Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes

Anton Mallasto
Department of Computer Science
University of Copenhagen
mallasto@di.ku.dk

Aasa Feragen
Department of Computer Science
University of Copenhagen
aasa@di.ku.dk

Abstract

We introduce a novel framework for statistical analysis of populations of non-degenerate Gaussian processes (GPs), which are natural representations of uncertain curves. This allows inherent variation or uncertainty in function-valued data to be properly incorporated in the population analysis. Using the 2-Wasserstein metric we geometrize the space of GPs with L^2 mean and covariance functions over compact index spaces. We prove existence and uniqueness of the barycenter of a population of GPs, as well as convergence of the metric and the barycenter of their finite-dimensional counterparts. This justifies practical computations. Finally, we demonstrate our framework through experimental validation on GP datasets representing brain connectivity and climate development. A MATLAB library for relevant computations will be published at `https://sites.google.com/view/antonmallasto/software`.

1 Introduction

Gaussian processes (GPs, see Fig. 1) are the counterparts of Gaussian distributions (GDs) over functions, making GPs natural objects to model uncertainty in estimated functions. With the rise of GP modelling and probabilistic numerics, GPs are increasingly used to model uncertainty in function-valued data such as segmentation boundaries [17, 19, 29], image registration [38] or time series [27]. Centered GPs, or covariance operators, appear as image features in computer vision [12,16,24,25] and as features of phonetic language structure [22]. A natural next step is therefore to analyze populations of GPs, where performance depends crucially on proper incorporation of inherent uncertainty or variation. This paper contributes a principled framework for population analysis of GPs based on Wasserstein, a.k.a. earth mover's, distances.

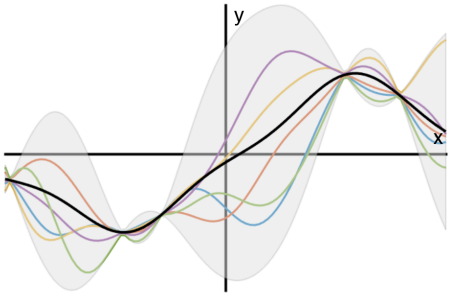

Figure 1: An illustration of a GP, with mean function (*in black*) and confidence bound (*in grey*). The colorful curves are sample paths of this GP.

The importance of incorporating uncertainty into population analysis is emphasized by the example in Fig. 2, where each data point is a GP representing the minimal temperature in the Siberian city Vanavara over the course of one year [9, 33]. A naïve way to compute its average temperature curve is to compute the per-day mean and standard deviation of the yearly GP mean curves. This is shown in the bottom right plot, and it is clear that the temperature variation is grossly underestimated,

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

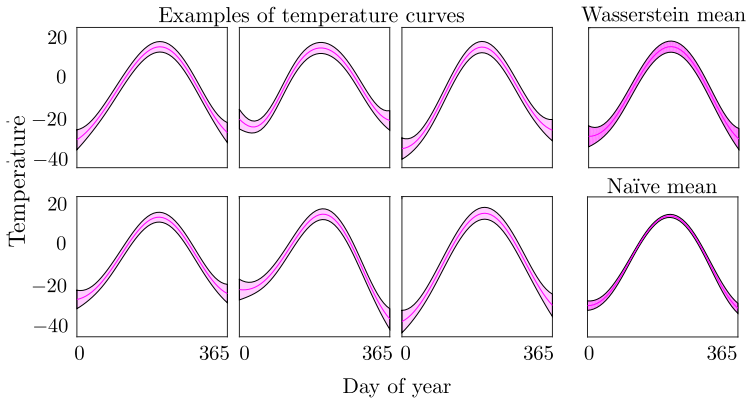

Figure 2: **Left:** Example GPs describing the daily minimum temperatures in a Siberian city (see Sec. 4). **Right top:** The mean GP temperature curve, computed as a Wasserstein barycenter. Note that the inherent variability in the daily temperature is realistically preserved, in contrast with the naïve approach. **Right bottom:** A naïve estimation of the mean and standard deviation of the daily temperature, obtained by taking the day-by-day mean and standard deviation of the temperature. All figures show a 95% confidence interval.

especially in the summer season. The top right figure shows the mean GP obtained with our proposed framework, which preserves a far more accurate representation of the natural temperature variation.

We propose analyzing populations of GPs by geometrizing the space of GPs through the *Wasserstein distance*, which yields a metric between probability measures with rich geometric properties. **We contribute** i) closed-form solutions for arbitrarily good approximation of the Wasserstein distance by showing that the 2-Wasserstein distance between two finite-dimensional GP representations converges to the 2-Wasserstein distance of the two GPs; and ii) a proof that the barycenter of a population of GPs exists, is unique, and can be approximated by its finite-dimensional counterpart.

We evaluate the Wasserstein distance in two applications. First, we illustrate the use of the Wasserstein distance for processing of uncertain white-matter trajectories in the brain segmented from noisy diffusion-weighted imaging (DWI) data using *tractography*. It is well known that the noise level and the low resolution of DWI images result in unreliable trajectories (*tracts*) [23]. This is problematic as the estimated tracts are e.g. used for surgical planning [8]. Recent work [17, 29] utilizes probabilistic numerics [28] to return *uncertain* tracts represented as GPs. We utilize the Wassertein distance to incorporate the estimated uncertainty into typical DWI analysis tools such as tract clustering [37] and visualization. Our second study quantifies recent climate development based on data from Russian meteorological stations using permutation testing on population barycenters, and supplies interpretability of the climate development using GP-valued kernel regression.

