamma_n(E_1) \le \limsup_{n\to\infty} \gamma_n(A_k) + \limsup_{n\to\infty} \gamma_n(E_1 \setminus A_k) \le \gamma(A_k) + \frac{C}{k} = \frac{C}{k},$$

where we have used the Portmanteau lemma again, $A_k \cap \mathrm{supp}(\gamma) = \varnothing$ and $\gamma_n(A) \le C\,\mathrm{Leb}(A)$. Consequently, for all k

$$\limsup_{n\to\infty}\left|\int g_{n,R}\,\mathrm{d}\gamma_n - \int g_R\,\mathrm{d}\gamma\right| \le \limsup_{n\to\infty}\left|\int_{\mathbb{R}^d\setminus\Omega} (g_{n,R} - g_R)\,\mathrm{d}\gamma_n\right| \le \frac{8R_\varepsilon C}{k} + 8(C+1)\varepsilon.$$

Letting $k \to \infty$, then incorporating the truncation error yields

$$\limsup_{n\to\infty}\left|\int g_n\,\mathrm{d}\gamma_n - \int g\,\mathrm{d}\gamma\right| \le 8(C+1)\varepsilon + 8\varepsilon.$$

The proof is complete upon noticing that ε is arbitrary. □

Our proof will now be complete if we show that the sequence γ_k generated by the algorithm satisfies the assumptions of the last proposition. First, we show that limits of the sequence are indeed regular.

Proposition 4 (Sequence has bounded density). *Let μ^i have density g^i for $i = 1, \dots, N$ and let γ_0 be a regular probability measure. Then the density of γ_1 is bounded by a constant $C_\mu = \min\{N^{d-1}\max_i \|g^i\|_\infty, N^d \min_i \|g^i\|_\infty\}$ that depends only on $\{\mu^1, \dots, \mu^N\}$.*

Proof. Let h_i be the density of γ_i. By the change of variables formula, for γ_0-almost any x

$$h_1\bigl(\mathbf{t}_{\gamma_0}^{\gamma_1}(x)\bigr) = \frac{h_0(x)}{\det \nabla \mathbf{t}_{\gamma_0}^{\gamma_1}(x)}; \qquad g^i\bigl(\mathbf{t}_{\gamma_0}^{\mu^i}(x)\bigr) = \frac{h_0(x)}{\det \nabla \mathbf{t}_{\gamma_0}^{\mu^i}(x)}.$$

Fiedler [30] shows that if B_1 and B_2 are $d \times d$ positive semidefinite matrices with eigenvalues $0 \le \alpha_i, \beta_i$, then

$$\det(B_1 + B_2) \ge \prod_{i=1}^{d} (\alpha_i + \beta_i).$$

The right-hand side contains 2^d nonnegative summands of which two are $\det B_1$ and $\det B_2$, and so we see that $\det(B_1 + B_2) \geq \det B_1 + \det B_2$. (One can show the stronger result $\sqrt[d]{\det(B_1 + B_2)} \geq \sqrt[d]{\det B_1} + \sqrt[d]{\det B_2}$.) Since $\nabla \mathbf{t}_{\gamma_0}^{\gamma_1}$ is an average of N $d \times d$ positive semidefinite matrices, we obtain

$$h_1\bigl(\mathbf{t}_{\gamma_0}^{\gamma_1}(x)\bigr) = \frac{N^d h_0(x)}{\det \sum \nabla \mathbf{t}_{\gamma_0}^{\mu^i}(x)} \leq \frac{N^d h_0(x)}{\sum \det \nabla \mathbf{t}_{\gamma_0}^{\mu^i}(x)} = N^d \left[\sum_{i=1}^N \frac{1}{g^i(\mathbf{t}_{\gamma_0}^{\mu^i}(x))}\right]^{-1}$$
$$\leq N^d \left[\sum_{i=1}^N \frac{1}{\|g^i\|_\infty}\right]^{-1}.$$

Let Σ be the set of points where this inequality holds; then $\gamma_0(\Sigma) = 1$. Hence,

$$\gamma_1\bigl(\mathbf{t}_{\gamma_0}^{\gamma_1}(\Sigma)\bigr) = \gamma_0\bigl[\bigl(\mathbf{t}_{\gamma_0}^{\gamma_1}\bigr)^{-1}\bigl(\mathbf{t}_{\gamma_0}^{\gamma_1}(\Sigma)\bigr)\bigr] \geq \gamma_0(\Sigma) = 1.$$

Thus γ_1-almost surely,

$$h_1 \leq N^d \left[\sum_{i=1}^N \frac{1}{\|g^i\|_\infty}\right]^{-1} \leq \min\Bigl\{N^{d-1} \max_i \|g^i\|_\infty, N^d \min_i \|g^i\|_\infty\Bigr\} = C_\mu.$$

For C_μ to be finite it suffices that $\|g^i\|_\infty$ be finite for some i. □

Our task is now essentially complete. All that remains is to show:

Corollary 5 (Limits are regular). *Every limit of the sequence generated by Algorithm 1 is absolutely continuous provided the density of μ^i is bounded for some i.*

Proof. Each γ_k ($k = 1, 2, \dots$) has a density that is bounded by the finite constant C_μ. For any open set O, $\liminf \gamma_k(O) \leq C_\mu \operatorname{Leb}(O)$, so any limit point γ of (γ_k) is such that $\gamma(O) \leq C_\mu \operatorname{Leb}(O)$ by the Portmanteau lemma. It follows that γ is absolutely continuous with density bounded by C_μ. We note that Agueh and Carlier [2] show that the density of the Fréchet mean is bounded by $N^d \min_i \|g^i\|_\infty \geq C_\mu$, a slightly weaker bound. □

Proof of Theorem 4. Let $E = \operatorname{supp}(\bar{\mu})$ and set $A^i = E^{\mathrm{den}} \cap \{x : \mathbf{t}_{\bar{\mu}}^{\mu^i}(x) \text{ is multivalued}\}$. By Corollary 6 $\bar{\mu}(A^i) = 1$. Choose $A = \bigcap_{i=1}^N A^i$ and apply Proposition 6. This proves the first assertion.

Now let $E^i = \operatorname{supp}(\mu^i)$ and set $B^i = (E^i)^{\mathrm{den}} \cap \{x : \mathbf{t}_{\mu^i}^{\bar{\mu}}(x) \text{ is univalued}\}$. Since μ^i is regular, $\mu^i(B^i) = 1$. Apply Proposition 6. If in addition $E^1 = \cdots = E^N$, then $\mu^i(B) = 1$ for $B = \cap B^i$. □

Proof of Corollary 3. The proof is very similar to that of Proposition 3. Define $\eta_j, \eta \in \mathcal{P}_2((\mathbb{R}^d)^{N+1})$ by

$$\eta_j = \bigl(\mathbf{t}_{\gamma_j}^{\mu^1}, \dots, \mathbf{t}_{\gamma_j}^{\mu^n}, \mathbf{i}\bigr)\#\gamma_j, \qquad \eta = \bigl(\mathbf{t}_{\gamma}^{\mu^1}, \dots, \mathbf{t}_{\gamma}^{\mu^n}, \mathbf{i}\bigr)\#\gamma.$$

We establish convergence of η_j to η. Since the optimal multicouplings are marginals of η_j and η their convergence follow. Let $h : (\mathbb{R}^d)^{N+1} \to \mathbb{R}$ be any continuous function such that

$$\left|h(t_1, \dots, t_N, y)\right| \le \frac{1}{N}\sum_{i=1}^{N} \|t_i\|^2 + \|y\|^2.$$

Define $g_j : \mathbb{R}^d \to \mathbb{R}$ by $g_j(x) = h(\mathbf{t}_{\gamma_j}^{\mu^1}, \dots, \mathbf{t}_{\gamma_j}^{\mu^n}, x)$ and analogously define g. By [70], Theorem 7.12, it suffices to show that (if this holds for h, it also holds for $a + bh$ with a, b scalars)

$$\int_{\mathbb{R}^{d(N+1)}} h \, \mathrm{d}\eta_j = \int_{\mathbb{R}^d} g_j \, \mathrm{d}\gamma_j(x) \quad \to \quad \int_{\mathbb{R}^d} g \, \mathrm{d}\gamma(x) = \int_{\mathbb{R}^{dN}} h \, \mathrm{d}\eta.$$

(In Proposition 3, we had $h = \|y - \bar{t}\|^2$.) Since h is continuous, we can modify g_n and g to be well-defined, finite and so that g be continuous γ-almost surely. Define R as in the proof of Proposition 3, $A_j = \{x : |g_j(x)| \ge 4R\}$, invoke (5.4) and translate the bound on h to a bound on $|g_j|$ to conclude that $\int_{A_j} |g_j| \, \mathrm{d}\gamma_j \le 4\varepsilon$. Carry out the same (now two-sided) truncation $g_{j,R}(x) = \max(-4R, \min(g_j(x), 4R))$ to obtain $|g_j - g_{j,R}| \le |g_j|\mathbf{1}\{|g_j| > 4R\}$, $|g_R| \le 4R$ and g_R is continuous γ-almost surely (see (5.5)). The rest can be done as in the proof of Proposition 3, since it did not depend on the precise form of g. □

5.4. Proofs of statements in Section 4.1

Proof of Theorem 5. Since λ is regular and T is injective with nonsingular derivative, $\Lambda = T\#\lambda$ is also regular by Lemma 5.5.3 in [7]. Moreover, Λ is supported on K because T takes values there. Consequently, the Fréchet mean of Λ is unique and supported itself on K; this is essentially a consequence of Corollary 2.9 in [5]. For tidiness, we provide the full details in Section 4 of the supplementary material [72].

