Original Version: February 2016
UpdatedVersion: August 2016

FRÉCHET MEANS IN WASSERSTEIN SPACE: GRADIENT DESCENT AND PROCRUSTES ANALYSIS

By Yoav Zemel and Victor M. Panaretos

Ecole Polytechnique Fédérale de Lausanne

Given a collection of absolutely continuous measures on $\mathbb{R}^d$, we consider the complementary problems of: determining their Fréchet mean with respect to the L^2-Wasserstein metric; and of constructing their optimal multi-coupling in the mean square sense. While both problems are of intrinsic probabilistic interest, they are also at the core of several statistical challenges in image analysis, multivariate dependence, nonparametric registration, and spatial statistics, to name but a few. And, although they admit simple and explicit solutions when $d = 1$, they have been elusive when $d \geq 2$. To tackle them, we exploit the relationship between the tangent bundle geometry of Wasserstein space and the class of optimal transport maps. This allows us to derive the gradient of the Fréchet functional, and to characterise Karcher means and their relation to Fréchet means. These results are then used to construct the empirical mean via gradient descent. We determine the optimal descent step size, and show that this renders descent equivalent to a classical Procrustes algorithm. The key advantage of this algorithm is that it only requires successive solutions to pairwise optimal transportation problems. We prove that the algorithm converges to a Karcher mean in Wasserstein distance, regardless of the starting point; and that the corresponding Procrustes registration maps converge uniformly. Motivated by the problem of registering warped multidimensional point processes, we also consider the case of discrete observation, where one only observes a finite sample or a point process from each measure. We construct regularised estimators that are consistent for the Fréchet mean, and uniformly consistent for the Procrustes registration maps.

1. Introduction. In nonparametric statistics for stochastic processes, one is concerned with making inferences on the law of a process $\{X(t) : t \in K\}$, on the basis of a sample of N realisations, $\{X_1, \dots, X_N\}$, often observed discretely, over a compact set $K \subset \mathbb{R}^d$. The case that has been most studied by far is that of *Functional Data Analysis* (Ramsay and Silverman [52], Horváth & Kokoszka [37], Hsing & Eubank [38]) where the process X is modelled as an element of a normed vector space, typically a separable Hilbert space with $d = 1$. This setting captures an extremely rich variety of practical problems, ranging from growth curves and electricity consumption, to DNA mechanics and computational linguistics (see Wang et al. [57] for a compendious overview). Still, there are important instances where a linear space is not the most natural framework to

[*] Research Supported by a European Research Council Starting Grant Award.

AMS 2000 subject classifications: Primary 62M30; 60G57; 60D05; secondary 62G07; 60G55

Keywords and phrases: Functional Data Analysis, Monge–Kantorovich problem, Optimal Transportation, Phase Variation, Point Process, Random Measure, Registration, Warping

model the law of a stochastic process $\{X(t) : t \in K\}$. A key such exception, with manifold manifestations in neurophysiology, imaging, and environmetrics, is that of processes that are best modelled as *distributions* over the domain $K \subset \mathbb{R}^d$, that is collections of *random measures* $\{\mu^1, \dots, \mu^N\}$ rather than *random functions* $\{X_1, \dots, X_N\}$ (Chiu et al. [22], Kallenberg [39]). The total intensity itself may vary from realisation to realisation, but this is often *not* the main source of variation to be modelled and understood. Rather, the key variational feature of interest to the statistician is frequently that of *deformation* (originating in *morphometrics*, Bookstein [18]). In fact, it is precisely deformations that allow for more natural interpolation, extrapolation, and averaging of images.

Deformation-based variation may be the only source of variation: in this case, every manifestation of the random function can be usually modelled as the deformation of a template, according to a randomly selected element of a class of transformations (e.g. Allassonnière et al. [5]). Or, it may be a first layer of variation, corresponding to a deformation of the coordinate system, followed by a second layer where further stochasticity is injected (for instance additive variation, if an image is observed with noise corruption, e.g. Amit et al. [8], or sampling-based variation, if one observes samples from the density corresponding to the deformed template, e.g. Panaretos & Zemel [49]). In the first case, recovering the template and the transformations is the main focus of the analysis. In the second case, the recovery of these two components corresponds to the problem of *registration*, where one selects a common coordinate system, thus removing the effects of warping/deformation, and then carries out a separate analysis of the additional level of variation.

In either case, it is crucial to be able to estimate the underlying template, which can be modelled as a Fréchet mean with respect to some metric structure; and to use the Fréchet mean to recover the deformation maps, that *register* the individual realisations $\{\mu^1, \dots, \mu^N\}$ to their Fréchet mean. Often, finding the Fréchet mean and finding the registration maps are interwoven problems. These interwoven problems generalise the concept of a Procrustes analysis (Gower [33]; Dryden & Mardia [25]), as carried out in shape theory: Euclidean configurations are replaced by measures, and the group of rigid motions (or similarities) is replaced by a class of deformations. Obviously, the methods and algorithms for estimating a mean and carrying out a registration/Procrustes analysis are inextricably linked with the geometry one considers for the measures, which can be a matter of modelling choice or of first principles. In this paper, we choose to study the problem of *averaging* and *registration* when the measures $\{\mu^1, \dots, \mu^N\}$ are normalised to a common intensity (assumed 1) and are viewed as elements of the L^2-Wasserstein space. We choose this setting as it can be seen to be the natural analogue of using $L^2[0, 1]$ for functions X, in the case of measures. This analogy is valid in a very strong sense when $K = [0, 1] \subset \mathbb{R}$. In this case, it can be seen that the L^2-Wasserstein metric induces the $L^2[0, 1]$ geometry on the *quantile functions* of the measures $\{\mu^1, \dots, \mu^N\}$. This allows one to explicitly determine the Fréchet mean, and use it as a template for *registration*, thus producing a Procrustes Analysis in the space of measures, which can be elegantly interpreted through the prism of probabilistic coupling (e.g. Bolstad et al. [16], Gallon et al. [30], Dupuy et al. [26]). Indeed, Panaretos & Zemel [49] show that the choice of the Wasserstein space is essentially *the canonical choice* when modelling deformations of measures and point processes on $[0,1]$, and exploit the connection to L^2 in order to determine Fréchet means, registration maps, and related central limit theorems. Similar arguments have also been made for $K \subseteq \mathbb{R}^d$, where the Wasserstein setting has been shown to possess desirable properties, at least for parametric families of random measures (see, e.g. Bigot & Klein [13], Agulló-Antolín et al. [3]).

In the language of Wasserstein space and optimal transportation, *averaging* is the problem of finding a Fréchet mean, whereas *registration* is the problem of constructing an optimal multicoupling. Both problems admit simple and explicit solutions in the case of measures on $\mathbb{R}$, owing to the flat nature of the Wasserstein space when $d = 1$. However, no explicit solutions are possible in general in the multivariate case, and the construction of algorithms yielding the Fréchet mean and/or the optimal multicoupling has been elusive. Progress has been made by considering restrictions and/or variants of the problem that allow relaxation to a tractable version (Boissard et al. [15], Bonneel et al. [17], Cuturi & Doucet [23]; see Section 3 for more details). The purpose of this paper is to constructively solve the Fréchet averaging *and* optimal multicoupling problem without such workarounds.

Our main contributions are:

1. We show how knowledge of the Fréchet mean gives an explicit solution to the optimal multicoupling problem (Section 3), by coupling each sample measure to the Fréchet mean (Theorem 1). This reduces the problem of multicoupling to determining the mean and the Procrustes maps mapping it to each sample measure.
2. We determine the gradient of the Fréchet functional (Section 4.2, Theorem 2), and characterise Karcher means via its zeroes (Corollary 1, Section 4.3). We also give criteria for determining when a Karcher mean is a Fréchet mean (Theorem 3).
3. We construct a gradient descent algorithm (Algorithm 1), find its optimal stepsize (Lemma 2), and show that with this stepsize, it is equivalent to a Procrustes algorithm (Section 5). This reduces the determination of the mean to the successive solution of pairwise optimal transport problems.
4. We provide a convergence analysis of the algorithm (Sections 5.2 and 5.3). In particular we prove that the gradient iterate converges to a Karcher mean in the Wasserstein metric (Theorem 4); and that the the induced transportation maps converge uniformly to the Procrustes maps (required for optimal mutlicoupling; Theorem 5).
5. We prove that our results are stable under discrete observation (Section 6). That is, if one does not observe the actual measures $\{\mu^1, \dots, \mu^N\}$, but random samples or point processes with these measures as distributions/intensities, we construct regularised nonparametric estimators of the Fréchet means and Procrustes maps, and prove that they are consistent as sample size increases (Theorems 6 and 7).

Before presenting our main results, we first provide a short introduction to Wasserstein space in Section 2. Our results are then developed in Sections 3 through 6. Section 7 gathers all proofs, for the sake of tidiness, and Section 6.4 studies several examples. An online Supplement [62] provides further details omitted from the main paper.

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 [56]): 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

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

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 [56]). 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, see e.g. Bickel & Freedman [11]).

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

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

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. But they can be assumed to be exactly optimal, i.e. $\mathbf{q}_i = \mathbf{t}_\mu^{\mu^i}$ as a matter of parsimony, and in any case without loss of generality. 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 ν.

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 nondecreasing 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 7.5), 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 [56, 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. Fréchet Means and Optimal Multicoupling.

The notion of a Fréchet mean (Fréchet [28]) 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 [40]; Le [44]), a term that describes stationary points of the sum of squares functional, when the latter is differentiable. See Kendall [41], and Kendall & Le [42] for an overview and a detailed review, respectively. In the context of Wasserstein space, a Fréchet mean of a collection of measures $\{\mu^1, \ldots, \mu^N\}$, is a minimiser of the Fréchet functional

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

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. Interestingly, Fréchet himself [29] considered the Wasserstein metric between probability measures on $\mathbb{R}$, and some refer to this as the *Fréchet distance* (e.g. Dowson & Landau [24]). In general, existence and uniqueness of a Fréchet mean can be subtle, but Agueh & Carlier [2] have shown that it *will uniquely exist* in the Wasserstein space, provided that some regularity is asserted. 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 & 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 & Święch [31]. 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 shows precisely how:

THEOREM 1 (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,

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

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 [32]). 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, similarly to "quantile renormalisation" in one dimension, e.g. Bolstad et al. [16], Gallon et al. [30]. The Procrustes maps can also be used in order to produce a Principal Component Analysis, capturing the main modes of deformation-based variation (Boissard et al. [15], Bigot et al. [12], Panaretos & Zemel [49]).

Clearly, the problem of determining the Fréchet mean and the problem of Procrustes analysis are complementary: solution of one immediately gives a solution to the other. This complementarity is exemplified in shape theory, where Procrustes analysis is used as the canonical method *precisely* to find the mean with respect to the Procrustes metric. Still, with the exception of the case $d = 1$, neither problem can be solved explicitly in the Wasserstein space. Important progress has been made by considering *work-arounds* that modify, restrict, or approximate the problem (Cuturi & Doucet [23]; Bonneel et al. [17]; Boissard et al. [15]), but one might ask for algorithms with general applicability and with provable convergence guarantees.

Our main contribution will be to provide a solution to *both* the determination of the Fréchet mean *and* the registration maps (i.e. the multi-coupling), avoiding such work-arounds. Our departure point is the Procrustean heuristic: in Procrustes analysis, one typically starts from an arbitrary template, and sequentially registers every observation in a pairwise fashion to that template; once all observations are registered, they are averaged, producing an updated template (Gower [33]; Dryden & Mardia [25, p. 90]). The same idea could be applied in the Wasserstein space, precisely in order to use the feasibility of the pairwise problem. However, there is a priori no guarantee that this approach would work (convergence of Procrustes algorithms is subtle even in finite dimensions, see for instance Le [44, 45] and Groisser [34]). The key will be to connect Procrustes analysis to gradient descent: this is done in the next section.

4. Wasserstein Geometry and the Gradient of the Fréchet Functional.

In this section, we determine the conditions for the Fréchet derivative of the Fréchet functional (Equation (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.