Related work. Multiple frameworks exist for comparing Gaussian distributions (GDs) represented by their covariance matrices, including the Frobenius, Fisher-Rao (affine-invariant), log-Euclidean and Wasserstein metrics. Particularly relevant to our work is the 2-Wasserstein metric on GDs, whose Riemannian geometry is studied in [32], and whose barycenters are well understood [1, 4].

A body of work exists on generalizing the aforementioned metrics to the infinite-dimensional covariance operators. As pointed out in [22], extending the affine-invariant and Log-Euclidean metrics is problematic as covariance operators are not compatible with logarithmic maps and their inverses are unbounded. These problems are avoided in [24, 25] by regularizing the covariance operators, but unfortunately, this also alters the data in a non-unique way. The Procrustes metric from [22] avoids this, but as it is, only defines a metric between covariance operators.

The 2-Wasserstein metric, on the other hand, generalizes naturally from GDs to GPs, does not require regularization, and can be arbitrarily well approximated by a closed form expression, making the computations cheap. Moreover, the theory of optimal transport [5, 6, 36] shows that the Wasserstein metric yields a rich geometry, which is further demonstrated by the previous work on GDs [32].

Structure. Prior to introducing the Wasserstein distance between GPs, we review GPs, their Hilbert space covariance operators and the corresponding Gaussian measures in Sec. 2. In Sec. 3 we introduce the Wasserstein metric and its barycenters for GPs and prove convergence properties of the metric and barycenters, when GPs are approximated by finite-dimensional GDs. Experimental validation is found in Sec. 4, followed by discussion and conclusion in Sec. 5.

2 Prerequisites

Gaussian processes and measures. A *Gaussian process* (GP) f is a collection of random variables, such that any finite restriction of its values $(f(x_i))_{i=1}^N$ has a joint Gaussian distribution, where $x_i \in X$, and X is the *index set*. A GP is entirely characterized by the pair

$$m(x) = \mathbb{E}\left[f(x)\right],\ k(x,x') = \mathbb{E}\left[(f(x)-m(x))(f(x')-m(x'))\right]\ , \tag{1}$$

where m and k are called the *mean function* and *covariance function*, respectively. We use the notation $f \sim \mathcal{GP}(m,k)$ for a GP f with mean function m and covariance function k. It follows from the definition that the covariance function k is symmetric and positive semidefinite. We say that f is *non-degenerate*, if k is positive definite. We will assume the GPs used to be non-degenerate.

GPs relate closely to *Gaussian measures* on Hilbert spaces. Given probability spaces (X,Σ_X,μ) and (Y,Σ_Y,ν), we say that the measure ν is a *push-forward* of μ if $\nu(A) = \mu(T^{-1}(A))$ for a measurable $T\colon X \to Y$ and any $A \in \Sigma_Y$. Denote this by $T_{\#}\mu = \nu$. A Borel measure μ on a separable Hilbert space $\mathcal{H}$ is a *Gaussian measure*, if its push-forward with respect to any non-zero continuous element of the dual space of $\mathcal{H}$ is a Gaussian measure on $\mathbb{R}$ (i.e., the push-forward gives a univariate Gaussian distribution). A Borel-measurable set B is a *Gaussian null set*, if $\mu(B) = 0$ for any Gaussian measure μ on X. A measure ν on $\mathcal{H}$ is *regular* if $\nu(B) = 0$ for any Gaussian null set B.

Covariance operators. Denote by $L^2(X)$ the space of L^2-integrable functions from X to $\mathbb{R}$. The covariance function k has an associated integral operator $K\colon L^2(X) \to L^2(X)$ defined by

$$[K\phi](x) = \int_X k(x,s)\phi(s)ds,\ \forall \phi \in L^2(X)\ , \tag{2}$$

called the *covariance operator* associated with k. As a by-product of the 2-Wasserstein metric on centered GPs, we get a metric on covariance operators. The operator K is Hilbert-Schmidt, self-adjoint, compact, positive, and of trace class, and the space of such covariance operators is a convex space. Furthermore, the assignment $k \mapsto K$ from $L^2(X \times X)$ to the covariance operators is an isometric isomorphism onto the space of positive Hilbert-Schmidt operators on $L^2(X)$ [7, Prop. 2.8.6]. This justifies us to write both $f \sim \mathcal{GP}(m,K)$ and $f \sim \mathcal{GP}(m,k)$.

Trace of an operator. The Wasserstein distance between GPs admits an analytical formula using traces of their covariance operators, as we will see below. Let $(\mathcal{H},\langle\cdot,\cdot\rangle)$ be a *separable* Hilbert space with the orthonormal basis $\{e_k\}_{k=1}^\infty$. Then the *trace* of a bounded linear operator T on $\mathcal{H}$ is given by

$$\text{Tr}\ T := \sum_{k=1}^{\infty} \langle Te_k, e_k\rangle\ , \tag{3}$$

which is absolutely convergent and independent of the choice of the basis if $\text{Tr}\ (T^*T)^{\frac{1}{2}} < \infty$, where T^* denotes the adjoint operator of T and $T^{\frac{1}{2}}$ is the square-root of T. In this case T is called a *trace class operator*. For positive self-adjoint operators, the trace is the sum of their eigenvalues.