In view of the preceding paragraph, it suffices to show that

$$\mathbb{E}d^2(\lambda, \Lambda) \le \mathbb{E}d^2(\theta, \Lambda), \qquad \theta \in \mathcal{P}_2(K).$$

As a gradient of a convex function, $T = \mathbf{t}_\lambda^\Lambda$ is optimal. Let ϕ be the convex potential of T, and define ϕ^* its Legendre transform. Then the pair $(\|x\|^2/2 - \phi, \|y\|^2/2 - \phi^*)$ is dual optimal. Invoking strong duality for λ and weak duality for θ, we find

$$d^2(\lambda, \Lambda) = \int_{\mathbb{R}^d} \left(\frac{1}{2}\|x\|^2 - \phi(x)\right) \mathrm{d}\lambda(x) + \int_{\mathbb{R}^d} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y);$$

$$d^2(\theta, \Lambda) \ge \int_{\mathbb{R}^d} \left(\frac{1}{2}\|x\|^2 - \phi(x)\right) \mathrm{d}\theta(x) + \int_{\mathbb{R}^d} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y).$$

By Fubini's theorem (see the supplementary material for a justification), we have

$$\mathbb{E}d^2(\lambda, \Lambda) = \int_{\mathbb{R}^d} \left(\frac{1}{2}\|x\|^2 - \mathbb{E}\phi(x)\right) \mathrm{d}\lambda(x) + \mathbb{E}\int_{\mathbb{R}^d} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y);$$

$$\mathbb{E}d^2(\theta,\Lambda)\ge\int_{\mathbb{R}^d}\left(\frac{1}{2}\|x\|^2-\mathbb{E}\phi(x)\right)\mathrm{d}\theta(x)+\mathbb{E}\int_{\mathbb{R}^d}\left(\frac{1}{2}\|y\|^2-\phi^*(y)\right)\mathrm{d}\Lambda(y).$$

The function $\mathbb{E}T$ is continuous (by the bounded convergence theorem and boundedness of K), so equals the identity for all $x\in K$. Again by Fubini's theorem (see the supplementary material), it follows that $\mathbb{E}\phi(x)=\|x\|^2/2$ for all $x\in K$, perhaps up to an additive constant. Since θ and λ are both supported on K, the integrals with respect to λ and θ vanish, and this completes the proof. □

As part of our proofs, we will need to control the Wasserstein distance between the regularised measures and their true counterparts:

Lemma 5. *The smooth measure $\widehat{\Lambda}_i$ defined by* (4.1) *satisfies*

$$d^2\left(\widehat{\Lambda}_i,\frac{\widetilde{\Pi}_i}{\widetilde{\Pi}_i(K)}\right)\le C_{\psi,K}\sigma^2\qquad \textit{if } \sigma\le 1 \textit{ and } \widetilde{\Pi}_i(K)>0,\tag{5.6}$$

where $C_{\psi,K}$ is a (*finite*) *constant that depends only on ψ and K.*

We prove the lemma in the supplementary material [72], Section 4.

Remark 5. There is no need for ψ to be isotropic: it is sufficient that merely

$$\delta_\psi(r)=\inf_{\|x\|\le r}\psi(x)>0,\qquad r>0,$$

which is satisfied as long as ψ is continuous and strictly positive.

We now remark that a trivial extension of [58], Lemma 3, yields:

Lemma 6 (Number of points per process is $O(\tau_n)$). *If $\tau_n/\log n\to\infty$, then there exists a constant $C_\Pi>0$, depending only on the distribution of the Π's, such that*

$$\liminf_{n\to\infty}\frac{\min_{1\le i\le n}\Pi_i^{(n)}(K)}{\tau_n}\ge C_\Pi\qquad \textit{almost surely}.$$

In particular, there are no empty point processes, so the normalisation is well-defined.

Proof of Theorem 6. The proof is very similar to that of Theorem 1 in Panaretos and Zemel [58], and we give the details in the supplementary material [72]. □

Proof of Theorem 7. The argument is considerably different than the case $d=1$ considered in [58], and brings into play the geometry of convex functions in $\mathbb{R}^d$. Let i be a fixed integer and for $n\ge i$ set

$$\mu_n=\widehat{\Lambda}_i;\qquad \nu_n=\widehat{\lambda}_n;\qquad \mu=\Lambda_i;\qquad \nu=\lambda;\qquad u_n=\widehat{T}_i^{-1};\qquad u=T_i^{-1}.$$

We wish to show that $u_n \to u$ uniformly on compact sets, using our knowledge that

$$\begin{cases} \mu_n \to \mu; \\ \nu_n \to \nu; \end{cases} \qquad u_n\#\mu_n = \nu_n; \qquad u\#\mu = \nu; \qquad u_n, u \text{ optimal.}$$

This follows from Proposition 6 below. To verify the conditions, notice that all the measures are supported on $K = E$, a compact and convex set. Furthermore μ_n, μ and ν all have strictly positive densities there, so their support is exactly K. Continuity of u on $\operatorname{int}(K)$ follows from the assumptions that T_i and T_i^{-1} are continuous. The finiteness in (5.7) follows from the compactness of K, and the uniqueness follows from the regularity of μ.

The same proposition can be applied to show convergence of $\widehat{T_i}$ to T_i uniformly on $\Omega \subseteq \operatorname{int}(K)$: one needs to reverse the roles of μ_n and ν_n and of μ to ν, and notice that ν too is regular, which guarantees the uniqueness in (5.7). □

Proof of Corollary 4. The square of the distance is

$$\int_K \bigl\| \widehat{T}_i^{-1}\bigl(T_i(x)\bigr) - x \bigr\|^2 \, \mathrm{d}\frac{\Pi_i}{\Pi_i(K)},$$

and this is well-defined (that is, $\Pi_i(K) > 0$) almost surely for n large enough by Lemma 6. Since $\lambda(\partial K) = 0$, almost surely there are no points on the boundary and the integral can be taken on the interior of K. Let $\Omega \subseteq \operatorname{int}(K)$ be compact and split the integral to Ω and its complement. Then

$$\int_{\operatorname{int}(K)\setminus\Omega} \bigl\| \widehat{T}_i^{-1}\bigl(T_i(x)\bigr) - x \bigr\|^2 \, \mathrm{d}\frac{\Pi_i}{\Pi_i(K)} \le d_K^2 \frac{\Pi_i(\operatorname{int}(K)\setminus\Omega)}{\tau_n} \frac{\tau_n}{\Pi_i(K)} \overset{\text{as}}{\to} d_K^2 \lambda\bigl(\operatorname{int}(K)\setminus\Omega\bigr),$$

by the law of large numbers. Since the interior of K can be written as a countable union of compact sets, the right-hand side can be made arbitrarily small by selection of Ω.

Let us now consider the integral on Ω. Since

$$\int_\Omega \bigl\| \widehat{T}_i^{-1}\bigl(T_i(x)\bigr) - x \bigr\|^2 \, \mathrm{d}\frac{\Pi_i}{\Pi_i(K)} \le \sup_{x\in\Omega} \bigl\| \widehat{T}_i^{-1}\bigl(T_i(x)\bigr) - x \bigr\|^2 = \sup_{y\in T_i(\Omega)} \bigl\| \widehat{T}_i^{-1}(y) - T_i^{-1}(y) \bigr\|^2$$

and $T_i(\Omega)$ is compact, we only need to show that it is included in $\operatorname{int}(K)$ in order to apply Theorem 7. Suppose towards contradiction that $y = T_i(x) \in \partial K$ for $x \in \operatorname{int}(K)$. Let $\alpha \in \mathbb{R}^d \setminus \{0\}$ with $\langle y, \alpha \rangle \ge \sup\langle K, \alpha \rangle$. Let $x' = x + t\alpha$ for $t > 0$ small enough such that $x' \in \operatorname{int}(K)$. Then $y' = T_i(x') \in K$, so that

$$0 \le \bigl\langle y' - y, x' - x \bigr\rangle = t\bigl\langle y' - y, \alpha \bigr\rangle.$$

Either condition in the statement of the corollary imply that $y' = y$, in contradiction to T_i being injective. □

5.5. Monotone operators, optimal transportation, stochastic convergence

This section contains the statements and proofs of analytical results needed in our proofs, culminating in Proposition 6. The latter is the backbone result needed for the proofs of Theorem 7, Theorem 3 (more precisely, Proposition 3) and Theorem 4. Rather than start with all the background definitions, we will define the necessary objects en route.

We shall follow the notation and terminology of Alberti and Ambrosio [3]. Let u be a set-valued function (or multifunction) on $\mathbb{R}^d$, that is, $u : \mathbb{R}^d \to 2^{\mathbb{R}^d}$. It is said that u is *monotone* if

$$\langle y_2 - y_1, x_2 - x_1 \rangle \geq 0 \qquad \text{whenever } y_i \in u(x_i) \ (i = 1, 2).$$

When $d = 1$, the definition reduces to u being a nondecreasing (set-valued) function. It is said that u is *maximal* if no points can be added to its graph while preserving monotonicity:

$$\bigl\{\bigl\langle y' - y, x' - x\bigr\rangle \geq 0 \text{ whenever } y \in u(x)\bigr\} \quad \Longrightarrow \quad y' \in u\bigl(x'\bigr).$$

We sometimes use the notation $(x, y) \in u$ to mean $y \in u(x)$. Note that $u(x)$ can be empty, even when u is maximal.

The relevance of monotonicity stems from the fact that subdifferentials of convex functions are monotone. That is, if $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is lower semicontinuous and convex (and not identically infinite), then $u = \partial\varphi$ is maximally monotone [3], Section 7, where

$$\partial\varphi(x) = \bigl\{y : \varphi(z) \geq \varphi(x) + \langle y, z - x\rangle \text{ for any } z\bigr\}$$

is the *subdifferential* of φ at x. Here $u(x) = \varnothing$ if $\varphi(x) = \infty$.

We will use extensively the continuity of u at points where it is univalued.