4.1. *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

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

This curve is known as McCann's interpolation (McCann [47], Villani [56]). 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])

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

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]:

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

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 $\left[\mathbf{i} + t(\mathbf{t}_{\mu^0}^{\mu^1} - \mathbf{i})\right]\#\mu^0$ 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.

4.2. *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

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

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

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

We have proven:

Theorem 2 (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*

$$F(\gamma) = \frac{1}{2N}\sum_{i=1}^N d^2(\gamma, \mu^i), \qquad \gamma \in \mathcal{P}_2(\mathbb{R}^d). \tag{4.1}$$

is Fréchet-differentiable, and its gradient satisfies

$$F'(\gamma) = -\frac{1}{N}\sum_{i=1}^N \log_\gamma(\mu^i) = -\frac{1}{N}\sum_{i=1}^N \left(\mathbf{t}_\gamma^{\mu_i} - \mathbf{i}\right). \tag{4.2}$$

4.3. *Karcher and Fréchet Means.* We can now characterise Karcher means, and also show that the empirical Fréchet mean must be sought amongst them:

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*

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

Furthermore, the Fréchet mean of $\{\mu^i\}$ is itself a Karcher mean, i.e. satisfies $F'(\mu) = 0$ μ-almost everywhere.

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 3 (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 3, then we conjecture the Fréchet functional F to in fact have a unique Karcher mean, coinciding with the Fréchet mean.*

5. Gradient Descent and Procrustes Analysis.

5.1. *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 follow the steepest descent of F given by the negative of the gradient:

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

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 2, 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*

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

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 [33]; Dryden & Mardia [25, p. 90]):

Algorithm 1 Gradient Descent via Procrustes Analysis

```
(A) Set a tolerance threshold $\epsilon > 0$.
(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 $\gamma_{j+1} = T_j \# \gamma_j$, i.e. push-forward $\gamma_j$ via $T_j$ to obtain $\gamma_{j+1}$.
(F) If $\|F'(\gamma_{j+1})\| < \epsilon$, 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).
```

(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 $\{\mu^1, \dots, \mu^N\}$, i.e. 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 optimal 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. [35], Chartrand et al. [20]). 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. 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. The gradient/Procrustes iteration is presented succinctly as Algorithm 1.

5.2. *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 [45], Goissard [34]) 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 4 (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:

COROLLARY 2 (Wasserstein Convergence of Gradient Descent). *Under the conditions of Theorem 4, 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 4 with Theorem 3 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 4 is elaborate, and is constructed via a series of intermediate results in a separate section (Section 7.3.1) 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.

5.3. *Uniform Convergence of Procrustes Maps.* 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 Theorem 1), 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 5 (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*

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

for any pair of compacta $\Omega_1\subseteq A$, $\Omega_2\subseteq B^i$, where the sequence $\mathbf{t}_{\mu^i}^{\gamma_j}$ and $\mathbf{t}_{\gamma_j}^{\mu^i}=\left(\mathbf{t}_{\mu^i}^{\gamma_j}\right)^{-1}$ 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 $B^1=\cdots=B^N$.

6. Discrete Observation and Registration of Warped Point Processes. In practice, it may be the case that we are unable to fully observe the measures $\{\mu^1,\dots,\mu^N\}$.

Instead, we may have access to discrete versions, $\{\mu'_1, \dots, \mu'_N\}$, where each μ'_i is a finite measure corresponding to a sample of size n_i from μ^i. More generally, we might have point processes $\{\widetilde{\Pi}_1, \dots, \widetilde{\Pi}_N\}$ on $\mathbb{R}^d$ with corresponding mean measures $\{\Lambda_1, \dots, \Lambda_N\}$, where we have switched to the notation $\Lambda_i = \mu^i$ that is perhaps more natural in the context of point process intensities. Two natural questions that extend the development in Section 3 are:

1. Is it possible to still construct a proxy for the Fréchet mean of $\{\Lambda_1, \dots, \Lambda_N\}$, based on $\{\widetilde{\Pi}_1, \dots, \widetilde{\Pi}_N\}$?
2. If $\{\widetilde{\Pi}_1, \dots, \widetilde{\Pi}_N\}$ are viewed as warped point processes, is there a way to register them by carrying out a Procrustes analysis?

In fact, the problem of registration of point processes within a functional data analysis context, number (2) above, is a topic of intense current research (Panaretos & Zemel [49], Wu & Srivastava [61], Wu et al. [58], Patriarca et al. [50], Cheng et al. [21], Lu & Marron [46], Hadjipantelis et al. [36]). Panaretos & Zemel [49] show that, in a certain sense, the two problems (1) and (2) are complementary, and the canonical solution to (2) requires the solution to (1), similarly to the discussion in Section 3. Progress so far, however, has been restricted to the one-dimensional case, driven in most cases by the modelling of neuronal firing times (Wu & Srivastava [59, 60]). Nevertheless, multidimensional situations are clearly of potential interest: rather than registering the firing times of a specific neuron over N individuals, one may wish to jointly register the firing times of d neurons per individual.

We will describe how both (1) and (2) can be solved, by means of the techniques introduced earlier in the paper.

6.1. *Discretely Observed Random Measures.* 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[1], and let $\{\Pi_1, \dots, \Pi_N\}$ be i.i.d point processes with intensity measure λ,

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

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] 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).

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 importance and canonicity of these two assumptions has been discussed in depth in Panaretos & Zemel [49, Section 3.3]. In brief, unbiasedness stipulates that the warp maps do not deform the *template* λ on average – in other words, the intensity measure $T_i\#\lambda$ of $\widetilde{\Pi}_i$ has Fréchet mean λ. Regularity assumes that the deformation maps are identifiable, by asking them to be optimal transport maps (in view of Brenier's characterisation). Notice that this is a reasonable assumption for an additional reason: it requires that orientation be preserved (for instance, if $K = [0,1]^2$ representing the time domain for the firing times of two neurons, it is not allowed that a warp map affect the lexicographic ordering of the firing times; see Section 7.5).

6.2. *Regularised Procrustes Registration.* Our strategy will be to estimate the unknown structural mean measure λ, and the 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. The following Proposition guarantees that the estimands considered are identifiable.

Proposition 2 (Identifiability). *Let Λ be a random probability measure on $\mathbb{R}^d$, $d \geq 1$. If Λ is almost surely regular and supported on a convex compact set K, the population Fréchet functional*

$$\gamma \mapsto \mathbb{E}\left[d^2(\Lambda,\gamma)\right],$$

is strictly convex over probability measures γ on $\mathbb{R}^d$, and has a (unique) minimum λ. Consequently, the optimal transportation map $T := \mathbf{t}_\lambda^\Lambda$ is uniquely defined almost surely.

Generalising the approach of Panaretos & Zemel [49], our procedure follows the steps:

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

$$\widehat{\Lambda}_i = \frac{1}{m}\sum_{j=1}^{m} \frac{\delta\{x_j\} * \psi_\sigma}{[\delta\{x_j\} * \psi_\sigma](K)}\bigg|_K, \tag{6.1}$$

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.

6.3. *Consistency.* 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 6.1. 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 $\sigma_n = \max_i \sigma_i^{(n)} \xrightarrow{\text{p}} 0$ then*

1. *For any i,*
$$d(\widehat{\Lambda}_i, \Lambda_i) \xrightarrow{\text{p}} 0;$$
2. *The estimator $\widehat{\lambda}_n$ is strongly consistent*
$$d(\widehat{\lambda}_n, \lambda) \xrightarrow{\text{as}} 0.$$

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 $\mathbb{E}\left[\Pi_1^{(1)}\right]^4 < \infty$, $\sum_n \tau_n^{-2} < \infty$ and $\sigma_n \xrightarrow{\text{as}} 0$ 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.*

Our next two results concern the (uniform) consistency of the Procrustes registration procedure. Though the results themselves parallel their one-dimensional counterparts, 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:

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

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

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

COROLLARY 3 (Consistency of Procrustes Registration). *Under the same conditions of Theorem 6, the registration procedure is consistent: for any i*

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

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

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 [43, 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.

6.4. *Illustrative Examples.* As an illustration, we implement Algorithm 1 in several scenarios for which pairwise optimal maps can be calculated explicitly at every iteration, allowing for fast computation without error propagation. More details on the calculations and properties of each individual scenario can be found in Section 3 of the supplement [62].

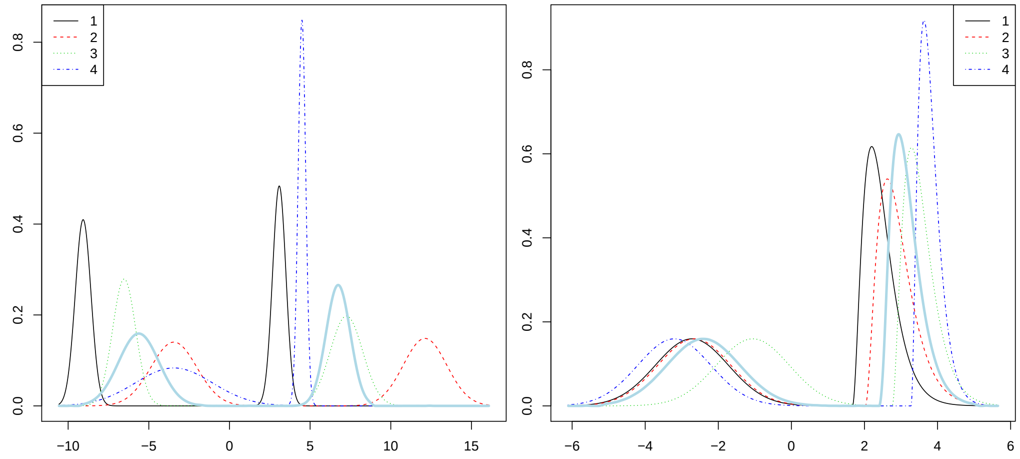


FIG 1. *Densities of bimodal Gaussian mixture (left) and a mixture of Gaussian with gamma (right), with the Fréchet mean density in light blue.*

6.4.1. *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

$$f^i(x) = \frac{1}{2}\phi\left(\frac{x - m_1^i}{\sigma_1^i}\right) + \frac{1}{2}\phi\left(\frac{x - m_2^i}{\sigma_2^i}\right), \tag{6.2}$$

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

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

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

$$f^i(x) = \frac{3}{5}\frac{\beta_i^3}{\Gamma(3)}(x - m_3^i)^2 e^{-\beta_i(x-3)} + \frac{2}{5}\phi(x - m_4^i), \tag{6.3}$$

with

$$\beta^i \sim Gamma(4, 1), \quad m_3^i \sim U[1, 4], \quad m_4^i \sim U[-4, -1].$$

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 $\mu^1, \ldots, \mu^N$ 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.

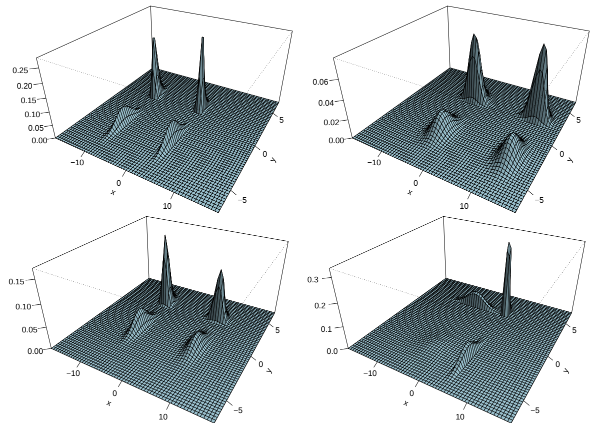


Fig 2. *Density plots of the four product measures of the measures in Figure 1.*

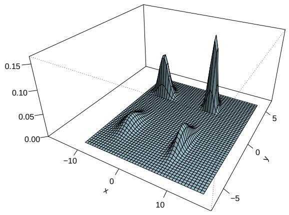


Fig 3. *Density plot of the Fréchet mean of the measures in Figure 2.*

6.4.2. *Independence.* We next take measures μ^i on $\mathbb{R}^2$, having independent marginal densities f_X^i as in (6.2), and f_Y^i as in (6.3). 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 [62] that, unsurprisingly, the Fréchet mean is a product measure.