The Wasserstein metric. The *Wasserstein metric* on probability measures derives from the optimal transport problem introduced by Monge and made rigorous by Kantorovich. The p-Wasserstein distance describes the minimal cost of transporting the unit mass of one probability measure into the unit mass of another probability measure, when the cost is given by a L^p distance [5, 6, 36].

Let (M,d) be a Polish space (complete and separable metric space) and denote by $\mathcal{P}_p(M)$ the set of all probability measures μ on M satisfying $\int_M d^p(x,x_0)d\mu(x) < \infty$ for some $x_0 \in M$. The p-Wasserstein distance between two probability measures $\mu,\nu \in \mathcal{P}_p(M)$ is given by

$$W_p(\mu,\nu) = \left(\inf_{\gamma\in\Gamma[\mu,\nu]} \int_{M\times M} d^p(x_1,x_2)d\gamma(x_1,x_2)\right)^{\frac{1}{p}},\ (x_1,x_2) \in M \times M, \tag{4}$$

where $\Gamma[\mu,\nu]$ is the set of joint measures on $M \times M$ with marginals μ and ν. Defined as above, W_p satisfies the properties of a metric. Furhermore, a minimizer in (4) is always achieved.

3 The Wasserstein metric for GPs

We will now study the Wasserstein metric with $p = 2$ between GPs. For GDs, this has been studied in [11, 14, 18, 21, 32].

From now on, assume that all GPs $f \sim \mathcal{GP}(m,k)$ are indexed over a compact $X \subset \mathbb{R}^n$ so that $\mathcal{H} := L^2(X)$ is separable. Furthermore, we assume $m \in L^2(X)$, $k \in L^2(X \times X)$, so that observations of f live almost surely in $\mathcal{H}$. Let $f_1 \sim \mathcal{GP}(m_1,k_1)$ and $f_2 \sim \mathcal{GP}(m_2,k_2)$ be GPs with associated covariance operators K_1 and K_2 , respectively. As the sample paths of f_1 and f_2 are in $\mathcal{H}$, they induce Gaussian measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathcal{H})$ on $\mathcal{H}$, as there is a 1-1 correspondence between GPs having sample paths almost surely on a $L^2(X)$ space and Gaussian measures on $L^2(X)$ [26].

The 2-Wasserstein metric between the Gaussian measures μ_1, μ_2 is given by [13]

$$W_2^2(\mu_1,\mu_2) = d_2^2(m_1,m_2) + \text{Tr}\,(K_1 + K_2 - 2(K_1^{\frac{1}{2}}K_2K_1^{\frac{1}{2}})^{\frac{1}{2}}), \tag{5}$$

where d_2 is the canonical metric on $L^2(X)$. Using this, we get the following definition

Definition 1. *Let f_1, f_2 be GPs as above, and the induced Gaussian measures of f_1 and f_2 be μ_1 and μ_2, respectively. Then, their squared 2-Wasserstein distance is given by*

$$W_2^2(f_1,f_2) := W_2^2(\mu_1,\mu_2) = d_2^2(m_1,m_2) + \mathit{Tr}\,(K_1 + K_2 - 2(K_1^{\frac{1}{2}}K_2K_1^{\frac{1}{2}})^{\frac{1}{2}})\,.$$

Remark 2. *Note that the case $m_1 = m_2 = 0$ defines a metric for the covariance operators K_1, K_2, as* (5) *shows that the space of GPs is isometric to the cartesian product of $L^2(X)$ and the covariance operators. We will denote this metric by $W_2^2(K_1,K_2)$. Furthermore, as GDs are just a subset of GPs, W_2^2 yields also the 2-Wasserstein metric between GDs studied in [11, 14, 18, 21, 32].*

Barycenters of Gaussian processes. Next, we define and study barycenters of populations of GPs, in a similar fashion as the GD case in [1].

Given a population $\{\mu_i\}_{i=1}^N \subset \mathcal{P}_2(\mathcal{H})$ and weights $\{\xi_i \geq 0\}_{i=1}^N$ with $\sum_{i=1}^N \xi_i = 1$, and $\mathcal{H}$ a separable Hilbert space, the solution $\bar{\mu}$ of the problem

$$(\mathcal{P}) \quad \inf_{\mu \in \mathcal{P}_2(\mathcal{H})} \sum_{i=1}^N \xi_i W_2^2(\mu_i,\mu),$$

is the *barycenter* of the population $\{\mu_i\}_{i=1}^N$ with *barycentric coordinates* $\{\xi_i\}_{i=1}^N$. The barycenter for GPs is defined to be the barycenter of the associated Gaussian measures.

We now state the main theorem of this section, which we will prove using Prop. 4 and Prop. 5 below.