Proposition 5 (Continuity at singletons). *Let u be a maximal monotone function, and suppose that $u(x) = \{y\}$ is a singleton. Then u is nonempty on some neighbourhood of x and it is continuous at x: if $x_n \to x$ and $y_n \in u(x_n)$, then $y_n \to y$.*

Proof. See [3], Corollary 1.3(4). Notice that this result implies that differentiable convex functions are continuously differentiable ([63], Corollary 25.5.1). □

It turns out that when u is univalued, monotonicity is a local property. To state the result in the general form that we shall use, we need to introduce the notion of points of Lebesgue density.

Let $B_r(y) = \{x : \|x - y\| < r\}$ for $r \geq 0$ and $y \in \mathbb{R}^d$. A point x_0 is of *Lebesgue density* of a measurable set $G \subseteq \mathbb{R}^d$ if for any $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that

$$\frac{\operatorname{Leb}(B_t(x_0) \cap G)}{\operatorname{Leb}(B_t(x_0))} > 1 - \varepsilon, \qquad 0 < t < t_\varepsilon.$$

We denote the set of points of Lebesgue density of G by G^{den}. Clearly, G^{den} lies between $\operatorname{int}(G)$ and $\overline{G}$. Stein and Shakarchi [68], Chapter 3, Corollary 1.5, show that almost any point of G is in G^{den}. By the Hahn–Banach theorem, $G^{\text{den}} \subseteq \operatorname{int}(\operatorname{conv}(G))$.

Lemma 7 (Density points and distance). *Let x_0 be a point of Lebesgue density of a measurable set $G \subseteq \mathbb{R}^d$. Then*

$$\delta(z) = \inf_{x\in G} \|z - x\| = o\big(\|z - x_0\|\big) \qquad \text{as } z \to x_0.$$

This result was given as an exercise in [68]; for completeness we provide a full proof in the supplementary material [72].

Lemma 8 (Local monotonicity). *Let u be a maximal monotone function such that $u(x_0) = \{y_0\}$. Suppose that x_0 is a point of Lebesgue density of a set G satisfying*

$$\big\langle y - y^*, x - x_0\big\rangle \geq 0 \qquad \forall x \in G \ \forall y \in u(x).$$

Then $y^* = y_0$. In particular, the result is true if the inequality holds on $G = O \setminus \mathcal{N}$ with $\varnothing \neq O$ open and $\mathcal{N}$ Lebesgue negligible.

Proof. Set $z_t = x_0 + t(y^* - y_0)$ for $t > 0$ small. It may be that $z_t \notin G$; but Lemma 7 guarantees existence of $x_t \in G$ with $\|x_t - z_t\|/t \to 0$. By Proposition 5 $u(x_t)$ is nonempty for t small enough. For $y_t \in u(x_t)$,

$$\begin{aligned} 0 &\leq \big\langle y_t - y^*, x_t - x_0\big\rangle = \big\langle y_t - y^*, x_t - z_t\big\rangle + \big\langle y_t - y^*, z_t - x_0\big\rangle \\ &= \big\langle y_t - y^*, x_t - z_t\big\rangle + t\big\langle y_t - y_0, y^* - y_0\big\rangle - t\big\|y^* - y_0\big\|^2. \end{aligned}$$

Rearrangement, division by $t > 0$ and application of the Cauchy–Schwarz inequality gives

$$\big\|y^* - y_0\big\|^2 \leq \|y_t - y_0\|\big\|y^* - y_0\big\| + t^{-1}\|x_t - z_t\|\big(\|y_t - y_0\| + \big\|y^* - y_0\big\|\big).$$

As $t \searrow 0$ the right-hand side vanishes, since $y_t \to y_0$ (Proposition 5) and $\|x_t - z_t\|/t \to 0$. It follows that $y^* = y_0$. □

This concludes the necessary discussion on monotone operators. We will now state some necessary results on optimal transportation maps, and specifically their convergence properties. Consider the following setting: let $\{\mu_n\}$, $\{\nu_n\}$ be two sequences of probability measures on $\mathbb{R}^d$ that converge weakly to μ and ν respectively. Let π_n be an optimal coupling between μ_n and ν_n having finite cost, which is supported on the graph of a subdifferential of a proper (not identically infinite) convex lower semicontinuous function φ_n [70], Chapter 2. The set-valued function $u_n = \partial\varphi_n$ that maps x to the subdifferential of φ_n at x is maximally monotone [3], Section 7. The appropriate functions for μ and ν will be denoted by φ and $u = \partial\varphi$ and the optimal coupling by π. This setting will be succinctly referred to by the equation

$$\begin{array}{ll} \mu_n \to \mu & \pi_n \text{ finite optimal for } \mu_n, \nu_n (u_n = \partial\varphi_n)\#\mu_n = \nu_n, \\ \nu_n \to \nu & \pi \text{ unique optimal for } \mu, \nu (u = \partial\varphi)\#\mu = \nu. \end{array} \tag{5.7}$$

We notice now that uniqueness of π and the stability of optimal transportation imply that π_n converge weakly to π (even if π_n is not unique); see Schachermayer and Teichmann [65], Theorem 3, or Cuesta-Albertos *et al.* [25], Theorem 3.2. This weak convergence will be used in the following form.

Lemma 9 (Portmanteau). *Weak convergence of Borel probability measures μ_k to μ on $\mathbb{R}^d$ is equivalent to any of the following conditions*:

(I) *for any open set G, $\liminf \mu_k(G) \geq \mu(G)$;*
(II) *for any closed set F, $\limsup \mu_k(F) \leq \mu(F)$;*
(III) *$\int h \, \mathrm{d}\mu_k \to \int h \, \mathrm{d}\mu$ for any bounded measurable h whose set of discontinuity points is a μ-null set.*

Proof. The equivalence with the first two conditions is classical and can be found in Billingsley [15], Theorem 2.1; for the third, see Pollard [61], Section III.2. □

We shall now translate this into convergence of u_n to u under certain regularity conditions.

Proposition 6 (Uniform convergence of optimal maps). *In the setting of Display (5.7), denote $E = \operatorname{supp}(\mu)$.*

Let Ω be a compact subset of E^{den} on which u is univalued, where E^{den} is the set of points of Lebesgue density of E. Then u_n converges to u uniformly on Ω: $u_n(x)$ is nonempty for all $x \in \Omega$ and all $n > N_\Omega$, and

$$\sup_{x\in\Omega} \sup_{y\in u_n(x)} \left\| y - u(x) \right\| \to 0, \qquad n \to \infty.$$

In particular, if u is univalued throughout $\operatorname{int}(E)$ (so that $\varphi \in C^1$ there), then uniform convergence holds for any compact $\Omega \subset \operatorname{int}(E)$.

Corollary 6 (Pointwise convergence μ-almost surely). *If in addition μ is absolutely continuous then $u_n(x) \to u(x)$ μ-almost surely.*

Proof. The set of points $x \in E$ for which $\Omega = \{x\}$ fails to satisfy the conditions of Proposition 6 is included in

$$\left(E \setminus E^{\text{den}}\right) \cup \left\{x \in \operatorname{int}\bigl(\operatorname{conv}(E)\bigr) : u(x) \text{ contains more than one point}\right\}.$$

(Since u is nonempty on $\operatorname{int}(\operatorname{conv}(E))$ by [3], Corollary 1.3(2).) Both sets are Lebesgue-negligible (see [3], Remark 2.3, for the latter), and μ is absolutely continuous. □

Remark 6. In the setting of Theorem 7, E is convex, μ is absolutely continuous, and u is univalued on $\operatorname{int}(E)$, so one can take any $\Omega \subseteq \operatorname{int}(E)$, without the need to introduce Lebesgue density. The more general statement of the proposition is used in the proof of Proposition 3, where we have no control on the support of γ or the regularity of the transport maps.

We split the proof of Proposition 6 into two steps: (1) Limit points of the graphs of u_n are in the graph of u (Lemma 11); (2) Points in the graphs of u_n stay in a bounded set (Proposition 7). Each of these points will be proven using one intermediate lemma.

Lemma 10 (Points in the limit graph are limit points). *Assume* (5.7). *For any* $x_0 \in \operatorname{supp}(\mu)$ *such that* $u(x_0) = \{y_0\}$ *is a singleton there exists a subsequence* $(x_{n_k}, y_{n_k}) \in u_{n_k}$ *that converges to* (x_0, y_0).

Proof. Since $u = \partial\varphi$ is a maximal monotone function [3], Section 7, that is univalued at x_0, it is continuous there (Proposition 5). This means that for any $\varepsilon > 0$ there exists $\delta > 0$ such that if $x \in B_\delta(x_0) = \{x : \|x - x_0\| < \delta\}$ then $u(x)$ is nonempty and if $y \in u(x)$, then $\|y - y_0\| < \varepsilon$. Take $\varepsilon_k \to 0$ and corresponding $\delta_k \to 0$, and set $B_k = B_{\delta_k}(x_0)$, $V_k = B_{\varepsilon_k}(y_0)$. Then $u(B_k) \subseteq V_k$, so

$$\pi(B_k \times V_k) = \pi\bigl\{(x, y) : x \in B_k, y \in u(x) \cap V_k\bigr\} = \pi\bigl\{(x, y) : x \in B_k, y \in u(x)\bigr\} = \mu(B_k) > 0,$$

because B_k is a neighbourhood of $x_0 \in \operatorname{supp}(\mu)$. Since $B_k \times V_k$ is open, we have by the Portmanteau lemma that $\pi_n(B_k \times V_k) > 0$ for n large. Consequently, there exists n_k such that

$$\pi_{n_k}(B_k \times V_k) > 0 \quad \text{and} \quad n_k \to \infty \qquad \text{as } k \to \infty.$$

