Electronic Journal of Statistics
Vol. 14 (2020) 2742–2772
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1725

Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding

François Bachoc

Institut de Mathématiques de Toulouse
e-mail: francois.bachoc@math.univ-toulouse.fr

Alexandra Suvorikova

Weierstrass Institut,
and IITP RAS
e-mail: suvorikova@wias-berlin.de

David Ginsbourger

Idiap Research Institute,
and IMSV, University of Bern
e-mail: david.ginsbourger@stat.unibe.ch

Jean-Michel Loubes

Institut de Mathématiques de Toulouse
e-mail: loubes@math.univ-toulouse.fr

Vladimir Spokoiny

Weierstrass Institute, and HU Berlin,
IITP RAS, and HSE Moscow
e-mail: spokoiny@wias-berlin.de

Abstract: In this work, we propose a way to construct Gaussian processes indexed by multidimensional distributions. More precisely, we tackle the problem of defining positive definite kernels between multivariate distributions via notions of optimal transport and appealing to Hilbert space embeddings. Besides presenting a characterization of radial positive definite and strictly positive definite kernels on general Hilbert spaces, we investigate the statistical properties of our theoretical and empirical kernels, focusing in particular on consistency as well as the special case of Gaussian distributions. A wide set of applications is presented, both using simulations and implementation with real data.

MSC 2010 subject classifications: Primary 60G15.
Keywords and phrases: Kernel methods, Wasserstein distance, Hilbert space embeddings.

Received March 2019.

1. Introduction

Gaussian process models are widely used in fields such as geostatistics, computer experiments and machine learning [44, 46]. In a nutshell, Gaussian process modelling consists in assuming for an unknown function of interest to be one realisation of a Gaussian process, or equivalently of a Gaussian random field indexed by the source space of the objective function, and is often cast as part of the Bayesian arsenal for non-parametric estimation in function spaces. For instance, in computer experiments, the input points of the function are simulation input parameters and the output values are quantities of interest obtained from simulation responses. Furthermore, there has been a huge amount of literature dealing with the use of Gaussian processes in machine learning over the last decade. We refer for instance to [44, 48] or [20] and references therein.

Gaussian process models heavily rely on the specification of a covariance function, or "kernel", that characterizes linear dependencies between values of the process at different observation points. In fact, the kernel, which can be seen as a similarity measure between locations in the index space, also induces a (pseudo-)metric on the index space often referred to as the "canonical metric associated with the kernel" via the variogram function of geostatisticians. A natural question for a given kernel is how those inherently associated notions of similarity/dissimilarity interplay with prescribed metrics on the index space. In Euclidean space, one often speaks of radial or isotropic kernel for those covariance functions that are explicitly depending on the Euclidean distance between points. Radial kernels with respect to other metrics have also been investigated, see e.g. kernels writing as functions of the ℓ^1 distance in multivariate Euclidean spaces [54].

In this paper we consider Gaussian processes indexed by distributions supported on $\mathbb{R}^p$, and we investigate ways to build positive definite kernels based on the Wasserstein distance. Distributional inputs can occur in a number of practical situations and exploring admissible kernels for using Gaussian process and related methods in this context is a pressing issue. Situations of that kind include the case of uncertain vector inputs to a vector-to-scalar deterministic function, but also a variety of other settings such as histogram inputs standing for instance for ratings from a panel of experts, compositional data in geosciences, or randomized strategies in a Bayesian game-theoretic framework.

In some situations, distribution-valued inputs may arise as a convenient way to describe complex objects and media, e.g. a number of physical simulations require maps or parameter fields as inputs, and in some cases it can be beneficial to reparametrize them so as to work with probability distributions. For instance in [25], the computer model [29] is studied, where the input simulation parameter consists of a set of disks located on a unit square $[0,1]^2$, modelling a material, for which a stress output is associated. A Gaussian process model on distributions enables to treat the input sets of disks as measures, and to model the stress outputs as stemming from a random field indexed by the input distributions.

In this framework, a natural aim is to construct covariance functions for Gaussian processes indexed by such inputs, that is constructing positive definite kernels on sets of probability measures.

The simplest method is perhaps to compare a set of parametric features built from the probability distributions, such as the mean or the higher moments. This approach is limited, as the effect of such parameters does not take the whole distribution into account. Specific positive definite kernels should instead be designed so as to take their entire distribution inputs into account. This issue has recently been considered in [39] or [31]. We aim at basing these kernels on the Wasserstein, or transport-based, distance which was shown to be relevant and insightful for comparing or studying distributions [53, 19, 42, 16].

This issue has been studied for the one dimensional case in [8] or in [51], using the specific expression of the Wasserstein distance in dimension 1. Yet this case uses the property of the optimal coupling with the uniform random variable which is very specific to the one dimensional case. The positive definite kernels provided in the one dimensional case are not necessarily positive definite any longer, when they are extended to higher dimensions, as we illustrate numerically in Section 5. We refer to Remark 4 for a specific discussion of the one-dimensional case.

In the general dimension case, in order to build a positive definite kernel from the Wasserstein distance, we associate to each input distribution its optimal transport map to a reference distribution. We then provide positive definite kernels on the Hilbert space corresponding to the inverses of these optimal transport maps. This results in a positive definite kernel for multidimensional distributions. As a reference distribution, we recommend to take the empirical Fréchet mean (or barycenter) of the distributions. We remark that the notion of Wasserstein barycenters and their use in machine learning and in statistics has been tackled recently, for instance, in [2, 13, 15]. Although computational aspects of optimal transports are a difficult issue, substantial work has been conducted to provide feasible algorithms to compute barycenters and optimal transport maps, see for instance [32, 52], or [42] and references therein. Thus our suggested procedure is feasible in practice, as is confirmed by our simulation results on synthetic data and on the data from the CASTEM computer model [29, 25].

We also present a characterization of all the continuous radial positive definite and strictly positive definite kernels on Hilbert spaces. This is carried out by showing that they coincide with continuous radial positive definite and strictly positive definite kernels on Euclidean spaces of arbitrary dimension, and by revisiting existing results for the Euclidean case [54]. In addition, we show that when considering parametric families of covariance functions for Gaussian processes on infinite dimensional Hilbert spaces, all the covariance parameters are microergodic in general. Microergodicity is an important concept for the asymptotic analysis of Gaussian processes [50, 55, 5]. More precisely, in a parametric family of covariance functions, a covariance parameter is said to be microergodic if, for two different values of it, the two corresponding Gaussian measures induced by the two covariance functions are orthogonal. If a covariance parameter

is not microergodic, this means that there can not exist any consistent estimator of it, and also that changing the value of this parameter will not have an asymptotic impact on the predictions of the Gaussian process values [50, 55]. On the contrary, if a covariance parameter is microergodic, then the value it takes is asymptotically important for prediction, and it is possible to construct consistent estimators of it.

We provide furthermore statistical results related to our positive definite kernel construction. We study the asymptotic closeness of the two kernels obtained by taking the empirical barycenter and the population barycenter as reference distributions. We obtain additional more quantitative results in the special case of Gaussian input distributions. We also discuss stationarity and universality.

In the aforementioned simulations, we compare the Gaussian process regression model obtained from our suggested positive definite kernels with the distribution regression procedure of [43]. The results show the benefit of our method.

The paper falls into the following parts. In Section 2 we recall some definitions on kernels and on the notion of optimal transport, Wasserstein distance and Wasserstein barycenter of distributions. We also provide our positive definite kernel construction. The analysis of radial positive definite kernels and Gaussian processes on Hilbert spaces is provided in Section 3. Section 4 is devoted to the statistical results related to our kernel construction. The simulation results are provided in Section 5. Conclusions are discussed in Section 6. The proofs are postponed to the appendix.

2. Construction of positive definite kernels for distributions with Hilbert space embedding and optimal transport

2.1. Some basic notions of optimal transport

In this paper we focus on Gaussian processes for which the input parameters are in $\mathcal{W}_2(\mathbb{R}^p) \subset \mathcal{P}(\mathbb{R}^p)$, where $\mathcal{P}(\mathbb{R}^p)$ is the set of distributions supported on $\mathbb{R}^p$ and $\mathcal{W}_2(\mathbb{R}^p)$ is the subset of $\mathcal{P}(\mathbb{R}^p)$ composed of distributions with finite second moments. To study such models, Gaussian processes must be defined over the set of distributions.

Let us recall that a Gaussian process $(Y_x)_{x\in E}$ indexed by a set E is entirely characterized by its mean and covariance functions. A covariance function is defined by $(x, y) \in E \times E \mapsto \mathrm{Cov}(Y_x, Y_y)$. In general, a symmetric function $K : E\times E \to \mathbb{R}$ is actually the covariance of a (square-integrable) random process if and only if it is a *positive definite kernel*, that is for every $x_1, \cdots, x_n \in E$ and $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \tag{2.1}$$

In this case we say that K is a *covariance kernel*. If the quadratic form (2.1) is always strictly positive when $x_1, \ldots, x_n$ are two-by-two distinct and $\lambda_1, \ldots, \lambda_n$ are not all zero, then we say that K is a *strictly positive definite kernel*. We

also say that K is a *conditionally negative definite kernel* if the quadratic form in (2.1) is non-positive when $\sum_{i=1}^{n} \lambda_i = 0$.

Classical examples of strictly positive definite kernels, when E is an Euclidean space, are the square exponential, Matérn and power exponential ones, that are detailed below in (3.2) to (3.4). They can be extended to the case where E is an Hilbert space, see Section 3. As discussed above, the core of this paper is dedicated to the case where E is $\mathcal{W}_2(\mathbb{R}^p)$ or a subset of it.

The notions of Wasserstein distance and optimal transport will be central to our construction of positive definite kernels with inputs in $\mathcal{W}_2(\mathbb{R}^p)$. Let us introduce them now (see also [53]). For two μ, ν in $\mathcal{W}_2(\mathbb{R}^p)$, we denote by $\Pi(\mu, \nu)$ the set of all probability measures π over the product set $\mathbb{R}^p \times \mathbb{R}^p$ with first (resp. second) marginal μ (resp. ν).

The transportation cost with quadratic cost function, or quadratic transportation cost, between these two measures μ and ν is defined as

$$\mathcal{T}_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\pi(x, y). \tag{2.2}$$

In the above display and throughout this paper, we let $\|\cdot\|$ be the Euclidean norm on any Euclidean space. This transportation cost allows to endow the set $\mathcal{W}_2(\mathbb{R}^p)$ with a metric by defining the quadratic Monge-Kantorovich, or quadratic Wasserstein distance between μ and ν as

$$W_2(\mu, \nu) = \mathcal{T}_2(\mu, \nu)^{1/2}. \tag{2.3}$$

A probability measure π in $\Pi(\mu, \nu)$ realizing the infimum in (2.2) is called an optimal coupling. A random vector (X_1, X_2) with distribution π in $\Pi(\mu, \nu)$ realizing this infimum is also called an optimal coupling.

Our aim is to base our suggested covariance functions on the notion of optimal transport, and in particular of Wasserstein distance and barycenter. Indeed, the Wasserstein distance has been shown to be a very useful tool in statistics and machine learning [42, 19].