Theorem 3. *Let $\{f_i\}_{i=1}^N$ be a population of GPs with $f_i \sim \mathcal{GP}(m_i,K_i)$, then the unique barycenter with barycentric coordinates $(\xi_i)_{i=1}^N$ is $\bar{f} \sim \mathcal{GP}(\bar{m},\bar{K})$, where $\bar{m}$ and $\bar{K}$ satisfy*

$$\bar{m} = \sum_{i=1}^N \xi_i m_i, \;\; \sum_{i=1}^N \xi_i \left(\bar{K}^{\frac{1}{2}}K_i\bar{K}^{\frac{1}{2}}\right)^{\frac{1}{2}} = \bar{K}.$$

Proposition 4. *Let $\{\mu_i\}_{i=1}^N \subset \mathcal{P}_2(\mathcal{H})$ and $\bar{\mu}$ be a barycenter with barycentric coordinates $(\xi_i)_{i=1}^N$. Assume μ_i is regular for some i, then $\bar{\mu}$ is the unique minimizer of $(\mathcal{P})$.*

Proof. We first show that the map $\nu \mapsto W_2^2(\mu,\nu)$ is convex, and strictly convex if μ is a regular measure. To see this, let $\nu_i \in \mathcal{P}_2(\mathcal{H})$ and $\gamma_i^* \in \Gamma[\mu,\nu_i]$ be the optimal transport plans between μ and ν_i for $i = 1,2$, then $\lambda\gamma_1^* + (1-\lambda)\gamma_2^* \in \Gamma[\mu, \lambda\nu_1 + (1-\lambda)\nu_2]$ for $\lambda \in [0,1]$. Therefore

$$\begin{aligned} W_2^2(\mu, \lambda\nu_1 + (1-\lambda)\nu_2) &= \inf_{\gamma \in \Gamma[\mu,\lambda\nu_1+(1-\lambda)\nu_2]} \int_{\mathcal{H}\times\mathcal{H}} d^2(x,y)d\gamma \\ &\leq \int_{\mathcal{H}\times\mathcal{H}} d^2(x,y)d(\lambda\gamma_1^* + (1-\lambda)\gamma_2^*) \\ &= \lambda W_2^2(\mu,\nu_1) + (1-\lambda)W_2^2(\mu,\nu_2), \end{aligned}$$

which gives convexity. Note that for $\lambda \in]0,1[$, the transport plan $\lambda\gamma_1^* + (1-\lambda)\gamma_2^*$ splits mass. Therefore it cannot be the unique optimal plan between μ and $(1-t)\nu_1 + t\nu_2$. As μ is regular, the optimal plan does not split mass, as it is induced by a map [3, Thm. 6.2.10], so we have strict convexity. From this follows the strict convexity of the object function in $(\mathcal{P})$. □

Next we characterize the barycenter in spirit of the finite-dinemsional case in [1, Thm. 6.1].

Proposition 5. *Let $\{f_i\}_{i=1}^N$ be a population of centered GPs, $f_i \sim \mathcal{GP}(0, K_i)$. Then $(\mathcal{P})$ has a unique solution $f \sim \mathcal{GP}(0, \bar{K})$, where $\bar{K}$ is the unique bounded self-adjoint positive linear operator satisfying*

$$F(K) := \sum_{i=1}^{N} \xi_i \left(K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}} = K. \tag{6}$$

Proof. First we show that (6) has a solution. Following the proof presented in [1, Thm. 6.1], let $\lambda_{\max}(K_i)$ be the maximum eigenvalue of K_i. Then pick β such that $\beta \geq \left(\sum_{i=1}^N \xi_i \sqrt{\lambda_{\max}(K_i)} \right)^2$ and define the convex set $K_\beta = \{K \mid \beta I \geq K > 0\}$, where $A \geq B$ denotes that the operator $A - B$ is positive.

Then note that the map F in (6) is a compact operator as the set of compact operators forms a two-sided ideal and is closed under taking the square-root, K_β is bounded, and so by the definition of a compact operator, $F(K_\beta)$ is contained in a compact set (the closure of the image). Finally, one can check that $\beta I \geq F(K) > 0$, so therefore by Schauder's fixed point theorem, there exists a solution for (6).

Next, we show that the solution to (6) is the barycenter. Let $\bar{K}$ satisfy (6) and $0 < \lambda_1, \lambda_2, ...$ be its eigenvalues with eigenfunctions $e_1, e_2, ...$ By [10, Prop. 2.2.] the transport map between μ and μ_k is given by

$$T_k(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{\langle x, e_j \rangle \langle (\bar{K}^{\frac{1}{2}} K_k \bar{K}^{\frac{1}{2}})^{\frac{1}{2}} e_j, e_i \rangle}{\lambda_i^{\frac{1}{2}} \lambda_j^{\frac{1}{2}}} e_i(x) \,, \tag{7}$$

for almost surely any $x \in \text{supp}(\mu)$ which equals the whole of $\mathcal{H}$ [34, Thm. 1].

Then one can check the identity $(\bar{K}^{\frac{1}{2}} K_k \bar{K}^{\frac{1}{2}})^{\frac{1}{2}} x = \bar{K}^{\frac{1}{2}} T_k \bar{K}^{\frac{1}{2}} x$, which gives

$$F(\bar{K})x = \sum_{k=1}^{N} \xi_k (\bar{K}^{\frac{1}{2}} K_k \bar{K}^{\frac{1}{2}})^{\frac{1}{2}} x = \sum_{k=1}^{N} \xi_k \bar{K}^{\frac{1}{2}} T_k \bar{K}^{\frac{1}{2}} x = \bar{K} x.$$