Since π_{n_k} is concentrated on the graph of u_{n_k}, it follows that there exist $(x_{n_k}, y_{n_k}) \in u_{n_k}$ with $\|x_{n_k} - x_0\| < \delta_k$ and $\|y_{n_k} - y_0\| < \varepsilon_k$. Hence $(x_{n_k}, y_{n_k}) \to (x_0, y_0)$. □

Lemma 11 (Limit points are in the limit graph). *Assume that* (5.7) *holds and denote* $E = \operatorname{supp}(\mu)$. *If a subsequence* $(x_{n_k}, y_{n_k}) \in u_{n_k}$ *converges to* (x_0, y^*), *where* x_0 *is a point of Lebesgue density of* E, *and* $u(x_0)$ *is a singleton, then* $y^* = u(x_0)$. *In particular, the statement is true if* $x_0 \in \operatorname{int}(E)$ *and* $u(x_0)$ *is a singleton.*

Proof. The set $\mathcal{N} \subseteq \mathbb{R}^d$ of points where u contains more than one element is Lebesgue negligible [3], Remark 2.3. There exists a neighbourhood V of x_0 on which u is nonempty (Proposition 5). Thus, x_0 is a point of Lebesgue density of $G = (E \cap V) \setminus \mathcal{N}$, and $u(x)$ is a singleton for every $x \in G$. Fix such an x and set $y = u(x)$. By Lemma 10 (applied to $\{u_{n_k}\}_{k=1}^\infty$ at x) there exist sequences $x'_{n_{k_l}} \to x$ and $y'_{n_{k_l}} \to y$ with $(x'_{n_{k_l}}, y'_{n_{k_l}}) \in u_{n_{k_l}}$. Consequently,

$$\langle y - y^*, x - x_0\rangle = \lim_{l\to\infty} \langle y'_{n_{k_l}} - y_{n_{k_l}}, x'_{n_{k_l}} - x_{n_{k_l}}\rangle \geq 0.$$

This holds for any $(x, y) \in u$ such that $x \in G$. Since x_0 is a point of Lebesgue density of G (and u is maximal), it follows from Lemma 8 that $y^* = u(x_0)$. □

Let $B_\varepsilon^\infty(x_0) = \{x : \|x - x_0\|_\infty < \varepsilon\}$ be the ℓ_∞ ball around x_0 and $\overline{B}_\varepsilon^\infty(x_0)$ its closure.

Lemma 12 (Continuity of convex hulls). *Let* $Z = \{z_i\} \subseteq \mathbb{R}^d$ *be a set of points whose convex hull,* $\operatorname{conv}(Z)$, *includes* $B_\rho^\infty(x_0)$ *and let* $\tilde{Z} = \{\tilde{z}_i\}$ *be a set of points such that* $\|\tilde{z}_i - z_i\|_\infty \leq \varepsilon$. *Then the convex hull of* $\tilde{Z}$ *includes* $B_{\rho-\varepsilon}^\infty(x_0)$.

For a proof, see the supplementary material [72].

Proposition 7 (Boundedness). *Suppose that* (5.7) *holds, and fix a compact* $\Omega \subseteq \operatorname{int}(\operatorname{conv}(\operatorname{supp}(\mu)))$. *Then for* $n > N(\Omega)$ *sufficiently large,* $u_n(x)$ *is nonempty for all* $x \in \Omega$ *and* $u_n(\Omega)$ *is bounded uniformly.*

Proof. Denote $E = \operatorname{supp}(\mu)$ and its convex hull by $F = \operatorname{conv}(E)$. There exists $\delta = \delta(\Omega) > 0$ such that the closed ℓ_∞-ball, $\overline{B}^\infty_{3\delta}(\Omega)$, is included in $\operatorname{int}(F)$. Cover Ω by a finite union of $B^\infty_\delta(\omega_j)$, and denote by Q be the finite set of vertices of $\bigcup_j \overline{B}^\infty_{3\delta}(\omega_j)$. Since Q is included in the convex hull of E, each point in Q can be written as a convex combination of elements of E. We conclude that there exists a finite set $Z = \{z_1, \dots, z_m\}$ of points in E whose convex hull includes $B^\infty_{3\delta}(\omega_j)$ for any j.

Let $B_i = B^\infty_\delta(z_i)$. Since B_i is an open neighbourhood of $z_i \in E = \operatorname{supp}(\mu)$, the Portmanteau lemma implies that when n is large, $\mu_n(B_i) > \varepsilon_i = \mu(B_i)/2$ for any $i = 1, \dots, m$. Let $\varepsilon = \min_i \varepsilon_i > 0$. Since $\{\nu_n\}$ is a tight sequence, there exists a compact set K_ε such that $\nu_n(K_\varepsilon) > 1 - \varepsilon$ for any integer n. In particular, there exist $x_{ni} \in B_i$ and $y_{ni} \in u_n(x_{ni})$ such that $y_{ni} \in K_\varepsilon$. Application of Lemma 12 to

$$\tilde{Z} = X_n = \{x_{n1}, \dots, x_{nm}\}$$

and noticing that by definition $\|x_{ni} - z_i\|_\infty \le \delta$ yields

$$\operatorname{conv}(X_n) = \operatorname{conv}\bigl(\{x_{n1}, \dots, x_{nm}\}\bigr) \supseteq B^\infty_{3\delta-\delta}(\omega_j) = B^\infty_{2\delta}(\omega_j) \qquad \text{for all } j.$$

For each $\omega \in \Omega$ there exists j such that $\|\omega - \omega_j\|_\infty \le \delta$, so that $\operatorname{conv}(X_n) \supseteq B^\infty_\delta(\omega) \supseteq B_\delta(\omega)$, since ℓ_2-balls are smaller than ℓ_∞-balls. Summarising: $\operatorname{conv}(X_n) \supseteq B_\delta(\Omega)$.

By [3], Lemma 1.2(4), it follows that for any $\omega \in \Omega$ and any $y_0 \in u_n(\omega)$,

$$\|y_0\| \le \frac{[\sup_{x,z \in X_n} \|x - z\|][\max_{x \in X_n} \inf_{y \in u_n(x)} \|y\|]}{d(\omega, \mathbb{R}^d \setminus \operatorname{conv}(X_n))} \le \frac{1}{\delta}\Bigl[\sup_{k,l} \|x_{nk} - x_{nl}\|\Bigr]\Bigl[\max_i \inf_{y \in u_n(x_{ni})} \|y\|\Bigr].$$

Now observe that the infimum at the right-hand side is bounded by $\|y_{ni}\| \le \sup_{y \in K_\varepsilon} \|y\|$. Furthermore, $\|x_{nk} - x_{nl}\| \le 2\sqrt{d}\delta + \|z_k - z_l\|$. Hence,

$$\forall \omega \in \Omega \ \forall y_0 \in u_n(\omega): \qquad \|y_0\| \le \frac{1}{\delta}\Bigl(2\sqrt{d}\delta + \max_{k,l} \|z_k - z_l\|\Bigr) \sup_{y \in K_\varepsilon} \|y\|,$$

and the right-hand side is independent of n. We may therefore conclude that for n large enough, $u_n(\Omega)$ stays in a compact set; it is nonempty by [3], Corollary 1.3(2). □

Proof of Proposition 6. By Proposition 7 when $n > N_\Omega$ is large, $u_n(x) \ne \varnothing$ for all $x \in \Omega$ and

$$\sup_{x \in \Omega} \sup_{y \in u_n(x)} \|y\| \le C_{\Omega,d} < \infty, \qquad n > N_\Omega,$$

where $C_{\Omega,d}$ is a constant that depends only on Ω (and the dimension d).

Suppose that the converse is true, and uniform convergence does not hold. Then there exist $\varepsilon > 0$ and subsequences $y_{n_k} \in u_{n_k}(x_{n_k})$ such that $x_{n_k} \in \Omega$ and

$$\bigl\|y_{n_k} - u(x_{n_k})\bigr\| > \varepsilon, \qquad k = 1, 2, \dots.$$

The x_{n_k}'s lie in the compact set Ω, whereas by Proposition 7 the y_{n_k}'s lie in the ball of radius $C_{\Omega,d}$ centred at the origin. Therefore, up to the extraction of a subsequence, we have $x_{n_k} \to x \in \Omega$ and

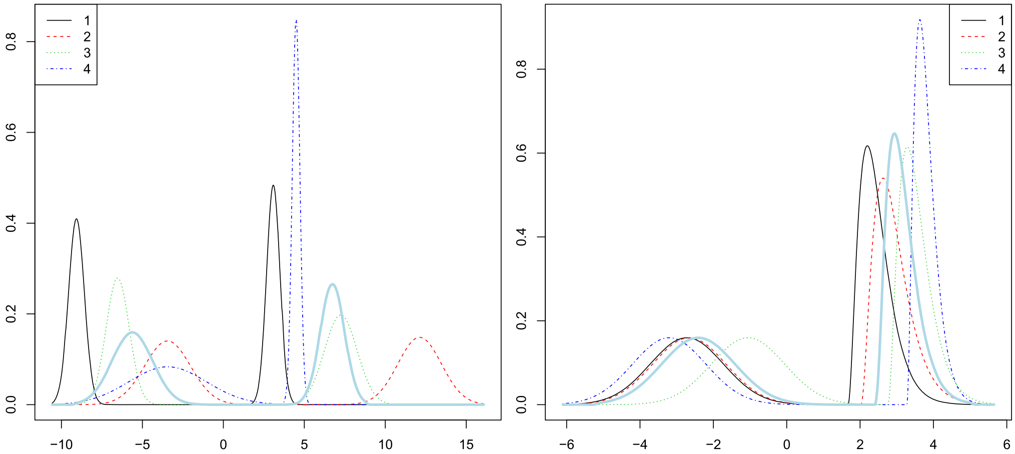


Figure 1. Densities of bimodal Gaussian mixtures (left) and Gaussian-gamma mixtures (right), with the corresponding Fréchet mean densities.