2.2. Construction of positive definite kernels by Hilbert space embedding of optimal transport maps

The class of positive definite kernels that we present here is based on the notion of optimal transport map, that we now introduce. Consider a reference distribution $\eta \in \mathcal{W}_2(\mathbb{R}^p)$, which will typically be chosen as a Wasserstein barycenter (see Section 2.3) and which is further discussed in Remark 3 below. For $\mu \in \mathcal{W}_2(\mathbb{R}^p)$, let $T_\mu : \mathbb{R}^p \to \mathbb{R}^p$ be the optimal transportation map defined by

$$T_{\mu\sharp}\mu = \eta$$

where $f_\sharp \pi = \pi \circ f^{-1}$ is the push-forward measure of a function f from a measure π, and

$$||\mathrm{id} - T_\mu||_{L^2(\mu)} = W_2(\mu, \eta).$$

Note that the map T_μ is uniquely defined when μ is absolutely continuous w.r.t. Lebesgue measure. Furthermore, T_μ is invertible from the support of μ to the support of η if also η is absolutely continuous.

Remark 1. *We point out that the existence of transportation maps that can be considered as gradients of convex functions is commonly referred to as Brenier's theorem and originated from Y. Brenier's work in the analysis and mechanics literature in [18]. Much of the current interest in transportation problems emanates from this area of mathematics. We conform to the common use of the name. It is worthwhile pointing out that a similar statement was established earlier independently in a probabilistic framework in [21], where the authors show the existence of an optimal transport map, for the quadratic cost, over Euclidean and Hilbert spaces, and prove the monotonicity of this optimal map in some sense (Zarantarello monotonicity).*

We are now in position to construct a positive definite kernel, by associating the transport map T_μ^{-1} to each distribution μ, and by using positive definite kernels on the Hilbert space $L^2(\eta)$, containing these transport maps. The following proposition provides the explicit kernel construction, and proves the positive definiteness and strict positive definiteness.

Proposition 1. *Let η be a continuous distribution in $\mathcal{W}_2(\mathbb{R}^p)$. Consider a positive definite kernel $\overline{K} : L^2(\eta) \times L^2(\eta) \to \mathbb{R}$. Consider the function K on the set of continuous distributions in $\mathcal{W}_2(\mathbb{R}^p)$ defined by*

$$K(\mu, \nu) = \overline{K}(T_\mu^{-1}, T_\nu^{-1}).$$

Then K is positive definite. Furthermore, if the function $\overline{K}$ above is strictly positive definite, then K is strictly positive definite.

Proof. We use the following classical mapping argument. For any $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and continuous distributions $\mu_1, \ldots \mu_n$,

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j K(\mu_i, \mu_j) = \sum_{i,j=1}^{n} \lambda_i \lambda_j \overline{K}\left(T_{\mu_i}^{-1}, T_{\mu_j}^{-1}\right) \geq 0 \tag{2.4}$$

because $\overline{K}$ is positive definite on the Hilbert space $L^2(\eta)$. This proves the first part of the proposition. Under the setting of the second part, if the $(\mu_i)_{i=1,\ldots,n}$ are two-by-two distinct, then the $(T_{\mu_i}^{-1})_{i=1,\ldots,n}$ are two-by-two distinct. Indeed $T_{\mu_i}^{-1} = T_{\mu_j}^{-1}$ implies $\mu_i = (T_{\mu_i}^{-1})_\sharp(T_{\mu_i\sharp}\mu_i) = (T_{\mu_i}^{-1})_\sharp\eta = (T_{\mu_j}^{-1})_\sharp\eta = (T_{\mu_j}^{-1})_\sharp(T_{\mu_j\sharp}\mu_j) = \mu_j$. Thus (2.4) is strictly positive when $\lambda_1, \ldots, \lambda_n$ are not all zero because $\overline{K}$ is strictly positive definite. □

From Proposition 1, any positive definite kernel on the Hilbert space $L^2(\eta)$ yields a corresponding positive definite kernel on the set of continuous distributions in $\mathcal{W}_2(\mathbb{R}^p)$. While, most classically, covariance functions operating on Euclidean spaces are considered, as can be seen in the general references discussed in the introduction, there exists a fair amount of work dedicated to covariance functions operating on (infinite dimensional) Hilbert spaces. One can for

instance mention the covariance function of the isonormal process [41]. Furthermore, covariance functions which inputs belong to the Hilbert space of square summable functions can be used for Gaussian process modelling of computer experiments with functional input parameters [40, 11]. In this paper, we will focus on covariance functions which depend on the Hilbert norm of the difference between their two inputs, when applying Proposition 1. These covariance functions are called radial covariance functions. Section 3 is dedicated to these radial covariance functions and the associated Gaussian processes. Before moving to this topic, we first conclude Section 2.2 with a few remarks and then address Wasserstein barycenters, in the aim of selecting the reference distribution η, in Section 2.3.

Remark 2. *Proposition 1 will still hold, even if T_μ^{-1} is not exactly the inverse of an optimal transport map. The only constraint for Proposition 1 to hold is that T_μ^{-1} is uniquely defined as a function of μ (and that the mapping $\mu \mapsto T_\mu^{-1}$ is injective, for the strict positive definiteness part of Proposition 1). Hence, in practice, we can use approximated optimal transport maps, and retain the positive definiteness, or strict positive definiteness, guarantee (see also Section 5).*

Remark 3. *Let us discuss the choice of the reference measure η. In the case where input distributions $\mu_1, \ldots, \mu_n$ are observed, we recommend to select their empirical barycenter as the reference distribution, $\eta = \overline{\mu}_n$ (see Section 2.3). If these distributions are realizations from a distribution $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ (see Section 2.3), the barycenter $\overline{\mu}$ of $\mathbb{P}$ (see Section 2.3) is also a good choice of a reference distribution, from a theoretical point of view. In the theoretical and numerical results in Sections 4 and 5, we use either the empirical barycenter $\overline{\mu}_n$ or its population counterpart $\overline{\mu}$ as the reference distribution η.*

Remark 4. *In the one dimensional case, it is actually possible to create covariance functions which values at $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$ are functions of $W_2(\mu, \nu)$ [8]. Indeed, in this case, using a covariance based on the Wasserstein distance amounts to using the following well-known optimal coupling (see [53]). For all $\mu \in \mathcal{P}(\mathbb{R})$ with finite second order moment, let*

$$Z_\mu := F_\mu^{-1}(U), \tag{2.5}$$

where F_μ^{-1} is defined as

$$F_\mu^{-1}(t) = \inf\{u, F_\mu(u) \geq t\},$$

and denotes the quantile function of the distribution μ, and where U is a uniform random variable on $[0, 1]$. The stochastic process given by $(Z_\mu)_{\mu \in \mathcal{W}_2(\mathbb{R})}$ can be seen as a non-Gaussian random field indexed by the set of distributions on the real line with finite second order moment. As such, its variogram

$$(\mu, \nu) \mapsto \mathbb{E}(Z_\mu - Z_\nu)^2 \tag{2.6}$$

defines a conditionally negative definite kernel, equal to $W_2^2(\mu, \nu)$ since the coupling (Z_μ) is optimal. This kernel can be used to construct families of covariance functions based on the one-dimensional Wasserstein distance, see [8].

In general dimension $p \geq 2$, however, this construction can not be extended and it is not clear which functions $F : \mathbb{R} \to \mathbb{R}$ are such that $(\mu, \nu) \mapsto F(W_2(\mu, \nu))$ are positive definite kernels. For instance, in Section 5 we provide simulations where the function $(\mu, \nu) \mapsto \exp(-W_2(\mu, \nu)^2)$ fails to be a positive definite kernel in the case $p = 2$ (while it is indeed a valid kernel when $p = 1$, see [8]). More precisely, we consider 900 distributions $\mu_1, \ldots, \mu_{900}$ on $\mathbb{R}^2$ that each are the uniform probability measure over the union of 10 disjoint disks. These 900 distributions model the presence of circular inclusions in materials and can serve as input parameters of the CASTEM simulation model (see Section 5.4). We show that the 900×900 matrix $(\exp(-W_2(\mu_i, \mu_j)^2))_{1 \leq i,j \leq 900}$ has strictly negative eigenvalues, implying that $(\mu, \nu) \mapsto \exp(-W_2(\mu, \nu)^2)$ indeed fails to be a positive definite kernel in the case $p = 2$.

2.3. Wasserstein barycenter and empirical barycenter

Two choices of reference distribution η that we advocate (see Remark 3) are based on the notion of Wasserstein barycenter, that we now introduce. When dealing with a collection of distributions $\mu_1, \ldots, \mu_n$, we can define a notion of variation of these distributions. For any $\nu \in \mathcal{W}_2(\mathbb{R}^p)$, set

$$\mathrm{Var}_{\mu_1,\ldots,\mu_n}(\nu) = \sum_{i=1}^{n} W_2^2(\nu, \mu_i).$$

Finding the distribution minimizing the variation of the distributions has been tackled by defining the notion of barycenter of distributions with respect to the Wasserstein distance in the seminal work of [2]. More precisely, given $p \geq 1$, the authors of [2] provide conditions to ensure the existence and uniqueness of the barycenter of the probability measures $(\mu_i)_{1 \leq i \leq n}$ with weights $(\lambda_i)_{1 \leq i \leq n}$, i.e. a minimizer of the following criterion

$$\nu \mapsto \sum_{i=1}^{n} \lambda_i W_2^2(\nu, \mu_i). \tag{2.7}$$

In the last years several works have studied the empirical properties of the barycenters and their applications to several fields. We refer for instance to [13, 15] and references therein. Hence the Wasserstein barycenter or Fréchet mean of distributions appears to be a meaningful feature to represent the mean behavior of a set of distributions.

This notion of Wasserstein barycenter has been recently extended to distributions defined in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$, that is the set of measures on $\mathcal{W}_2(\mathbb{R}^p)$ such that the corresponding (random) Wasserstein distance to any fixed distribution in $\mathcal{W}_2(\mathbb{R}^p)$ has finite variance. Let $\mathbb{P}$ be a distribution in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ and consider $\mu_1, \ldots, \mu_n$ i.i.d probabilities drawn according to the distribution $\mathbb{P}$. In this framework, the Wasserstein distance between a distribution $\mathbb{P}$ on $\mathcal{W}_2(\mathbb{R}^p)$ and a Dirac distribution δ_ν on $\mathcal{W}_2(\mathbb{R}^p)$ at a measure $\nu \in \mathcal{W}_2(\mathbb{R}^p)$ is defined as

$$W_2(\mathbb{P}, \delta_\nu) = \left(\int W_2^2(\nu, \mu) d\mathbb{P}(\mu) \right)^{1/2}. \tag{2.8}$$

If $\widetilde{\mu}$ is a random distribution obeying law $\mathbb{P}$, this corresponds to

$$W_2(\mathbb{P}, \delta_\nu) = \left(\mathbb{E}_{\{\widetilde{\mu}\sim\mathbb{P}\}} W_2^2(\widetilde{\mu}, \nu)\right)^{1/2}.$$

Note that we use the same notations for the Wasserstein distances between two distributions in $\mathcal{W}_2(\mathbb{R}^p)$ and between two distributions on distributions in $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ (one of the two being a Dirac measure). The space $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ inherits the properties of the space $\mathcal{W}_2(\mathbb{R}^p)$ and is a good choice for considering asymptotic properties of Wasserstein barycenteric sequences.

