A Gaussian Process Regression Model for Distribution Inputs

François Bachoc, Fabrice Gamboa, Jean-Michel Loubes and Nil Venet

Abstract

Monge-Kantorovich distances, otherwise known as Wasserstein distances, have received a growing attention in statistics and machine learning as a powerful discrepancy measure for probability distributions. In this paper, we focus on forecasting a Gaussian process indexed by probability distributions. For this, we provide a family of positive definite kernels built using transportation based distances. We provide a probabilistic understanding of these kernels and characterize the corresponding stochastic processes. We prove that the Gaussian processes indexed by distributions corresponding to these kernels can be efficiently forecast, opening new perspectives in Gaussian process modeling.

Index Terms

Gaussian process, Positive definite kernel, Kriging, Monge-Kantorovich distance, Fractional Brownian motion

The four authors are affiliated to the Institute of Mathematics of Toulouse, Université Paul Sabatier, Toulouse, France. NV is also affiliated to, and completely funded by CEA. E-mail: (francois.bachoc, jean-michel.loubes, fabrice.gamboa,nil.venet)@math.univ-toulouse.fr. An overview of the main results of this article was described in the conference article [24].

A Gaussian Process Regression Model for Distribution Inputs

I. Introduction

ORIGINALLY used in spatial statistics (see for instance [1] and references therein), Kriging has become very popular in many fields such as machine learning or computer experiment, as described in [2]. It consists in predicting the value of a function at some point by a linear combination of observed values at different points. The unknown function is modeled as the realization of a random process, usually Gaussian, and the Kriging forecast can be seen as the posterior mean, leading to the optimal linear unbiased predictor of the random process.

Gaussian process models rely on the definition of a covariance function that characterizes the correlations between values of the process at different observation points. As the notion of similarity between data points is crucial, *i.e.* close location inputs are likely to have similar target values, covariance functions are the key ingredient in using Gaussian processes, since they define nearness or similarity. In order to obtain a satisfying model one need to chose a covariance function (*i.e.* a positive definite kernel) that respects the structure of the index space of the dataset. Continuity of the covariance is a minimal assumption, as one may ask for additional properties such as stationarity or stationary increments with respect to a distance. These stronger assumptions allow to obtain a model where the correlations between data points depend on the distance between them.

First used in Support Vector (see for instance [3]), positive definite kernels are nowadays used for a wide range of applications. There is a huge statistical literature dealing with the construction and properties of kernel functions over $\mathbb{R}^d$ for $d \geq 1$ (we refer for instance to [4] or [5] and references therein). Yet the construction of kernels with adequate properties on more complex spaces is still a growing field of research (see for example [6], [7], [8]).

Within this framework, we tackle the problem of forecasting a process indexed by one-dimensional distributions. Our motivations come from a variety of applied problems: in the classical ecological inference problem (see [9]), outputs are not known for individual inputs but for groups, for which the distribution of a covariate is known. This situation happens for instance in political studies, when one wants to infer the correlation between a vote and variables such as age, gender or wealth level, from the distributions of these covariates in different states (see for example [9]). The problem of causal inference can also be considered in a distribution learning setting (see [10]).

As [11] remarks, learning on distribution inputs offers two important advantages in the big data era. By bagging together individual inputs with similar outputs, one reduces the size of a dataset and anonymizes the data. Doing so results on learning on the distribution of the inputs in the bags.

Another application arises in numerical code experiments, when the prior knowledge of the input conditions may not be an exact value but rather a set of acceptable values that will be modeled using a prior distribution. Hence we observe output values for such probability distributions and want to forecast the process for other ones. A similar application of distribution inputs for numerical code experiments is given by non-negative functional inputs. We give a detailed example of this situation in Section II.

Several approaches already exist to deal with distribution inputs regression. An important class of methods relies on some notion of divergence between distributions (see [12]–[14]). Other methods have been proposed, such as kernel mean embedding [11] and kernel ridge regression methods [15]. In this paper we focus on Gaussian process regression method.

The first issue when considering Gaussian process regression for distribution inputs is to define a covariance function, which will allow to compare the similarity between probability distributions. Several approaches can be considered here. The simplest method is 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 do not take into account the whole shape of the law. Specific kernels should be designed in order to map distributions into a reproducing kernel Hilbert space in which the whole arsenal of kernel methods can be extended to probability measures. This issue has recently been considered in [16] or [17].

In the past few years, transport based distances such as the Monge-Kantorovich or Wasserstein distance have become a growing way to assess similarity between probability measures and are used for numerous applications in learning and forecast problems. Since such distances are defined as a cost to transport one distribution to the other one, they appear to be a very relevant way to measure similarities between probability measures. Details on Wasserstein distances and their links with optimal transport problems can be found in [18]. Applications in statistics are developed in [19], [20], [21] while kernels have been developed in [17] or [22].

In this paper, we construct covariance functions in order to obtain Gaussian processes indexed by probability measures. We provide a class of covariances which are functions of the Monge-Kantorovich distance, corresponding to stationary Gaussian processes. We also give covariances corresponding to the fractional Brownian processes indexed by probability distributions, which have stationary increments with respect to the Monge-Kantorovich distance. Furthermore we show original nondegeneracy results for these kernels. Then, in this framework, we focus on the selection of a stationary covariance kernel in a parametric model through maximum likelihood. We prove the consistency and asymptotic normality of the covariance parameter estimators. We then consider the Kriging of such Gaussian processes. We prove the asymptotic accuracy of the Kriging prediction under the estimated covariance parameters. In simulations, we show the strong benefit of the studied kernels, compared to more standard kernels operating on finite dimensional projections of the distributions. In addition, we show in the simulations that the Gaussian process model suggested in this article is significantly more accurate that the kernel smoothing based predictor of [23]. Our results consolidate the idea that the Monge-Kantorovich distance is an efficient tool to assess variability between distributions, leading to sharp predictions of the outcome of a Gaussian process with distribution-type inputs.

The paper falls into the following parts. In Section III we recall generalities on the Wasserstein space, covariance kernels and stationarity of Gaussian processes. Section IV is devoted to the construction and analysis of an appropriate kernel for probability measures on $\mathbb{R}$. Asymptotic results on the estimation of the covariance function and properties of the prediction of the associated Gaussian process are presented in Section V. Section VI is devoted to numerical applications while the proofs are postponed to the appendix.

II. An applicative case from nuclear safety

The research that conducted to this article have been partially funded by CEA, and is motivated by a nuclear safety application, which we detail here.

A standard problem for used fissile storage process is the axial burn up analysis of fuel pins [25]. In this case study, fuel pins may be seen as one-dimensional curves $X : [0,1] \to \mathbb{R}^+$ [26]. These curves correspond to the axial irradiation profiles for fuel in transportation or storage packages which define the neutronic reactivity of the systems. From a curve X, corresponding to a given irradiation profile, it is then possible to compute the resulting neutron multiplication factor $k_{eff}(X)$ by numerical simulation [27]. It can be insightful, for profiles with a given total irradiation $\int_0^1 X(t)dt$, to study the impact of the shape of the irradiation curve X on the multiplication factor $k_{eff}(X)$. This type of study can be addressed by considering k_{eff} as a realization of a Gaussian process indexed by one-dimensional distributions.

III. Generalities

In this section we recall some basic definitions and properties of the Wasserstein spaces and of covariance kernels.

a) The Monge-Kantorovich distance: Let us consider the set $\mathcal{W}_2(\mathbb{R})$ of probability measures on $\mathbb{R}$ with a finite moment of order two. For two μ, ν in $\mathcal{W}_2(\mathbb{R})$, we denote by $\Pi(\mu,\nu)$ the set of all probability measures π over the product set $\mathbb{R}\times\mathbb{R}$ 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

$$\mathcal{T}_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2 \, d\pi(x,y). \tag{1}$$

This transportation cost allows to endow the set $\mathcal{W}_2(\mathbb{R})$ with a metric by defining the quadratic Monge-Kantorovich, or quadratic Wasserstein distance between μ and ν as

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

A probability measure π in $\Pi(\mu,\nu)$ realizing the infimum in (1) is called an optimal coupling. This vocabulary transfers to a random vector (X_1, X_2) with distribution π. We will call $\mathcal{W}_2(\mathbb{R})$ endowed with the distance W_2 the Wasserstein space.

We will consider on several occasion the collection of random variables $(F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}$, where F_μ^{-1} defined as

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

denotes the quantile function of the distribution μ, and U is an uniform random variable on $[0,1]$. For every $\mu,\nu \in \mathcal{W}_2(\mathbb{R})$, the random vector $((F_\mu^{-1}(U)),(F_\nu^{-1}(U)))$ is an optimal coupling (see [18]). Notice that the random variable $F_\mu^{-1}(U)$ does not depend on ν, so that $(F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}$ is an optimal coupling between every distribution of $\mathcal{W}_2(\mathbb{R})$.

More details on Wasserstein distances and their links with optimal transport problems can be found in [28] or [18] for instance.

b) Covariance kernels: Let us recall that the law of a Gaussian random process $(X(x))_{x\in E}$ indexed by a set E is entirely characterized by its mean and covariance functions

$$M : x \mapsto \mathbb{E}(X(x))$$

and

$$K : (x,y) \mapsto \mathrm{Cov}(X(x)X(y))$$

(see $e.g.$ [29]).

A function K is actually the covariance of a random process if and only if it is a *positive definite kernel*, that is to say for every $x_1,\cdots,x_n \in E$ and $\lambda_1,\cdots,\lambda_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} \lambda_i\lambda_j K(x_i,x_j) \geq 0. \tag{3}$$

In this case we say that K is a *covariance kernel.*

On the other hand, any function can be chosen as the mean of a random process. Hence without loss of generality we focus on centered random processes in Section IV.