By noting that $\bar{K}^{\frac{1}{2}}$ is bijective, we get

$$\bar{K}^{\frac{1}{2}} \left(\sum_{k=1}^{N} \xi_k T_k \bar{K}^{\frac{1}{2}} x \right) = \bar{K} x \Rightarrow \sum_{k=1}^{N} \xi_k T_k \bar{K}^{\frac{1}{2}} x = \bar{K}^{\frac{1}{2}} x \overset{y := \bar{K}^{\frac{1}{2}} x}{\Rightarrow} \sum_{k=1}^{N} \xi_k T_k y = y, \ \forall y \in \mathcal{H}.$$

Therefore, by (3⇒1) in Proposition 3.8 in [1] we are done (replacing measures vanishing on small sets on $\mathbb{R}^n$ by regular measures on $\mathcal{H}$, the proof carries over). Also, by Proposition 4, this is the unique barycenter. □

Proof of Theorem 3. Use Prop. 5, the properties of a barycenter in a Hilbert space, and that the space of GPs is isometric to the cartesian product of $L^2(X)$ and the covariance operators. □

Remark 6. *For the practical computations of barycenters of GDs approximating GPs, to be discussed below, a fixed-point iteration scheme with a guarantee of convergence exists [4, Thm. 4.2].*

Convergence properties. Now, we show that the 2-Wasserstein metric for GPs can be approximated arbitrarily well by the 2-Wasserstein metric for GDs. This is important, as in real-life we observe finite-dimensional representations of the covariance operators.

Let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis for $L^2(X)$. Then we define the GDs given by restrictions m_{in} and K_{in} of m_i and K_i, $i = 1, 2$, on $V_n = \text{span}(e_1, ..., e_n)$ by

$$m_{in}(x) = \sum_{k=1}^{n} \langle m_i, e_k \rangle e_k(x),\ K_{in}\phi = \sum_{k=1}^{n} \langle \phi, e_k \rangle K_i e_k,\ \forall \phi \in V_n,\ \forall x \in X\ , \tag{8}$$

and prove the following:

Theorem 7. *The 2-Wasserstein metric between GDs on finite samples converges to the Wasserstein metric between GPs, that is, if $f_{in} \sim \mathcal{N}(m_{in}, K_{in})$, $f_i \sim \mathcal{GP}(m_i, K_i)$ for $i = 1, 2$, then*

$$\lim_{n\to\infty} W_2^2(f_{1n}, f_{2n}) = W_2^2(f_1, f_2).$$

By the same argument, it also follows that $W_2^2(\cdot, \cdot)$ is continuous in both arguments in operator norm topology.

Proof. $K_{in} \to K_i$ in operator norm as $n \to \infty$. Because taking a sum, product and square-root of operators are all continuous with respect to the operator norm, it follows that

$$K_{1n} + K_{2n} - 2(K_{1n}^{\frac{1}{2}} K_{2n} K_{1n}^{\frac{1}{2}})^{\frac{1}{2}} \to K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}.$$

Note that for any sequence $A_n \to A$ with convergence in operator norm, we have

$$|\text{Tr } A - \text{Tr } A_n| \leq \sum_{k=1}^{\infty} |\langle (A - A_n)e_k, e_k \rangle| \overset{\text{Cauchy-Schwarz}}{\leq} \sum_{k=1}^{\infty} \|(A - A_n)e_k\|_{L^2} \overset{\text{MCT}}{\to} 0\ , \tag{9}$$

as $\lim_{n\to\infty} \sup_{v \in L^2_\omega(X)} \|(A - A_n)v\|_{L^2} = 0$ due to the convergence in operator norm. Here MCT stands for the monotone convergence theorem. Thus we have

$$\begin{aligned} W_2^2(f_{1n}, f_{2n}) &= d_2^2(m_{1n}, m_{2n}) + \text{Tr } (K_{1n} + K_{2n} - 2(K_{1n}^{\frac{1}{2}} K_{2n} K_{1n}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &\overset{n\to\infty}{\to} d_2^2(m_1, m_2) + \text{Tr } (K_1 + K_2 - 2(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= W_2^2(f_1, f_2). \end{aligned}$$

□

The importance of Proposition 7 is that it justifies computations of distances using finite representations of GPs as approximations for the infinite-dimensional case.

Next, we show that we can also approximate the barycenter of a population of GPs by computing the barycenters of populations of GDs converging to these GPs.

Theorem 8. *The barycenter of a population of GPs varies continuously, that is, the map $(f_1, ..., f_N) \mapsto \bar{f}$ is continuous in the operator norm. Especially, this implies that the barycenter $\bar{f}_n$ of the finite-dimensional restrictions $\{f_{in}\}_{i=1}^N$ converges to $\bar{f}$.*

First, we show that if $f_i \sim \mathcal{GP}(m_i, K_i)$ and $\bar{f} = \mathcal{GP}(\bar{m}, \bar{K})$, then that the map $(K_1, ..., K_N) \mapsto \bar{K}$ is continuous. Continuity of $(m_1, ..., m_N) \mapsto \bar{m}$ is clear.

Let K be a covariance operator, denote its maximal eigenvalue by $\lambda_{\max}(K)$. Note that this map is well-defined, as K is also bounded, normal operator, thus $\lambda_{\max}(K) = \|K\|_{op} < \infty$ holds. Now let $\mathbf{a} = (K_1, ..., K_N)$ be a population of covariance operators, denote i^{th} as $\mathbf{a}(i) = K_i$, then define the continuous function β and correspondence (a set valued map) Φ as follows

$$\beta : \mathbf{a} \mapsto \left(\sum_{i=1}^{N} \xi_i \sqrt{\lambda_{max}(\mathbf{a}(i))} \right)^2,\ \Phi : \mathbf{a} \mapsto K_{\beta(\mathbf{a})} = \{K \in \text{HS}(\mathcal{H}) \mid \beta(\mathbf{a})I \geq K \geq 0\}.$$

Recall that β and Φ were already implicitly used in the proof of Proposition 6.