We define (if it exists) the Wasserstein barycenter of $\mathbb{P}$ as a probability measure $\overline{\mu}$ in $\mathcal{W}_2(\mathbb{R}^p)$ such that

$$\int W_2^2(\overline{\mu}, \mu) d\mathbb{P}(\mu) = \inf\left\{\int W_2^2(\nu, \mu) d\mathbb{P}(\mu),\ \nu \in \mathcal{W}_2(\mathbb{R}^p)\right\}.$$

First, we point out that the notion of barycenter developed in (2.7) also corresponds to the barycenter of the atomic probability $\mathbb{P}$ on the Wasserstein space, defined by

$$\mathbb{P} = \sum_{i=1}^{n} \lambda_i \delta_{\mu_i}.$$

We also recall some facts on the Wasserstein barycenter that are used in the rest of the paper. The following theorem from [3] guarantees the existence and uniqueness of this barycenter under some assumptions.

Theorem 1 (Existence of a Wasserstein Barycenter, [3])**.** *Let $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$. Assume that every distribution in the support of $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^p$. Then there exists a unique distribution $\overline{\mu} \in \mathcal{W}_2(\mathbb{R}^p)$ such that*

$$\int W_2^2(\overline{\mu}, \mu) d\mathbb{P}(\mu) = \inf_{\nu\in\mathcal{W}_2(\mathbb{R}^p)} \left\{\int W_2^2(\nu, \mu) d\mathbb{P}(\mu)\right\}$$

or, in other words,

$$\overline{\mu} = \operatorname*{argmin}_{\nu\in\mathcal{W}_2(\mathbb{R}^p)} \left\{\int W_2^2(\nu, \mu) d\mathbb{P}(\mu)\right\}. \tag{2.9}$$

Using the expression (2.8), we can see that Theorem 1 can be reformulated as stating the existence of the metric projection of $\mathbb{P}$ onto the subset of $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ composed of Dirac measures.

Consider a sample of i.i.d random distributions μ_i, $i = 1, \ldots, n$, drawn from the distribution $\mathbb{P}$ and set $\overline{\mu}$ to be the barycenter of $\mathbb{P}$. Let for fixed n, $\overline{\mu}_n$ be the empirical barycenter of the $\mu_1, \ldots, \mu_n$, defined as

$$\sum_{i=1}^{n} \lambda_i W_2^2(\overline{\mu}_n, \mu_i) = \inf\left\{\sum_{i=1}^{n} \lambda_i W_2^2(\nu, \mu_i),\ \nu \in \mathcal{W}_2(\mathbb{R}^p)\right\},$$

with $\lambda_1 = \ldots = \lambda_n = 1$. This empirical barycenter exists and is unique as soon as one of the μ_i is absolutely continuous w.r.t Lebesgue measure in $\mathbb{R}^p$. This result follows immediately from Proposition 6 in [34].

The following theorem, from [34], states that under uniqueness assumption the empirical Wasserstein barycenter $\overline{\mu}_n$ converges to the population Wasserstein barycenter $\overline{\mu}$.

Theorem 2 (Consistency of empirical barycenter, [34])**.** *Assume that $\mathbb{P}$ belongs to $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$ and that its barycenter is unique. Let $\mu_1, \ldots, \mu_n$ be independently drawn from $\mathbb{P}$ and let $\overline{\mu}_n$ be defined as above. Then the empirical barycenter $\overline{\mu}_n$ is a.s.-consistent, that is*

$$\lim_{n\to\infty} W_2(\overline{\mu}, \overline{\mu}_n) = 0, \quad a.s.$$

The above consistency theorem for the empirical barycenter will be useful in Section 4, where we will compare asymptotically two versions of our positive definite kernel construction in Proposition 1: one where the reference measure is the empirical barycenter and one where the reference measure is the population barycenter.

As announced in Section 2.3, the next section is now dedicated to radial covariance functions on Hilbert spaces and their associated Gaussian processes.

3. Radial kernels and associated Gaussian processes on Hilbert spaces

We consider a real Hilbert space H with inner product $\langle \cdot, \cdot \rangle_H$ and norm $\|\cdot\|_H$. In this section, we consider radial covariance functions on $H \times H$, that is covariance functions K defined by $K(h_1, h_2) = F(\|h_1 - h_2\|_H)$ for $h_1, h_2 \in H$ and $F : [0, \infty) \to \mathbb{R}$. In the sequel, we use the short notation $F(\|\cdot - \cdot\|_H)$ for a radial covariance function on $H \times H$.

In Section 3.1, we present a characterization of all the continuous functions F such that $F(\|\cdot - \cdot\|_H)$ is positive definite (resp. strictly positive definite) on $H \times H$. Specific examples that can readily be used in practice are provided in (3.2) to (3.4).

More precisely, in Propositions 2 and 3, we explain that $F(\|\cdot - \cdot\|_H)$ is a (strictly) positive definite kernel on any Hilbert space H, if and only if it is a (strictly) positive definite kernel when $H = \mathbb{R}^d$ for any $d \in \mathbb{N}$. Thanks to these results, in Proposition 4, we revisit classical results on radial positive definite functions on $\mathbb{R}^d$ [54], by showing that when F is continuous, $F(\|\cdot - \cdot\|_H)$ is strictly positive definite if and only if $F(\sqrt{\cdot})$ is completely monotone and if and only if F is an integral of negative square exponential functions with respect to a finite measure.

Then, Section 3.2 is dedicated to the microergodicity of covariance parameters of families of radial covariance functions on Hilbert spaces. We show in Theorem 3 that when H is of infinite dimension, virtually all covariance parameters are microergodic when considering Gaussian processes on bounded sets.

3.1. Characterization of radial positive definite kernels

We consider kernels $K : H \times H \to \mathbb{R}$ of the form

$$K(u, v) = F(\|u - v\|_H), \tag{3.1}$$

for $u, v \in H$. We call them radial kernels. The next proposition shows that F provides a positive definite kernel on any Hilbert space H if and only if it does so on finite dimensional Euclidean spaces.

Proposition 2. *Let $F : \mathbb{R}^+ \to \mathbb{R}$. Then the two following statements are equivalent.*

1. *For any $d \in \mathbb{N}$, the kernel $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by $K_d(x, y) = F(\|x - y\|)$ for $x, y \in \mathbb{R}^d$ is positive definite.*
2. *For any Hilbert space H, the kernel K of the form (3.1) is positive definite.*

Next, we provide a similar characterization of the strict positive definitness property.

Proposition 3. *Let $F : \mathbb{R}^+ \to \mathbb{R}$. Then the two following statements are equivalent.*

1. *For any $d \in \mathbb{N}$, the kernel $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by $K_d(x, y) = F(\|x - y\|)$ for $x, y \in \mathbb{R}^d$ is strictly positive definite.*
2. *For any Hilbert space H, the kernel K of the form (3.1) is strictly positive definite.*

In the case where F is continuous, we can use the existing work on radial kernels on $\mathbb{R}^d$ (see e.g. [54]) to further characterize the functions F providing strictly positive definite kernels in (3.1). In this view, we call a function $f : [0, \infty) \to \mathbb{R}$ completely monotone if it is C^∞ on $(0, \infty)$, continuous at 0 and satisfies $(-1)^\ell f^{(\ell)}(r) \geq 0$ for $r > 0$ and $\ell \in \mathbb{N}$.

Proposition 4. *Let $F : \mathbb{R}^+ \to \mathbb{R}$ be continuous. Then the following statements are equivalent.*

1. *For any Hilbert space H, the kernel $K : H \times H \to \mathbb{R}$ of the form (3.1), defined by $K(u, v) = F(\|u - v\|_H)$, is strictly positive definite.*
2. *$F(\sqrt{\cdot})$ is completely monotone on $[0, \infty)$ and not constant.*
3. *There exists a finite nonnegative Borel measure ν on $[0, \infty)$ that is not concentrated at zero, such that*

$$F(t) = \int_{\mathbb{R}} e^{-ut^2} \nu(du).$$

Proof. The proposition is a direct consequence of Proposition 3 and Theorem 7.14 in [54]. □

Remark 5. *We remark that the version of Proposition 4 where strict positive definiteness is replaced by positive definiteness, and where $F(.)$ does not have*

to be non-constant, is provided by Schoenberg [47] (see also [9]). Compared to these two references, Proposition 4 enables to provide a characterization of strict positive definiteness on Hilbert spaces. In addition, the proof of Proposition 4 is here very short, after Propositions 2 and 3 are established. These two latter propositions and their short proofs provide an useful interpretation of the relationship between (strict) positive definiteness on Euclidean spaces and on Hilbert spaces.

The previous proposition ensures that the following choices of F can be used in (3.1) to provide strictly positive definite covariance functions on H. The square exponential covariance function is given by

$$F_{\sigma^2,\ell}(t) = \sigma^2 e^{-(t/\ell)^2}, \tag{3.2}$$

with $\sigma^2, \ell \in (0, \infty)$. The Matérn covariance function is given by

$$F_{\sigma^2,\ell,\nu}(t) = \frac{\sigma^2 (t/\ell)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(t/\ell), \tag{3.3}$$

where $\sigma^2, \ell, \nu \in (0, \infty)$, where Γ is the Gamma function and K_ν is the modified Bessel function of the second kind [50, 35]. Finally, the power exponential function

$$F_{\sigma^2,\ell,s}(t) = \sigma^2 \exp(-(t/\ell)^s), \tag{3.4}$$

where $\sigma^2, \ell \in (0, \infty)$ and $s \in (0, 2]$, satisfies the condition of Proposition 4 (see e.g. [8]).

One can also remark that, while Mercer's theorem has become classic for continuous positive definite kernels on compact sets of $\mathbb{R}^d$ [54], a similar construction has not been shown to exist on bounded subsets of Hilbert spaces in infinite dimension. This can be considered as a structural difficulty when tackling Gaussian processes on infinite dimensional Hilbert spaces. On the other hand, we now show that infinite dimensional Hilbert spaces provide more space, so to speak, that enable to distinguish between distinct covariance functions in a more stringent way. More precisely, we show next that, when considering parametric sets of covariance functions, virtually all the covariance parameters are microergodic.

3.2. Microergodicity results

Let H be a Hilbert space. Consider a set of functions $\{F_\theta; \theta \in \Theta\}$, with $F_\theta : \mathbb{R}^+ \to \mathbb{R}$ for $\theta \in \Theta$ and with $\Theta \subset \mathbb{R}^q$. To F_θ we associate the radial covariance function on $H \times H$ defined by $K_\theta(h_1, h_2) = F_\theta(\|h_1 - h_2\|_H)$ for $h_1, h_2 \in H$.