Positive definite kernels are closely related to negative definite kernels. A function $K : E\times E \to \mathbb{R}$ is said to be a *negative definite kernel* if for every $x \in E$,

$$K(x,x) = 0 \tag{4}$$

and for every $x_1,\cdots,x_n \in E$ and $c_1,\cdots,c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$,

$$\sum_{i,j=1}^{n} c_ic_j K(x_i,x_j) \leq 0. \tag{5}$$

Example The variogram $(x, y) \mapsto \mathbb{E}(X(x) - X(y))^2$ of any random field X is a negative definite kernel.

If the inequality (3) (resp. (5)) is strict as soon as not every λ_i (resp. c_i) is null and the x_i are two by two distinct, a positive definite (resp. negative definite) kernel is said to be *nondegenerate*. Nondegeneracy of a covariance kernel is equivalent to the fact that every covariance matrix built with K is invertible. We will say that a Gaussian random process is nondegenerate if its covariance function is a nondegenerate kernel. Nondegeneracy is is usually a desirable condition for Kriging, since the forecast is built using the inverse of the covariance matrix of the observations. In addition, Gaussian process models with degenerate kernels have structural restrictions that can prevent them for being flexible enough. We give the nondegeneracy of the fractional Brownian motion indexed by the Wasserstein space in Section IV.

c) Stationarity: Stationarity is a property of random processes that is standard in the Kriging literature. Roughly speaking, a stationary random process behaves in the same way at every point of the index space. It is also an enjoyable property for technical reasons. In particular it is a key assumption for the proofs of the properties we give in Section V.

We say that a random process X indexed by a metric space (E, d) is *stationary* if it has constant mean and for every isometry g of the metric space we have

$$\operatorname{Cov}(X(g(x)), X(g(y))) = \operatorname{Cov}(X(x), X(y)). \tag{6}$$

Let us notice in particular that if the covariance of a random process is a function of the distance, equation (6) is verified. This is the assumption we make in Section V.

One can also found the assumption of stationarity for the increments of a random process. Many classical random processes have stationarity increments, such as the fractional Brownian motion. We prove the existence of fractional Brownian motion indexed by the Wasserstein space in Section IV.

We will say that X has *stationary increments* starting in $o \in E$ if X is centred, $X(o) = 0$ almost surely, and for every isometry g we have

$$\operatorname{Cov}\left(X(g(x)) - X(g(o))\right) = \operatorname{Cov}\left(X(x) - X(o)\right). \tag{7}$$

Notice that the variance of a random process with stationary increments increases as the input get far from the origin point o.

Let us remark that the definitions we gave are usually called “in the wide sense”, in contrast with stationarity definitions “in the strict sense”, which asks for the law of the process (or its increments) to be invariant under the action of the isometries, and not only the first and second moments. Since we are only dealing with Gaussian processes those definitions coincide.

d) Isometries of the Wasserstein space: Since we are interested in processes indexed by the Wasserstein space with stationarity features, let us recall a few facts about isometries of the Wasserstein space $\mathcal{W}_2(\mathbb{R})$, that is to say maps $i : \mathcal{W}_2(\mathbb{R}) \to \mathcal{W}_2(\mathbb{R})$ that preserve the Wasserstein distance.

Trivial isometries come from isometries of $\mathbb{R}$: to any isometry $g : \mathbb{R} \to \mathbb{R}$, we can associate an isometry $g_\# : \mathcal{W}_2(\mathbb{R}) \to \mathcal{W}_2(\mathbb{R})$ that maps any measure $\mu \in \mathcal{W}_2(\mathbb{R})$ to the measure

$$g_\#(\mu) : A \mapsto \mu(g^{-1}(A)).$$

Stationarity of a random process with regard to these trivial isometries is an interesting feature, since it means that the statistical properties of the outputs do not change when we apply an isometry to the real line.

However let us mention that not every isometry of the Wasserstein space is trivial. In particular, mapping every distribution to its symmetric regarding its expectancy defines an isometry of $\mathcal{W}_2(\mathbb{R})$. We refer to [30] for a complete description of the isometries of the Wasserstein space.

IV. Gaussian process models for distribution inputs

In this section we give covariance kernels on the space of probability distributions on the real line. This allows for modeling and Gaussian process regression of datasets with distribution inputs.

We start in Section IV-A by giving a generalization of the seminal fractional Brownian motion to distributions inputs endowed with the Wasserstein distance.

Then, in Section IV-B we give Gaussian processes that are stationary with respect to the Wasserstein distance on the inputs.

A. Fractional Brownian motion with distribution inputs

We first consider the family of *fractional Brownian kernels*

$$\begin{aligned} K^{H,\mu_0}&(\mu, \nu) \\ &= \frac{1}{2}\left(W_2^{2H}(\mu_0, \mu) + W_2^{2H}(\mu_0, \nu) - W_2^{2H}(\mu, \nu)\right), \end{aligned} \tag{8}$$

where $0 < H \leq 1$ and $\mu_0 \in \mathcal{W}_2(\mathbb{R})$ are fixed.

Note that these kernels are obtained by taking the covariances of the classical fractional Brownian motions and replacing the distance $|t - s|$ between two times $s, t \in \mathbb{R}$ by the Wasserstein distance $W_2(\mu, \nu)$ between two distribution inputs. The measure $\mu_0 \in \mathcal{W}_2(\mathbb{R})$ plays the role of the origin $0 \in \mathbb{R}$.

Theorem IV.1. *For every $0 \leq H \leq 1$ and a given $\mu_0 \in \mathcal{W}_2(\mathbb{R})$ the function K^{H,μ_0} defined by* (8) *is a covariance function on $\mathcal{W}_2(\mathbb{R})$. Furthermore K^{H,μ_0} is nondegenerate if and only if* $0 < H < 1$.

The Gaussian process $(X(\mu))_{\mu \in \mathcal{W}_2(\mathbb{R})}$ such that

$$\begin{cases} \mathbb{E}\, X(\mu) = 0, \\ \operatorname{Cov}(X(\mu), X(\nu)) = K^{H,\mu_0}(\mu, \nu) \end{cases} \tag{9}$$

is the *H-fractional Brownian motion with index space $\mathcal{W}_2(\mathbb{R})$ and origin in μ_0*. It inherits properties from the classical fractional Brownian motion.

It is easy to check that the output at the origin measure μ_0 is zero, $X(\mu_0) = 0$ almost surely. Furthermore

$$\mathbb{E}(X(\mu) - X(\nu))^2 = W_2^{2H}(\mu, \nu), \tag{10}$$

from which we deduce that $(X(\mu))_{\mu\in\mathcal{W}_2(\mathbb{R})}$ has stationary increments, which means that the statistical properties of $X(\mu)-X(\nu)$ are the same as those of $X(g(\mu))-X(g(\nu))$ for every isometry g of the Wasserstein space.

The fractional Brownion motion is well known for its parameter H governing the regularity of the trajectories: small values of H correspond to very irregular trajectories while greater values give steadier paths. Moreover for $H>1/2$ the process exhibits long-range dependence (see [31]).

From the modelling point of view, it is interesting to consider the following process: consider $(X(\mu))_{\mu\in\mathcal{W}_2(\mathbb{R})}$ the H-fractional Brownian motion with origin in δ_0 the Dirac measure at 0, f a real-valued function and define

$$Y(\mu) := X(\bar{\mu}) + f(m(\mu)), \tag{11}$$

where $\bar{\mu}$ denotes the centred version of μ. We then have, using $X(\delta_0)=0$ almost surely and (10):

$$\begin{aligned}\operatorname{Var}(Y(\mu)) = \mathbb{E}(X(\bar{\mu}))^2 = \mathbb{E}(X(\bar{\mu})-X(\delta_0))^2 = W_2^{2H}(\bar{\mu},\delta_0)\\ = (\operatorname{Var}(\mu))^H.\end{aligned}$$

Hence the mean of the output $Y(\mu)$ is a function of the mean of the input distribution μ and its dispersion is an increasing function of the dispersion of μ. This is a valuable property when modeling a function $\mu\mapsto g(\mu)$ as a Gaussian Process realization $\mu\mapsto Y(\mu)$, when it is believed that the range of possible values for g increases with the variance of the input mu.

Let us further notice that for $f=id$ and $H=1$ we have

$$\mathbb{E}(Y(\mu)) = \mathbb{E}(F_\mu^{-1}(U))$$

and

$$\operatorname{Cov}(Y(\mu),Y(\nu)) = \operatorname{Cov}(F_\mu^{-1}(U),F_\mu^{-1}(U)),$$

where F_μ^{-1} denotes the quantile function of the distribution μ, and U is an uniform random variable on $[0,1]$. In some sense, Y is in this case the Gaussian process that mimics the statistical properties of the optimal coupling $(F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}$ (see Section III a).

From now on (with the exception of Section A from the appendix where we prove Theorem IV.1) we will focus on stationary processes, which are more adapted to learning tasks on distributions where there is no a priori reason to associate different dispersion properties to the outputs corresponding to different distribution inputs.

B. Stationary processes

We now construct Gaussian processes which are stationary with respect to the Wasserstein distance.

Theorem IV.2. *For every completely monotone function F and $0<H\leq 1$ the function*

$$(\mu,\nu)\mapsto F\left(W_2^{2H}(\mu,\nu)\right) \tag{12}$$

is a covariance function on $\mathcal{W}_2(\mathbb{R})$. Furthermore a Gaussian random process with constant mean and covariance (12) *is stationary with respect to the Wasserstein distance.*

We recall that a $\mathcal{C}^\infty$ function $F:\mathbb{R}^+\to\mathbb{R}^+$ is said to be *completely monotone* if for every $n\in\mathbb{N}$ and $x\in\mathbb{R}^+$,

$$(-1)^n F^{(n)}(x)\geq 0.$$

Here $F^{(n)}$ denotes the derivative of order n of F. The prototype of a completely monotone fuction is $x\mapsto e^{-\lambda x}$, for any positive λ. Furthermore F is completely monotone if and only if it is the Laplace transform of a positive measure μ_F with finite mass on $\mathbb{R}^+$, that is to say

$$F(x) = \int_{\mathbb{R}^+} e^{-\lambda x} d\mu_F(\lambda).$$

Other examples of completely monotone functions include $x^{-\lambda}$ for positive values of λ and $\log\left(1+\frac{1}{x}\right)$.