6.4.3. *Common Copulas.* Let μ^i be a measure on $\mathbb{R}^2$ with density

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

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 (6.2), f_Y^i as in (6.3), 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 density, having four clearly distinguished peaks. Figure 8(b), depicting the resulting Procrustes maps, allows for a clearer interpretation:

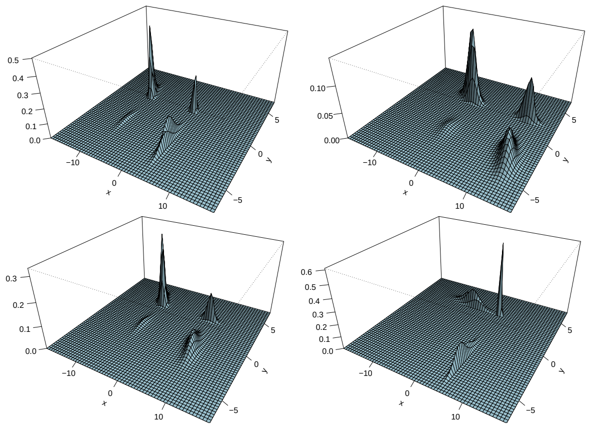


FIG 4. *Density plots of four measures in $\mathbb{R}^2$ with Frank copula of parameter -8.*

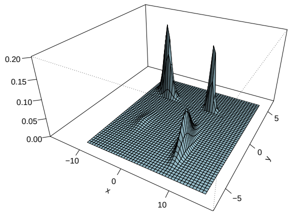


FIG 5. *Density plot of the Fréchet mean of the measures in Figure 4.*

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. [15] tackle the problem of finding the Fréchet mean in such a setting, by means of the *iterated barycentre*. In the supplementary material [62] 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, admissibility 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.

6.4.4. *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 [24]; Olkin and Pukelsheim [48])

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

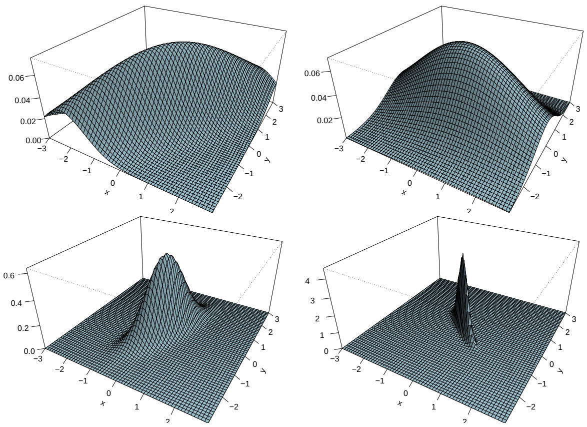


FIG 6. *Density plot of four Gaussian measures in* $\mathbb{R}^2$*.*

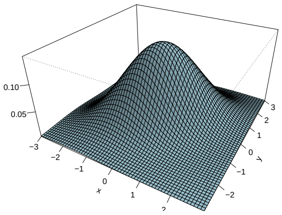
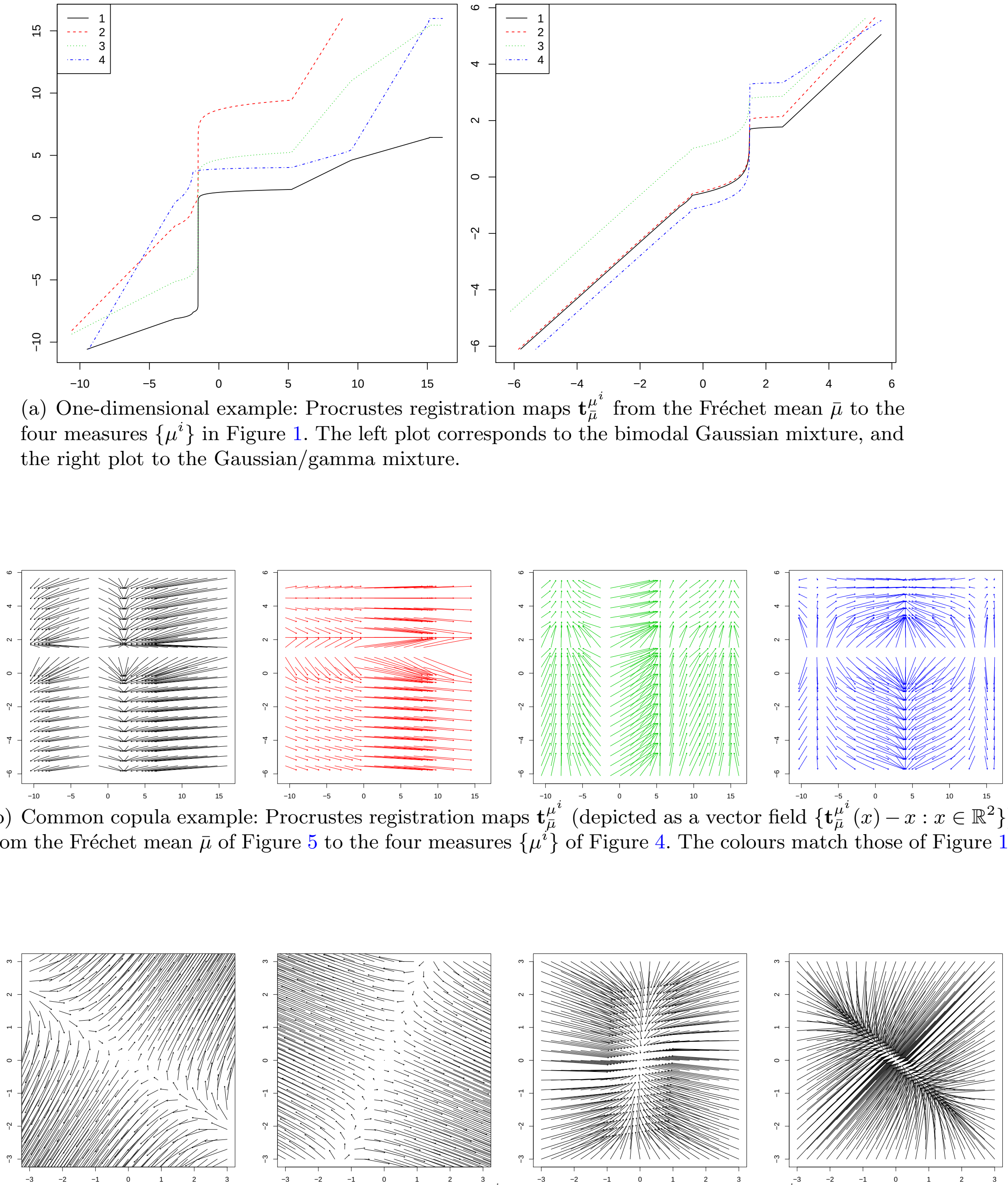


FIG 7. *Density plot of the Fréchet mean of the measures in Figure 6.*

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 [62] we prove that γ_k must converge to the unique Fréchet mean, which is also a Gaussian measure.

Notice that the Gaussian measures $\{\mu^i\}$ will be compatible if $S_iS_j = S_jS_i$, 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 $S_i \sim \mathrm{Wishart}(I_2, 2)$, 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.

6.4.5. *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



(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).

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

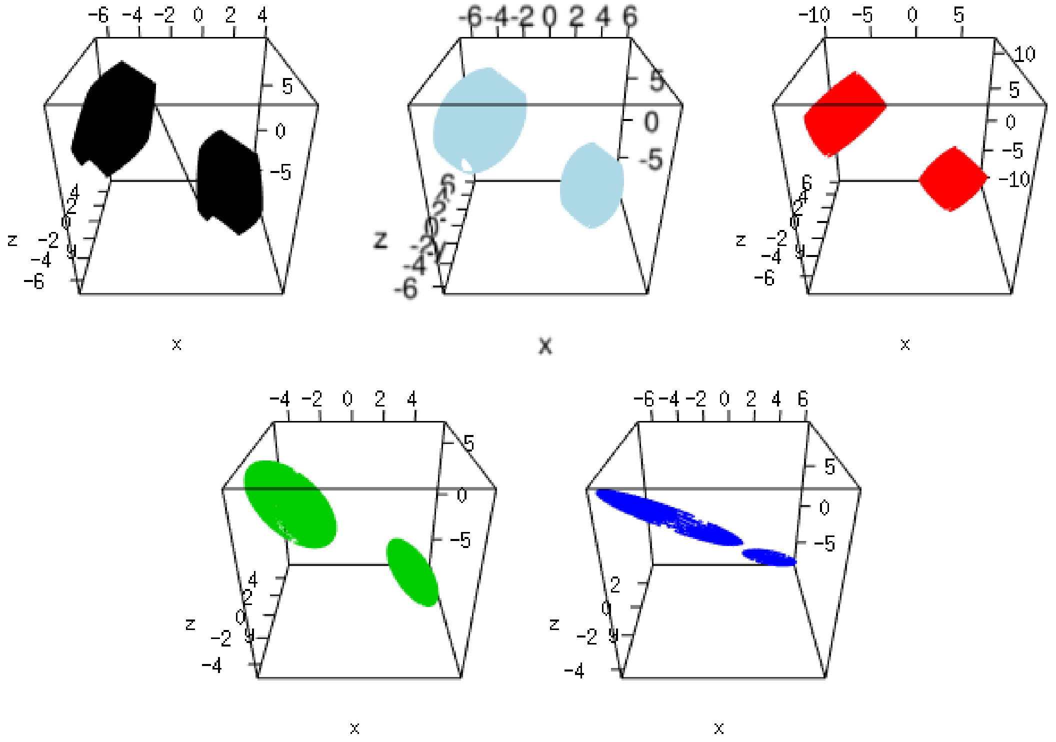


Fig 9. *The set* $\{v \in \mathbb{R}^3 : g^i(v) = 0.0003\}$ *for* $i = 1$ *(black), the Fréchet mean (light blue),* $i = 2, 3, 4$ *in red, green and dark blue respectively.*

with columns U_1, U_2, U_3 and let μ^i have density

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

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 (6.2) 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 [62] 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 the 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.

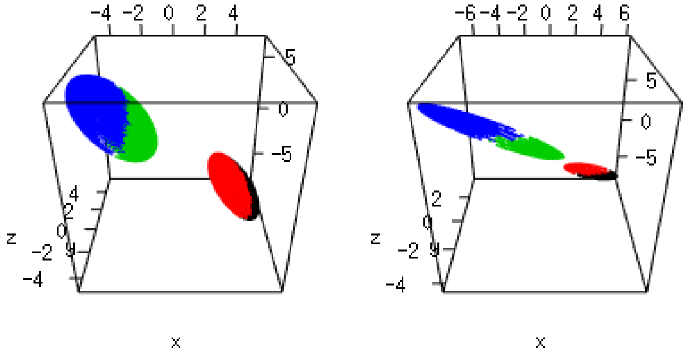


Fig 10. *The set* $\{v \in \mathbb{R}^3 : g^i(v) = 0.0003\}$ *for* $i = 3$ *(left) and* $i = 4$ *(right), with each of the four different inverses of the bimodal density* f^i *corresponding to a colour.*

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. [35] or Chartrand et al. [20] to obtain the N optimal maps at each iteration. 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. 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 non-trivial. For tidiness, we will state and prove these results separately at the end of this section (Section 7.5), developing our main results first, and referring to the analytical background when necessary.