We want to show that this correspondence is continuous in order to put the Maximum theorem to use. A correspondence $\Phi : A \to B$ is *upper hemi-continuous* at $a \in A$, if all convergent sequences $(a_n) \in A$, $(b_n) \in \Phi(a_n)$ satisfy $\lim_{n\to\infty} b_n = b$, $\lim_{n\to\infty} a_n = a$ and $b \in \Phi(a)$. The correspondence is

lower hemi-continuous at $a \in A$, if for all convergent sequences $a_n \to a$ in A and any $b \in \Phi(a)$, there is a subsequence a_{n_k}, so that we have a sequence $b_k \in \Phi(a_{n_k})$ which satisfies $b_k \to b$. If the correspondence is both upper and lower hemi-continuous, we say that it is *continuous*. For more about the Maximum theorem and hemi-continuity, see [2].

Lemma 9. *The correspondence $\Phi : \mathbf{a} \mapsto K_{\beta(\mathbf{a})}$ is continuous as correspondence.*

Proof. First, we show the correspondence is lower hemi-continuous. Let $(\mathbf{a}_n)_{n=1}^{\infty}$ be a sequence of populations of covariance operators of size N, that converges $\mathbf{a}_n \to \mathbf{a}$. Use the shorthand notation $\beta_n := \beta(\mathbf{a}_n)$, then $\beta_n \to \beta_\infty := \beta(\mathbf{a})$, and let $\mathbf{b} \in \Phi(\mathbf{a}) = K_{\beta_\infty}$.

Pick subsequence $(\mathbf{a}_{n_k})_{k=1}^{\infty}$ so that $(\beta_{n_k})_{k=1}^{\infty}$ is increasing or decreasing. If it was decreasing, then $K_{\beta_\infty} \subseteq K_{\beta_{n_k}}$ for every n_k. Thus the proof would be finished by choosing $\mathbf{b}_k = \mathbf{b}$ for every k. Hence assume the sequence is increasing, so that $K_{\beta_{n_k}} \subseteq K_{\beta_{n_{k+1}}}$. Now let $\gamma(t) = (1-t)\mathbf{b}_1 + t\mathbf{b}$, where $\mathbf{b}_1 \in K_{\beta_1}$, and let t_{n_k} be the solution to $(1-t)\beta_1 + t\beta_\infty = \beta_{n_k}$, then $\mathbf{b}_k := \gamma(t_{n_k}) \in K_{\beta_{n_k}}$ and $\mathbf{b}_k \to \mathbf{b}$.

For upper hemicontinuity, assume that $\mathbf{a}_n \to \mathbf{a}$, $\mathbf{b}_n \in K_{\beta_n}$ and that $\mathbf{b}_n \to \mathbf{b}$. Then using the definition of Φ, we get the positive sequence $\langle(\beta_n I - \mathbf{b}_n)x, x\rangle \geq 0$ indexed by n, then by continuity and the positivity of this sequence it follows that

$$0 \leq \lim_{n\to\infty} \langle(\beta_n I - \mathbf{b}_n)x, x\rangle = \langle(\beta_\infty I - \mathbf{b})x, x\rangle.$$

One can check the criterion $\mathbf{b} \geq 0$ similarly, and so we are done. □

Proof of Theorem 8. Now let $\mathbf{a} = (K_1, ..., K_n)$, $\mathbf{f}(K, \mathbf{a}) := \sum_{i=1}^{N} \xi_i W_2^2(K, K_i)$ and $F(K) := \sum_{i=1}^{N} \xi_i (K^{\frac{1}{2}} K_i K^{\frac{1}{2}})^{\frac{1}{2}}$, then the unique minimizer $\bar{K}$ of $\mathbf{f}$ is the fixed point of F. Furthermore, the closure $\mathrm{cl}(F(K_{\beta(\mathbf{a})}))$ is compact, $\mathbf{a} \mapsto \mathrm{cl}(F(K_{\beta(\mathbf{a})}))$ is a continuous correspondence as the closure of composition of two continuous correspondence. Additionally, we know that $\bar{K} \in \mathrm{cl}(F(K_{\beta(a)}))$, so applying the maximum theorem, we have shown that the barycenter of a population of covariance operators varies continuously, i.e. the map $(K_1, ..., K_N) \mapsto \bar{K}$ is continuous, finishing the proof. □

4 Experiments

We illustrate the utility of the Wasserstein metric in two different applications: Processing of uncertain white-matter tracts estimated from DWI, and analysis of climate development via temperature curve GPs.

Experimental setup. The white-matter tract GPs are estimated for a single subject from the Human Connectome Project [15, 31, 35], using probabilistic shortest-path tractography [17]. See the supplementary material for details on the data and its preprocessing. From daily minimum temperatures measured at a set of 30 randomly sampled Russian metereological stations [9, 33], GP regression was used to estimate a GP temperature curve per year and station for the period $1940 - 2009$ using maximum likelihood parameters. All code for computing Wasserstein distances and barycenters was implemented in MATLAB and ran on a laptop with 2,7 GHz Intel Core i5 processor and 8 GB 1867 MHz DDR3 memory. On the temperature GP curves (represented by 50 samples), the average runtime of the 2-Wasserstein distance computation was 0.048 ± 0.014 seconds (estimated from 1000 pairwise distance computations), and the average runtime of the 2-Wasserstein barycenter of a sample of size 10 was 0.69 ± 0.11 seconds (estimated from 200 samples).