Example Applying theorem IV.2 with the completely monotone functions $e^{-\lambda x}$ we obtain the stationary covariance kernels

$$e^{-\lambda W_2^{2H}(\mu,\nu)}, \tag{13}$$

for every $\lambda>0$ and $0<H\leq 1$.

These kernels are generalizations to distribution inputs of the kernels of the form $e^{-\lambda\|x-y\|^{2H}}$ on $\mathbb{R}^d$, which are classical in spatial statistics and machine learning. In particular setting $H=1/2$ gives the family of *Laplace kernels*, and $H=1$ the family of *Gaussian kernels*.

At this point we have obtained enough covariance functions to consider parametric models that fit practical datasets. Section V addresses the question of the selection of the best covariance kernel amongst a parametric family of stationary kernels, together with the prediction of the associated Gaussian process. In Section VI we carry out simulations with the following parametric model, which is directly derived from (13):

$$\left\{K_{\sigma^2,\ell,H} = \sigma^2 e^{-\frac{W_2^{2H}}{\ell}},\ (\sigma^2,\ell,H)\in C\times C'\times[0,1]\right\}, \tag{14}$$

where $C, C'\subset(0,\infty)$ are two compact sets.

C. Ideas of proof

Theorems IV.1 and IV.2 are direct corollaries of the following result:

Theorem IV.3. *The function W_2^{2H} is a negative definite kernel if and only if $0\leq H\leq 1$. Furthermore, it is nondegenerate if and only if $0<H<1$.*

One can find in [17] a proof of the negative definiteness of the kernel W_2^{2H} restricted to absolutely continuous distributions in $\mathcal{W}_2(\mathbb{R})$. The proof given here holds for any distribution of $\mathcal{W}_2(\mathbb{R})$, and we provide the nondegeneracy property of the kernel.

In short (see Appendinx A for a detailed proof), we consider $H=1$ and the optimal coupling (see Section III a)

$$(Z(\mu))_{\mu\in\mathcal{W}_2(\mathbb{R})} := (F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}, \tag{15}$$

where F_μ^{-1} is the quantile function of the distribution μ and U is an uniform random variable on $[0,1]$. This coupling can be seen as a (non-Gaussian !) random field indexed by $\mathcal{W}_2(\mathbb{R})$. As such, its variogram

$$(\mu,\nu) \mapsto \mathbb{E}(Z(\mu)-Z(\nu))^2 \tag{16}$$

is a negative definite kernel. Furthermore it is equal to $W_2^2(\mu,\nu)$ since the coupling $(Z(\mu))$ is optimal (see (1)). The proof ends with the use of the following classical lemma:

Lemma IV.4. *If K is a negative definite kernel then K^H is a negative definite kernel for every $0 \leq H \leq 1$.*

See $e.g.$ [32] for a proof Lemma IV.4.

Remark In [7], Istas defines the fractional index of a metric space E endowed with a distance d by

$$\beta_E := \sup\left\{\beta > 0 \mid d^\beta \text{ is negative definite}\right\}. \tag{17}$$

One of the interpretation of the fractional index is that $\beta_E/2$ it is the maximal regularity for a fractional Brownian motion indexed by (E,d): indeed the H-fractional Brownian motion indexed by a metric space exists if and only if $H \leq \beta_E/2$. For instance, the fractional exponent of the Euclidean spaces $\mathbb{R}^n$ is equal to 2, while the fractional index of the spheres $\mathbb{S}^n$ is only 1. Recall that an H-fractional Brownian motion have more regular paths and exhibits long-distance correlation for large values of H. In a non-rigorous way, the fractional index can be seen as some measure of the difficulty to construct long-distance correlated random field indexed by the space (E,d).

It is in general a difficult problem to find the fractional index of a given space. Theorem IV.3 states that the fractional exponent $\beta_{\mathcal{W}_2(\mathbb{R})}$ of the Wasserstein space is equal to 2.

V. Model selection and Gaussian process regression

A. Maximum Likelihood and prediction

Let us consider a Gaussian process Y indexed by $\mathcal{W}_2(\mathbb{R})$, with zero mean function and unknown covariance function K_0. Most classically, it is assumed that the covariance function K_0 belongs to a parametric set of the form

$$\{K_\theta; \theta \in \Theta\}, \tag{18}$$

with $\Theta \subset \mathbb{R}^p$ and where K_θ is a covariance function and θ is called the covariance parameter. Hence we have $K_0 = K_{\theta_0}$ for some true parameter $\theta_0 \in \Theta$.

For instance, considering the fractional Brownian motion kernel given in (8), we can have $\theta = (\sigma^2, H)$, $\Theta = (0,\infty) \times (0,1]$ and $K_\theta = \sigma^2 K^{H,\eta}$, where η is fixed in $\mathcal{W}_2(\mathbb{R})$. In this case, the covariance parameters are the order of magnitude parameter σ^2 and the regularity parameter H.

Typically, the covariance parameter θ is selected from a data set of the form $(\mu_i, y_i)_{i=1,\dots,n}$, with $y_i = Y(\mu_i)$. Several techniques have been proposed for constructing an estimator $\hat{\theta} = \hat{\theta}(\mu_1, y_1, ..., \mu_n, y_n)$, in particular maximum likelihood (see e.g. [33]) and cross validation [34]–[36]. In this paper, we shall focus on maximum likelihood, which is widely used in practice and has received a lot of theoretical attention.

Maximum Likelihood is based on maximizing the Gaussian likelihood of the vector of observations $(y_1, ..., y_n)$. The estimator is $\hat{\theta}_{ML} \in \operatorname{argmin} L_\theta$ with

$$L_\theta = \frac{1}{n}\ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{19}$$

where $R_\theta = [K_\theta(\mu_i, \mu_j)]_{1\leq i,j\leq n}$

Given the maximum likelihood estimator $\hat{\theta}_{ML}$, the value $Y(\mu)$, for any input $\mu \in \mathcal{W}_2(\mathbb{R})$, can be predicted by plugging (see for instance in [33]) $\hat{\theta}_{ML}$ in the conditional expectation (or posterior mean) expression for Gaussian processes. More precisely, $Y(\mu)$ is predicted by $\hat{Y}_{\hat{\theta}_{ML}}(\mu)$ with

$$\hat{Y}_\theta(\mu) = r_\theta^t(\mu) R_\theta^{-1} y \tag{20}$$

and

$$r_\theta(\mu) = \begin{bmatrix} K_\theta(\mu,\mu_1) \\ \vdots \\ K_\theta(\mu,\mu_n) \end{bmatrix}.$$

Note that $\hat{Y}_\theta(\mu)$ is the conditional expectation of $Y(\mu)$ given $y_1, ..., y_n$, when assuming that Y is a centered Gaussian process with covariance function K_θ.

B. Asymptotic properties

In this section, we aim at showing that some of the asymptotic results of the Gaussian process literature, which hold for Gaussian processes indexed by $\mathbb{R}^d$, can be extended to Gaussian processes indexed by $\mathcal{W}_2(\mathbb{R})$. To our knowledge, this extension has not been considered before.

For a Gaussian process indexed by $\mathbb{R}^d$, two main asymptotic frameworks are under consideration: fixed-domain and increasing-domain asymptotics [33]. Under increasing-domain asymptotics, as $n \to \infty$, the observation points $x_1, ..., x_n \in \mathbb{R}^d$ are so that $\min_{i\neq j} ||x_i - x_j||$ is lower bounded. Under fixed-domain asymptotics, the sequence (or triangular array) of observation points $(x_1, ..., x_n)$ become dense in a fixed bounded subset of $\mathbb{R}^d$. To be specific, for a Gaussian process indexed by $\mathbb{R}$, a standard increasing-domain framework would be given by $x_i = i$ for $i \in \mathbb{N}$, while a standard fixed-domain framework would be given by, for $n \in \mathbb{N}$, $x_i = i/n$ for $i = 1, ..., n$.

Let us now briefly review the existing results for Gaussian processes indexed by $\mathbb{R}^d$. Typically, under increasing-domain asymptotics, the true covariance parameter θ_0 is estimated consistently by maximum likelihood, with asymptotic normality [37]–[42]. Also, predicting with the estimated covariance parameter $\hat{\theta}$ is asymptotically as good as predicting with θ_0 [41].

Under fixed-domain asymptotics, there are cases where some components of the true covariance parameter θ_0 can not be consistently estimated [33], [43]. Nevertheless, these components which can not be estimated consistently do not have an asymptotic impact on prediction [44]–[46]. Some results on prediction with estimated covariance parameters are available in [47]. Also, asymptotic properties of maximum likelihood estimators are obtained in [48]–[52].

We remark, finally, that the above increasing-domain asymptotic results hold for fairly general classes of covariance functions, while fixed-domain asymptotic results currently have to be derived for specific covariance functions and on a case-by-case basis.

For this reason, in this paper, we focus on extending some of the above increasing-domain asymptotic results to Gaussian processes indexed by $\mathcal{W}_2(\mathbb{R})$. Indeed, this will enable us to obtain a fair amount of generality with respect to the type of covariance functions considered.