Let $h_0 \in H$ and $0 < L < \infty$ be fixed and let $\overline{\mathcal{B}}_{2,L} = \{h \in H; ||h - h_0||_H \leq L\}$. Let $\overline{F} = \mathbb{R}^{\overline{\mathcal{B}}_{2,L}}$ be the set of functions from $\overline{\mathcal{B}}_{2,L}$ to $\mathbb{R}$. Let $\mathcal{F}$ be the cylinder sigma algebra on $\overline{F}$ generated by the functions $f \mapsto (f(h_1), \ldots, f(h_r))$ for any $r \in \mathbb{N}$ and $h_1, \ldots, h_r \in H$. For any $\theta \in \Theta$, let $\mathbb{P}_\theta$ be the measure on $(\overline{F}, \mathcal{F})$ equal to the law of a Gaussian process on $\overline{\mathcal{B}}_{2,L}$ with mean function zero and

covariance function $(h_1, h_2) \mapsto F_\theta(||h_1 - h_2||_H)$. Then, following [50], we say that the covariance parameter θ is microergodic if, for any $\theta_1, \theta_2 \in \Theta$ with $\theta_1 \neq \theta_2$, the measures $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are orthogonal, that is there exists $\mathcal{A} \in \mathcal{F}$ such that $\mathbb{P}_{\theta_1}(A) = 1$ and $\mathbb{P}_{\theta_2}(A) = 0$.

In the most classical case where $H = \mathbb{R}^d$, microergodicity is an important concept. Indeed, it is a necessary condition for consistent estimators of θ to exist under fixed-domain asymptotics [50], and a fair amount of work has been devoted to showing microergodicity or non-microergodicity of parameters, for various models of covariance functions [50, 55, 5]. Typically, when $H = \mathbb{R}^d$ there are several standard sets of functions $\{F_\theta; \theta \in \Theta\}$ for which θ is not microergodic. A classical example is the set $\{F_{\sigma^2,\ell,\nu}\}$ of the form (3.3) [55].

In contrast, we now show that, under very mild assumptions, all covariance parameters θ are microergodic when H has infinite dimension.

Theorem 3. *Assume that H has infinite dimension. Assume that there does not exist $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$, such that $t \mapsto F_{\theta_1}(t) - F_{\theta_2}(t)$ is constant on $[0, 2L]$. Then the covariance parameter θ is microergodic.*

In Theorem 3, the condition on the parametric family $\{F_\theta; \theta \in \Theta\}$ holds for all the commonly used families of functions F_θ that are used to construct covariance functions on $\mathbb{R}^d$ as in Proposition 3. These commonly used families include notably the Matérn covariance functions and the power exponential covariance functions that are introduced above. They also include the generalized Wendland covariance functions and the spherical covariance functions [12, 1].

Hence, Theorem 3 shows that it is possible that consistent estimators exist for θ, in many parametric models of covariance functions of the form (3.1), for infinite dimensional Hilbert spaces.

Let us conclude this section by putting Theorem 3 into perspective with the case of finite-dimensional input spaces with increasing dimension. For $d \in \mathbb{N}$ and $u_0 \in \mathbb{R}^d$, write $\overline{\mathcal{B}}_{d,2,L} = \{u \in \mathbb{R}^d, ||u - u_0|| \leq L\}$. Consider the set of covariance functions $\{K_{d,\theta}; \theta \in \Theta\}$ on $\overline{\mathcal{B}}_{d,2,L}$ defined by $K_{d,\theta}(u_1, u_2) = F_\theta(||u_1 - u_2||)$. In the case where $H = \mathbb{R}^d$, θ is microergodic if, for each $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$, there exists a measurable set A of functions from $\overline{\mathcal{B}}_{d,2,L}$ to $\mathbb{R}$, for which $P_{d,\theta_1}(A) = 0$ and $P_{d,\theta_2}(A) = 1$, where $P_{d,\theta}$ is the distribution of a Gaussian process on $\overline{\mathcal{B}}_{d,2,L}$ with mean zero and covariance function $K_{d,\theta}$. Let us say in this case that θ is microergodic for the dimension d. We then have the following lemma.

Lemma 1. *Assume that θ is microergodic for the dimension $d_1 \in \mathbb{N}$ and let $d_2 \in \mathbb{N}$, $d_2 \geq d_1$. Then θ is microergodic for the dimension d_2.*

Hence, for families of functions $\{F_\theta : \mathbb{R}^+ \to \mathbb{R}\}$ applied to Euclidean distances in dimension d, a higher dimension d increases the possiblity that θ is microergodic for the dimension d. In agreement with this fact, Theorem 3 can be interpreted as follows: when d is infinite, θ is always microergodic.

4. Statistical properties of our suggested positive definite kernels on distributions

4.1. General consistency properties

Here, we consider the case where n i.i.d. random continuous distributions $\mu_1, \ldots, \mu_n$ are observed, from a distribution $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^p))$. Hence, two possible reference distributions for our suggested construction of Proposition 1 are the empirical barycenter $\overline{\mu}_n$ of $\mu_1, \ldots, \mu_n$ and the barycenter $\overline{\mu}$ of $\mathbb{P}$. We now show that these two reference points will asymptotically give the same kernel when n is large.

For $\mu \in \mathcal{W}_2(\mathbb{R}^p)$, let $T_\mu, T_{\mu,n} : \mathbb{R}^p \to \mathbb{R}^p$ be the optimal transportation maps defined by

$$T_{\mu\sharp}\mu = \overline{\mu} \quad , \quad T_{\mu,n\sharp}\mu = \overline{\mu}_n$$

and

$$||\mathrm{id} - T_\mu||_{L^2(\mu)} = W_2(\mu, \overline{\mu}) \quad , \quad ||\mathrm{id} - T_{\mu,n}||_{L^2(\mu)} = W_2(\mu, \overline{\mu}_n).$$

Let also, for $i = 1, \ldots, n$, $T_i = T_{\mu_i}$ and $T_{i,n} = T_{\mu_i,n}$.

We remark that, because of the assumption on $\mathbb{P}$, both the barycenter and the empirical barycenter are absolutely continuous w.r.t Lebesgue measure on $\mathbb{R}^p$. Hence, $T_1, \ldots, T_n$ and $T_{1,n}, \ldots, T_{n,n}$ are uniquely defined. For $F : \mathbb{R}^+ \to \mathbb{R}$, we let

$$K_n(\mu, \nu) = F(\|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|^2_{L^2(\overline{\mu}_n)}) \tag{4.1}$$

be the empirical kernel and

$$K(\mu, \nu) = F(\|T_\mu^{-1} - T_\nu^{-1}\|^2_{L^2(\overline{\mu})}) \tag{4.2}$$

be the theoretical kernel. We now prove that the empirical kernel K_n provides a good approximation of the kernel K. We will use the consistency property of Theorem 2, stating that the empirical barycenter is a consistent estimate for $\overline{\mu}$.

Proposition 5 (Consistency of kernel)**.** *Let F in (4.1) and (4.2) be continuous. The empirical kernel is a good approximation of the true covariance kernel in the sense that, for any two fixed absolutely continuous measures μ and ν in $\mathcal{W}_2(\mathbb{R}^p)$, we have*

$$K_n(\mu, \nu) \xrightarrow{\mathbb{P}\text{-}a.s.} K(\mu, \nu),$$

when n goes to infinity.

Proof. Using the continuity of the function F, it is enough to show that

$$\|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|^2_{L^2(\mu_n)} - \|T_\mu^{-1} - T_\nu^{-1}\|^2_{L^2(\overline{\mu})} \xrightarrow{\mathbb{P}\text{-a.s.}} 0.$$

Lemma 3, whose proof is presented in the Appendix, leads to the result. □

In the next Corollary, we show that the consistency result in Proposition 5 implies that the conditional means and variances based on the empirical kernel asymptotically coincide with those based on the true kernel.

Corollary 1. *Let $N \in \mathbb{N}$ and let $\nu_1, \ldots, \nu_N, \nu$ be fixed absolutely continuous measures in $\mathcal{W}_2(\mathbb{R}^p)$. Let $y = (y_1, \ldots, y_N)^\top$ be fixed in $\mathbb{R}^N$. Set $R = (K(\nu_i, \nu_j))_{1 \leq i,j \leq N}$, with the notation (4.2), and assume that R is invertible. Let $Y = (Y_\mu)$ be a Gaussian process with zero mean function and covariance function given by (4.2). Then*

$$\mathbb{E}(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) = r_\nu^\top R^{-1} y$$

with $r_\nu = (K(\nu, \nu_1), \ldots, K(\nu, \nu_N))^\top$. Let

$$\mathbb{E}_n(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) = r_{\nu,n}^\top R_n^{-1} y$$

with $r_{\nu,n} = (K_n(\nu, \nu_1), \ldots, K_n(\nu, \nu_N))^\top$ and $R_n = (K_n(\nu_i, \nu_j))_{1 \leq i,j \leq N}$, with the notation (4.1). Also

$$\mathrm{Var}(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) = K(\nu, \nu) - r_\nu^\top R^{-1} r_\nu$$

and we let

$$\mathrm{Var}_n(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) = K_n(\nu, \nu) - r_{\nu,n}^\top R_n^{-1} r_{\nu,n}.$$

Then, as $n \to \infty$,

$$\mathbb{E}_n(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) \xrightarrow{\mathbb{P}\text{-a.s.}} \mathbb{E}(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N)$$
$$\mathrm{Var}_n(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N) \xrightarrow{\mathbb{P}\text{-a.s.}} \mathrm{Var}(Y_\nu | Y_{\nu_1} = y_1, \ldots, Y_{\nu_N} = y_N).$$

Proof. The Corollary is a direct consequence of the facts that N is fixed as $n \to \infty$ and that R is invertible. □

Remark that consistency results similar to the one of Corollary 1 could be obtained, for other statistical or machine learning applications of the empirical and theoretical kernels in (4.1) and (4.2). In particular, the estimated Hilbert-Schmidt independence criterion (HSIC) [27] enables to test if two random variables on $\mathcal{W}_2(\mathbb{R}^p)$ are independent. The estimated HSIC can be computed from a kernel on $\mathcal{W}_2(\mathbb{R}^p)$ (or a subset of it) and from two samples from the two random variables. Thus, the kernels (4.1) and (4.2) enable to test if two samples of distributions come from independent variables. In this setting, a convergence result similar as Corollary 1 holds. Indeed, the two samples being fixed, it is clear from the expression of the estimated HSIC (see (3) in [27]) that the estimated HSIC computed from the empirical kernel (4.1) converges to the one computed from the theoretical kernel (4.2).

Another standard application of positive definite kernels consists in the support vector machines binary classifiers [28, 10]. The classifier function obtained from support vector machines is defined by parameters that are solution of a convex optimization problem. The corresponding objective function depends continuously on the kernel and on the data set. Hence, similarly as above, Corollary 1 can be extended to show that the classifier obtained from the empirical kernel (4.1) converges to the one obtained from the theoretical kernel (4.2).

These types of convergence results also show that statistical applications of our suggested kernels are robust to the choice of the reference measure, in the sense that the statistical results depend continuously on the choice of this reference measure.

Finally, the above arguments can be applied to other methods involving kernels, see the examples in [28, 10]. Furthermore, we find our introduced kernel construction technique to be of potential use for modelling Gaussian processes on a family of Hi-C interaction matrices [33], as it suggests a convenient method of information analysis related to gene expression. We also consider this subject, and related consistency results, as a possible direction for further research.