$y_{n_k} \to y$. By Lemma 11, $y = u(x)$. But u is continuous at x (Proposition 5), whence

$$\varepsilon < \left\| y_{n_k} - u(x_{n_k}) \right\| \le \| y_{n_k} - y \| + \left\| y - u(x) \right\| + \left\| u(x) - u(x_{n_k}) \right\| \to 0, \qquad k \to \infty,$$

a contradiction. □

6. Some examples

As an illustration, we implement Algorithm 1 in several settings for which pairwise optimal maps can be calculated explicitly at every iteration, allowing for fast computation without error propagation. Indeed, these settings allow for stronger convergence statements to be made on a case-by-case basis. More details on the calculations and properties of each individual scenario can be found in Section 3 of the supplement [72].

6.1. The case $d = 1$

When the measures are supported on the real line, the optimal maps have the explicit expression given in Equation (2.1) and one may apply Algorithm 1 starting from one of these measures. Figure 1 plots $N = 4$ univariate densities and the Fréchet mean yielded by the algorithm in two different scenarios. At the left, the densities were generated as

$$f^i(x) = \frac{1}{2}\phi\left(\frac{x - m_1^i}{\sigma_1^i}\right) + \frac{1}{2}\phi\left(\frac{x - m_2^i}{\sigma_2^i}\right), \tag{6.1}$$

with ϕ the standard normal density, and the parameters generated independently as

$$m_1^i \sim U[-13,-3], \qquad m_2^i \sim U[3,13], \qquad \sigma_1^i, \sigma_2^i \sim \text{Gamma}(4,4).$$

At the right of Figure 1, we used a mixture of a shifted gamma and a Gaussian:

$$f^i(x) = \frac{3}{5}\frac{\beta_i^3}{\Gamma(3)}\big(x - m_3^i\big)^2 e^{-\beta_i(x-3)} + \frac{2}{5}\phi\big(x - m_4^i\big), \tag{6.2}$$

with

$$\beta^i \sim \text{Gamma}(4,1), \qquad m_3^i \sim U[1,4], \qquad m_4^i \sim U[-4,-1].$$

The resulting Fréchet mean density for both settings is shown in thick light blue, and can be seen to capture the bimodal nature of the data. Even though the Fréchet mean of Gaussian mixtures is not a Gaussian mixture itself, it is approximately so, provided that the peaks are separated enough. Figure 8(a) shows the Procrustes maps pushing the Fréchet mean $\bar{\mu}$ to the measures $\mu^1, \dots, \mu^N$ in each case. If one ignores the "middle part" of the x axis, the maps appear (approximately) affine for small values of x and for large values of x, indicating how the peaks are shifted. In the middle region, the maps need to "bridge the gap" between the different slopes and intercepts of these affine maps.

6.2. Independence

We next take measures μ^i on $\mathbb{R}^2$, having independent marginal densities f_X^i as in (6.1), and f_Y^i as in (6.2). Figure 2 shows the density plot of $N = 4$ such measures, constructed as the product of the measures from Figure 1. One can distinguish the independence by the "parallel" structure of the figures: for every pair (y_1, y_2), the ratio $g(x, y_1)/g(x, y_2)$ does not depend on x (and vice versa, interchanging x and y). Figure 3 plots the density of the resulting Fréchet mean. We observe that the Fréchet mean captures the four peaks, and their location. Furthermore, the parallel nature of the figure is preserved in the Fréchet mean. Indeed, we prove in the supplement [72] that, unsurprisingly, the Fréchet mean is a product measure.

6.3. Common copulas

Let μ^i be a measure on $\mathbb{R}^2$ with density

$$g^i(x,y) = c\big(F_X^i(x), F_Y^i(y)\big) f_X^i(x) f_Y^i(y),$$

where f_X^i and f_Y^i are random densities on the real line with distribution functions F_X^i and F_Y^i, and c is a copula density. Figure 4 shows the density plot of $N = 4$ such measures, with f_X^i generated as in (6.1), f_Y^i as in (6.2), and c is the Frank(-8) copula density, while Figure 5 plots the density of the Fréchet mean obtained. (For ease of comparison, we use the same realisations of the densities that appear in Figure 1.) The Fréchet mean can be seen to preserve the shape of the

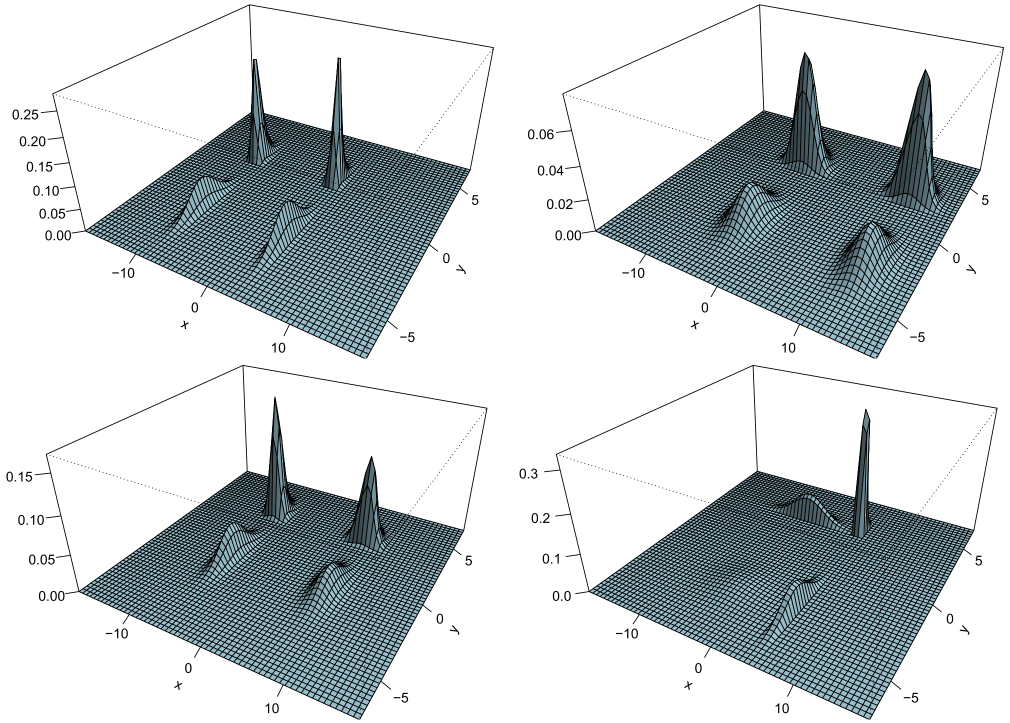


Figure 2. Density plots of the four product measures of the measures in Figure 1.

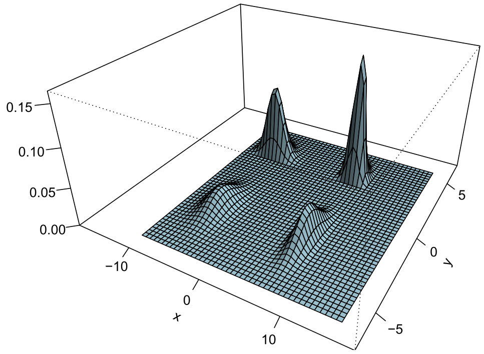


Figure 3. Density plot of the Fréchet mean of the measures in Figure 2.

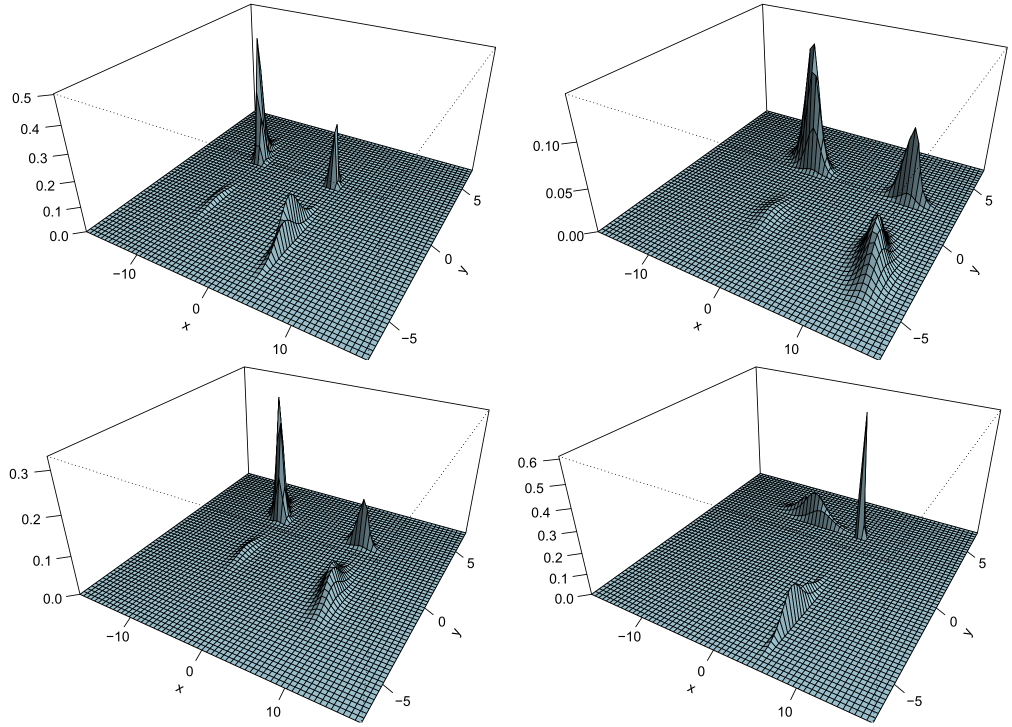


Figure 4. Density plots of four measures in $\mathbb{R}^2$ with Frank copula of parameter -8.

density, having four clearly distinguished peaks. Figure 8(b), depicting the resulting Procrustes maps, allows for a clearer interpretation: for instance the leftmost plot (in black) shows more clearly that the map splits the mass around $x = -2$ to a much wider interval; and conversely a very large amount mass is sent to $x \approx 2$. This rather extreme behaviour matches the peak of the density of μ^1 located at $x = 2$.