We thus extend the contributions of [41] in the case of Gaussian processes with probability distribution inputs. In the rest of the section, we first list and discuss technical conditions for the asymptotic results. Then, we show the consistency and asymptotic normality of maximum likelihood and show that predictions from the maximum likelihood estimator are asymptotically as good as those obtained from the true covariance parameter. In Section V-C, we study an explicit example, for which all the technical conditions can be satisfied. All the proofs are postponed to the appendix. At the end of Section V-C, we discuss the novelty of these proofs, compared to those of the literature, and especially those in [41].

The technical conditions for this section are listed below.

Condition V.1. *We consider a triangular array of observation points $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ so that for all $n \in \mathbb{N}$ and $1 \leq i \leq n$, μ_i has support in $[i, i+L]$ with a fixed $L < \infty$.*

Condition V.2. *The model of covariance functions $\{K_\theta, \theta \in \Theta\}$ satisfies*

$$\forall \theta \in \Theta,\ K_\theta(\mu, \nu) = F_\theta\left(W_2(\mu, \nu)\right),$$

with $F_\theta : \mathbb{R}^+ \to \mathbb{R}$ and

$$\sup_{\theta \in \Theta} |F_\theta(t)| \leq \frac{A}{1 + |t|^{1+\tau}}$$

with a fixed $A < \infty$, $\tau > 1$.

Condition V.3. *We have observations $y_i = Y(\mu_i)$, $i = 1, \cdots, n$ of the centered Gaussian Process Y with covariance function K_{θ_0} for some $\theta_0 \in \Theta$.*

Condition V.4. *The sequence of matrices $R_\theta = (K_\theta(\mu_i, \mu_j))_{1 \leq i,j \leq n}$ satisfies*

$$\lambda_{\inf}(R_\theta) \geq c$$

for a fixed $c > 0$, where $\lambda_{\inf}(R_\theta)$ denotes the smallest eigenvalue of R_θ.

Condition V.5. $\forall \alpha > 0$,

$$\liminf_{n \to \infty} \inf_{\|\theta - \theta_0\| \geq \alpha} \frac{1}{n} \sum_{i,j=1}^{n} \left[K_\theta(\mu_i, \mu_j) - K_{\theta_0}(\mu_i, \mu_j)\right]^2 > 0.$$

Condition V.6. $\forall t \geq 0$, *$F_\theta(t)$ is continuously differentiable with respect to θ and we have*

$$\sup_{\theta \in \Theta} \max_{i=1,\cdots,p} \left| \frac{\partial}{\partial \theta_i} F_\theta(t) \right| \leq \frac{A}{1 + t^{1+\tau}},$$

with A, τ as in Condition V.2.

Condition V.7. $\forall t \geq 0$, *$F_\theta(t)$ is three times continuously differentiable with respect to θ and we have, for $q \in \{2, 3\}$, $i_1 \cdots i_q \in \{1, \cdots p\}$,*

$$\sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_q}} F_\theta(t) \right| \leq \frac{A}{1 + t^{1+\tau}}.$$

Condition V.8. $\forall (\lambda_1 \cdots, \lambda_p) \neq (0, \cdots, 0)$,

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i,j=1}^{n} \left(\sum_{k=1}^{p} \lambda_k \frac{\partial}{\partial \theta_k} K_{\theta_0}\left(\mu_i, \mu_j\right) \right)^2 > 0.$$

Condition V.1 mimics the increasing-domain asymptotic framework discussed above for vectorial inputs. In particular, the observation measures μ_i and μ_j yield a large Wasserstein distance when $|i - j|$ is large.

Condition V.2 entails that all the covariance functions under consideration are stationary in the sense that the covariance between μ and ν depends only on the distance $W_2(\mu, \nu)$. Stationarity is also assumed when considering increasing-domain asymptotics for Gaussian processes indexed by $\mathbb{R}^d$ [37]–[42]. Hence, we remark that the asymptotic results of the present section do not apply to the covariance functions of fractional Brownian motion in (8). On the other hand, these results apply to the power exponential covariance functions in (14).

Condition V.2 also imposes that the covariance functions in the parametric model decrease fast enough with the Wasserstein distance. This condition is standard in the case of vector inputs, and holds for instance for the covariance functions in (14).

Condition V.3 means that we address the well-specified case [34], [35], where there is a true covariance parameter θ_0 to estimate.

Condition V.4 is technically necessary for the proof techniques of this paper. This condition holds whenever the covariance model satisfies, for all $\theta \in \Theta, w \geq 0$, $F_\theta(w) = \bar{F}_\theta(w) + \delta_\theta \mathbf{1}_{\{w=0\}}$, where $\bar{F}_\theta$ is a continuous covariance function and where $\inf_{\theta \in \Theta} \delta_\theta > 0$. This situation corresponds to Gaussian processes observed with Gaussian measure errors, or to Gaussian processes with very small scale irregularities, and is thus representative of a significant range of practical applications.

In the case where F_θ is continuous (which usually means that we have exact observations of a Gaussian process with continuous realizations), then Condition V.4 implies that

$$\inf_{n \in \mathbb{N}, i \neq j = 1, \ldots, n} W_2(\mu_i, \mu_j) > 0. \tag{21}$$

For a large class of Gaussian processes indexed by $\mathbb{R}^d$, it has been shown that the condition in (21) (with W_2 replaced by the Euclidean distance) is also sufficient for Condition V.4 [41], [53]. The proof relies on the Fourier transform on $\mathbb{R}^d$. For Gaussian processes indexed by $\mathcal{W}_2(\mathbb{R})$, one could expect the condition in (21) to be sufficient to guarantee V.4 in many cases, although, to our knowledge, obtaining rigorous proofs in this direction is an open problem.

Condition V.5 means that there is enough information in the triangular array $\{\mu_1, ..., \mu_n\}$ to differentiate between the covariance functions K_{θ_0} and K_θ, when θ is bounded away from θ_0. We believe that Condition V.5 can be checked for specific explicit instances of the triangular array $\{\mu_1, ..., \mu_n\}$, as it involves an explicit sum of covariance values.

Conditions V.6 and V.7 are standard regularity and asymptotic decorrelation conditions for the covariance model. They hold, in particular, for the power exponential covariance model of (14).

Finally, Condition V.8 is interpreted as an asymptotic local linear independence of the p derivatives of the covariance function, around θ_0. Since this condition involves an explicit sum of covariance function derivatives, we believe that it can be checked for specific instances of the triangular array $\{\mu_1, ..., \mu_n\}$.

We now provide the first result of this section, showing that the maximum likelihood estimator is asymptotically consistent.

Theorem V.9. *Let $\hat{\theta}_{ML}$ be as in* (19). *Under Conditions V.1 to V.5, we have as $n \to \infty$*

$$\hat{\theta}_{ML} \xrightarrow{\mathbb{P}} \theta_0.$$

In the next theorem, we show that the maximum likelihood estimator is asymptotically Gaussian. In addition, the rate of convergence is $\sqrt{n}$, and the asymptotic covariance matrix M_{ML}^{-1} of $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ (that may depend on n) is asymptotically bounded and invertible, see (22).

Theorem V.10. *Let M_{ML} be the $p \times p$ matrix defined by*

$$(M_{ML})_{i,j} = \frac{1}{2n} Tr\left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right),$$

with R_θ as in (19). *Under Conditions V.1 to V.8 we have*

$$\sqrt{n} M_{ML}^{1/2} \left(\hat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_n).$$

Furthermore,

$$0 < \liminf_{n\to\infty} \lambda_{min}(M_{ML}) \le \limsup_{n\to\infty} \lambda_{max}(M_{ML}) < +\infty. \tag{22}$$

In the next theorem, we show that, when using the maximum likelihood estimator, the corresponding predictions of the values of Y are asymptotically equal to the predictions using the true covariance parameter θ_0. Note that, in the increasing-domain framework considered here, the mean square prediction error is typically lower-bounded, even when using the true covariance parameter. Indeed, this occurs in the case of Gaussian processes with vector inputs, see Proposition 5.2 in [41].

Theorem V.11. *Under Conditions V.1 to V.8 we have*

$$\forall \mu \in \mathcal{W}_2(\mathbb{R}), \ \left| \hat{Y}_{\hat{\theta}_{ML}}(\mu) - \hat{Y}_{\theta_0}(\mu) \right| = o_{\mathbb{P}}(1),$$

with $\hat{Y}_\theta(\mu)$ as in (20).

C. An example

In this section, we provide an explicit example of triangular array of probability measures for which Conditions V.5 and V.8 are satisfied. We consider random probability measures $(\mu_i)_{i\in\mathbb{N}}$ which are independent and identically distributed (up to support shifts to satisfy condition V.1). We then show that Conditions V.5 and V.8 are satisfied almost surely. The motivation for studying shifted independent and identically distributed random probability measures is that this this model is simple to describe and can generate a large range of sequences $\{\mu_1, ..., \mu_n\}$.

Proposition V.12. *Assume that Conditions V.2, V.6 and V.7 hold.*

Assume that for $\theta \neq \theta_0$, F_θ and F_{θ_0} are not equal everywhere on $\mathbb{R}^+$. Assume that there does not exist $(\lambda_1, ..., \lambda_p) \neq (0, ..., 0)$ so that $\sum_{i=1}^p (\partial/\partial\theta_i) F_{\theta_0}$ is the zero function on $\mathbb{R}^+$.

Let $(Z_i)_{i\in\mathbb{Z}}$ be independent and identically distributed Gaussian processes on $\mathbb{R}$ with continuous trajectories. Assume that Z_0 has mean function 0 and covariance function C_0. Assume that $C_0(u, v) = C_0(u', v')$ whenever $v - u = v' - u'$ and let $C_0(u, v) = C_0(u - v)$ for ease of notation. Let $\hat{C}_0(w) = \int_{\mathbb{R}} C_0(t) e^{\mathrm{i}wt} dt$ with $\mathrm{i}^2 = -1$. Assume that $\hat{C}_0(w)|w|^{2q}$ is bounded away from 0 and ∞ as $|w| \to \infty$, for some fixed $q \in (0, \infty)$.