7.1. *Proofs of Statements in Section 3.*

Proof of Theorem 1. 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

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

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

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

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

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

Thus ξ' is optimal. □

7.2. *Proofs of Statements in Section 4.*

Proof of Corollary 1. The characterisation of Karcher means is immediate from Theorem 2. 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 [62]. □

Proof of Theorem 3. The result exploits Caffarelli's regularity theory for Monge–Ampère equations. In the first case, by Theorem 4.14(iii) in Villani [56] 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 a theorem of Caffarelli [19], 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 in $\mathcal{P}_2(X)$ (see the proof of Proposition 2). □

7.3. *Proofs of Statements in Section 5.*

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

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

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 $A = \cup A_i$, 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 $|\mathrm{det}\nabla\mathbf{t}_{\gamma_0}^{\mu^i}| > 0$ 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

$$2NF(\gamma_0) = \sum_{i=1}^N d^2(\gamma_0, \mu^i) = \sum_{i=1}^N \int_{\mathbb{R}^d} \|S_i - \mathbf{i}\|^2 \,\mathrm{d}\gamma_0 = \sum_{i=1}^N \langle W_i, W_i\rangle = \sum_{i=1}^N \|W_i\|^2, \tag{7.1}$$

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

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

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 (7.1) 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. □

7.3.1. *Proof of Theorem 4.* 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

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

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

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

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 [56], it suffices to show that

$$\lim_{R\to\infty} \sup_{j\in\mathbb{N}} \int_{\{x:\|x\|>R\}} \|x\|^2 \, \mathrm{d}\gamma_j(x) = 0. \tag{7.2}$$

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

$$\int_{\mathbb{R}^d} \|x\|^3 \, \mathrm{d}\mu^i(x) \le M(3), \qquad i = 1, \dots, N. \tag{7.3}$$

In Section 2 of the supplementary material [62] we show that (7.2) holds even if (7.3) does not.

For any $j \ge 1$ it holds that

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

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

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

and (7.2) 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 5, 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 [4] 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 (7.2) by [56, Theorem 7.12]. It is easy to see that this implies the uniform absolute continuity

$$\forall \epsilon > 0 \exists \delta > 0 \forall j \geq 1 \forall A \subseteq \mathbb{R}^d \text{ Borel}: \quad \gamma_j(A) \leq \delta \quad \Longrightarrow \int_A \|x\|^2 \, \mathrm{d}\gamma_j(x) < \epsilon. \tag{7.4}$$

The δ's can be chosen in such a way that (7.4) holds true for the finite collection $\{\mu^1, \dots, \mu^N\}$ as well. Fix $\epsilon > 0$, set $\delta = \delta_\epsilon$ as in (7.4), and let $A_n = \{x : g_n(x) \geq 4R\}$, where $R = R_\epsilon \geq 1$ is such that (using (7.2))

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

The bound

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

implies that

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

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

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

so that $\gamma_n(A_n) < \delta$. We use this in conjunction with (7.4) to bound

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

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

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

The analogous truncated function g_R satisfies

$$0 \le g_R(x) \le 4R \quad \forall x \in \mathbb{R}^d \quad \text{and} \quad \{x : g_R \text{ is continuous }\} \text{ is of } \gamma\text{-full measure.} \tag{7.5}$$

Let $E = \mathrm{supp}(\gamma)$. Proposition 6 (Section 7.5) implies pointwise convergence of $\mathbf{t}_{\gamma_n}^{\mu^i}(x)$ to $\mathbf{t}_{\gamma}^{\mu^i}(x)$ for any $i = 1, \dots, N$ and any $x \in E \setminus \mathcal{N}$, where $\mathcal{N} = \cup_{i=1}^{N} \mathcal{N}^i$ and

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

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 $\gamma_n(\mathbb{R}^d \setminus K_\epsilon) < \epsilon/R$ for all n. Clearly, $g_{n,R} \to g_R$ on $E' = K_\epsilon \cap E \setminus \mathcal{N}$, and by Egorov's theorem (valid as $\mathrm{Leb}(E') \le \mathrm{Leb}(K_\epsilon) < \infty$), there exists a Borel set $\Omega = \Omega_\epsilon \subseteq E'$ on which the convergence is uniform, and $\mathrm{Leb}(E' \setminus \Omega) < \epsilon/R$. 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 (7.5) and the Portmanteau lemma (Lemma 9, Section 7.5). 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 $\mathcal{N} \cup (E' \setminus \Omega) \cup (\mathbb{R}^d \setminus E) \cup (\mathbb{R}^d \setminus K_\epsilon)$, 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 $\gamma_n(\mathbb{R}^d \setminus K_\epsilon) < \epsilon/R$ that

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

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, $A_k \cap \mathrm{supp}(\gamma) = \emptyset$ and $\gamma_n(A) \le C\mathrm{Leb}(A)$. Consequently, for all k

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

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

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

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 $i = 1, \ldots, N$ 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 $\{\mu^1, \ldots, \mu^N\}$.*

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

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

Fiedler [27] 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

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

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

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

Thus γ_1-almost surely,

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

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 4 (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, $\liminf\gamma_k(O)\leq C_\mu\mathrm{Leb}(O)$, so any limit point γ of (γ_k) is such that $\gamma(O)\leq C_\mu\mathrm{Leb}(O)$ 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 5. Let $E=\mathrm{supp}(\bar\mu)$ and set $A^i=E^{\mathrm{den}}\cap\{x:\mathbf{t}_{\bar\mu}^{\mu^i}(x)\text{ is multivalued}\}$. By Corollary 5 $\bar\mu(A^i)=1$. Choose $A=\cap_{i=1}^N A^i$ and apply Proposition 6. This proves the first assertion.

Now let $E^i=\mathrm{supp}(\mu^i)$ 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 $E^1=\dots=E^N$ then $\mu^i(B)=1$ for $B=\cap B^i$. □

7.4. *Proofs of Statements in Section 6.*

PROOF OF PROPOSITION 2. This is essentially a consequence of Corollary 2.9 in Álvarez–Esteban et al. [6], and we provide the details in Section 4 of the supplementary material [62]. □

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* (6.1) *satisfies*

$$d^2\left(\widehat{\Lambda}_i, \frac{\widetilde{\Pi}_i}{\widetilde{\Pi}_i(K)}\right) \le C_{\psi,K}\sigma^2 \qquad \textit{if } \sigma \le 1 \qquad \textit{and} \qquad \widetilde{\Pi}_i(K) > 0, \tag{7.6}$$

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

We prove the lemma in the supplementary material [62, Section 4].

REMARK 4. *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 [49, 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*

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

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

PROOF OF THEOREM 6. The proof is very similar to the proof of Theorem 1 in Panaretos and Zemel [49], and we give the details in the supplementary material [62]. □

PROOF OF THEOREM 7. The argument is considerably different than the case $d = 1$ considered in [49], 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

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

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 $\mathrm{int}(K)$ follows from the assumptions that T_i and T_i^{-1} are continuous. The finiteness in (7.7) 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 $\Omega \subseteq \mathrm{int}(K)$: 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 (7.7). □

Proof of Corollary 3. The square of the distance is

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

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 $\Omega \subseteq \mathrm{int}(K)$ be compact and split the integral to Ω and its complement. Then

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

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

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

and $T_i(\Omega)$ is compact, we only need to show that it is included in $\mathrm{int}(K)$ in order to apply Theorem 7. Suppose towards contradiction that $y = T_i(x) \in \partial K$ for $x \in \mathrm{int}(K)$. 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 $x' \in \mathrm{int}(K)$. Then $y' = T_i(x') \in K$, so that

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

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

7.5. *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 4 (more precisely, Proposition 3) and Theorem 5. 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 [4]. 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

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

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:

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

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 [4, Section 7], where

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

is the *subdifferential* of φ at x. Here $u(x) = \emptyset$ 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 [4, Corollary 1.3(4)]. Notice that this result implies that differentiable convex functions are continuously differentiable [53, 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 $\epsilon > 0$ there exists $t_\epsilon > 0$ such that

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

We denote the set of points of Lebesgue density of G by G^{den}. Clearly, G^{den} lies between $\mathrm{int}(G)$ and $\overline{G}$. Stein and Shakarchi [55, Chapter 3, Corollary 1.5] show that almost any point of G is in G^{den}. By the Hahn–Banach theorem, $G^{\mathrm{den}} \subseteq \mathrm{int}(\mathrm{conv}(G))$.

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*

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

This result was given as an exercise in [55]; for completeness we provide a full proof in the supplementary material [62].

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*

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

Then $y^* = y_0$. In particular, the result is true if the inequality holds on $G = O \setminus \mathcal{N}$ with $\emptyset \neq O$ 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 $y_t \in u(x_t)$,

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

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

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

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 [56, 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 [4, 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

$$\begin{array}{llll}\mu_n \to \mu & \pi_n \text{ finite} & \text{optimal for } \mu_n, \nu_n & (u_n = \partial\varphi_n)\#\mu_n = \nu_n \\ \nu_n \to \nu & \pi \text{ unique} & \text{optimal for } \mu, \nu & (u = \partial\varphi)\#\mu = \nu.\end{array} \tag{7.7}$$

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 [54, Theorem 3]. 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, $\liminf \mu_k(G) \ge \mu(G)$;
(II) for any closed set F, $\limsup \mu_k(F) \le \mu(F)$;
(III) $\int h \, d\mu_k \to \int h \, d\mu$ 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 [14, Theorem 2.1]; for the third, see Pollard [51, 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 (7.7), denote $E = \mathrm{supp}(\mu)$.*

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

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

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

Corollary 5 (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

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

(Since u is nonempty on $\mathrm{int}(\mathrm{conv}(E))$ by [4, Corollary 1.3(2)].) Both sets are Lebesgue-negligible (see [4, Remark 2.3] for the latter), and μ is absolutely continuous. □

Remark 5. *In the setting of Theorem 7, E is convex, μ is absolutely continuous, and u is univalued on $\mathrm{int}(E)$, so one can take any $\Omega \subseteq \mathrm{int}(E)$, 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 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 (7.7). For any $x_0 \in \mathrm{supp}(\mu)$ 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 [4, Section 7] that is univalued at x_0, it is continuous there (Proposition 5). This means that for any $\epsilon > 0$ 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 $\|y - y_0\| < \epsilon$. Take $\epsilon_k \to 0$ and corresponding $\delta_k \to 0$, and set $B_k = B_{\delta_k}(x_0)$, $V_k = B_{\epsilon_k}(y_0)$. Then $u(B_k) \subseteq V_k$, so

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

because B_k is a neighbourhood of $x_0 \in \mathrm{supp}(\mu)$. 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

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

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 $\|y_{n_k} - y_0\| < \epsilon_k$. Hence $(x_{n_k}, y_{n_k}) \to (x_0, y_0)$. □

Lemma 11 (Limit points are in the limit graph). *Assume that (7.7) holds and denote $E = \mathrm{supp}(\mu)$. 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 $x_0 \in \mathrm{int}(E)$ 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 [4, 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 $B_\epsilon^\infty(x_0) = \{x : \|x - x_0\|_\infty < \epsilon\}$ be the ℓ_∞ ball around x_0 and $\overline{B}_\epsilon^\infty(x_0)$ its closure.

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

For a proof, see the supplementary material [62].

Proposition 7 (Boundedness). *Suppose that (7.7) holds, and fix a compact $\Omega \subseteq \mathrm{int}(\mathrm{conv}(\mathrm{supp}(\mu)))$. 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 $E = \mathrm{supp}(\mu)$ and its convex hull by $F = \mathrm{conv}(E)$. There exists $\delta = \delta(\Omega) > 0$ such that the closed ℓ_∞-ball, $\overline{B}_{3\delta}^\infty(\Omega)$, is included in $\mathrm{int}(F)$. Cover Ω by a finite union of $B_\delta^\infty(\omega_j)$, and denote by Q be the finite set of vertices of $\cup_j \overline{B}_{3\delta}^\infty(\omega_j)$. 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 $B_{3\delta}^\infty(\omega_j)$ for any j.

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

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

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

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

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

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

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

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 [4, Corollary 1.3(2)]. □

Proof of Proposition 6. By Proposition 7 when $n > N_\Omega$ is large, $u_n(x) \neq \emptyset$ 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 $\epsilon > 0$ and subsequences $y_{n_k} \in u_{n_k}(x_{n_k})$ such that $x_{n_k} \in \Omega$ and

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

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

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

a contradiction. □

8. 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, $\{x^i_m\}_{m=1}^M$. Define as in Anderes et al. [9] the set

$$S = \frac{1}{N}\left\{x^1_{m_1} + \dots + x^N_{m_N} : 1 \le m_i \le M, \quad i = 1, \dots, N\right\}$$

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.