The first three scenarios are examples of situations where the measures $\{\mu^i\}$ are *compatible with each other* in the sense that $\mathbf{t}_{\mu^j}^{\mu^k} \circ \mathbf{t}_{\mu^i}^{\mu^j} = \mathbf{t}_{\mu^i}^{\mu^k}$. Boissard *et al.* [16] tackle the problem of finding the Fréchet mean in such a setting, by means of the *iterated barycentre*. In the supplementary material [72] we show that Algorithm 1 will always converges to the Fréchet mean, provided the initial point γ_0 is compatible with $\{\mu^i\}$ (for instance, if $\gamma_0 = \mu^i$). In fact, we show that convergence is established after a single iteration of the algorithm. Since optimal maps are gradients of convex potentials, they must have positive definite derivatives. Under regularity conditions, compatibility is essentially equivalent to the commutativity of the $d \times d$ matrices $\nabla \mathbf{t}_{\mu^j}^{\mu^k}(\mathbf{t}_{\mu^i}^{\mu^j}(x))$ and $\nabla \mathbf{t}_{\mu^i}^{\mu^j}(x)$ for μ^i-almost any x. We next discuss examples where this condition fails.

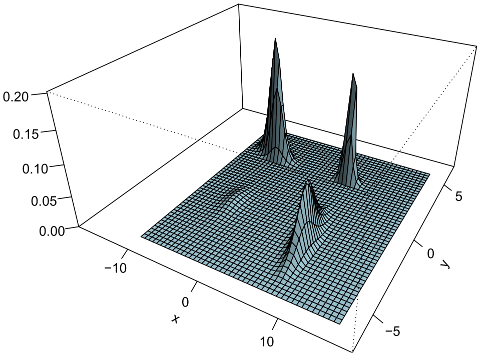


Figure 5. Density plot of the Fréchet mean of the measures in Figure 4.

6.4. Gaussian measures

Suppose that each μ^i follows a non-degenerate multivariate Gaussian distribution with mean 0 and covariance matrix S_i. The optimal maps are known to be linear and admit the explicit formula (Dowson and Landau [28]; Olkin and Pukelsheim [57])

$$\mathbf{t}_i^j = S_j^{1/2}\bigl[S_j^{1/2} S_i S_j^{1/2}\bigr]^{-1/2} S_j^{1/2}.$$

If the initial point γ_0 is another Gaussian measure with covariance matrix Γ_0, then by the linearity of the maps one sees that $\gamma_k \sim \mathcal{N}(0, \Gamma_k)$ for some positive definite Γ_k. Thus, one can calculate the optimal maps at each iteration; in the supplement [72] we prove that γ_k must converge to the unique Fréchet mean, which is also a Gaussian measure. This example is also studied independently in Álvarez-Esteban *et al.* [6], Section 4, where an alternative proof can be found. Our proof is shorter and arguably simpler, but the proof in [6] shows the additional property that the traces of the matrix iterates are monotonically increasing.

Notice that the Gaussian measures $\{\mu^i\}$ will be compatible if $S_i S_j = S_j S_i$, but they might well fail to be. Thus, the algorithm does not converge in one step. We observed, however, rapid convergence of the iterates of Algorithm 1 to the Fréchet mean, even for rather large values of N and d. Figure 6 shows density plots of $N = 4$ centred Gaussian measures on $\mathbb{R}^2$ with covariances $S_i \sim \text{Wishart}(I_2, 2)$, and Figure 7 shows the density of the resulting Fréchet mean. In this particular example, the algorithm needed 11 iterations starting from the identity matrix. The corresponding Procrustes registration maps are displayed in Figure 8(c). It is apparent from the figure that these maps are linear, and after a more careful reflection one can be convinced that their average is the identity. The four plots in the figure are remarkably different, in accordance with the measures themselves having widely varying condition numbers and orientations; μ^3 and more so μ^4 are very concentrated, so the registration maps "sweep" the mass towards zero. In contrast, the registration maps to μ^1 and μ^2 spread the mass out away from the origin.

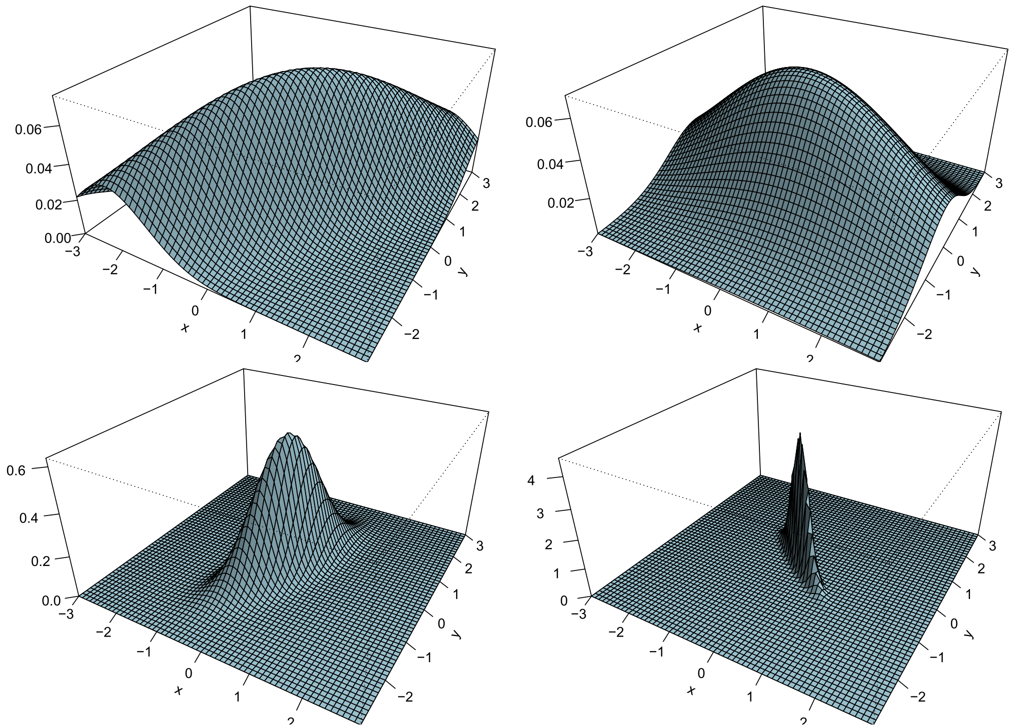


Figure 6. Density plot of four Gaussian measures in $\mathbb{R}^2$.

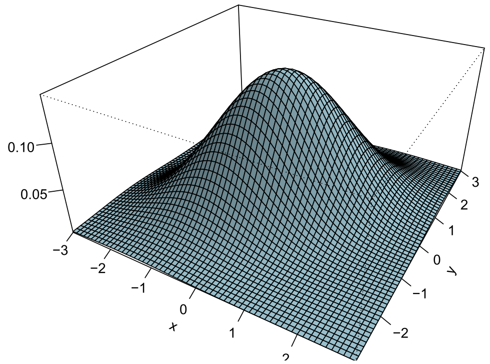
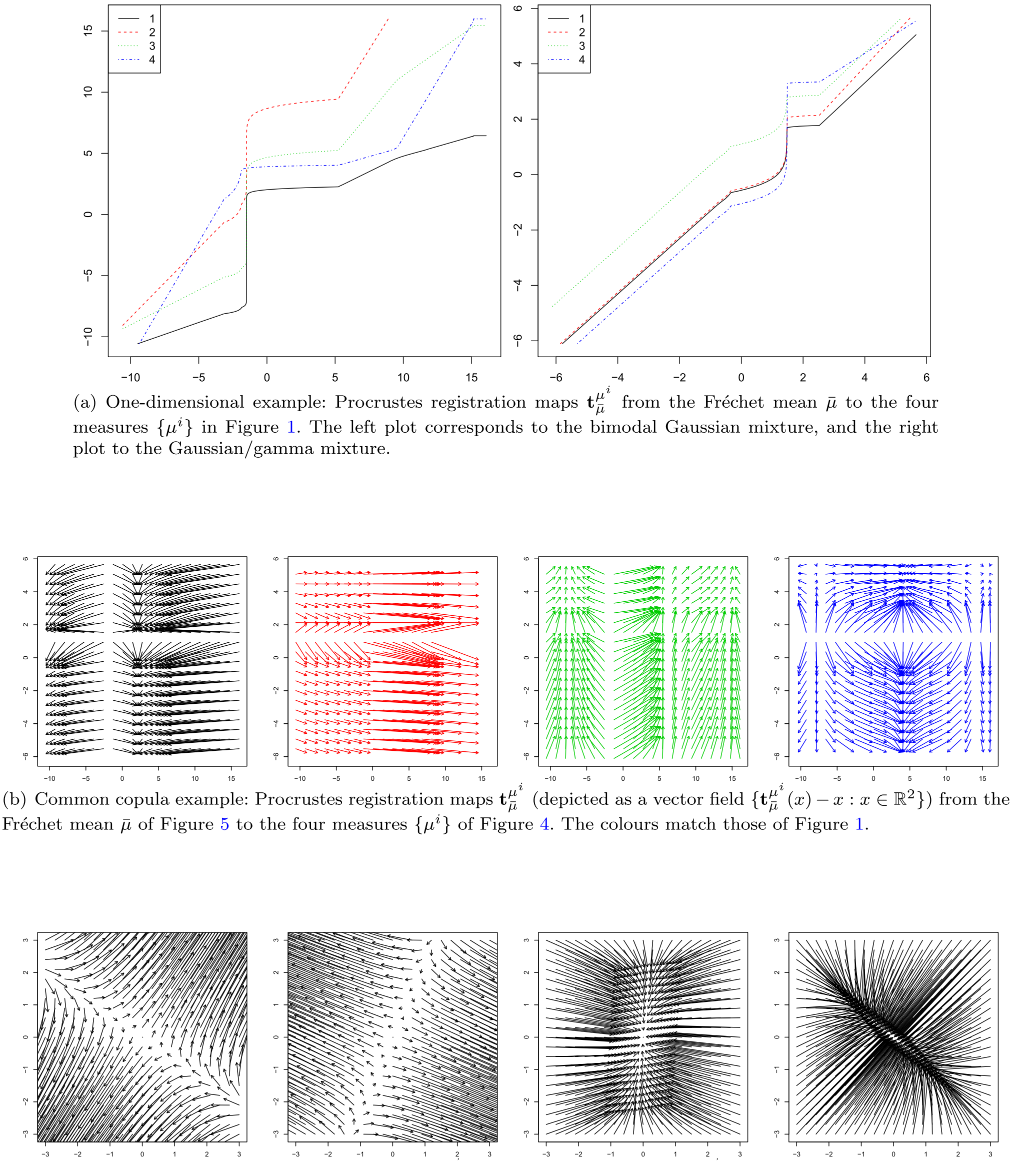


Figure 7. Density plot of the Fréchet mean of the measures in Figure 6.