Let $L > 1$ be fixed. For $i \in \mathbb{Z}$, let $f_i : \mathbb{R} \to \mathbb{R}^+$ be defined by $f_i(t) = \exp(Z_i(t - i))/M_i$ if $t \in [i, i + L]$ and $f_i(t) = 0$ else, where $M_i = \int_i^{i+L} \exp(Z_i(t - i)) dt$. Let μ_i be the measure with probability density function f_i. Then, almost surely, with the sequence of random probability measures $\{\mu_1, ..., \mu_n\}$, Conditions V.5 and V.8 hold.

In Proposition V.12, the identifiability assumptions on $\{F_\theta\}$ are very mild, and hold for instance for the power exponential model in (14).

In Proposition V.12, the random probability measures have probability density functions obtained from exponentials of realizations of Gaussian processes. Hence, these measures have a non-parametric source of randomness, and can take flexible forms. Several standard covariance functions on $\mathbb{R}$ satisfy the conditions in Proposition V.12, in particular the Matérn covariance functions (see e.g. [33]).

We remark that, in the context of Proposition V.12, when $F_\theta(w) = \bar{F}_\theta(w) + \delta_\theta \mathbf{1}_{\{w=0\}}$, with $\bar{F}_\theta$ a continuous covariance function and $\inf_{\theta\in\Theta} \delta_\theta > 0$, as described when discussing Condition V.4, then Conditions V.1 to V.8 hold so that Theorems V.9, V.10 and V.11 hold. If however F_θ is continuous, then Condition V.4 almost surely does not hold since $L > 1$ (as there will almost surely be pairs of distributions μ_i, μ_i, $i \neq j$, with arbitrarily small $W_2(\mu_i, \mu_j)$). Nevertheless, when $L < 1$, it can be shown that Proposition V.12 still holds when, in the conditions of this proposition on $\{F_\theta\}$, $\mathbb{R}^+$ is replaced by $\cup_{i=1}^\infty [i - L, i + L]$. Also, as discussed above, when $L < 1$, the condition in (21) is satisfied and one could expect Condition V.4 to hold.

We conclude this section by discussing the corresponding proofs (in the appendix). These proofs can be divided into two groups. In the first group (proofs of Theorems V.9, V.10 and

V.11 and of Proposition A.7) we show that the arguments in [41] can be adapted and extended to the setting of the present article. The main innovations in this first group compared to [41] are that we allow for triangular arrays of observation points, and are not restricted to the specific structure of observation points of [41].

The proofs of the second group (proofs of Lemma A.4 and Proposition V.12) are specific to Gaussian processes with distribution inputs and are thus original for the most part. In particular, in the proof of Proposition V.12, we show that, for two measures obtained by taking exponentials of Gaussian processes, the corresponding random Wasserstein distance has maximal distribution support. In this aim, we use equivalence of Gaussian measure tools and specific technical manipulations of the Wasserstein distance.

VI. Simulation study

We now compare the Gaussian process model suggested in the present paper, with various models for predicting scalar outputs corresponding to distributional inputs. Among the covariance functions introduced in this paper, we shall focus on the power-exponential model (14), since its covariance functions are stationary with respect to the Wasserstein distance. We will not consider the fractional Brownian motion model (8), since it imposes to choose a "zero distribution", from which the variance increases with the distance. While this feature is relevant in some applications (for instance in finance), it is not natural in the simulation examples adressed here.

A. *Comparison with projection-based covariance functions*

In this section, we focus on Gaussian process models for prediction. We compare the covariance functions (14) of this paper, operating directly on the input probability distributions, to more classical covariance functions operating on projections of these probability measures on finite dimensional spaces.

1) Overview of the simulation procedure: We address the input-output map given by, for a distribution ν on $\mathbb{R}$,

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}},$$

where $m_k(\nu) = \int_{\mathbb{R}} x^k d\nu(x)$.

We first simulate independently $n = 100$ learning distributions $\nu_1, ..., \nu_{100}$ as follows. First, we sample uniformly $\mu_i \in [0.3, 0.7]$ and $\sigma_i \in [0.001, 0.2]$, and compute f_i, the density of the Gaussian distribution with mean μ_i and variance σ_i^2. Then, we generate the function g_i with value $f_i(x) \exp(Z_i(x))$, $x \in [0, 1]$, where Z_i is a realization of a Gaussian process on $[0, 1]$ with mean function 0 and Matérn $5/2$ covariance function with parameters $\sigma = 1$ and $\ell = 0.2$ (see e.g. [54] for the expression of this covariance function). Finally, ν_i is the distribution on $[0, 1]$ having density $g_i / (\int_0^1 g_i)$. In Figure 1, we show the density functions of 10 of these n sampled distributions. From the figure, we see that the learning distributions keep a relatively strong underlying two dimensional structure, driven by the randomly generated means and standard deviations. At the same time, because of the random perturbations generated with the Gaussian processes Z_i, these distributions are not restricted in a finite-dimensional space, and can exhibit various degrees of asymmetries.

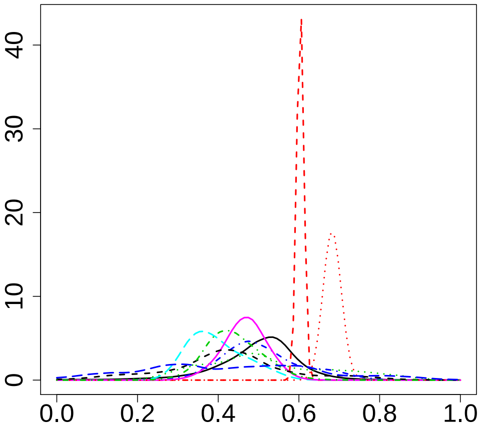

Fig. 1. Probability density functions of 10 of the randomly generated learning distributions for the simulation study.

From the learning set $(\nu_i, F(\nu_i))_{i=1,\dots,n}$, we fit three Gaussian process models, which we call "distribution", "Legendre" and "PCA", and for which we provide more details below. Each of these three Gaussian process models provide a conditional expectation function

$$\nu \to \hat{F}(\nu) = \mathbb{E}(F(\nu) | F(\nu_1), ..., F(\nu_n))$$

and a conditional variance function

$$\nu \to \hat{\sigma}^2(\nu) = \mathrm{var}(F(\nu) | F(\nu_1), ..., F(\nu_n)).$$

We then evaluate the quality of the three Gaussian process models on a test set of size $n_t = 500$ of the form $(\nu_{t,i}, F(\nu_{t,i}))_{i=1,\dots,n_t}$, where the $\nu_{t,i}$ are generated in the same way as the ν_i above. We consider the two following quality criteria. The first one is the root mean square error (RMSE),

$$RMSE^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i}) \right)^2,$$

which should be minimal. The second one is the confidence interval ratio (CIR) at level $\alpha \in (0, 1)$,

$$CIR_\alpha = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{1} \left\{ \left| F(\nu_{t,i}) - \hat{F}(\nu_{t,i}) \right| \leq q_\alpha \hat{\sigma}(\nu_{t,i}) \right\},$$

with q_α the $\left(\frac{1}{2} + \frac{\alpha}{2}\right)$ quantile of the standard normal distribution. The CIR_α criterion should be close to α.

2) Details on the Gaussian process models: The "distribution" Gaussian process model is based on the covariance functions discussed before, operating directly on probability distributions. In this model, the Gaussian process has mean function zero and a covariance function of the form

$$K_{\sigma^2, \ell, H}(\nu_1, \nu_2) = \sigma^2 \exp\left(- \frac{W_2(\nu_1, \nu_2)^{2H}}{\ell} \right).$$

We call the covariance parameters $\sigma^2 > 0$, $\ell > 0$ and $H \in [0,1]$ the variance, correlation length and exponent. These parameters are estimated by maximum likelihood from the training set $(\nu_i, F(\nu_i))_{i=1,...,n}$, which yields the estimates $\hat{\sigma}^2, \hat{\ell}, \hat{H}$. Finally, the Gaussian process model for which the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed is a Gaussian process with mean function zero and covariance function $K_{\hat{\sigma}^2,\hat{\ell},\hat{H}}$.

The "Legendre" and "PCA" Gaussian process models are based on covariance functions operating on finite-dimensional linear projections of the distributions. These projection-based covariance functions are used in the literature, in the general framework of stochastic processes with functional inputs, see e.g. [55], [56]. For the "Legendre" covariance function, for a distribution ν with density f_ν and support $[0,1]$, we compute, for $i = 0, ..., o-1$

$$a_i(\nu) = \int_0^1 f_\nu(t) p_i(t) dt,$$

where p_i is the $i-th$ normalized Legendre polynomial, with $\int_0^1 p_i^2(t)dt = 1$. The integer o is called the order of the decomposition. Then, the covariance function operates on the input vector $(a_0(\nu), ..., a_{o-1}(\nu))$ and is of the form

$$K_{\sigma^2,\ell_0,...,\ell_{o-1},H}(\nu_1,\nu_2) = \sigma^2 \exp\left(-\left\{\sum_{i=0}^{o-1}\left[\frac{|a_i(\nu_1)-a_i(\nu_2)|}{\ell_i}\right]\right\}^H\right).$$

The covariance parameters $\sigma^2 \geq 0, \ell_0 > 0, ..., \ell_{o-1} > 0, H \in (0,1]$ are estimated by maximum likelihood, from the learning set $(a_0(\nu_i), ..., a_{o-1}(\nu_i), F(\nu_i))_{i=1,...,n}$. Finally, the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed as for the "distribution" Gaussian process model.