SUPPLEMENTARY MATERIAL

Online Supplement: "Fréchet Means in Wasserstein space: Gradient Descent and Procrustes Analysis"
(URL to go here). The online supplement contains more details on the simulation experiments, additional discussion, as well as those proofs that were omitted from the main paper.

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 also thank the referees and the editors for their constructive comments, technical remarks, and suggested improvements.

References.

[1] B. Afsari, R. Tron, and R. Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *Society for Industrial and Applied Mathematics*, 43:2:904–924, 2011.

[3] M. Agulló-Antolín, J. A. Cuesta-Albertos, H. Lescornel, and J.-M. Loubes. A parametric registration model for warped distributions with wasserstein?s distance. *Journal of Multivariate Analysis*, 135:117–130, 2015.

[4] G. Alberti and L. Ambrosio. A geometrical approach to monotone functions in $\mathbb{R}^n$. *Math. Z.*, 230(2):259–316, 1999.

[5] S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

[6] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):358–375, 2011.

[7] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics. ETH Zürich (closed). Springer London, Limited, 2nd edition, 2008.

[8] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86(414):376–387, 1991.

[9] E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, pages 1–21, 2016.

[10] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[11] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.

[12] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the Wasserstein space. *arXiv preprint arXiv:1307.7721*, 2013.

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

[14] P. Billingsley. *Convergence of probability measures*, volume 137. John Wiley&Sons Inc., New York, 2nd edition, 1999.

[15] E. Boissard, T. Le Gouic, J.-M. Loubes, et al. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[16] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[17] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[18] F. L. Bookstein. *Morphometric tools for landmark data: geometry and biology.* Cambridge University Press, 1997.

[19] L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[20] R. Chartrand, B. Wohlberg, K. Vixie, and E. Bollt. A gradient descent solution to the Monge–Kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

[21] W. Cheng, I. L. Dryden, D. B. Hitchcock, H. Le, et al. Analysis of spike train data: Classification and Bayesian alignment. *Electronic Journal of Statistics*, 8(2):1786–1792, 2014.

[22] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications.* John Wiley & Sons, 2013.

[23] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the International Conference on Machine Learning 2014, JMLR W&CP*, 32(1):685–693, 2014.

[24] D. Dowson and B. Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[25] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*, volume 4. J. Wiley Chichester, 1998.

[26] J.-F. Dupuy, J.-M. Loubes, and E. Maza. Non parametric estimation of the structural expectation of a stochastic increasing function. *Statistics and Computing*, 21(1):121–136, 2011.

[27] M. Fiedler. Bounds for the determinant of the sum of hermitian matrices. *Proceedings of the American Mathematical Society*, pages 27–31, 1971.

[28] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

[29] M. Fréchet. Sur la distance de deux lois de probabilité. *C. R. Acad. Sci. Paris*, 244(6):689–692, 1957.

[30] S. Gallón, J.-M. Loubes, and E. Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical biosciences*, 242(2):129–142, 2013.

[31] W. Gangbo and A. Święch. Optimal maps for the multidimensional Monge–Kantorovich problem. *Communications on pure and applied mathematics*, 51(1):23–45, 1998.

[32] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285–339, 1991.

[33] J. C. Gower. Generalized Procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[34] D. Groisser. On the convergence of some Procrustean averaging algorithms. *Stochastics An International Journal of Probability and Stochastic Processes*, 77(1):31–60, 2005.

[35] E. Haber, T. Rehman, and A. Tannenbaum. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing*, 32(1):197–211, 2010.

[36] P. Z. Hadjipantelis, J. A. Aston, H.-G. Müller, J. Moriarty, et al. Analysis of spike train data:

A multivariate mixed effects model for phase and amplitude. *Electronic Journal of Statistics*, 8(2):1797–1807, 2014.

[37] L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

[38] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

[39] O. Kallenberg. *Random measures*. Academic Press, New York, 1986.

[40] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

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

[42] W. S. Kendall, H. Le, et al. Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics*, 25(3):323–352, 2011.

[43] S. Krantz. *Convex Analysis*. Textbooks in Mathematics. CRC Press, 2014.

[44] H. Le. Mean size-and-shapes and mean shapes: a geometric point of view. *Advances in Applied Probability*, pages 44–55, 1995.

[45] H. Le. Locating Fréchet means with application to shape spaces. *Advances in Applied Probability*, pages 324–338, 2001.

[46] X. Lu, J. S. Marron, et al. Analysis of spike train data: Comparison between real and the simulated data. *Electronic Journal of Statistics*, 8(2):1793–1796, 2014.

[47] R. J. McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[48] I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[49] V. M. Panaretos, Y. Zemel, et al. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

[50] M. Patriarca, L. M. Sangalli, P. Secchi, S. Vantini, et al. Analysis of spike train data: An application of k-mean alignment. *Electronic Journal of Statistics*, 8(2):1769–1775, 2014.

[51] D. Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

[52] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

[53] R. T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, NJ, 1970.

[54] W. Schachermayer and J. Teichmann. Characterization of optimal transport plans for the Monge–Kantorovich problem. *Proceedings of the American Mathematical Society*, 137:519–529, 2009.

[55] E. M. Stein and R. Shakarchi. *Real Analysis: Measure Theory, Integration & Hilbert Spaces*. Princeton University Press, 2005.

[56] C. Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Society, 2003.

[57] J.-L. Wang, J.-M. Chiou, and H.-G. Mueller. Review of Functional Data Analysis. *ArXiv e-prints*, July 2015.

[58] W. Wu, N. G. Hatsopoulos, A. Srivastava, et al. Introduction to neural spike train data for phase-amplitude analysis. *Electronic Journal of Statistics*, 8(2):1759–1768, 2014.

[59] W. Wu and A. Srivastava. An information-geometric framework for statistical inferences in the neural spike train space. *Journal of computational neuroscience*, 31(3):725–748, 2011.

[60] W. Wu and A. Srivastava. Estimating summary statistics in the spike-train space. *Journal of computational neuroscience*, 34(3):391–410, 2013.

[61] W. Wu, A. Srivastava, et al. Analysis of spike train data: Alignment and comparisons using the extended Fisher–Rao metric. *Electronic Journal of Statistics*, 8(2):1776–1785, 2014.

[62] Y. Zemel and V. M. Panaretos. Supplement to: "Fréchet means in Wasserstein space: Gradient descent and Procrustes analysis". 2016.

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

SUPPLEMENT TO: "FRÉCHET MEANS IN WASSERSTEIN SPACE: GRADIENT DESCENT AND PROCRUSTES ANALYSIS"

By Yoav Zemel and Victor M. Panaretos

Ecole Polytechnique Fédérale de Lausanne

1. Introduction. This supplement includes four Sections. Section 2 contains the proof that no further requirement except for finite second moments is needed for the convergence of the algorithm presented in the article. Section 3 contains further details and theoretical results pertaining to the simulation scenarios described in Section 6.4 of the article. Finally, Section 4 contains the proofs that do not appear in the article itself, and Section 5 concludes with some additional discussion.

2. A complete proof of Lemma 4 in the article. In this section we show that condition (7.3) in the article is not needed for (7.2) to hold. The idea is that (7.2) only requires a tiny bit more than finite second moments, and that is provided in Lemma 2.2. Throughout this section, all functions are assumed nonnegative (possibly infinite-valued) and defined on $[0,\infty)$ unless explicitly stated otherwise. We write $f(x)\in\omega(g(x))$ or $f\in\omega(g)$ if $f(x)/g(x)\to\infty$ as $x\to\infty$.

Lemma 2.1. *Let f be integrable. Then there exists a continuous nondecreasing function $g\in\omega(1)$ such that fg is integrable.*

Proof. Set $F(x)=\int_x^\infty f(t)\,\mathrm{d}t$ and $g(x)=[F(x)]^{-1/2}$. Then a change of variables gives

$$\int_0^\infty f(x)g(x)\,\mathrm{d}x=\int_0^\infty f(x)[F(x)]^{-1/2}\,\mathrm{d}x=\int_0^{F(0)}u^{-1/2}\,\mathrm{d}u=2\sqrt{\|f\|_1}<\infty,$$

and $g(x)\to\infty$ because $F(x)\to 0$ as $x\to\infty$ by dominated convergence. □

Lemma 2.2. *Let X be a random variable with $\mathbb{E}X^2<\infty$. Then there exists a convex nondecreasing function $H(x)\in\omega(x^2)$ such that $\mathbb{E}H(X)<\infty$.*

Proof. Since

$$\infty>\mathbb{E}X^2=\int_0^\infty\mathbb{P}(X^2>t)\,\mathrm{d}t,$$

there exists a function g as in Lemma 2.1 such that

$$\infty>\int_0^\infty\mathbb{P}(X^2>t)g(t)\,\mathrm{d}t=\int_0^\infty\mathbb{P}(X^2>G^{-1}(u))\,\mathrm{d}u=\int_0^\infty\mathbb{P}(G(X^2)>u)\,\mathrm{d}u=\mathbb{E}G(X^2),$$

where G is the primitive of g and $G(0)=0$. The properties of g imply that G is convex and invertible, and that for $y<x$,

$$G(x)\geq \int_y^x g(t)\,\mathrm{d}t \geq \int_y^x g(y)\,\mathrm{d}t = (x-y)g(y),$$

which, combined with $g(y)\to\infty$ as $y\to\infty$, yields

$$\liminf_{x\to\infty}\frac{G(x)}{x}\geq g(y)\to\infty, \qquad y\to\infty,$$

so that $G(x)\in\omega(x)$. The function $H(x)=G(x^2)$ then has all the desired properties. □

Proposition 2.3. *Equation (7.2) of the article holds if merely*

$$\int_{\mathbb{R}^d}\|x\|^2\,\mathrm{d}\mu^i(x)<\infty, \qquad i=1,\dots,N.$$

Proof. Let $X_i=\|Z^i\|$ where $Z^i\sim\mu^i$. Then there exist functions g^i as in Lemma 2.1 with

$$\int_0^\infty \mathbb{P}(X_i^2>t)g^i(t)\,\mathrm{d}t<\infty, \qquad i=1,\dots,N.$$

The same holds with g^i replaced by $g=\min_i g^i$, which is still continuous, nondecreasing and divergent. Setting H as in Lemma 2.2, we see that $H(x)\in\omega(x^2)$ and

$$M^i=\mathbb{E}H(X^i)=\int_{\mathbb{R}^d}H(\|x\|)\,\mathrm{d}\mu^i(x)<\infty, \qquad i=1,\dots,N.$$

Convexity of H and $\|\cdot\|$ combined with monotonicity of H yield

$$\int_{\mathbb{R}^d}H(\|x\|)\,\mathrm{d}\gamma_j(x)=\int_{\mathbb{R}^d}H\left(\left\|\frac{1}{N}\sum_{i=1}^N \mathbf{t}_{\gamma_{j-1}}^{\mu^i}(x)\right\|\right)\mathrm{d}\gamma_{j-1}(x)$$
$$\leq \frac{1}{N}\sum_{i=1}^N\int_{\mathbb{R}^d}H(\|\mathbf{t}_{\gamma_{j-1}}^{\mu^i}(x)\|)\,\mathrm{d}\gamma_{j-1}(x)=\frac{1}{N}\sum_{i=1}^N\int_{\mathbb{R}^d}H(\|x\|)\,\mathrm{d}\mu^i(x)\leq M,$$

where $M=\sum_{i=1}^N M^i/N$. This implies that for any $R>0$ and any $j>0$,

$$\int_{\{x:\|x\|>R\}}\|x\|^2\,\mathrm{d}\gamma_j(x)\leq \sup_{y>R}\frac{y^2}{H(y)}\int_{\{x:\|x\|>R\}}H(\|x\|)\,\mathrm{d}\gamma_j(x)\leq M\sup_{y>R}\frac{y^2}{H(y)},$$

and (7.2) follows because $H(y)\in\omega(y^2)$. □

3. Details for the illustrative examples in Section 6.4. In this section we provide further details for finding the optimal maps in the examples of Section 6.4 in the article and theoretical results about the Fréchet mean and the behaviour of the algorithm. Throughout this section, $\mu^1,\dots,\mu^N$ are given measures and γ_0 is the initial point of Algorithm 1. We begin with two lemmas regarding compatibility of the measures as defined in Section 6.4.