4.2. Universality

Note that when considering a kernel K, a natural property to be studied would be its universality. Actually, a kernel is said to be universal on $\Omega \subset \mathcal{W}_2(\mathbb{R}^p)$ as soon as the space generated by its linear combinations $\mu \in \Omega \mapsto \sum_{i=1}^n \alpha_i K(\mu, \mu_i) \in \mathbb{R}$ can generate all continuous functions on Ω. The general form (4.2) of the kernel may provide universal kernels under regularity assumptions on the transportation maps T_i. More precisely, injectivity and continuity are required as pointed out in [38] to get a universal kernel. In some particular cases, it is possible to obtain such results. In the case of Gaussian distributions, the transport map is linear and thus it entails the universality of the kernel in this case. In [23], Proposition 1.4.1, derived from Theorem 1.1 from [24], some conditions for the continuity of the transportation maps are provided but regularity of transportation maps in general dimensions is a difficult issue. It has received a lot of attention in the last years see for instance [45]. These types of conditions in [23] can not be guaranteed in a very general framework but could only be studied for very particular classes of distributions, leading to too restrictive cases, which are not at the heart of this paper.

4.3. Specific properties for Gaussian distributions

In some special cases, the optimal transportation maps can be written explicitly. Unfortunately, this holds only for a particular class of admissible transformations. An example of explicit calculations is given by a family of Gaussian distributions. Let $\mathcal{F} = \{\mathcal{N}(0, S)\}_S \subset E$, with E being all centred Gaussian distributions with non-degenerated covariance matrices w.l.g. supported on $\mathbb{R}^d$. Further we assume the covariance matrices in $\mathcal{F}$ to be random: $S \overset{iid}{\sim} \mathbb{P}$. This setting is equivalent to the definition of some distribution $\mathbb{P}$ over $\mathcal{F}$. We denote as $\overline{\mu} = \mathcal{N}(0, \overline{S})$ the unique population barycenter of $\mathbb{P}$.

Let $\{\mu_i\}_{i=1,\dots,n}$ be a family of observed random Gaussian distributions with zero mean and non-degenerated covariance S_i: $\mu_i = \mathcal{N}(0, S_i)$, $S_i \sim \mathbb{P}$. An empirical barycenter is recovered uniquely: $\overline{\mu}_n = \mathcal{N}(0, \overline{S}_n)$ with $\overline{S}_n$ a solution of the following fixed-point equation $\overline{S}_n = \frac{1}{n} \sum \left(S_i^{1/2} \overline{S}_n S_i^{1/2}\right)^{1/2}$. This result is

well known and has been described in many papers, see for instance the seminal work [2]. The solution can be obtained by an iterative method presented in [4].

The Gaussian setting allows to write explicitly a formula for the optimal transport map T_i between μ_i and the population barycenter $\overline{\mu} = \mathcal{N}(0, \overline{S})$ and its inverse:

$$T_i = S_i^{-1/2}\left(S_i^{1/2}\overline{S}S_i^{1/2}\right)^{1/2}S_i^{-1/2}, \quad T_i^{-1} = \overline{S}^{-1/2}\left(\overline{S}^{1/2}S_i\overline{S}^{1/2}\right)^{1/2}\overline{S}^{-1/2}.$$

In this case, we can compute the distance between the transport maps in $L^2(\overline{\mu})$ using the expression in (4.3) $\|T_i^{-1} - T_j^{-1}\|^2_{L^2(\overline{\mu})}$, as the distance is the variance of a linear transform of a Gaussian random variable:

$$\|T_i^{-1} - T_j^{-1}\|^2_{L^2(\overline{\mu})} = \left\|\overline{S}^{-1/2}\left[\left(\overline{S}^{1/2}S_i\overline{S}^{1/2}\right)^{1/2} - \left(\overline{S}^{1/2}S_j\overline{S}^{1/2}\right)^{1/2}\right]^{1/2}\right\|_F^2. \quad (4.3)$$

Here and in what follows we use $\|\cdot\|_F$ to denote the Frobenius norm. The same expression holds for $\left\|T_{i,n}^{-1} - T_{j,n}^{-1}\right\|^2_{L^2(\overline{\mu}_n)}$ by replacing the barycenter by its empirical counterpart. We can see that in this case the kernel amounts to compute a natural distance between the two distributions μ_i and μ_j obtained by the scale deformations $S_i^{1/2}X$ and $S_j^{1/2}X$ of a Gaussian random variable $X \sim \mathcal{N}(0, Id)$. This distance is then used through any kernel which provides some insights on a proper notion of covariance between processes indexed by these two distributions.

We point out that in the Gaussian case, the rate of convergence of the covariance estimates can be made precise.

Proposition 6. *Let $\mathcal{F}$ be s.t. $\mathbb{E}_{S\sim\mathbb{P}}tr(S) < \infty$. Let M_n and M be respectively the empirical and true $N \times N$ covariance matrices of a Gaussian process constructed from the kernels K_n and K using a fixed grid $\mathcal{N}(0, \Sigma_1), \ldots, \mathcal{N}(0, \Sigma_N)$, and the reference empirical barycenter $\overline{\mu}_n$, and the population one $\overline{\mu}$ of $\mathbb{P}$ on $\mathcal{F}$, respectively. The kernels K_n and K are constructed as in (4.1) and (4.2). Then there exists a finite constant C such that with high probability*

$$\|M_n - M\|_F^2 \leq C\frac{N^2}{n}.$$

Finally, since we are dealing with Gaussian distributions, it is possible to understand the stationarity property of the kernel. The following proposition illustrates that in the Gaussian case the kernel is indeed invariant with respect to orthogonal transformations.

Proposition 7. *Let U be some predefined orthogonal matrix, and set ϕ_U be a deterministic map, that sends any $\mathcal{N}(0, S)$ to $\mathcal{N}(0, USU^T)$. For any $i = 1, \ldots, n$ denote as $T_{i,\phi}$ the optimal transportation map $T_{i,\phi\sharp}\phi_U\left(\mathcal{N}(0, S_i)\right) = \phi_U\left(\mathcal{N}(0, \overline{S})\right)$. Then it holds*

$$\left\|T_{i,\phi}^{-1} - T_{j,\phi}^{-1}\right\|_{L^2(\phi_U(\overline{\mu}))} = \left\|T_i^{-1} - T_j^{-1}\right\|_{L^2(\overline{\mu})}. \quad (4.4)$$

Equality (4.4) ensures the stationarity of the kernels under application of transformation ϕ_U.

5. Numerical simulations

5.1. Computational aspects

In practice, finding analytical representations of optimal transportation maps is a difficult issue, especially if the dimension of the problem grows. A possible solution consists in approximating an optimal transportation map by its empirical counterpart. Let μ_m and ν_m be empirical measures sampled from μ and ν respectively. Then the optimal Monge map $T_\sharp \mu = \nu$ can be replaced by $T^m_\sharp \mu_m = \nu_m$, see e.g. [19] or [14]. In this case, the problem of finding T^m is reduced to the solution of assignment problem with quadratic cost and can be solved by the *adagio* R-package by [17].

In dimension $p = 2$ or $p = 3$, it is also possible to represent the distributions by their matrices of probability weights on regular grids. Optimal transport maps can then be approximated, by means of various numerical procedures [36, 26, 37]. In our practical implementations, we tend to use the packages [49] and [30], with the R programming language.

5.2. Numerical study of the kernel consistency on a subspace of Gaussian measures

In what follows we present some simulations to highlight the consistency of the empirical kernel obtained in the Gaussian case from the empirical barycenter. We consider a distribution supported on a family $\mathcal{F}$ of 100000 centred Gaussians on $\mathbb{R}^d$ with covariance $S_i = A_i A_i'$, with $i = 1, \ldots, 100000$, where $A_i = (a_{jk})_{1 \leq j,k \leq d}, a_{jk} \sim_{i.i.d.} \text{Unif}[5, 15]$. In these experiments we consider $d = (4, 7, 15, 30)$.

Hence, for $d = 4, 7, 15, 30$ we construct kernels on the set $E \subset \mathcal{W}_2(\mathbb{R}^d)$ composed of Gaussian distributions with mean vector zero. Our data set is composed of realizations from the distribution $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$ that is an average of 100000 Dirac distributions on the 100000 Gaussian distributions $\mathcal{N}(0, S_1), \ldots, \mathcal{N}(0, S_{100000})$ described above. We compute the true barycenter $\mathcal{N}(0, \overline{S})$ of $\mathbb{P}$ for which the whole $\mathcal{F}$ is used, while $\overline{S}_n$ is computed as a Wasserstein-mean of a random n-sample taken with replacement from $\mathcal{F}$. Let M and M_n be the covariance matrices, obtained from the kernels K and K_n constructed using (3.2) with parameters $l = \sigma = 1$ on a grid of $N = 30$ randomly selected measures from $\mathcal{F}$.

Table 1 illustrates the mean approximation error rate $\|M_n - M\|_F$ for the cases $n = (20, 140, 260, 380, 500, 620)$.

As expected, we can see the convergence of the empirical kernel towards the theoretical one in all cases.

5.3. Prediction experiments on synthetic data

We consider the following simulations for the 2 dimensional case. We simulate 100 random two-dimensional Gaussian distributions split into a training sample

TABLE 1
Error: $\|M_n - M\|_F$ for centred Gaussians on $\mathbb{R}^d$.

	$n = 20$	$n = 140$	$n = 260$	$n = 380$	$n = 500$	$n = 620$
$d = 4$	1.52	0.69	0.16	0.29	0.24	0.14
$d = 7$	2.08	0.59	0.17	0.19	0.11	0.14
$d = 15$	0.91	0.12	0.09	0.08	0.05	0.05
$d = 30$	0.90	0.13	0.05	0.03	0.04	0.02

of 50 and a test sample of 50. Both mean vectors and covariance matrices are chosen randomly. The mean vector follows a uniform distribution over $[0.2, 0.8]^2$. The covariance matrix is isotropic and the standard deviation is uniform over $[0.01^2, 0.02^2]$. The value of the random field Y for a Gaussian distribution μ, given by its mean $(m_1, m_2)^T$ and variance σ^2, is given by

$$Y(\mu) = \frac{(m_1 - m_2^2)}{1 + \sigma}.$$

We then carry out our suggested Gaussian process model, based on the kernels suggested in Proposition 1. Hence, we construct kernels on the set $E \subset \mathcal{W}_2(\mathbb{R}^2)$ composed of Gaussian distributions with mean vector in $[0.2, 0.8]^2$ and covariance matrix of the form $\sigma^2 \mathrm{Id}$ with $\sigma \in [0.01^2, 0.02^2]$. Our data set is composed of realizations from the distribution $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^2))$ that corresponds to sampling the mean vector uniformly in $[0.2, 0.8]^2$ and sampling the standard deviation uniformly in $[0.01^2, 0.02^2]$.

Optimal transport maps T_μ^{-1}, from the barycenter to the Gaussian measures μ, are calculated using the package [49] and barycenters are calculated using the package [30] with parameter $\lambda = 20$ to balance computational time and similarity between the penalized transport and the optimal transport without regularization.

More precisely, the Gaussian distributions are discretized over a grid of 50×50 cells on $[0, 1]^2$. The Gaussian distributions are thus approximated by discrete distributions on the grid. We remark that the package [49] does not exactly provide deterministic transport maps. Indeed, the probability mass of a given input grid point can be split and mapped to several output grid points. Numerically, in this case, we transport all the probability mass of the input grid point to the output grid point that is assigned the most mass by the package [49]. Hence, to each discretized input Gaussian measure μ, we associate a transport map T_μ^{-1} from the barycenter that is an approximation of the inverse of the optimal transport map from μ to the barycenter. Nevertheless, since the mapping from μ to T_μ^{-1} is uniquely defined in our procedure, Remark 2 applies and we are guaranteed to obtain positive definite kernels.