For the "PCA" covariance function, we discretize each of the n probability density functions f_{ν_i} to obtain n vectors $v_i = (f_{\nu_i}(j/(d-1)))_{j=0,...,d-1}$, with $d = 100$. Then, we let $w_1, ..., w_o$ be the first o principal component vectors of the set of vectors $(v_1, ..., n_n)$. For any distribution ν with density f_ν, we associate its projection vector $(a_1(\nu), ..., a_o(\nu))$ defined as

$$a_i(\nu) = \frac{1}{d}\sum_{j=0}^{d-1} f_\nu(j/(d-1))(w_i)_j.$$

This procedure corresponds to the numerical implementation of functional principal component analysis presented in Section 2.3 of [57]. Then, the covariance function in the "PCA" case operates on the input vector $(a_1(\nu), ..., a_o(\nu))$. Finally, the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed as for the "Legendre" Gaussian process model.

model	RMSE	$CIR_{0.9}$
"distribution"	0.094	0.92
"Legendre" order 5	0.49	0.92
"Legendre" order 10	0.34	0.89
"Legendre" order 15	0.29	0.91
"PCA" order 5	0.63	0.82
"PCA" order 10	0.52	0.87
"PCA" order 15	0.47	0.93

TABLE I
Values of different quality criteria for the "distribution", "Legendre" and "PCA" Gaussian process models. The "distribution" Gaussian process model is based on covariance functions operating directly on the input distributions, while "Legendre" and "PCA" are based on linear projections of the input distributions on finite-dimensional spaces. For "Legendre" and "PCA", the order value is the dimension of the projection space. The quality criteria are the root mean square error (RMSE) which should be minimal and the confidence interval ratio ($CIR_{0.9}$) which should be close to 0.9. The "distribution" Gaussian process model clearly outperforms the two other models.

3) Results: In Table I we show the values of the RMSE and $CIR_{0.9}$ quality criteria for the "distribution", "Legendre" and "PCA" Gaussian process models. From the values of the RMSE criterion, the "distribution" Gaussian process model clearly outperforms the two other models. The RMSE of the "Legendre" and "PCA" models slightly decreases when the order increases, and stay well above the RMSE of the "distribution" model. Note that with orders 10 and 15, despite being less accurate, the "Legendre" and "PCA" models are significantly more complex to fit and interpret than the "distribution" model. Indeed these two models necessitate to estimate 12 and 17 covariance parameters, against 3 for the "distribution" model. The maximum likelihood estimation procedure thus takes more time for the "Legendre" and "PCA" models than for the "distribution" model. We also remark that all three models provide appropriate predictive confidence intervals, as the value of the $CIR_{0.9}$ criterion is close to 0.9. Finally, "Legendre" performs slightly better than "PCA".

Our interpretation for these results is that, because of the nature of the simulated data $(\nu_i, F(\nu_i))$, working directly on distributions, and with the Wasserstein distance, is more appropriate than using linear projections. Indeed, in particular, two distributions with similar means and small variances are close to each other with respect to both the Wasserstein distance and the value of the output function F. However, if the ratio between the two variances is large, the probability density functions of the two distributions are very different from each other, with respect to the L^2 distance. Hence, linear projections based on probability density functions is inappropriate in the setting considered here.

B. *Comparison with the kernel regression procedure of [23]*

In this section, we compare the "distribution" method of Table I which is suggested in the present article, with the "kernel regression" procedure of [23]. This procedure consists in predicting $f(P) \in \mathbb{R}$, with $P \in \mathcal{W}_2(\mathbb{R})$, from $\hat{P}, \hat{P}_1, ..., \hat{P}_n, f(P_1), ..., f(P_n)$ where $\hat{P}, \hat{P}_1, ..., \hat{P}_n$ are estimates of $P, P_1, ..., P_n \in \mathcal{W}_2(\mathbb{R})$ obtained from sample values of $P, P_1, ..., P_n$. In [23], $\hat{P}, \hat{P}_1, ..., \hat{P}_n$ correspond to kernel smoothing estimates of probability density functions constructed from the sample values. Then, the prediction $\hat{f}(\hat{P})$ of $f(P)$ is obtained by a weighted average of $f(P_1), ..., f(P_n)$ where the weights are computed by applying a kernel to the distances $D(\hat{P}, \hat{P}_1), ..., D(\hat{P}, \hat{P}_n)$. The distances suggested in [23] are the L^1 distances between the estimated probability density functions. We remark that there is no estimate of the prediction error $f(P) - \hat{f}(\hat{P})$ in [23], which is a downside compared to the Gaussian process model considered in this paper.

An interesting feature of the setting of [23] is that the input P of the function value $f(P)$ is not observed. Only a sample from P is available (this is the "two-stage sampling" difficulty described in [15], which arises in various applications) We shall demonstrate in this section that Gaussian process models can accommodate with this constraint. The idea is that $f(\hat{P})$ differs from $f(P)$, and that this difference can be modeled by adding a nugget variance parameter to the Gaussian process model. More precisely, the covariance functions we shall study in this section are

$$K_{\sigma^2,\ell,H,\delta}(\nu_1,\nu_2) \\ = \sigma^2 \exp\left(-\frac{W_2(\nu_1,\nu_2)^{2H}}{\ell}\right) + \delta \mathbf{1}\{W_2(\nu_1,\nu_2)=0\}, \quad (23)$$

where $\delta \geq 0$ is an additional covariance parameter, which can also be estimated in the maximum likelihood procedure. Apart from this modification of the covariance model, we carry out the Gaussian process model computation as in Section VI-A, with always $W_2(P,Q)$ replaced by $W_2(\check{P},\check{Q})$, where $\check{P},\check{Q}$ are the empirical distributions corresponding to the available sample values from P,Q.

We first reproduce the "skewness of Beta" example of [23]. In this example $n = 275$ distributions $P_1,...,P_n$ are randomly and independently generated for the learning set. We have that $P_i = B_{a_i}$ is the Beta distribution with parameters (a_i, b) where a_i is uniformly distributed on $[3,20]$ and $b = 3$. The test set consists in $n_t = 50$ distributions $P_{t,1},...,P_{t,n_t}$ generated independently in the same way. The function to predict is defined by $f(P_a) = [2(b-a)(a+b+1)^{1/2}]/[(a+b+2)(ab)^{1/2}]$ and corresponds to the skewness of the Beta distribution. For each distribution, 500 sample values are available. For the "kernel regression" procedure, we used the same settings (kernel, bandwidth selection, training and validation sets...) as in [23].

The predictions obtained by the "distribution" and "kernel regression" procedures are presented in Figure 2. We observe that both methods perform equally well. The prediction errors are small, and are essentially due to to the fact that we only observe random samples from the distributions. [We have repeated the simulation of Figure 2 with $5,000$ sample values instead of 500, and the predicted values have become visually equal to the true values.] Our conclusion on this "skewness of Beta" example is that the setting is here very favourable (the input space of distributions is one-dimensional and 275 observations of the function are available) so that both methods have similar good performances.

Next, we repeat the "distribution" and "kernel regression" procedures on the same setting as in Table I (except that each input and predictand distribution is only observed indirectly, through 500 sample values from it). The prediction results, based on the same criteria as in Table I are given in Table II. We observe that the RMSE prediction criterion for the "distribution" model is deteriorated compared to Table I. This is due to the fact that the distributions are not observed exactly anymore. The $CIR_{0.9}$ criterion is equal to 0.91 for the "distribution" Gaussian process model. Hence, thanks to the addition of the nugget variance parameter, the Gaussian process model is able to take into account the additional uncertainty due to the random samples of the unobserved distributions, and to yield appropriate conditional variances.

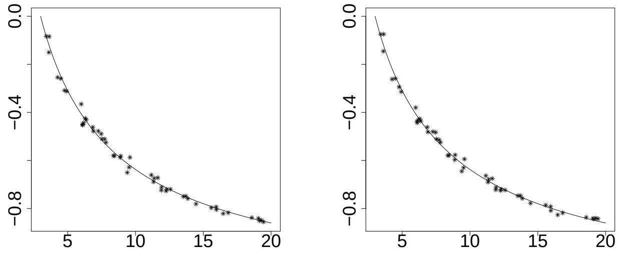

Fig. 2. Comparison of the "distribution" Gaussian process model of this paper (left) with the "kernel regression" procedure (right) for the "skewness of Beta" example. We predict the skewness of the Beta distribution (y-axis) from samples obtained from Beta distributions with parameter $(a, 3)$ with $a \in [3, 20]$ (x-axis). The true skewness is in plain line and the predictions are the dots. Both methods perform equally well.

model	RMSE	$CIR_{0.9}$
"distribution"	0.21	0.91
"kernel regression"	0.93	

TABLE II
SAME SETTING AS IN TABLE I, EXCEPT THAT THE INPUT AND PREDICTAND DISTRIBUTIONS ARE ONLY OBSERVED INDIRECTLY, THROUGH SAMPLE VALUES FROM THEM. THE "DISTRIBUTION" MODEL SUGGESTED IN THIS PAPER CLEARLY OUTPERFORMS THE "KERNEL REGRESSION" PROCEDURE.

We also observe that the RMSE pediction criterion is much larger for the "kernel regression" procedure. Hence, in this more challenging scenario (the input-space of distributions is non-parametric and only 100 learning function values are available), the "distribution" Gaussian process model become strongly preferable. In our opinion, this is because the Wasserstein distance is here more relevant than distances between probability density functions (as discussed for Table I). Also, Gaussian process prediction has benefits compared to prediction with weighted kernel averages. In particular, Gaussian process predictions come with a probabilistic model and have optimality properties under this model.