Lemma 3.1 (Compatibility and Convergence). *If $\mathbf{t}_{\mu^1}^{\mu^i}\circ\mathbf{t}_{\gamma_0}^{\mu^1}=\mathbf{t}_{\gamma_0}^{\mu^i}$ and $\mathbf{t}_{\mu^1}^{\mu^j}\circ\mathbf{t}_{\mu^i}^{\mu^1}=\mathbf{t}_{\mu^i}^{\mu^j}$ (in the relevant L^2 spaces) for all i and all j, then Algorithm 1 converges after a single step.*

Proof. For all i, j and k we have $\mathbf{t}_{\mu^j}^{\mu^k}\circ\mathbf{t}_{\mu^i}^{\mu^j}=\mathbf{t}_{\mu^i}^{\mu^k}$, so that the optimal maps are admissible, and

$$\gamma_1=\left[\frac{1}{N}\sum_{i=1}^N\mathbf{t}_{\gamma_0}^{\mu^i}\right]\#\gamma_0=\left[\frac{1}{N}\sum_{i=1}^N\mathbf{t}_{\mu^1}^{\mu^i}\circ\mathbf{t}_{\gamma_0}^{\mu^1}\right]\#\gamma_0=\left[\frac{1}{N}\sum_{i=1}^N\mathbf{t}_{\mu^1}^{\mu^i}\right]\#\mu^1.$$

Boissard et al. [4] show that this is indeed the Fréchet mean. □

When $d=1$, all (diffuse) measures are compatible with each other, and Algorithm 1 converges after one step. Generally, the algorithm requires the calculation of N pairwise optimal maps, and this can be reduced to $N-1$ if $\gamma_0=\mu^1$. This is the same computational complexity as the calculation of the iterated barycentre proposed in [4].

Measures on $\mathbb{R}^d$ that have a common dependence structure are compatible with each other. More precisely, we say that $C:[0,1]^d\to[0,1]$ is a *copula* if there exists a random vector U with $U[0,1]$ margins and such that

$$\mathbb{P}(U_1\le u_1,\dots,U_d\le u_d)=C(u_1,\dots,u_d),\qquad u_i\in[0,1].$$

In other words, a copula is the restriction to $[0,1]^d$ of the probability distribution function of some d-dimensional random variable with uniform margins. See, for example, Nelsen [6] for an overview. Given a measure μ on $\mathbb{R}^d$ with distribution function G and marginal distribution functions G_j, the copula associated with μ is a copula such that

$$G(a_1,\dots,a_d)=\mu((-\infty,a_1]\times\cdots\times(-\infty,a_d])=C(G_1(a_1),\dots,G_d(a_d)).$$

This equation defines C uniquely if each marginal G_i is continuous, which we shall assume for simplicity. (If some G_i is discontinuous then C might not be unique, but it always exists, see [6, Chapter 2].)

Lemma 3.2 (Compatibility and Copulae). *Let $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$ be regular. Then μ and ν have the same associated copula if and only if $\mathbf{t}_\mu^\nu$ takes the separable form*

$$\mathbf{t}_\mu^\nu(x_1,\dots,x_d)=(T_1(x_1),\dots,T_d(x_d)),\qquad T_i:\mathbb{R}\to\mathbb{R}. \tag{3.1}$$

Proof. If μ and ν have the same copula then

$$G(G_1^{-1}(u_1),\dots,G_d^{-1}(u_d))=C(u_1,\dots,u_d)=F(F_1^{-1}(u_1),\dots,F_d^{-1}(u_d)),$$

where $G_j^{-1}(u_j)$ is any number satisfying $G_j(G_j^{-1}(u_j))=u_j$ (such numbers exist because G_j is surjective), and similarly for F_j^{-1}. Consequently, $F(x_1,\dots,x_d)=G(T_1(x_1),\dots,T_d(x_d))$ with $T_j=G_j^{-1}\circ F_j$. It follows that $\nu=(T_1,\dots,T_d)\#\mu$, and this map is optimal, hence equals $\mathbf{t}_\mu^\nu$, because the T_j's are nondecreasing.

One proves the converse implication similarly: if $\mathbf{t}_\mu^\nu$ takes this form, then each T_j needs to be nondecreasing. Since it must push F_j forward to G_j, we have $T_j=G_j^{-1}\circ F_j$, and this yields the above equality for the copula. □

It is easy to see that if the optimal maps between each μ^i and each μ^j are of the form (3.1), then $\{\mu^i\}$ are compatible with other. This follows from this property holding for each marginal, and the possibility of working with the marginals separately; it has already been observed by Boissard et al. [4, Proposition 4.1]. This explains why the algorithm converges in one iteration for the example with the Frank copula.

Next, we give a convergence analysis for the Gaussian example.

Theorem 3.3 (Convergence in Gaussian case). *Let $\mu^i\sim\mathcal{N}(0,S_i)$ for S_i positive definite, and let the initial point $\gamma_0=\mathcal{N}(0,\Gamma_0)$ for positive definite Γ_0. Then the sequence of iterates generated by Algorithm 1 converges to the unique Fréchet mean of $(\mu^1,\dots,\mu^N)$.*

Proof. We first observe that for any centred measure μ with covariance matrix S,

$$d^2(\mu,\delta_0)=\mathrm{tr}S,$$

where δ_0 is a dirac mass at the origin. (This follows from the singular value decomposition of S.) Next, each iteration stays (centred) Gaussian, say $\mathcal{N}(0,\Gamma_k)$, because the optimal maps are linear; and since the iterates are absolutely continuous (Lemma 1), each Γ_k is nonsingular.

Proposition 4 implies that $\det\Gamma_k$ is bounded below uniformly; on the other hand,

$$0\le\mathrm{tr}\Gamma_k=d^2(\gamma_k,\delta_0)$$

is bounded uniformly, because $\{\gamma_k\}$ stays in a Wasserstein-compact set by Lemma 4. Let $C_1=\inf_k\det\Gamma_k>0$ and $C_2=\sup_k\mathrm{tr}\Gamma_k<\infty$. Then each eigenvalue λ of Γ_k is nonnegative, bounded above by C_2, and satisfies

$$C_1\le\det\Gamma_k\le\lambda C_2^{d-1}\qquad\Longrightarrow\qquad\lambda\ge C_1C_2^{1-d}=C_3>0.$$

The matrices Γ_k stay in a bounded set, and each limit point Γ is positive definite because $x^t\Gamma x\ge C_3\|x\|^2$ for all $x\in\mathbb{R}^d$. Each limit point γ of γ_k is a Karcher mean by Theorem 4, and the limit must follow a $\mathcal{N}(0,\Gamma)$ distribution with Γ (nonsingular) limit point of Γ_k (e.g., by Lehmann–Scheffé's theorem). Since $F'(\gamma)=0$ everywhere on $\mathbb{R}^d$, γ is the Fréchet mean by the discussion after Corollary 1. Every limit of γ_k is the Fréchet mean and the sequence is compact, so γ_k must converge to the Fréchet mean. □

In order to deal with the last example of Section 6.4, we need two more results. The first involves coupling measures of dimensions greater than one, while the second shows the equivariance of the Fréchet mean with respect to rotations.

Invoking the *independence copula* $C(u_1, \dots, u_d) = u_1 \dots u_d$, a special case of Lemma 3.2 above is when the marginals of μ and ν are independent. In this independence case, it is possible in fact to replace the marginals by measures of arbitrary dimension:

Lemma 3.4. *Let $\mu^1, \dots, \mu^N$ and $\nu^1, \dots, \nu^N$ be regular measures in $\mathcal{P}_2(\mathbb{R}^{d_1})$ and $\mathcal{P}_2(\mathbb{R}^{d_2})$ with (unique) Fréchet means μ and ν respectively. Then the independent coupling $\mu \otimes \nu$ is the Fréchet mean of $\mu^1 \otimes \nu^1, \dots, \mu^N \otimes \nu^N$.*

By induction (or a straightforward modification of the proof), one can show that the Fréchet mean of $(\mu^i \otimes \nu^i \otimes \rho^i)$ is $\mu \otimes \nu \otimes \rho$, and so on. While we are confident this result should already be known, we could not find a reference, and thus we provide a full proof for completeness.

Proof. Agueh and Carlier [1, Proposition 3.8] show that there exist convex lower semicontinuous potentials ψ_i on $\mathbb{R}^{d_1}$ and φ_i on $\mathbb{R}^{d_2}$ whose gradients push μ forward to μ^i and ν to ν^i respectively, and such that

$$\frac{1}{N}\sum_{i=1}^N \psi_i^*(x) \le \frac{\|x\|^2}{2}, \quad x \in \mathbb{R}^{d_1}; \qquad \frac{1}{N}\sum_{i=1}^N \varphi_i^*(y) \le \frac{\|y\|^2}{2}, \quad y \in \mathbb{R}^{d_2},$$

with equality μ- and ν-almost surely respectively. It is easy to see that the extensions $\tilde{\psi}_i(x,y) = \psi_i(x)$ and $\tilde{\varphi}_i(x,y) = \varphi_i(y)$ defined on $\mathbb{R}^{d_1+d_2}$ are convex lower semicontinuous functions whose sum ϕ_i is a convex function satisfying

$$\phi_i^*(x,y) = (\tilde{\psi}_i + \tilde{\varphi}_i)^*(x,y) = \psi_i^*(x) + \varphi_i^*(y).$$

Clearly $\nabla\phi_i\#(\mu^i \otimes \nu^i) = \mu \otimes \nu$ and

$$\frac{1}{N}\sum_{i=1}^N \phi_i^*(x,y) \le \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} = \frac{\|(x,y)\|^2}{2}, \quad (x,y) \in \mathbb{R}^{d_1+d_2},$$

with equality $\mu\otimes\nu$-almost surely. By the same Proposition 3.8 in [1], $\mu\otimes\nu$ is the Fréchet mean. □

Lemma 3.5. *If μ is the Fréchet mean of the regular measures $\mu^1, \dots, \mu^N$, one with bounded density, and U is orthogonal, then $U\#\mu$ is the Fréchet mean of $U\#\mu^1, \dots, U\#\mu^N$.*

Proof. Bonneel et al. sketch a proof of this statement in [5, Proposition 1], and it also appears implicitly in Boissard et al. [4, Proposition 4.1]; we give an alternative argument here.

If $x \mapsto \varphi(x)$ is convex, then $x \mapsto \varphi(U^{-1}x)$ is convex with gradient $U\nabla\varphi(U^{-1}x)$ at (almost all) x and conjugate $x \mapsto \varphi^*(U^{-1}x)$. If φ_i are convex potentials with $\nabla\varphi_i\#\mu = \mu^i$, then $\nabla(\varphi_i \circ U^{-1})$ pushes $U\#\mu$ forward to $U\#\mu^i$ and by [1, Proposition 3.8]

$$\frac{1}{N}\sum_{i=1}^N (\varphi_i \circ U^{-1})^*(Ux) = \frac{1}{N}\sum_{i=1}^N \varphi_i^*(x) \le \frac{\|x\|^2}{2} = \frac{\|Ux\|^2}{2}$$

with equality for μ-almost any x. A change of variables $y = Ux$ shows that the set of points y such that $\sum(\varphi_i \circ U^{-1})^*(y) < N\|y\|^2/2$ is $(U\#\mu)$-negligible, completing the proof. □

We apply these results in the context of the simulated example in Section 6.4 in the article. If $Y = (y_1, y_2, y_3) \sim \mu^i$, then the random vector $(x_1, x_2, x_3) = X = U^{-1}Y$ has joint density

$$f^i(x_3)\exp\left[-\frac{(x_1,x_2)(\Sigma^i)^{-1}\binom{x_1}{x_2}}{2}\right]\frac{1}{2\pi\sqrt{\det\Sigma^i}},$$

so the probability law of X is $\rho^i \otimes \nu^i$ with ρ^i centred Gaussian with covariance matrix Σ^i and ν^i having density f^i on $\mathbb{R}$. By Lemma 3.4, the Fréchet mean of $(U^{-1}\#\mu^i)$ is the product measure of that of (ρ^i) and that of (ν^i); by Lemma 3.5, the Fréchet mean of (μ^i) is therefore

$$U\#(\mathcal{N}(0,\Sigma) \otimes f), \qquad f = F', \quad F^{-1}(q) = \frac{1}{N}\sum_{i=1}^N F_i^{-1}(q), \quad F_i(x) = \int_{-\infty}^x f^i(s)\,\mathrm{d}s,$$

where Σ is the Fréchet–Wasserstein mean of $\Sigma_1, \dots, \Sigma_N$.