The kernel we choose is K_θ given by

$$K_\theta(\mu, \nu) := \theta_1^2 * \exp(-\theta_2 \left\| T_\mu^{-1} - T_\nu^{-1} \right\|_{L^2(\overline{\mu})}^{\theta_3}) + \theta_4 1_{\left\| T_\mu^{-1} - T_\nu^{-1} \right\|_{L^2(\overline{\mu})} = 0}$$

for $\theta_1 \in [0.05, 10]$, $\theta_2 \in [0.01, 10]$, $\theta_3 \in [0.5, 2]$ and $\theta_4 \in [10^{-5}, 1]$. We will use the kernel with the parameters chosen to maximize the likelihood but also

TABLE 2
Prediction results for Gaussian simulations.

	RMSE	Q^2	CI Coverage
Kernel Smoothing	0.15	0.61	NA
Gaussian process	0.10	0.81	0.87
Gaussian process CV	0.10	0.81	0.88

parameters chosen to minimize the sum of the cross-validation square errors [6, 7]. For cross-validation, the total variance parameter $\theta_1^2 + \theta_4$ is estimated as suggested in [6].

We compare our kernel methods with the kernel smoothing procedure of [43]. This procedure consists in predicting $Y(\mu) \in \mathbb{R}$ by a weighted average of $Y(\mu_1), \ldots, Y(\mu_{50})$ where the weights are computed by applying a kernel to the distances $D(\mu, \mu_1), \ldots, D(\mu, \mu_{50})$ where D, as suggested in [43] is the L^1 distances between the probability density functions. The kernel is the triangular kernel as in [43], and its bandwidth is selected by minimizing an empirical mean square error based on sample splitting (see [43]). We remark that there is no estimate of the prediction error $Y(\mu) - \widehat{Y}(\mu)$ which is a downside compared to the Gaussian process model considered in this paper.

We present hereafter in Table 2 the results obtained, with 50 observations and 50 values to be predicted. We study the Root Mean Square Error (RMSE) of the form

$$\sqrt{\frac{1}{50}\sum_{i=1}^{50}(\widehat{Y}_i - Y_i)},$$

where the Y_i are the values to be predicted and the $\widehat{Y}_i$ are the predictions. We also study the Q^2 criterion which is equal to $1 - \text{RMSE}^2/\text{var}$, where var is the empirical variance of the values to be predicted. Finally we study the Confidence Interval Coverage (CIC) which corresponds to the frequency of the event that the predicted value belongs to the 90% confidence interval from the Gaussian process model. From the table, one observes that the GP process model based on the kernel we suggest provides a better accuracy, catching better the variability of the underlying process.

5.4. Experiments on real data: stress response to traction for materials in nuclear safety

We focus on a computer code called CASTEM code (see [29]) from the French Atomic Energy Commission (CEA) designed to calculate equivalent stresses on biphasic materials subjected to uni-axial traction. The system is modelled as a unit square containing $m = 10$ circular inclusions, all with the same radius R at random locations associated to a numerical value which is the stress response. The simulations are performed in two dimensions over $[0, 1]^2$ and the radius is $R = 0.0564$. The input of the code is composed of m disks located at m points $\{c_1, \ldots, c_m\}$ while the stress response is a scalar numerical value provided by

TABLE 3
Prediction of the CASTEM code output.

	RMSE	Q^2	CI Coverage
Kernel Smoothing	0.96	0.03	NA
Gaussian process	0.93	0.10	0.92
Gaussian process CV	0.92	0.11	1

the CASTEM code. As pointed out in [25], finding a proper distance between the inputs to forecast the stress is a very difficult task.

In this framework, we propose to consider each input as a uniform distribution μ on the union of the disks. For all the inputs $i = 1, \ldots, n$, we let $c^{(i)} = (c_1^{(i)}, \ldots, c_m^{(i)})$ be the vector of dimension $2m$ composed by the m centers of the disks and we let $D_j^{(i)}$ be the disk with center $c_j^{(i)}$ and radius R. Then we let μ_i be the uniform distribution over $\cup_{j=1}^m D_j^{(i)}$. Then the stress is considered as a Gaussian random field indexed by the μ_i's.

As previously, to compute the barycenter, we use the package provided in [30]. We use a grid over $[0,1]^2$ that discretizes the set into 50×50 cells. The uniform distribution on the set of disks is evaluated onto these cells and is approximated by a discrete distribution that is considered as an image. The optimal transport maps from the distributions to the barycenters are calculated using [49], similarly as in Section 5.3. We provide a comparison with the kernel smoothing procedure also as in Section 5.3.

The results are presented in Table 3 in the same way as in Table 2. In Table 3, the methods use 500 ouputs of the CASTEM code and predict 400 other outputs.

The total 900 outputs of the Castem code correspond to 900 distribution inputs that were generated randomly and independently. To generate a distribution input, letting $D(c, R)$ be the disk with center $c \in [0,1]^2$ and radius R, we sample c_1 uniformly on $[R, 1-R]^2$, then we sample c_2 uniformly on $[R, 1-R]^2 \backslash D(c_1, 2R)$, then we sample c_3 uniformly on $[R, 1-R]^2 \backslash (D(c_1, 2R) \cup D(c_2, 2R))$ and so on until c_m. Hence, we construct kernels on the set $E \subset \mathcal{W}_2(\mathbb{R}^2)$ of all the uniform distributions on $[0,1]^2$, which support is the union of ten disks of radius R that are included in $[0,1]^2$ and non-overlapping. Our data set is composed of realizations from the distribution $\mathbb{P} \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^2))$ that corresponds to the sampling mechanism described above.

From the poor Q^2 scores observed in Table 3 for all the methods, forecasting the CASTEM code appears to be a very hard task. Indeed, the inputs are very complex. Yet the method proposed in this work provides some improvements with respect to the state of the art method from [43]. We point out that cross validation of the parameters for the Gaussian process provides a very small improvement of the prediction but at the expense of overly large confidence intervals.

We remark that the kernel we provide is a positive definite kernel as required to use the Gaussian process modelling framework. Using directly a kernel by computing the exponential of minus the square W_2 Wasserstein distance between the distributions does not lead to a positive definite kernel. Actually Figure 1

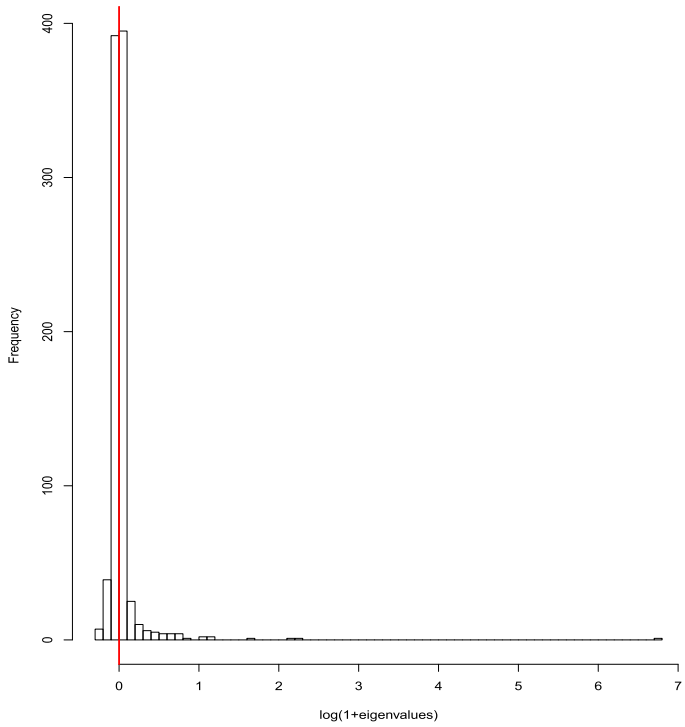

FIG 1. *Distribution of the eigenvalues of the* 900×900 *matrix obtained by the kernel of the form* $\exp(-W_2^2(\mu,\nu))$. *Many eigenvalues are negative, which shows that this kernel is not positive definite.*

shows the repartition of the eigenvalues of the 900×900 covariance matrix based on this kernel. We observe that many eigenvalues are negative (before the red line in the figure where we plot the logarithm of 1 plus the eigenvalues).

6. Conclusion and future directions

In this work, we have provided a theoretical way to use Wasserstein barycenters in order to define general positive definite kernels using optimal transportation maps. Considering the distance between the optimal transportation maps provides a natural way to quantify correlations between the values of a process indexed by distributions and provides a generalization to the multi-dimensional case of the work in [8].

Our suggested positive definite kernels then enable one to use the whole arsenal of kernel methods for statistical and machine learning applications. In Section 5, we provide numerical results related to regression and predictive confidence intervals, with Gaussian processes. In Section 4, we detail two other applications: the HSIC for independence test and support vector machines for binary classification.

Using a barycenter requires the distributions to be drawn according to the same measure over the set of distributions. This restricts the framework of the study to the case where the Gaussian process is defined on the support of this measure. For applications, this does not play a too important feature since inputs are often simulated according to a specified distribution. Yet for theoretical issues, this sets the frame of this study to the infill asymptotic framework and not the increasing-domain one. In this case, we have proved that parameters of families of covariance functions on multidimensional distributions, indexed by the distance between optimal transport maps, are microergodic. In order to obtain our microergodicity results, we show that, for all natural parametric families of radial covariance functions on Hilbert spaces, the covariance parameters are microergodic. This opens the perspective of studying the statistical consistency of specific estimators of these parameters. These consistency results would be relevant not only for statistical applications of Gaussian processes indexed by multidimensional distributions, but also in applications of Gaussian processes indexed by Hilbert spaces in general, for instance with functional inputs in computer experiments [40, 11]. We have also discussed how our microergodicity results enable to interpret that, for families of radial covariance functions, informally, the statistical estimation of the covariance parameters becomes easier as the dimension of the input space grows.

Finally contrary to the one-dimensional case, computational issues arise naturally when the Wasserstein distance is required. Hence the computation of a barycenter with respect to the Wasserstein distance is a difficult optimization program, unless the distributions are Gaussian, leading to tractable computations as shown in Section 4. Yet this idea of linearization around the barycenter to obtain a valid covariance kernel could be used and generalized to regularized Wasserstein distance using methods proposed in [22] for instance to provide a more tractable way of building kernels.

Appendix A: Proofs

Proof of Propositions 2 and 3

Proof. For both propositions, only the fact that 1. implies 2. needs to be proved. Let us now do this.