White-matter tract processing. The *inferior longitudinal fasiculus* is a white-matter bundle which splits into two separate bundles. Fig. 3 (top) shows the results of agglomerative hierarchical clustering of the GP tracts using average Wasserstein distance. The per-cluster Wasserstein barycenter can be used to represent the tracts; its overlap with the individual GP mean curves is shown in Fig. 3 (bottom).

The individual GP tracts are visualized via their mean curves, but they are in fact a population of GPs. To confirm that the two clusters are indeed different also when the covariance function is taken into account, we perform a permutation test for difference between per-cluster Wasserstein barycenters, and already with 50 permutations we observe a p-value of $p = 0.0196$, confirming that the two clusters are significantly different at a 5% significance level.

Quantifying climate change. Using the Wasserstein barycenters we perform nonparametric kernel regression to visualize how yearly temperature curves evolve with time, based on the Russian yearly temperature GPs. Fig. 4 shows snapshots from this evolution, and a continuous movie version `climate.avi` is found in the supplementary material. The regressed evolution indicates an increase in overall temperature as we reach the final year 2009. To quantify this observation, we perform a permutation test using the Wasserstein distance between population Wasserstein barycenters to compare the final 10 years 2000-2009 with the years 1940-1999. Using 50 permutations we obtain a p-value of 0.0392, giving significant difference in temperature curves at a 95% confidence level.

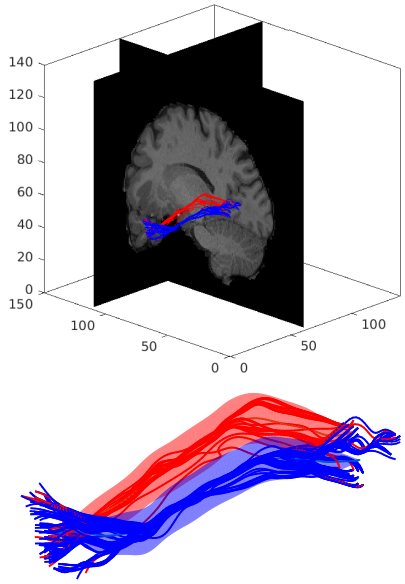

Figure 3: **Top:** The mean functions of the individual GPs, colored by cluster membership, in the context of the corresponding T1-weighted MRI slices. **Bottom:** The tract GP mean functions and the cluster mean GPs with 95% confidence bounds.

Significance. Note that the state-of-the-art in tract analysis as well as in functional data analysis would be to ignore the covariance of the estimated curves and treat the mean curves as observations. We contribute a framework to incorporate the uncertainty into the population analysis – but why would we want to retain uncertainty? In the white-matter tracts, the GP covariance represents spatial uncertainty in the estimated curve trajectory. The individual GPs represent connections between different endpoints. Thus, they do not represent observations of the exact same trajectory, but rather of distinct, nearby trajectories. It is common in diffusion MRI to represent such sets of estimated trajectories by a few prototype trajectories for visualization and comparative analysis; we obtain prototypes through the Wasserstein barycenter. To correctly interpret the spatial uncertainty, e.g. for a brain surgeon [8], it is crucial that the covariance of the prototype GP represents the covariances of the individual GPs, and not smaller. If you wanted to reduce uncertainty by increasing sample size, you would need more images, not more curves – because the noise is in the image. But more images are not usually available. In the climate data, the GP covariance models natural temperature variation, *not* measurement noise. Increasing the sample size decreases the error of the temperature distribution, but should not decrease this natural variation (i.e. the covariance).

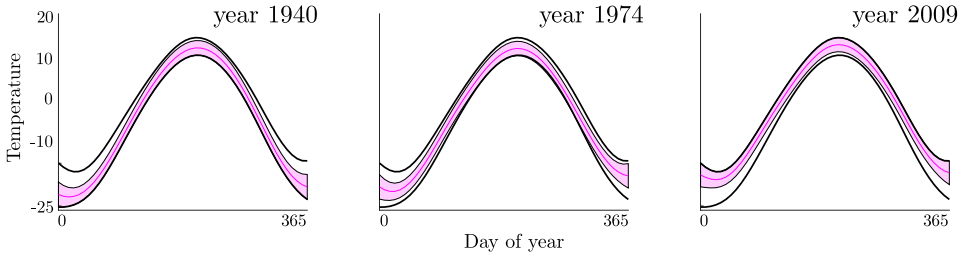

Figure 4: Snapshots from the kernel regression giving yearly temperature curves 1940-2009. We observe an apparent temperature increase which is confirmed by the permutation test.

5 Discussion and future work

We have shown that the Wasserstein metric for GPs is both theoretically and computationally well-founded for statistics on GPs: It defines unique barycenters, and allows efficient computations through finite-dimensional representations. We have illustrated its use in two different applications: Processing of uncertain estimates of white-matter trajectories in the brain, and analysis of climate development via GP representations of temperature curves. We have seen that the metric itself is discriminative for clustering and permutation testing, and we have seen how the GP barycenters allow truthful interpretation of uncertainty in the white matter tracts and of variation in the temperature curves.