Starting at an initial point $\gamma_0 = U\#(\mathcal{N}(0,\Sigma_0) \otimes \nu_0)$, with ν_0 having continuous distribution F_{ν_0}, the optimal maps are $U \circ \mathbf{t}_0^i \circ U^{-1} = \nabla(\varphi_0^i \circ U^{-1})$ with

$$\mathbf{t}_0^i(x_1, x_2, x_3) = \begin{pmatrix} \mathbf{t}_{\Sigma_0}^{\Sigma^j}(x_1, x_2) \\ F_j^{-1} \circ F_{\nu_0}(x_3) \end{pmatrix}$$

the gradients of the convex function

$$\varphi_0^i(x_1, x_2, x_3) = (x_1, x_2)\mathbf{t}_{\gamma_0}^{\Sigma^i}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \int_0^{x_3} F_j^{-1}(F_{\nu_0}(s))\,\mathrm{d}s,$$

where we identify $\mathbf{t}_{\gamma_0}^{\Sigma^i}$ with the positive definite matrix $(\Sigma^i)^{1/2}[(\Sigma^i)^{1/2}\Sigma_0(\Sigma^i)^{1/2}]^{-1/2}(\Sigma^i)^{1/2}$ that pushes forward $\mathcal{N}(0,\Sigma_0)$ to $\mathcal{N}(0,\Sigma^i)$. Due to the one-dimensionality, the algorithm finds the third component of the rotated measures after one step, but the convergence of the Gaussian component requires further iterations.

4. Proofs omitted from the article.

Proof of Corollary 1, Section 4.3. The characterisation of Karcher means follows immediately from Theorem 2. Now suppose that $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is regular and $F'(\mu) \neq 0 \in L^2(\mu)$. The function $S = N^{-1}\sum_{i=1}^N \mathbf{t}_\mu^{\mu^i}$ is a gradient of a convex function and

$$\lim_{\nu\to\mu} \frac{F(\nu) - F(\mu) + \langle S - \mathbf{i}, \mathbf{t}_\mu^\nu - \mathbf{i}\rangle_{L^2(\mu)}}{\|\mathbf{t}_\mu^\nu - i\|_{L^2(\mu)}} = \lim_{\nu\to\mu} \frac{F(\nu) - F(\mu) + \int \langle S(x) - x, \mathbf{t}_\mu^\nu(x) - x\rangle \,\mathrm{d}\mu(x)}{d(\nu,\mu)} = 0.$$

By assumption $W = S - \mathbf{i} \neq 0 \in L^2(\mu)$. The measure $\nu_s = [\mathbf{i} + s(W - \mathbf{i})]\#\mu$ with $s \in (0,1)$ is such that $d(\nu_s,\mu) = s\|W\|_{L^2(\mu)}$ and

$$0 = \lim_{s\to 0^+} \frac{F(\nu_s) - F(\mu) + \int\langle W(x), sW(x)\rangle \,\mathrm{d}\mu(x)}{s\|W\|_{L^2(\mu)}} = \lim_{s\to 0^+} \frac{F(\nu_s) - F(\mu)}{s\|W\|_{L^2(\mu)}} + \|W\|_{L^2(\mu)}.$$

This means that when s is small enough, $F(\nu_s) < F(\mu)$, so μ cannot be the minimiser of F. Since $\bar\mu$ has to be regular [1, Proposition 5.1], necessity of $F'(\bar\mu) = 0$ is proven. □

4.1. *Proofs of statements from Section 6 of the article.*

Proof of Proposition 2. Write $M(\gamma) = \mathbb{E}[d^2(\Lambda,\gamma)]$. We first establish (weak) convexity. Indeed, for given measures γ and ρ and $0 < t < 1$,

$$\begin{aligned} tM_\omega(\gamma) + (1-t)M_\omega(\rho) &= t\int_{\mathbb{R}^d\times\mathbb{R}^d} (x-y)^2 \,\mathrm{d}\pi_{\omega,\gamma}(x,y) + (1-t)\int_{\mathbb{R}^d\times\mathbb{R}^d} (x-y)^2 \,\mathrm{d}\pi_{\omega,\rho}(x,y) \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} (x-y)^2 \,\mathrm{d}[t\pi_{\omega,\gamma} + (1-t)\pi_{\omega,\rho}], \end{aligned}$$

where $\pi_{\omega,\gamma}$ is the optimal coupling between $\Lambda = \Lambda(\omega)$ and γ. The measure $t\pi_{\omega,\gamma} + (1-t)\pi_{\omega,\rho}$ is a coupling between Λ and $t\gamma + (1-t)\rho$, and this shows that M_ω is convex without any regularity assumptions on Λ. To upgrade to strict convexity when Λ is regular, observe firstly that M is finite on the set of probability measures supported on K. If Λ is regular, then optimal measures are supported on graphs of functions:

$$\begin{aligned} \pi_{\omega,\gamma}(A\times B) &= \Lambda(A\cap T_1^{-1}(B)) \\ \pi_{\omega,\rho}(A\times B) &= \Lambda(A\cap T_2^{-1}(B)) \\ \pi_{\omega,t\gamma+(1-t)\rho}(A\times B) &= \Lambda(A\cap T_3^{-1}(B)) \\ [t\pi_{\omega,\gamma} + (1-t)\pi_{\omega,\rho}](A\times B) &= t\Lambda(A\cap T_1^{-1}(B)) + (1-t)\Lambda(A\cap T_2^{-1}(B)). \end{aligned}$$

The measure $t\pi_{\omega,\gamma} + (1-t)\pi_{\omega,\rho}$ is supported on the graph of two functions, T_1 and T_2. It can only be optimal if it is supported on the graph of one function, and this will only happen if $T_1 = T_2$, Λ-almost surely, that is, if $\gamma = \rho$. (See [3, Corollary 2.9] for a rigorous proof.) We can thus conclude that

$$\Lambda \text{ regular} \quad\Longrightarrow\quad M \text{ strictly convex.}$$

Since M was already shown to be weakly convex in any case, it follows that

$$\mathbb{P}(\Lambda \text{ regular}) > 0 \quad\Longrightarrow\quad M \text{ strictly convex.}$$

Now we turn to the existence of a solution (once existence is established, uniqueness will follow from strict convexity). Let $\mathrm{proj}_K : \mathbb{R}^d \to K$ denote the projection onto the set K, which is well-defined since K is closed and convex, and of course satisfies

$$\|x-y\| \geq \|x - \mathrm{proj}_K(y)\|, \qquad x \in K, \quad y \in \mathbb{R}^d.$$

Since Λ is concentrated on K, the above inequality holds Λ-almost surely with respect to x. Let T be the optimal map from Λ to γ (a proper map almost surely, as argued above). Observe that

$$d^2(\Lambda,\gamma) = \int_K \|T(x)-x\|^2 \,\mathrm{d}\Lambda \geq \int^K \|\mathrm{proj}_K(T(x)) - x\|^2 \,\mathrm{d}\Lambda \geq d^2(\Lambda, \mathrm{proj}_K\#\gamma),$$

since $(\mathrm{proj}_K \circ T)\#\Lambda = \mathrm{proj}_K\#(T\#\Lambda) = \mathrm{proj}_K\#\gamma$. This measure is concentrated on K, and taking expectations gives $M(\gamma) \geq M(\mathrm{proj}_K\#\gamma)$. Hence, the infimum of M equals the infimum of M on $\mathcal{P}(K)$, the collection of probability measures supported on K (or else, we could project all the remaining mass to K to reduce the total cost further). The restriction of M to $\mathcal{P}(K)$ is a continuous functional on a compact set (measures whose support is contained in a common compactum are a compact set in Wasserstein space), and existence follows. □

Proof of Lemma 5. It is assumed that $\psi(z) = \psi_1(\|z\|)$ with ψ_1 non-increasing, strictly positive and

$$\int_{\mathbb{R}^d} \psi(z)\,\mathrm{d}z = 1 = \int_{\mathbb{R}^d} \|z\|^2 \psi(z)\,\mathrm{d}z.$$

Let $\Psi(A) = \int_A \psi(x)\,\mathrm{d}x$ be the corresponding probability measure and recall that $\psi_\sigma(x) = \sigma^{-d}\psi(x/\sigma)$ for $\sigma > 0$.

For $y \in K$ set $\tilde{\mu}_y = \delta\{y\} * \psi_\sigma$ and its restricted renormalized version $\mu_y = (1/\tilde{\mu}_y(K))\tilde{\mu}_y|_K$, so that $\widehat{\Lambda}_i = (1/m)\sum_{j=1}^m \mu_{x_j}$, and it is assumed that $m \geq 1$ and $x_j \in K$ (because $\Lambda_i(K) = 1$).

One way (certainly not optimal, unless $m = 1$) to couple $\widehat{\Lambda}_i$ with $\widetilde{\Pi}_i/m$ is to send the $1/m$ mass of μ_{x_j} to x_j. This gives

$$d^2(\widehat{\Lambda}_i, \tilde{\Pi}_i/m) \leq \frac{1}{m}\sum_{j=1}^m d^2(\mu_{x_j}, \delta\{x_j\}) = \frac{1}{m}\sum_{j=1}^m \frac{1}{\tilde{\mu}_{x_j}(K)} \int_K \|x - x_j\|^2 \psi_\sigma(x - x_j)\,\mathrm{d}x.$$

However, for an arbitrary $y \in K$,

$$\frac{1}{\tilde{\mu}_y(K)} \int_K \|x-y\|^2 \psi_\sigma(x-y)\,\mathrm{d}x = \frac{1}{\tilde{\mu}_y(K)} \sigma^2 \int_{(K-y)/\sigma} \|z\|^2 \psi(z)\,\mathrm{d}z.$$

The last displayed integral is bounded by 1. Hence, we seek a lower bound, uniformly in y and σ, for

$$\tilde{\mu}_y(K) = \int_K \psi_\sigma(x-y)\,\mathrm{d}x = \int_{(K-y)/\sigma} \psi(x)\,\mathrm{d}x = \Psi\left(\frac{K-y}{\sigma}\right).$$

Since $K-y$ is a convex set that contains the origin, the collection of sets $\{\sigma^{-1}(K-y)\}_{\sigma>0}$ is increasing as $\sigma \searrow 0$. Consequently, if $\sigma \leq 1$,

$$\Psi\left(\frac{K-y}{\sigma}\right) \geq \Psi(K-y) = \int_{K-y} \psi(x)\,\mathrm{d}x \geq \int_{K-y} \psi_1(d_K)\,\mathrm{d}x = \psi_1(d_K)\mathrm{Leb}(K) > 0.$$

Here $d_K = \sup\{\|x-y\| : x, y \in K\}$ is the (finite) diameter of K, and we have used the monotonicity of ψ_1.

It follows that for $C_{\psi,K} = [\psi_1(d_K)\mathrm{Leb}(K)]^{-1} < \infty$ (depending only on ψ and K),

$$d^2(\mu_y, \delta_y) \leq C_{\psi,K}\sigma^2, \qquad y \in K, \quad \sigma \leq 1. \tag{4.1}$$

Since the bound is uniform in y, the proof is complete. In the context of Remark 4 in the article, one simply needs to replace the term $\psi_1(d_K)$ in $C_{\psi,K}$ by $\delta_\psi(d_K)$. □

REMARK 1. *The upper bound* (4.1) *can be easily seen to be tight (up to constant): if* $y = 0$ *and* $K = [0,1]^d$, *then* $\sigma^{-2}d^2(\mu_y, \delta_y) \to 2^d$ *as* $\sigma \searrow 0$.