(a) One-dimensional example: Procrustes registration maps $\mathbf{t}_{\bar{\mu}}^{\mu^i}$ from the Fréchet mean $\bar{\mu}$ to the four measures $\{\mu^i\}$ in Figure 1. The left plot corresponds to the bimodal Gaussian mixture, and the right plot to the Gaussian/gamma mixture.

(b) Common copula example: Procrustes registration maps $\mathbf{t}_{\bar{\mu}}^{\mu^i}$ (depicted as a vector field $\{\mathbf{t}_{\bar{\mu}}^{\mu^i}(x) - x : x \in \mathbb{R}^2\}$) from the Fréchet mean $\bar{\mu}$ of Figure 5 to the four measures $\{\mu^i\}$ of Figure 4. The colours match those of Figure 1.

(c) Gaussian example: Procrustes registration maps $\mathbf{t}_{\bar{\mu}}^{\mu^i}$ (depicted as a vector field $\{\mathbf{t}_{\bar{\mu}}^{\mu^i}(x) - x : x \in \mathbb{R}^2\}$) from the Fréchet mean $\bar{\mu}$ of Figure 7 to the four measures $\{\mu^i\}$ of Figure 6. The order corresponds to that of Figure 6 (left to right and top to bottom).

Figure 8. Procrustes registration maps for the one-dimensional, common copula, and Gaussian examples.

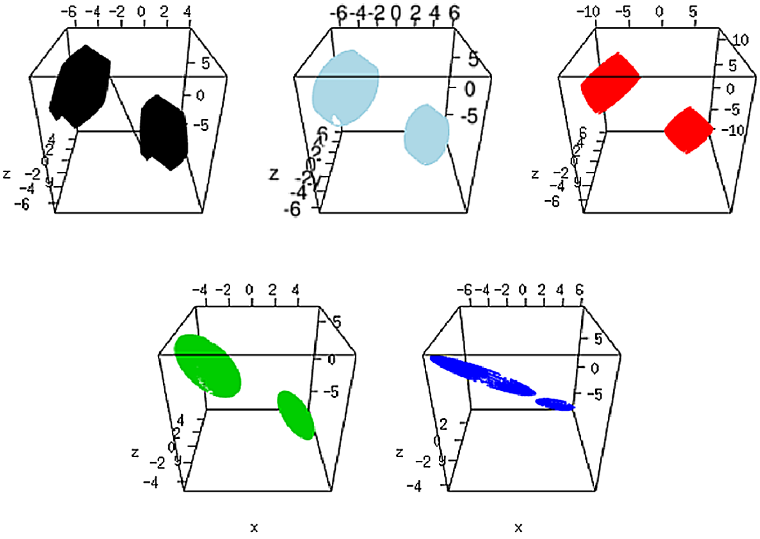


Figure 9. The set $\{v \in \mathbb{R}^3 : g^i(v) = 0.0003\}$ for $i = 1$ (top left), the Fréchet mean (top middle), $i = 2, 3, 4$ (top right, bottom left and bottom right respectively).

6.5. Partially Gaussian trivariate measures

We now apply Algorithm 1 in a situation that entangles two of the previous settings. Let U be a 3×3 real orthogonal matrix with columns U_1, U_2, U_3 and let μ^i have density

$$g^i(y_1, y_2, y_3) = g^i(y) = f^i\left(U_3^t y\right) \frac{1}{2\pi\sqrt{\det S^i}} \exp\left[-\frac{(U_1^t y, U_2^t y)(S^i)^{-1}\binom{U_1^t y}{U_2^t y}}{2}\right],$$

with f^i bounded density on the real line and $S^i \in \mathbb{R}^{2\times 2}$ positive definite. We simulated $N = 4$ such densities with f^i as in (6.1) and $S^i \sim \text{Wishart}(I_2, 2)$. We apply Algorithm 1 to this collection of measures and find their Fréchet mean (in Section 3 of the supplementary material [72] we provide precise details on how the optimal maps were calculated). Figure 9 shows level set of the resulting densities for some specific values. The bimodal nature of f^i implies that for most values of a, $\{x : f^i(x) = a\}$ has four elements. Hence the level sets in the figures are unions of four separate parts, with each peak of f^i contributing two parts that form together the boundary of an ellipsoid in $\mathbb{R}^3$ (see Figure 10). The principal axes of these ellipsoids and their position in $\mathbb{R}^3$ differ between the measures, but the Fréchet mean can be viewed as an average of those in some sense.

In terms of orientation (principal axes) of the ellipsoids, the Fréchet mean is most similar to μ^1 and μ^2, whose orientations are similar to one another.

In the most general examples, one might not be able to analytically obtain the optimal maps at each iteration. In such situations, one needs to resort to numerical schemes such as Benamou and Brenier [10], Haber *et al.* [39] or Chartrand *et al.* [23] to obtain the N optimal maps at each

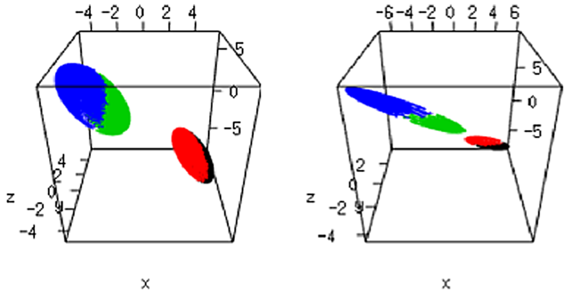


Figure 10. The set $\{v \in \mathbb{R}^3 : g^i(v) = 0.0003\}$ for $i = 3$ (left) and $i = 4$ (right).

iteration (see the concluding remarks for further discussion about numerical issues). Usually such schemes are iterative themselves, so one must take care in managing propagation of errors resulting from using approximate rather than exact transport maps.

7. Concluding remarks

While the algorithm and the convergence analysis in this work were discussed in the context of absolutely continuous measures, it is worth mentioning the possibility of applying it to discrete measures in some special cases. Specifically, suppose that each measure μ^i is uniform on a set of M distinct points, $\{x_m^i\}_{m=1}^M$. Define as in Anderes *et al.* [9] the set

$$S = \frac{1}{N}\bigl\{x_{m_1}^1 + \cdots + x_{m_N}^N : 1 \le m_i \le M, i = 1, \ldots, N\bigr\}$$

of averages of choices of points from the supports of $\{\mu^i\}$. Let γ_0 be an initial measure, uniform on M distinct points as well. There exist optimal maps (not necessarily unique) from γ_0 to each μ^i, and they can be averaged to yield γ_1. If $|S| = M^N$ (that is, the collection $\{x_m^i\}$ satisfies a general-position-type condition), then γ_1 will be concentrated on M points as well, and one may carry out further iterations. A conceptual problem with this application is that the Fréchet functional is not differentiable at discrete measures, so Algorithm 1 can no longer be viewed as gradient descent (but can still be seen as Procrustes averaging). Also, the Fréchet mean itself may fail to be unique. In simulations, we observed very rapid convergence of this iteration to a Karcher mean, but the specific limit depended quite heavily on the initial point, and was usually not a Fréchet mean. For problems of moderate size, one can recast the problem of minimising the Fréchet functional as a linear program [9] and find an exact Fréchet mean. In fact, Anderes *et al.* [9] treat the more general problem where the measures are supported on a different number of points and are not constrained to be uniform on their supports.

An important issue more generally is that of efficient approximate numerical schemes for calculating Fréchet means in Wasserstein space. This is a very active field of research with a rapidly-growing literature (both in numerical analysis and in computer science), and a detailed survey is far beyond the scope of this paper. If one is content with an *approximate* solution, then there are several approaches suggested in the literature. Indicatively, let us mention Bonneel *et al.* [19] who use a tomographic perspective to reduce the problem to 1-dimensional computations; Carlier, Oberman and Oudet [22] who use nonsmooth optimisation techniques to solve a discretised version of the dual problem; Oberman and Ruan [56] exploit the sparsity of optimal plans to reduce the size of the linear program to a tractable one.