Let $f_1, \ldots, f_n$ in H and consider the matrix $\widetilde{C} = (\langle f_i, f_j \rangle_H)_{1 \leq i,j \leq n}$. This matrix is a Gram matrix in $\mathbb{R}^{n \times n}$ hence there exists a non negative diagonal matrix D and an orthogonal matrix P such that

$$\widetilde{C} = PDP' = PD^{1/2}D^{1/2}P'.$$

Let $e_1, \ldots, e_n$ be the canonical basis of $\mathbb{R}^n$. Then

$$e_i \widetilde{C} e_j' = u_i u_j'$$

where $u_i = e_i P D^{1/2}$. Note that the u_i's are vectors in $\mathbb{R}^n$ that depend on the $f_1, \ldots, f_n$. By polarization, we hence get that $\langle f_i, f_j \rangle_H = \langle u_i, u_j \rangle$ where $\langle \cdot, \cdot \rangle$

denotes the usual scalar product on $\mathbb{R}^n$. Hence we get that for any elements $f_1, \dots, f_n$ in H there are $u_1, \dots, u_n$ in $\mathbb{R}^n$ such that $\|f_i - f_j\|_H = \|u_i - u_j\|$. So any covariance matrix that can be written as $(F(\|f_i - f_j\|_H))_{i,j}$ can be seen as a covariance matrix $(F(\|u_i - u_j\|))_{i,j}$ on $\mathbb{R}^n$ and inherits its properties. The invertibility and non-negativity of this covariance matrix entail the invertibility and non-negativity of the first one, which proves the results. □

Proof of Theorem 3

Proof. Without loss of generality, we can assume that $h_0 = 0 \in H$. Let $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$. Then, there exists $t^* \in [0, L]$ such that $F_{\theta_1}(0) - F_{\theta_1}(2t^*) \neq F_{\theta_2}(0) - F_{\theta_2}(2t^*)$.

For any $n \in \mathbb{N}$, let $e_1, \dots, e_n \in H$ satisfy $\langle e_i, e_j \rangle_H = \mathbf{1}_{i=j}$. Consider the $2n$ elements $(f_1, \dots, f_{2n})$ made by the pairs $(-t^* e_i, t^* e_i)$ for $i = 1, \dots, n$. Consider a Gaussian process Y on $\overline{\mathcal{B}}_{2,L}$ with mean function zero and covariance function K_{θ_1}. Then, the Gaussian vector $Z = (Y(f_i))_{i=1,\dots,2n}$ has covariance matrix C given by

$$C_{i,j} = \begin{cases} F_{\theta_1}(0), & \text{if } i = j \\ F_{\theta_1}(2t^*), & \text{if } i \text{ odd and } j = i+1 \\ F_{\theta_1}(2t^*), & \text{if } i \text{ even and } j = i-1 \\ F_{\theta_1}(\sqrt{2}t^*), & \text{else.} \end{cases}$$

Hence, we have $C = D + M$ where M is the matrix with all components equal to $F_{\theta_1}(\sqrt{2}t^*)$ and where D is block diagonal, composed of n blocks of size 2×2, with each block equal to

$$B = \begin{pmatrix} F_{\theta_1}(0) - F_{\theta_1}(\sqrt{2}t^*) & F_{\theta_1}(2t^*) - F_{\theta_1}(\sqrt{2}t^*) \\ F_{\theta_1}(2t^*) - F_{\theta_1}(\sqrt{2}t^*) & F_{\theta_1}(0) - F_{\theta_1}(\sqrt{2}t^*) \end{pmatrix}.$$

Hence, in distribution, $Z = Q + E$, with Q and E independent, $Q = (z, \dots, z)$ where $z \sim \mathcal{N}(0, F_{\theta_1}(\sqrt{2}t^*))$ and where the n pairs (E_{2k+1}, E_{2k+2}), $k = 0, \dots, n-1$, are independent, with distribution $\mathcal{N}(0, B)$. Hence, with $\overline{Z}_1 = (1/n) \times \sum_{k=0}^{n-1} Z_{2k+1}$, $\overline{Z}_2 = (1/n) \sum_{k=0}^{n-1} Z_{2k+2}$ and $\overline{E} = (1/n) \sum_{k=0}^{n-1} (E_{2k+1}, E_{2k+2})^t$, we have

$$\begin{aligned} \widehat{B} &:= \frac{1}{n} \sum_{i=0}^{n-1} \begin{pmatrix} Z_{2i+1} - \overline{Z}_1 \\ Z_{2i+2} - \overline{Z}_2 \end{pmatrix} \begin{pmatrix} Z_{2i+1} - \overline{Z}_1 \\ Z_{2i+2} - \overline{Z}_2 \end{pmatrix}^t \\ &= \frac{1}{n} \sum_{i=0}^{n-1} \begin{pmatrix} E_{2i+1} \\ E_{2i+2} \end{pmatrix} \begin{pmatrix} E_{2i+1} \\ E_{2i+2} \end{pmatrix}^t - \overline{E}\overline{E}^t \\ &\to^p_{n \to \infty} B. \end{aligned}$$

Hence, there exists a subsequence $n' \to \infty$ such that, almost surely, $\widehat{B} \to B$ as $n' \to \infty$. For $i, j = 1, 2$, let us write $\widehat{B}_{i,j}$ for the element i, j of the 2×2 matrix $\widehat{B}$. Then, almost surely, $\widehat{B}_{1,1} - \widehat{B}_{1,2} \to F_{\theta_1}(0) - F_{\theta_1}(2t^*)$ as $n' \to \infty$. Write

$\widehat{B}_{1,1} - \widehat{B}_{1,2} = \widehat{\Delta}_{n'}(Y(f_1), \ldots, Y(f_{2n'}))$, where $\widehat{\Delta}_{n'}$ is a deterministic function from $\mathbb{R}^{2n'}$ to $\mathbb{R}$. Then, the set

$$A = \left\{ g \in \overline{F}; \widehat{\Delta}_{n'}\left(g(f_1), \ldots, g(f_{2n'})\right) \to_{n' \to \infty} F_{\theta_1}(0) - F_{\theta_1}(2t^*) \right\}$$

satisfies $P_{\theta_1}(A) = 1$. With the same arguments, we can show $P_{\theta_2}(B) = 1$, where

$$B = \left\{ g \in \overline{F}; \widehat{\Delta}_{n''}\left(g(f_1), \ldots, g(f_{2n''})\right) \to_{n'' \to \infty} F_{\theta_2}(0) - F_{\theta_2}(2t^*) \right\}$$

where n'' is a subsequence extracted from n'. Since $A \cap B = \varnothing$, it follows that $P_{\theta_2}(A) = 0$. Hence, θ is microergodic. □

Proof of Lemma 1

Proof. For $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$, let A_{d_1} be the set described before the lemma such that $P_{d_1,\theta_1}(A_{d_1}) = 0$ and $P_{d_1,\theta_2}(A_{d_1}) = 1$. Let now A_{d_2} be the set of functions defined by $A_{d_2} = \{f : \overline{\mathcal{B}}_{d_2,2,L} \to \mathbb{R}; P_{d_1}(f) \in A_{d_1}\}$. Here $P_{d_1}(f)$ is the function g from $\overline{\mathcal{B}}_{d_1,2,L}$ to $\mathbb{R}$ such that $g(x) = f((x,0))$, where $(x,0) \in \mathbb{R}^{d_2}$ has its subvector of first d_1 coefficients equal to x and its other coefficients equal to zero. Then, when Z is a Gaussian process on $\overline{\mathcal{B}}_{d_2,2,L}$ with mean zero and covariance function $K_{d_2,\theta}$, one can see that $P_{d_1}(Z)$ is a Gaussian process on $\overline{\mathcal{B}}_{d_1,2,L}$ with mean zero and covariance function $K_{d_1,\theta}$. Consequently, $P_{d_2,\theta_1}(A_{d_2}) = 0$ and $P_{d_2,\theta_2}(A_{d_2}) = 1$. Hence, θ is microergodic for the dimension d_2. □

For the proof of Proposition 5

Recall that the empirical barycenters $(\overline{\mu}_n)_n$ are a sequence of continuous measures converging to $\overline{\mu}$ in 2-Wasserstein distance: $W_2(\overline{\mu}_n, \overline{\mu}) \to 0$ as $n \to \infty$ and $R_{n\sharp}\overline{\mu} = \overline{\mu}_n$ with $W_2(\overline{\mu}, \overline{\mu}_n) = ||R_n||_{L^2(\overline{\mu})}$.

Lemma 2. *Fix some distribution ν absolutely continuous with respect to Lebesgue measure and let $T = T_\nu$ and $T_n = T_{\nu,n}$. Then it holds a.s.*

$$\left\|T - T_n\right\|^2_{L^2(\nu)} \longrightarrow 0, \text{ as } n \to \infty.$$

Proof. Fix n s.t. $W_2(\overline{\mu}_n, \overline{\mu}) = \varepsilon_n$. Consider $\left\|\mathrm{id} - R_n \circ T\right\|_{L^2(\nu)}$. By change of variables and triangle inequality one obtains

$$\begin{aligned}\left\|\mathrm{id} - R_n \circ T\right\|_{L^2(\nu)} = \left\|T^{-1} - R_n\right\|_{L^2(\overline{\mu})} &\leq \left\|T^{-1} - \mathrm{id}\right\|_{L^2(\overline{\mu})} + \left\|R_n - \mathrm{id}\right\|_{L^2(\overline{\mu})} \\ &\leq W_2(\nu, \overline{\mu}) + \varepsilon_n \leq W_2(\nu, \overline{\mu}_n) + 2\varepsilon_n.\end{aligned}$$

Since T_n is the optimal transport map from ν to μ_n we recall that $W_2(\nu, \overline{\mu}_n) = \left\|\mathrm{id} - T_n\right\|_{L^2(\nu)}$. So due to the arbitrary choice of n it follows

$$\left| \left\|\mathrm{id} - R_n \circ T\right\|_{L^2(\nu)} - \left\|\mathrm{id} - T_n\right\|_{L^2(\nu)} \right| \underset{n \to \infty}{\longrightarrow} 0. \tag{A.1}$$

Now we are ready to prove, that $\|T_n - T\|_{L^2(\nu)} \overset{n\to\infty}{\longrightarrow} 0$. Assume the claim is wrong:

$$T_n \overset{n\to\infty}{\longrightarrow} T_1, \quad R_n \circ T \overset{n\to\infty}{\longrightarrow} T_2, \quad \|T_1 - T_2\| > \varepsilon.$$

Thus

$$\|\mathrm{id} - T_n\|_{L^2(\nu)} \overset{n\to\infty}{\longrightarrow} \|\mathrm{id} - T_1\|_{L^2(\nu)}, \quad \|\mathrm{id} - R_n \circ T\|_{L^2(\nu)} \overset{n\to\infty}{\longrightarrow} \|\mathrm{id} - T_2\|_{L^2(\nu)},$$

which contradicts to (A.1) □

The next lemma is a key ingredient in the proof of the fact that the true kernel can be replaced by its empirical counterpart.