C. Numerical complexity of the method

Our method inherits the numerical complexity of Gaussian Process Regression in more classical settings. Given a learning dataset $(\mu_i, y_i)_{i=1}^N$ the complexity of the Kriging method is given by the inversion of the covariance matrix $K_\theta(\mu_i,\mu_j)_{i,j=1}^N$, which is in $O(N^3)$ number of operations. The Wasserstein distances between every pair of μ_i need also to be evaluated, which costs $O(N^2 q)$ operations, where q is the size of the sampling of the distributions.

Each prediction is then obtained by a vector product in $O(N)$ operations, while the computation of the conditional variance at some outputs is obtained in $O(N^2)$.

The $O(N^3)$ cost of the overall method makes it challenging to use on very large datasets, however on moderately large datasets its good performances makes it an interesting choice, and in particular a preferable choice the other methods it was compared to in this simulation study.

For the sake of illustration, we remark that it took around 9 seconds to carry out our whole suggested Gaussian process procedure, in the case of Table I, and around 30 seconds in the case of Table II.

See also [58] for a discussion of the covariance tapering method to reduce the numerical cost of Gaussian Process Regression.

VII. Conclusion

We provided a new approach to learning with distribution inputs. Its strength relies on the existence of positive definite kernels on the distribution space, which enables the use of Gaussian process models and kernel learning methods. In particular, we generalized the seminal models that are the fractional Brownian motion and the power exponential stationary processes, to distribution inputs. The kernels we use are functions of the Wasserstein distance, which has proven its efficiency as a discrepancy measure between distributions in numerous applications. Our method requires only the distributions inputs to have a second order moment, which allows the simultaneous handling of very heterogeneous data, such as absolutely continuous distributions, deterministic inputs and empirical distributions, which is particularly important when only a sample of the input distributions is known.

Focusing on Gaussian process regression with stationary covariance functions, we proved that our method extends this classical tool to distribution inputs. In particular, we gave generalization of state of the art asymptotic results to our setting. As in vector input Kriging, the overall numerical complexity of the method is in $O(n^3)$, where n is the size of the dataset, which is more costly than other distribution regression methods (such as the kernel regression procedure from [23]), however our numerical simulations suggest that our method gives better prediction. Furthermore Kriging comes with an error estimation in the form of the conditional variance of the Gaussian process, which is an important guarantee in practice.

On the down side, the methods we use to prove the positive definiteness of our kernels are tightly related to the existence of an optimal coupling between every distribution, which existence is specific to dimension one. It is an important problem for numerous applications to give learning methods for multidimensional distributions. Hence, it would be valuable to obtain kernels based on the the multidimensional Wasserstein space. This would require an other approach that the one used in the present paper, and constitutes an interesting problem for further research.

Appendix
Proofs

A. Proofs for Section IV

Proof of Theorem IV.3. We start with the negative definiteness. For any $\mu \in \mathcal{W}_2(\mathbb{R})$ we denote by F_μ^{-1} the quantile function associated to μ. It is well known that given a uniform random variable U on $[0,1]$, $F_\mu^{-1}(U)$ is a random variable with law μ, and furthermore for every $\mu,\nu \in \mathcal{W}_2(\mathbb{R})$:

$$W_2^2(\mu,\nu) = \mathbb{E}\left(F_\mu^{-1}(U) - F_\nu^{-1}(U)\right)^2, \tag{24}$$

that is to say the coupling of μ and ν given by the random vector $(F_\mu^{-1}(U), F_\nu^{-1}(U))$ is optimal. Consider now $\mu_1,\cdots,\mu_n \in \mathcal{W}_2(\mathbb{R})$ and $c_1,\cdots,c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$. We have

$$\begin{aligned}
&\sum_{i,j=1}^n c_i c_j W_2^2(\mu_i,\mu_j)\\
&= \sum_{i,j=1}^n c_i c_j\, \mathbb{E}\left(F_{\mu_i}^{-1}(U) - F_{\mu_j}^{-1}(U)\right)^2\\
&= \sum_{i,j=1}^n c_i c_j\, \mathbb{E}\left(F_{\mu_i}^{-1}(U)\right)^2 + \sum_{i,j=1}^n c_i c_j\, \mathbb{E}\left(F_{\mu_j}^{-1}(U)\right)^2\\
&\quad - 2\sum_{i,j=1}^n c_i c_j\, \mathbb{E}\left(F_{\mu_i}^{-1}(U)F_{\mu_j}^{-1}(U)\right).
\end{aligned}$$

Using $\sum_{i=1}^n c_i = 0$ the first two sums vanish and we obtain

$$\begin{aligned}
&\sum_{i,j=1}^n c_i c_j W_2^2(\mu_i,\mu_j)\\
&= -2\sum_{i,j=1}^n c_i c_j\, \mathbb{E}\left(F_{\mu_i}^{-1}(U)F_{\mu_j}^{-1}(U)\right)\\
&= -2\,\mathbb{E}\left(\sum_{i=1}^n c_i F_{\mu_i}^{-1}(U)\right)^2 \le 0,
\end{aligned}$$

which proves that W_2^{2H} is a negative definite kernel for $0 \le H \le 1$.

Let us now consider $H > 1$. Using (1) it is clear that for every $x,y \in \mathbb{R}$, $W_2(\delta_x,\delta_y) = |x-y|$. It is well known (see *e.g* [7]) that $|x-y|^{2H}$ is not a negative definite kernel on $\mathbb{R}$ for $H > 1$, hence the same is true for W_2^{2H}.

Let us now prove the nondegeneracy of the kernel: the idea of the proof is to consider $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$ endowed with the product distance

$$d((\mu,s),(\nu,t)) = \left(W_2(\mu,\nu)^2 + |s-t|^2\right)^{1/2}.$$

We assume the degeneracy of the kernel W_2^{2H} on $\mathcal{W}_2(\mathbb{R})$ and deduce that d^{2H} is not negative definite on $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$, in contradiction with the following Lemma, from which we postpone the proof:

Lemma A.1. *The function d^{2H} is a negative definite kernel if and only if $0 \le H \le 1$.*

Let us fix $0 < H < 1$ and assume that W_2^{2H} is degenerate. There exists $\mu_1,\cdots,\mu_n \in \mathcal{W}_2(\mathbb{R})$ and $c_1,\cdots,c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$ and

$$\sum_{i,j=1}^n c_i c_j W_2^{2H}(\mu_i,\mu_j) = 0. \tag{25}$$

In $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$ we now consider the points $P_i = (\mu_i, 0)$ for $1 \le i \le n$ and $P_{n+1} = (\mu_n, \varepsilon)$ with $\varepsilon > 0$. We also set $c_i' = c_i$

for every $1 \leq i \leq n-1$ and $c'_n = c'_{n+1} = c_n/2$. Notice that we have

$$\sum_{i=1}^{n+1} c'_i = 0.$$

Now

$$\begin{aligned} &\sum_{i,j=1}^{n+1} c'_i c'_j d^{2H}(P_i, P_j) \\ =& \sum_{i,j=1}^{n-1} c'_i c'_j d^{2H}(P_i, P_j) + 2 \sum_{i=1}^{n-1} c'_i c'_n d^{2H}(P_i, P_n) \\ &+ 2 \sum_{i=1}^{n-1} c'_i c'_{n+1} d^{2H}(P_i, P_{n+1}) + 2 c'_n c'_{n+1} d^{2H}(P_n, P_{n+1}). \end{aligned}$$

We now use

$$\begin{aligned} d^{2H}(P_i, P_{n+1}) =& \left(W_2(\mu_i, \mu_n)^2 + \varepsilon^2\right)^H \\ =& W_2(\mu_i, \mu_n)^{2H} + O\left(\varepsilon^2\right) \end{aligned}$$

to obtain

$$\begin{aligned} &\sum_{i,j=1}^{n+1} c'_i c'_j d^{2H}(P_i, P_j) \\ =& \sum_{i,j=1}^{n-1} c_i c_j W_2^{2H}(\mu_i, \mu_j) + 2 \sum_{i=1}^{n-1} c_i \frac{c_n}{2} W_2^{2H}(\mu_i, \mu_n) \\ &+ 2 \sum_{i=1}^{n-1} c_i \frac{c_n}{2} W_2^{2H}(\mu_i, \mu_n) + \frac{c_n^2}{2} \varepsilon^{2H} + O\left(\varepsilon^2\right) \\ =& \sum_{i,j=1}^{n-1} c_i c_j W_2^{2H}(\mu_i, \mu_j) + 2 \sum_{i=1}^{n-1} c_i c_n W_2^{2H}(\mu_i, \mu_n) \\ &+ \frac{c_n^2}{2} \varepsilon^{2H} + O\left(\varepsilon^2\right) \\ =& \sum_{i,j=1}^{n} c_i c_j W_2^{2H}(\mu_i, \mu_j) + \frac{c_n^2}{2} \varepsilon^{2H} + O\left(\varepsilon^2\right). \end{aligned}$$

Finally using (25) and $H < 1$ we obtain

$$\sum_{i,j=1}^{n+1} c'_i c'_j d^{2H}(P_i, P_j) = \frac{c_n^2}{2} \varepsilon^{2H} + o\left(\varepsilon^{2H}\right),$$

which is positive for ε small enough. This shows that d^{2H} is not negative definite, in contradiction with Lemma A.1. In the end W_2^{2H} is nondegenerate for every $0 < H < 1$.

We now use the same argument as in the end of the proof of Theorem IV.3. Since $W_2^{2H}(\delta_x, \delta_y) = |x-y|^{2H}$ and $|x-y|^2$ and $|x-y|^0$ are degenerate kernels on $\mathbb{R}$, W_2^0 and W_2^2 are degenerate kernels.

□

Proof of Lemma A.1. For $H = 1$ we have

$$d^2((\mu, s), (\nu, t)) = W_2(\mu, \nu)^2 + |s-t|^2$$

hence d^2 is negative definite as the sum of two negative definite kernels. From Lemma IV.4 we get that d^{2H} is a negative definite kernel for every $0 \leq H \leq 1$.