Another line of research involves *entropic regularisation*, where one adds an entropy term to the definition of the Wasserstein distance. This leads to a strictly convex problem that is far better behaved than the original problem. Though its solution no longer yields the actual mean, it can be thought of as a regularised surrogate Fréchet mean. In this direction, Cuturi and Doucet [26] employ differentiability properties and carry out what could be thought of as a “gradient descent”, a discrete analogue of Algorithm 1; Benamou *et al.* [11] exploit the structure of the constraints as an intersection of convex sets by means of iterating Bregman projections that can be evaluated efficiently. Solomon *et al.* [66] extend this idea to the manifold setup, by convoluting with a heat kernel; and Cuturi and Peyré [27] employ the regularisation at the level of the dual, rather than the primal, problem. Recently, Rolet, Cuturi and Peyré [64] employed this technique in the context of dictionary learning; and Bonneel, Peyré and Cuturi [18] define a sort of “barycentric convex hull” of given histograms and show how to project a new histogram onto that convex hull.

Acknowledgements

This research was supported by a European Research Council Starting Grant Award to Victor M. Panaretos. Part of this work grew out of work presented at the Mathematical Biosciences Institute (Ohio State University), during the “Statistics of Time Warping and Phase Variation” Workshop, November 2012. We wish to acknowledge the stimulating environment offered by the Institute. We wish to warmly thank Prof. Clément Hongler for several useful discussions. We are also very thankful to two reviewers and an associate editor for their detailed and constructive feedback.

Supplementary Material

Fréchet means and Procrustes analysis in Wasserstein space (DOI: 10.3150/17-BEJ1009SUPP; .pdf). The online supplement contains more details on the examples, additional technical material, as well as those proofs that were omitted from the main paper.

References

[1] Afsari, B., Tron, R. and Vidal, R. (2013). On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM J. Control Optim.* **51** 2230–2260. MR3057324

[2] Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** 904–924. MR2801182

[3] Alberti, G. and Ambrosio, L. (1999). A geometrical approach to monotone functions in $\mathbf{R}^n$. *Math. Z.* **230** 259–316. MR1676726

[4] Allassonnière, S., Amit, Y. and Trouvé, A. (2007). Towards a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 3–29.

[5] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2011). Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** 358–375.

[6] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.* **441** 744–762.

[7] Ambrosio, L., Gigli, N. and Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed. London: Springer.

[8] Amit, Y., Grenander, U. and Piccioni, M. (1991). Structural image restoration through deformable templates. *J. Amer. Statist. Assoc.* **86** 376–387.

[9] Anderes, E., Borgwardt, S. and Miller, J. (2016). Discrete Wasserstein barycenters: Optimal transport for discrete data. *Math. Methods Oper. Res.* 1–21.

[10] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84** 375–393. MR1738163

[11] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L. and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.* **37** A1111–A1138.

[12] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* 1196–1217.

[13] Bigot, J., Gouet, R., Klein, T. and López, A. (2013). Geodesic PCA in the Wasserstein space. Preprint. Available at arXiv:1307.7721.

[14] Bigot, J. and Klein, T. (2012). Consistent estimation of a population barycenter in the wasserstein space. ArXiv e-prints.

[15] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. New York: Wiley. MR1700749

[16] Boissard, E., Le Gouic, T., Loubes, J.-M. et al. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli* **21** 740–759.

[17] Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

[18] Bonneel, N., Peyré, G. and Cuturi, M. (2016). Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. Graph.* **35** 71–1.

[19] Bonneel, N., Rabin, J., Peyré, G. and Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision* **51** 22–45. MR3300482

[20] Bookstein, F.L. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge: Cambridge Univ. Press. MR1469220

[21] Caffarelli, L.A. (1992). The regularity of mappings with a convex potential. *J. Amer. Math. Soc.* **5** 99–104.

[22] Carlier, G., Oberman, A. and Oudet, É. (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal.* **49** 1621–1642.

[23] Chartrand, R., Wohlberg, B., Vixie, K.R. and Bollt, E.M. (2009). A gradient descent solution to the Monge–Kantorovich problem. *Appl. Math. Sci.* (*Ruse*) **3** 1071–1080. MR2524965

[24] Chiu, S.N., Stoyan, D., Kendall, W.S. and Mecke, J. (2013). *Stochastic Geometry and Its Applications*. New York: Wiley.

[25] Cuesta-Albertos, J.A., Matrán, C. and Tuero-Diaz, A. (1997). Optimal transportation plans and convergence in distribution. *J. Multivariate Anal.* **60** 72–83.

[26] Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. *Proceedings of the International Conference on Machine Learning* 2014, *JMLR W&CP* **32** 685–693.

[27] Cuturi, M. and Peyré, G. (2016). A smoothed dual approach for variational Wasserstein problems. *SIAM J. Imaging Sci.* **9** 320–343. MR3466197

[28] Dowson, D. and Landau, B. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.* **12** 450–455.

[29] Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester: Wiley. MR1646114

[30] Fiedler, M. (1971). Bounds for the determinant of the sum of Hermitian matrices. *Proc. Amer. Math. Soc.* 27–31.

[31] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10** 215–310. MR0027464

[32] Fréchet, M. (1957). Sur la distance de deux lois de probabilité. *C. R. Math. Acad. Sci. Paris* **244** 689–692.

[33] Freitag, G. and Munk, A. (2005). On Hadamard differentiability in k-sample semiparametric models – With applications to the assessment of structural relationships. *J. Multivariate Anal.* **94** 123–158.

[34] Gallón, S., Loubes, J.-M. and Maza, E. (2013). Statistical properties of the quantile normalization method for density curve alignment. *Math. Biosci.* **242** 129–142. MR3068678

[35] Gangbo, W. and Święch, A. (1998). Optimal maps for the multidimensional Monge–Kantorovich problem. *Comm. Pure Appl. Math.* **51** 23–45.

[36] Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 285–339.

[37] Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika* **40** 33–51. MR0405725

[38] Groisser, D. (2005). On the convergence of some Procrustean averaging algorithms. *Stochastics* **77** 31–60.

[39] Haber, E., Rehman, T. and Tannenbaum, A. (2010). An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM J. Sci. Comput.* **32** 197–211.

[40] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Chichester: Wiley. MR3379106

[41] Huckemann, S., Hotz, T. and Munk, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20** 1–58. MR2640651

[42] Huckemann, S. and Ziezold, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.* 299–319.

[43] Kallenberg, O. (1986). *Random Measures*, 4th ed. Berlin: Akademie-Verlag. MR0854102

[44] Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. MR0442975

[45] Kendall, W.S. (2010). A survey of Riemannian centres of mass for data. In *Proceedings* 59*th ISI World Statistics Congress*.

[46] Kendall, W.S. and Le, H. (2011). Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Braz. J. Probab. Stat.* **25** 323–352. MR2832889

[47] Krantz, S. (2014). *Convex Analysis*. *Textbooks in Mathematics*. Boca Raton: CRC Press.

[48] Le, H. (1998). On the consistency of procrustean mean shapes. *Adv. in Appl. Probab.* 53–63.

[49] Le, H. (2001). Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.* 324–338.

[50] Le, H.L. (1995). Mean size-and-shapes and mean shapes: A geometric point of view. *Adv. in Appl. Probab.* **27** 44–55. MR1315576

[51] Le Gouic, T. and Loubes, J.-M. (2016). Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields* 1–17.

[52] McCann, R.J. (1997). A convexity principle for interacting gases. *Adv. Math.* **128** 153–179.

[53] Molchanov, I. and Zuyev, S. (2002). Steepest descent algorithms in a space of measures. *Stat. Comput.* **12** 115–123.

[54] Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 223–241.

[55] Munk, A., Paige, R., Pang, J., Patrangenaru, V. and Ruymgaart, F. (2008). The one-and multi-sample problem for functional data with application to projective shape analysis. *J. Multivariate Anal.* **99** 815–833.

[56] Oberman, A.M. and Ruan, Y. (2015). An efficient linear programming method for optimal transportation. Preprint. Available at arXiv:1509.03668.

[57] Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* **48** 257–263.

[58] Panaretos, V.M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.* **44** 771–812.

[59] Pass, B. (2013). Optimal transportation with infinitely many marginals. *J. Funct. Anal.* **264** 947–963. MR3004954

[60] Patrangenaru, V. and Ellingson, L. (2016). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. Boca Raton, FL: CRC Press. MR3444169

[61] Pollard, D. (2012). *Convergence of Stochastic Processes*. New York: Springer Science & Business Media.

[62] Rippl, T., Munk, A. and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109.

[63] Rockafellar, R.T. (1970). *Convex Analysis*. *Princeton Mathematical Series*, **28**. Princeton, NJ: Princeton Univ. Press. MR0274683

[64] Rolet, A., Cuturi, M. and Peyré, G. (2016). Fast dictionary learning with a smoothed Wasserstein loss. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* (A. Gretton and C.C. Robert, eds.). *Proceedings of Machine Learning Research* **51** 630–638. Cadiz, Spain.

[65] Schachermayer, W. and Teichmann, J. (2009). Characterization of optimal transport plans for the Monge–Kantorovich problem. *Proc. Amer. Math. Soc.* **137** 519–529. MR2448572

[66] Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T. and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.* **34** 66.

[67] Sommerfeld, M. and Munk, A. (2016). Inference for empirical Wasserstein distances on finite spaces. Preprint. Available at arXiv:1610.03287.

[68] Stein, E.M. and Shakarchi, R. (2005). *Real Analysis*: *Measure Theory*, *Integration*, *and Hilbert Spaces*. *Princeton Lectures in Analysis* **3**. Princeton, NJ: Princeton Univ. Press. MR2129625

[69] Tameling, C., Sommerfeld, M. and Munk, A. (2017). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. Preprint. Available at arXiv:1707.00973.

[70] Villani, C. (2003). *Topics in Optimal Transportation* **58**. Providence: AMS.

[71] Wang, W., Slepčev, D., Basu, S., Ozolek, J.A. and Rohde, G.K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vis.* **101** 254–269. MR3021062

[72] Zemel, Y. and Panaretos, V.M. (2019). Supplement to "Fréchet means and Procrustes analysis in Wasserstein space." DOI:10.3150/17-BEJ1009SUPP.

[73] Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44** 2281–2321. MR3546451

Received January 2017 and revised November 2017