Lemma 3. *Consider two fixed absolutely continuous measures μ and ν in $\mathcal{W}_2(\mathbb{R}^p)$. We have a.s.*

$$\left| \|T_\mu^{-1} - T_\nu^{-1}\|^2_{L^2(\overline{\mu})} - \|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|^2_{L^2(\overline{\mu}_n)} \right| \longrightarrow 0, \text{ as } n \to \infty.$$

Proof. Consider $\|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|_{L^2(\overline{\mu}_n)}$. Change of variables and triangle inequality yield

$$\begin{aligned} &\|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|_{L^2(\overline{\mu}_n)} = \|T_{\mu,n}^{-1} \circ R_n - T_{\nu,n}^{-1} \circ R_n\|_{L^2(\overline{\mu})} \\ &\leq \|T_{\mu,n}^{-1} \circ R_n - T_\mu^{-1}\|_{L^2(\overline{\mu})} + \|T_{\nu,n}^{-1} \circ R_n - T_\nu^{-1}\|_{L^2(\overline{\mu})} + \|T_\mu^{-1} - T_\nu^{-1}\|_{L^2(\overline{\mu})}. \end{aligned}$$

Therefore one obtains

$$\begin{aligned} &\|T_{\mu,n}^{-1} - T_{\nu,n}^{-1}\|_{L^2(\overline{\mu}_n)} - \|T_\mu^{-1} - T_\nu^{-1}\|_{L^2(\overline{\mu})} \\ &\leq \|T_{\nu,n}^{-1} \circ R_n - T_\nu^{-1}\|_{L^2(\overline{\mu})} + \|T_{\mu,n}^{-1} \circ R_n - T_\mu^{-1}\|_{L^2(\overline{\mu})} \overset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

where the last relation holds due to Lemma 2. □

Proof of Proposition 6

Proof. Actually using Lemma A.2 together with Theorem 2.2 in [32], we obtain, with R_n defined as before Lemma 2, that

$$\|R_n - \mathrm{Id}\|_{L^2(\overline{\mu})} = O_P\left(\frac{1}{\sqrt{n}}\right),$$

and that the empirical transportation maps can be linearized as

$$T_{i,n}^{-1} = T_i^{-1} + D(\overline{S}_n - \overline{S}) + o(\|\overline{S}_n - \overline{S}\|_F),$$

where D is a linear self-adjoint bounded operator acting on the space of symmetric matrices. Here for $i = 1, \ldots, N$, T_i and $T_{i,n}$ are as in Section 4.3 but with S_i replaced by Σ_i. Use the following decomposition

$$\|T_{i,n}^{-1} \circ R_n - T_i^{-1}\|_{L^2(\overline{\mu})} \leq \|T_i^{-1} \circ R_n - T_i^{-1} + (T_{i,n}^{-1} - T_i^{-1}) \circ R_n\|_{L^2(\overline{\mu})}$$

$$\leq \|T_i^{-1} \circ R_n - T_i^{-1}\|_{L^2(\overline{\mu})} + \|(T_{i,n}^{-1} - T_i^{-1}) \circ R_n\|_{L^2(\overline{\mu})}$$
$$\leq O_P\left(\frac{1}{\sqrt{n}}\right).$$

This entails that $\|T_{i,n}^{-1} - T_{j,n}^{-1}\|_{L^2(\overline{\mu}_n)} - \|T_i^{-1} - T_j^{-1}\|_{L^2(\overline{\mu})}$ is also of order $\frac{1}{\sqrt{n}}$ since

$$\left\|T_{i,n}^{-1} - T_{j,n}^{-1}\right\|_{L^2(\overline{\mu}_n)} - \left\|T_i^{-1} - T_j^{-1}\right\|_{L^2(\overline{\mu})}$$
$$\leq \left\|T_{i,n}^{-1} \circ R_n - T_i^{-1}\right\|_{L^2(\overline{\mu})} + \left\|T_{j,n}^{-1} \circ R_n - T_j^{-1}\right\|_{L^2(\overline{\mu})}.$$

Since for all $(i,j) \in \{1,\ldots,N\}^2$,

$$M_n(i,j) = F(\|T_{i,n}^{-1} - T_{j,n}^{-1}\|^2_{L^2(\overline{\mu}_n)})$$

as soon as F is continuously differentiable with bounded derivative, then we get that for a finite constant

$$\sum_{i,j=1}^{N} |M_n(i,j) - M(i,j)|^2 \leq N^2 \sup_{i,j} |M_n(i,j) - M(i,j)|^2 \leq C\frac{N^2}{n},$$

which concludes the proof. □

Proof of Proposition 7

Proof. Note, that for any orthogonal matrix U the following set of inequalities hold:

$$W_2^2\left(\mathcal{N}(0,S),\mathcal{N}(0,Q)\right) := \mathrm{tr}(S) + \mathrm{tr}(Q) - 2\mathrm{tr}\left(Q^{1/2}SQ^{1/2}\right)^{1/2}$$
$$= W_2^2\left(\mathcal{N}(0,USU^T),\mathcal{N}(0,UQU^T)\right)$$
$$= W_2^2\left(\phi_U(\mathcal{N}(0,S)),\phi_U(\mathcal{N}(0,Q))\right).$$

Thus, map ϕ_U preserves 2-Wasserstein distance. Equality (4.4) follows from (4.3) by substituting S_i, S_j, and $\overline{S}$ by US_iU^T, US_jU^T, and $U\overline{S}U^T$ respectively. □

Acknowledgments

The work of Alexandra Suvorikova in Sections 2 and 4 is funded by Russian Science Foundation grant No. 18-71-10108. The work of Vladimir Spokoiny in Section 5 is supported by the Russian Science Foundation grant No. 19-71-30020. Vladimir Spokoiny thanks the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689). Jean-Michel Loubes thanks the AI interdisciplinary institute ANITI, grant agreement N. ANR-19-PI3A-0004 under the French investing for the future PIA3 program. DG's contributions have taken place within the Swiss National Science Foundation project number 178858. The authors would like to thank Drs.

Jean Baccou and Frédéric Pérales (respectively at LIMAR and LPTM, Institut de Radioprotection et de Sûreté Nucléaire, Saint-Paul-lès-Durance, France) for the CASTEM data set, and Dr. Clément Chevalier (now with the Swiss Statistical Office, Neuchâtel, Switzerland) who has been involved in investigations on this data set in the framework of the ReDICE consortium. Finally, the authors are very grateful to the anonymous referees for their constructive comments, that led to an improvement of the paper, in terms of content and exposition.

References

[1] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Norwegian Computing Center, 1997.

[2] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. MR2801182

[3] Pedro C. Alvarez-Esteban, Eustasio del Barrio, Juan Antonio Cuesta-Albertos, and Carlos Matrán. Wide consensus for parallelized inference. *arXiv preprint* arXiv:1511.05350, 2015.

[4] Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan Antonio Cuesta-Albertos, and Carlos Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. MR3491556

[5] Ethan Anderes. On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics*, 38:870–893, 2010. MR2604700

[6] François Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013. MR3064023

[7] François Bachoc. Asymptotic analysis of covariance parameter estimation for gaussian processes in the misspecified case. *Bernoulli*, 24(2):1531–1575, 2018. MR3706801

[8] François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Nil Venet. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, 64(10):6620–6637, 2018. MR3860751

[9] B.J.C. Baxter. Positive definite functions on Hilbert space. *East Journal on Approximations*, 10(3):269–274, 2004. MR2076887

[10] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011. MR2239907

[11] José Betancourt, François Bachoc, Thierry Klein, Déborah Idier, Rodrigo Pedreros, and Jérémy Rohmer. Gaussian process metamodeling of functional-input code for coastal flood hazard assessment. *Reliability Engineering and System Safety, forthcoming.* `https://hal.archives-ouvertes.fr/hal-01998727/`, 2019.

[12] Moreno Bevilacqua, Tarik Faouzi, Reinhard Furrer, and Emilio Porcu. Estimation and prediction using generalized Wendland covariance functions

under fixed domain asymptotics. *The Annals of Statistics*, 47(2):828–856, 2019. MR3909952

[13] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018. MR3872127

[14] Melf Boeckel, Vladimir Spokoiny, and Alexandra Suvorikova. Multivariate Brenier cumulative distribution functions and their application to nonparametric testing. *arXiv preprint* arXiv:1809.04090, 2018.

[15] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015. MR3338645

[16] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. MR3300482

[17] Hans Werner Borchers. adagio: Discrete and global optimization routines. URL `http://CRAN.R-project.org/package=adagio`, 2016.

[18] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. MR1100809

[19] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017. MR3611491

[20] Nello Cristianini and John Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

[21] Juan Antonio Cuesta and Carlos Matrán. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.*, 17(3):1264–1276, 1989. ISSN 0091-1798. MR1009457

[22] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[23] Eustasio del Barrio, Juan Antonio Cuesta-Albertos, Marc Hallin, and Carlos Matrán. Center-outward distribution functions, quantiles, ranks, and signs in $\mathbb{R}^d$. *arXiv e-prints* arXiv:1806.01238, Jun 2018.

[24] Alessio Figalli. On the continuity of center-outward distribution and quantile functions. *Nonlinear Analysis*, 177:413–421, 2018. MR3886582

[25] David Ginsbourger, Jean Baccou, Clément Chevalier, and Frédéric Perales. Design of computer experiments using competing distances between set-valued inputs. In *mODa 11-Advances in Model-Oriented Design and Analysis*, pages 123–131. Springer, 2016.

[26] Carsten Gottschlich and Dominic Schuhmacher. The shortlist method for fast computation of the Earth mover's distance and finding optimal solutions to transportation problems. *PloS One*, 9(10):e110214, 2014.

[27] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005. MR2255909

[28] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009. MR2722294

[29] http://wwwcast3m.cea.fr. Cast3m software.

[30] Marcel Klatt. *Regularized Wasserstein Distances and Barycenters*, 2018. URL https://cran.r-project.org/web/packages/Barycenter/Barycenter.pdf. R package version 1.3.1.

[31] Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[32] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures-Wasserstein barycenters. *arXiv preprint* arXiv:1901.00226, 2019.

[33] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75, 2015.

[34] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017. MR3663634

[35] Wei-Liem Loh. Estimating the smoothness of a gaussian random field from irregularly spaced data via higher-order quadratic variations. *The Annals of Statistics*, 43(6):2766–2794, 2015. MR3405611

[36] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*, volume 2. Springer, 1984. MR2423726

[37] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[38] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006. MR2274454

[39] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[40] Thomas Muehlenstaedt, Jana Fruth, and Olivier Roustant. Computer experiments with functional inputs and scalar outputs by a norm-based approach. *Statistics and Computing*, 27:1083–1097, 2017. MR3627564

[41] David Nualart. *The Malliavin Calculus and Related Topics*, volume 1995. Springer. MR1344217

[42] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[43] Barnabas Poczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. Distribution-free distribution regression. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 507–515, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

[44] Carl Edward Rasmussen and Chris K.I. Williams. *Gaussian Processes for*

Machine Learning. The MIT Press, Cambridge, 2006. MR2514435
[45] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkäuser, NY, pages 99–102, 2015. MR3409718
[46] Thomas J. Santner, Brian J. Williams, and William Notz. *The Design and Analysis of Computer Experiments*. Springer, New York, 2003. MR3887662
[47] Isaac J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4):811–841, 1938. MR1503439
[48] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.
[49] Dominic Schuhmacher, Björn Bähre, Carsten Gottschlich, Valentin Hartmann, Florian Heinemann, and Bernhard Schmitzer. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, 2019. URL https://cran.r-project.org/package=transport. R package version 0.11-0.
[50] Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999. MR1697409
[51] Bui Thi Thien Trang, Jean-Michel Loubes, Laurent Risser, and Patricia Balaresque. Distribution regression model with a reproducing kernel Hilbert space approach. *Communications in Statistics – Theory and Methods*, pages 1–23, 2019.
[52] César A. Uribe, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Angelia Nedić. Distributed computation of Wasserstein barycenters over networks. In *2018 IEEE 57th Annual Conference on Decision and Control (CDC)*, 2018. Accepted, arXiv:1803.02933.
[53] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2009. MR2459454
[54] Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004. MR2131724
[55] Hao Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99:250–261, 2004. MR2054303