Future work includes more complex learning algorithms, starting with preprocessing tools such as PCA [30], and moving on to supervised predictive models. This includes a better understanding of the potentially Riemannian structure of the infinite-dimensional Wasserstein space, which would enable us to draw on existing results for learning with manifold-valued data [20].

The Wasserstein distance allows the inherent uncertainty in the estimated GP data points to be appropriately accounted for in every step of the analysis, giving truthful analysis and subsequent interpretation. This is particularly important in applications where uncertainty or variation is crucial: Variation in temperature is an important feature in climate change, and while estimated white-matter trajectories are known to be unreliable, they are used in surgical planning, making uncertainty about their trajectories a highly relevant parameter.

6 Acknowledgements

This research was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The authors would also like to thank Mads Nielsen for valuable discussions and supervision.

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] C. Aliprantis and K. Border. Infinite dimensional analysis: a hitchhiker's guide. *Studies in Economic Theory*, 4, 1999.

[3] P. Álvarez-Esteban, E. Del Barrio, J. Cuesta-Albertos, C. Matrán, et al. Uniqueness and approximate computation of optimal incomplete transportation plans. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 47, pages 358–375. Institut Henri Poincaré, 2011.

[4] 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*, 441(2):744–762, 2016.

[5] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[6] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[7] W. Arveson. *A short course on spectral theory*, volume 209. Springer Science & Business Media, 2006.

[8] J. Berman. Diffusion MR tractography as a tool for surgical planning. *Magnetic resonance imaging clinics of North America*, 17(2):205–214, 2009.

[9] O. Bulygina and V. Razuvaev. Daily temperature and precipitation data for 518 russian meteorological stations. *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee*, 2012.

[10] J. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Diaz. On lower bounds for the l^2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, 9(2):263–283, 1996.

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

[12] M. Faraki, M. T. Harandi, and F. Porikli. Approximate infinite-dimensional region covariance descriptors for image classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1364–1368. IEEE, 2015.

[13] M. Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[14] C. R. Givens, R. M. Shortt, et al. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[15] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome project. *Neuroimage*, 80:105–124, 2013.

[16] M. Harandi, M. Salzmann, and F. Porikli. Bregman divergences for infinite dimensional covariance matrices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1003–1010, 2014.

[17] S. Hauberg, M. Schober, M. Liptrot, P. Hennig, and A. Feragen. A random Riemannian metric for probabilistic shortest-path tractography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 597–604. Springer, 2015.

[18] M. Knott and C. S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[19] M. Lê, J. Unkelbach, N. Ayache, and H. Delingette. GPSSI: Gaussian process for sampling segmentations of images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 38–46. Springer, 2015.

[20] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst. Geodesic convolutional neural networks on Riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

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

[22] D. Pigoli, J. A. Aston, I. L. Dryden, and P. Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014.

[23] S. Pujol, W. Wells, C. Pierpaoli, C. Brun, J. Gee, G. Cheng, B. Vemuri, O. Commowick, S. Prima, A. Stamm, et al. The DTI challenge: toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery. *Journal of Neuroimaging*, 25(6):875–882, 2015.

[24] M. H. Quang and V. Murino. From covariance matrices to covariance operators: Data representation from finite to infinite-dimensional settings. In *Algorithmic Advances in Riemannian Geometry and Applications*, pages 115–143. Springer, 2016.

[25] M. H. Quang, M. San Biagio, and V. Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In *Advances in Neural Information Processing Systems*, pages 388–396, 2014.

[26] B. S. Rajput. Gaussian measures on Lp spaces, $1 \leq p < \infty$. *Journal of Multivariate Analysis*, 2(4):382–403, 1972.

[27] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A*, 371(1984):20110550, 2013.

[28] M. Schober, D. K. Duvenaud, and P. Hennig. Probabilistic ODE solvers with Runge-Kutta means. In *Advances in neural information processing systems*, pages 739–747, 2014.

[29] M. Schober, N. Kasenburg, A. Feragen, P. Hennig, and S. Hauberg. Probabilistic shortest path tractography in DTI using Gaussian Process ODE solvers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–272. Springer, 2014.

[30] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015.

[31] S. Sotiropoulos, S. Moeller, S. Jbabdi, J. Xu, J. Andersson, E. Auerbach, E. Yacoub, D. Feinberg, K. Setsompop, L. Wald, et al. Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE. *Magnetic resonance in medicine*, 70(6):1682–1689, 2013.

[32] A. Takatsu et al. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[33] R. Tatusko and J. A. Mirabito. *Cooperation in climate research: An evaluation of the activities conducted under the US-USSR agreement for environmental protection since 1974*. National Climate Program Office, 1990.

[34] N. Vakhania. The topological support of Gaussian measure in Banach space. *Nagoya Mathematical Journal*, 57:59–63, 1975.

[35] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The wu-minn Human Connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

[36] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[37] D. Wassermann, L. Bloy, E. Kanterakis, R. Verma, and R. Deriche. Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. *NeuroImage*, 51(1):228–241, 2010.

[38] X. Yang and M. Niethammer. Uncertainty quantification for LDDMM using a low-rank Hessian approximation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 289–296. Springer, 2015.