PROOF OF THEOREM 6. Convergence in probability in part (1) follows as in [7, pp. 793–794], using (7.6). For convergence almost surely, let $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$. A straightforward generalisation of the argument in [7, pp. 794–795] gives

$$\mathbb{P}\left(\frac{\widetilde{\Pi}_i((-\infty,a])}{\tau_n} - \Lambda_i((-\infty,a]) \to 0\right) = 1,$$

where for $-\infty \leq a_i \leq b_i \leq \infty$ $(i = 1, \ldots, d)$ we denote

$$(a,b] = (a_1, b_1] \times \cdots \times (a_d, b_d].$$

Consequently

$$\mathbb{P}\left(\frac{\widetilde{\Pi}_i((-\infty,a])}{\tau_n} - \Lambda_i((-\infty,a]) \to 0 \text{ for any } a \in \mathbb{Q}^d\right) = 1.$$

For a general $a \in \mathbb{R}^d$ there exist sequences $a^k \nearrow a \swarrow b^k$ with $a^k, b^k \in \mathbb{Q}^d$ (that is, $a_i^k \nearrow a_i \swarrow b_i^k$ for any coordinate i). Since for any k

$$\frac{\widetilde{\Pi}_i((-\infty,a])}{\tau_n} - \Lambda_i((-\infty,a]) \leq \frac{\widetilde{\Pi}_i((-\infty,b^k])}{\tau_n} - \Lambda_i((-\infty,b^k]) + \Lambda_i((-\infty,b^k]) - \Lambda_i((-\infty,a]),$$

it follows that with probability one

$$\limsup_{n\to\infty} \frac{\widetilde{\Pi}_i((-\infty,a])}{\tau_n} - \Lambda_i((-\infty,a]) \leq \Lambda_i\left((-\infty,b^k] \setminus (-\infty,a]\right) \to 0, \qquad k \to \infty,$$

as the sequence of sets at the right-hand side converges monotonically to the empty set.

Similarly, with probability one

$$\liminf_{n\to\infty}\frac{\widetilde{\Pi}_i((-\infty,a])}{\tau_n}-\Lambda_i((-\infty,a])\geq\Lambda_i\left((-\infty,a]\setminus(-\infty,a^k]\right)\to\Lambda_i((-\infty,a]\setminus(-\infty,a)),$$

and the right-hand side vanishes because Λ_i is assumed absolutely continuous (the set $(-\infty,a]\setminus(-\infty,a)$ is union of d $d-1$-dimensional rays). Specifying $a=\infty$ shows that almost surely $\widetilde{\Pi}_i(K)/\tau_n\to 1$ and we conclude that almost surely $\widetilde{\Pi}_i/\widetilde{\Pi}_i(K)\to\Lambda_i$ weakly. Further, $d(\widehat{\Lambda}_i,\widetilde{\Pi}_i/N_i)\to 0$ since $\sigma_n\to 0$ by Lemma 5.

We sketch the main ideas of the proof of (2); more details can be found in [7, pp. 795–797]. We wish to show that

$$\widehat{M}_n(\gamma)=\frac{1}{n}\sum_{i=1}^n d^2(\widehat{\Lambda}_i,\gamma)\to\mathbb{E}d^2(\Lambda_i,\gamma)=M(\gamma),\quad\text{uniformly in }\gamma.$$

In order to do this we write

$$\widehat{M}_n(\gamma)-M(\gamma)=\left[\widehat{M}_n(\gamma)-M_n(\gamma)\right]+\left[M_n(\gamma)-M(\gamma)\right],$$

where we introduce the empirical Fréchet functional

$$M_n(\gamma)=\frac{1}{n}\sum_{i=1}^n d^2(\Lambda_i,\gamma).$$

Since for any three probability measures on K it holds that

$$d(\mu,\nu)\leq\sqrt{\sup_{\gamma\in P(K^2)}\int_{K^2}\|x-y\|^2\,\mathrm{d}\gamma(x,y)}\leq\sqrt{\sup_{x,y\in K}\|x-y\|^2}=d_K<\infty;$$
$$|d^2(\mu,\rho)-d^2(\nu,\rho)|=|d(\mu,\rho)+d(\nu,\rho)||d(\mu,\rho)-d(\nu,\rho)|\leq 2d_Kd(\mu,\nu),$$

we see that

$$\sup_{\gamma\in P(K)}|\widehat{M}_n(\gamma)-M_n(\gamma)|\leq\frac{2d_K}{n}\sum_{i=1}^n d\left(\widehat{\Lambda}_i,\Lambda_i\right)=\frac{2d_K}{n}\sum_{i=1}^n X_{ni}=2d_K\overline{X}_n.$$

Each X_{ni} is a function of T_i, $\Pi_i^{(n)}$ and $\sigma_i^{(n)}$, and $0\leq X_{ni}\leq d_K$. If $\sigma_i^{(n)}$ is a function of $\widetilde{\Pi}_i^{(n)}=T_i\#\Pi_i^{(n)}$ only, then X_{ni} are iid across i. Part (1) shows that $X_{n1}\to 0$ in probability and by the bounded convergence theorem $\mathbb{E}\overline{X}_n=\mathbb{E}X_{n1}\xrightarrow{p}0$ and therefore the above expression converges to 0 in probability. In general, L^1-convergence of random variables does not guarantee convergence almost surely. As we deal with averages, however, almost sure convergence can be established: let $Y_{ni}=X_{ni}-\mathbb{E}X_{ni}\in[-d_K,d_K]$. Then Y_{ni} are mean zero iid random variables, so that

$$\mathbb{P}\left(|\overline{X}_n-\mathbb{E}\overline{X}_n|>\epsilon\right)=\mathbb{P}\left(\overline{Y}_n^4>\epsilon^4\right)\leq\frac{n\mathbb{E}\left[Y_{n1}^4\right]+3n(n-1)\mathbb{E}\left[Y_{n1}^2\right]}{\epsilon^4n^4}\leq\frac{3\max(d_K^4,d_K^2)}{\epsilon^4n^2}.$$

By the Borel–Cantelli lemma, $|\overline{X}_n - \mathbb{E}\overline{X}_n| \overset{as}{\to} 0$, hence $\overline{X}_n \overset{as}{\to} 0$.

If the smoothing is not carried out independently across trains, then X_{ni} may be correlated across i. In that case, one can introduce the functional $M_n^*(\gamma) = n^{-1}\sum_{i=1}^n d^2\left(\widetilde{\Pi}_i/N_i, \gamma\right)$ and proceed as in [7]. For M_n^* to be well-defined one may use Lemma 6 and that requires $\tau_n/\log n \to \infty$.

Finally, observe that by the strong law of large numbers $M_n(\gamma) \overset{as}{\to} M(\gamma)$ for all $\gamma \in P(K)$. That the convergence is uniform follows from the equicontinuity of the collection $\{M_n\}_{n=1}^\infty$ (they are $2d_K$-Lipschitz). We have thus established

$$\sup_{\gamma\in P(K)} |\widehat{M}_n(\gamma) - M(\gamma)| \overset{as}{\to} 0, \qquad n\to\infty.$$

By standard arguments, the minimiser $\widehat{\lambda}_n$ of $\widehat{M}_n$ converges to the minimiser λ^* of M, since the latter is unique by Proposition 2. But $\lambda^* = \lambda$ by the hypothesis. □

4.2. *Proofs of statements from Section 7.5 of the article.*

Proof of Lemma 7. For any $1 > \epsilon > 0$ there exists $0 < t_\epsilon$ such that for $t < t_\epsilon$,

$$\frac{\mathrm{Leb}(B_t(x_0)\cap G)}{\mathrm{Leb}(B_t(x_0))} > 1 - \epsilon^d.$$

Fix z such that $t = t(z) = \|z - x_0\| < t_\epsilon$. The intersection of $B_t(x_0)$ with $B_{2\epsilon t}(z)$ includes a ball of radius ϵt centred at $y = (1-\epsilon)z$, so that

$$\frac{\mathrm{Leb}(B_t(x_0)\cap B_{2\epsilon t}(z))}{\mathrm{Leb}(B_t(x_0))} \geq \frac{\mathrm{Leb}(B_{\epsilon t}(y))}{\mathrm{Leb}(B_t(x_0))} = \epsilon^d.$$

It follows that $G\cap B_{2\epsilon t}(z)$ is nonempty. In other words: for any $\epsilon > 0$ there exists t_ϵ such that if $\|z - x_0\| < t_\epsilon$, then there exists $x\in G$ with $\|z - x\| \leq 2\epsilon t(z) = 2\epsilon\|z - x_0\|$. This means precisely that $\delta(z) = o(\|z - x_0\|)$ as $z\to x_0$. □

Proof of Lemma 12. Assume $\epsilon < \rho$ (there is nothing to prove otherwise). Take a corner of the ℓ_∞ ball of radius $\rho' < \rho$ around x_0,

$$y = x_0 + \rho'(e_1,\dots,e_d), \qquad e_d\in\{\pm 1\},$$

and write $y = \sum a_i z_i$ as a (finite) convex combination of elements of Z. Then $\tilde{y} = \sum a_i\tilde{z}_i \in \mathrm{conv}(\tilde{Z})$ is such that $\|\tilde{y} - y\|_\infty \leq \epsilon$. It follows that $\tilde{y}$ lies at the same quadrant as y with each coordinate larger in absolute value than $\rho' - \epsilon$. In other words, $\tilde{y}$ is "more extreme" than the corner

$$x_0 + (\rho' - \epsilon)(e_1,\dots,e_d)$$

of the ℓ_∞-ball $B^\infty_{\rho'-\epsilon}(x_0)$. Since this is true for all the corners, $\mathrm{conv}(\tilde{Z}) \supseteq B_{\rho'-\epsilon}(x_0)$ for any $\rho' < \rho$. Now let $\rho' \nearrow \rho$ to conclude. □

5. Further Discussion. During the review process, a referee brought to our attention independent parallel work by Álvarez-Esteban et al. [2], that had concurrently been submitted for publication to a different journal. An Associate Editor suggested that we offer a brief comparison. Álvarez-Esteban et al. [2] also independently arrived at the same algorithm for determination of a Fréchet mean, and also showed convergence to a local minimum. Their motivation of the algorithm and their proof of convergence differ substantially from ours. Indeed, our algorithm is motivated by the geometry of the Wasserstein space, and is obtained as a gradient descent; while the one in [2] is motivated as a fixed point iteration through the special case of Gaussian measures, where it is known [1] that the Fréchet mean is the unique solution to a certain matrix equation. Also, rather than directly use the geometry of monotone operators in $\mathbb{R}^d$, the authors of [2] take advantage of an almost-sure representation result on the optimal transportation maps in order to prove convergence of their algorithm. Other key differences distinguishing our own contributions include the determination of the gradient of the Fréchet functional (Theorem 2), the criterion for determining when a local minimum is the global Fréchet mean (Theorem 3), and the parallels to Procrustes analysis (Section 5.1).

More importantly, [2] focusses on the algorithm itself, but does not make contact with the problem of optimal multicoupling and its applications to registration, which are arguably the main *statistical* reasons motivating the determination of a Fréchet mean. In contrast, these are a core part of our development and results, as we address the optimal multicoupling problem (Theorem 1), prove uniform convergence of the Procrustes maps required for multicoupling (Theorem 5), and give a detailed treatment of the statistical problem of nonparametric registration of multidimensional point processes, including asymptotic theory (Section 6).

References.

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *Society for Industrial and Applied Mathematics*, 43:2:904–924, 2011.

[2] P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, to appear.

[3] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):358–375, 2011.

[4] E. Boissard, T. Le Gouic, J.-M. Loubes, et al. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[5] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[6] R. B. Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013.

[7] V. M. Panaretos, Y. Zemel, et al. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

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