For $H > 1$ we notice that $d^{2H}(\mu, x)(\mu, y) = |x-y|^{2H}$ and use again the fact that $|x-y|^{2H}$ is not a negative definite kernel to conclude that d^{2H} is not negative definite. □

Proof of Theorem IV.1. The fact that (8) are covariance kernels is a direct consequence of Theorem IV.3 and the following Schoenberg Theorem (which is proven in [32]):

Theorem A.2 (Schoenberg)**.** *Given a set X, two functions $K, R : X \times X \to \mathbb{R}$, and $o \in X$ such that for every $x, y \in X$,*

$$K(x, x) = 0$$

and

$$R(x, y) = K(x, o) + K(y, o) - K(x, y),$$

the function R is a positive definite kernel if and only if K is a negative definite kernel.

We now prove the degeneracy: let $X = (X(\mu))_{\mu \in \mathcal{W}_2(\mathbb{R})}$ denote the H-fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$ with origin in σ. Assume X is degenerate: there exist $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$ and $\mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R})$ such that

$$\sum_{i=1}^{n} \lambda_n X(\mu_n) = 0 \text{ almost surely.}$$

Since $X(\sigma) = 0$ almost surely, setting $\mu_{n+1} = \sigma$ and $\lambda_{n+1} = -\sum_{i=1}^n \lambda_i$, it is clear that

$$\sum_{i=1}^{n+1} \lambda_n X(\mu_n) = 0 \text{ almost surely,}$$

which implies

$$\sum_{i,j=1}^{n+1} \lambda_i \lambda_j W_2^{2H}(\mu_i, \mu_j) = \mathbb{E}\left(\sum_{i=1}^{n+1} \lambda_n X(\mu_n)\right)^2 = 0.$$

Since $\sum_{i=1}^{n+1} \lambda_i = 0$ this shows that W_2^{2H} is degenerate, in contradiction with Theorem IV.3. Therefore X is nondegenerate for every $0 < H < 1$.

The degeneracy of the 0-fractional and the 2-fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$ is a direct consequence from the degeneracy of W_2^0 and W_2^2.

□

Proof of Theorem IV.2. The fact that (12) are covariance kernels is a direct consequence of Theorem IV.3 and the following Schoenberg Theorem (which proof can be found in [32]):

Theorem A.3 (Schoenberg)**.** *Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a completely monotone function, and K a negative definite kernel. Then $(x, y) \mapsto F(K(x, y))$ is a positive definite kernel.*

Furthermore as a function of the distance W_2, (12) is obviously invariant under the action of any isometry of $\mathcal{W}_2(\mathbb{R})$, so that the second claim holds. □

B. Proofs for Section V-B

Proof of Theorem V.9. We have $\hat{\theta}_{ML} \in \operatorname{argmin} L_\theta$ with

$$L_\theta = \frac{1}{n}\ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y.$$

From Lemma A.5 we have that

$$\sup_{\theta\in\Theta} \lambda_{\max}(R_\theta) \quad \text{and} \quad \sup_{\theta\in\Theta} \max_{i=1,\cdots,p} \lambda_{\max}\left(\frac{\partial}{\partial\theta_i} R_\theta\right)$$

are bounded as $n \to \infty$. Hence we can proceed as in the beginning of the proof of Proposition 3.1 in [41] to obtain

$$\sup_{\theta\in\Theta} \|L_\theta - \mathbb{E}(L_\theta)\| = o_{\mathbb{P}}(1). \tag{26}$$

Following again the proof of Proposition 3.1 in [41] we obtain the existence of a positive a such that

$$\mathbb{E}(L_\theta) - \mathbb{E}(L_{\theta_0}) \geq a|R_\theta - R_{\theta_0}|^2,$$

with $|\Lambda|^2 = (1/n)\sum_{i,j=1}^n \Lambda_{i,j}^2$.

Hence from Condition V.5 and (26) we have $\forall \alpha > 0$,

$$\mathbb{P}\left(\left\|\hat{\theta}_{ML} - \theta_0\right\| \geq \alpha\right) \underset{n\to\infty}{\longrightarrow} 0$$

and so

$$\hat{\theta}_{ML} \xrightarrow[n\to\infty]{\mathbb{P}} \theta_0.$$

□

Proof of Theorem V.10. From Lemma A.5 and Condition V.4 we have for every $n \in \mathbb{N}$, $\left|(M_{ML})_{i,j}\right| \leq B$ for a fixed $B < \infty$.

In addition, for any $\lambda_1, \cdots, \lambda_p \in \mathbb{R}$ such that $\sum_{i=1}^p \lambda_i^2 = 1$,

$$\begin{aligned}
&\sum_{i,j=1}^p \lambda_i \lambda_j \left(M_{ML}\right)_{i,j} \\
&= \frac{1}{2n} Tr\left(R_{\theta_0}^{-1}\left(\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\theta_i}\right) R_{\theta_0}^{-1}\left(\sum_{j=1}^p \lambda_j \frac{\partial R_{\theta_0}}{\theta_j}\right)\right) \\
&= \frac{1}{2}\left|R_{\theta_0}^{-1/2}\left(\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\partial\theta_i}\right) R_{\theta_0}^{-1/2}\right|^2 \\
&\geq C^2 \left|\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\partial\theta_i}\right|^2
\end{aligned}$$

with a fixed $C > 0$, since for every n

$$\lambda_{\min}\left(R_{\theta_0}^{-1}\right) = \frac{1}{\lambda_{\max}(R_{\theta_0})} \geq C > 0$$

from Lemma A.5. Hence from Condition V.8 we obtain

$$\liminf_{n\to\infty} \lambda_{\min}(M_{ML}) > 0.$$

Hence (22) is proved. Let us now assume that

$$\sqrt{n} M_{ML}^{1/2}\left(\hat{\theta}_{ML} - \theta_0\right) \cancel{\xrightarrow[n\to\infty]{\mathcal{L}}} \mathcal{N}(0, I_n). \tag{27}$$

Then there exists a bounded measurable function $g : \mathbb{R}^p \to \mathbb{R}$, $\xi > 0$ and a subsequence n' such that along n' we have

$$\left|\mathbb{E}\left[g\left(\sqrt{n} M_{ML}^{1/2}(\hat{\theta}_{ML} - \theta_0)\right)\right] - \mathbb{E}(g(U))\right| \geq \xi,$$

with $U \sim \mathcal{N}(0, I_p)$.

In addition, by compactness, up to extracting another subsequence we can assume that

$$M_{ML} \underset{n\to\infty}{\to} M_\infty,$$

where M_∞ is a symmetric positive definite matrix.

Now the remaining of the proof is similar to the proof of Proposition 3.2 in [41]. We have

$$\frac{\partial}{\partial\theta_i} L_\theta = \frac{1}{n}\left(Tr\left(R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta_i}\right) - y^t R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta_i} R_\theta^{-1} y\right).$$

Hence, exactly as in the proof of Proposition D.9 in [41] we can show

$$\sqrt{n}\frac{\partial}{\partial\theta_i} L_{\theta_0} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 4M_\infty).$$

Let us compute

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i\partial\theta_j} L_{\theta_0} &= \frac{1}{n} Tr\left(-R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial\theta_i} R_{\theta_0}^{-1}\frac{R_{\theta_0}}{\partial\theta_j} + R_{\theta_0}^{-1}\frac{\partial^2 R_{\theta_0}}{\partial\theta_i\partial\theta_j}\right) \\
&+ \frac{1}{n} y^t\left(2R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial\theta_i} R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial\theta_j} R_{\theta_0}^{-1} - R_{\theta_0}^{-1}\frac{\partial^2 R_{\theta_0}}{\partial\theta_i\partial\theta_j} R_{\theta_0}^{-1}\right) y.
\end{aligned}$$

We have

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} L_{\theta_0}\right) = 2M_{ML},$$

and from Condition V.4 and Lemma A.6,

$$\operatorname{Var}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} L_{\theta_0}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Hence

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} L_{\theta_0} \xrightarrow[n\to\infty]{\mathbb{P}} 2M_\infty.$$

Moreover, $\dfrac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} L_\theta$ can be written as

$$\frac{1}{n} Tr(A_\theta) + \frac{1}{n} y^t B_\theta y,$$

where A_θ and B_θ are sums of products of the matrices R_θ^{-1} or $\dfrac{\partial}{\partial\theta_{i_1}} \cdots \dfrac{\partial}{\partial\theta_{i_q}} R_\theta$ with $q \in \{0, \cdots, 3\}$ and $i_1, \cdots, i_q \in \{1, \cdots p\}$.

Hence from Condition V.4 and from Lemmas A.5 and A.6 we have

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_j} L_\theta\right\| = O_{\mathbb{P}}(1).$$

Following exactly the proof of Proposition D.10 in [41] we can show that

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow[n'\to\infty]{\mathcal{L}} \mathcal{N}(0, M_\infty^{-1}).$$

Moreover since $M_{ML} \underset{n\to\infty}{\to} M_\infty$ we have

$$\sqrt{n} M_{ML}^{1/2}(\hat{\theta}_{ML} - \theta_0) \xrightarrow[n'\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p).$$

This is in contradiction with (27) and concludes the proof.

□

Proof of Theorem V.11. From Theorem V.9 it is enough to show for $i = 1, \cdots, p$ that

$$\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\hat{Y}_\theta(\mu)\right| = O_{\mathbb{P}}(1).$$

From a version of Sobolev embedding theorem (see Theorem 4.12, part I, case A in [59]), there exists a finite constant A_Θ depending only on Θ such that

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\hat{Y}_\theta(\mu)\right| &\le A_\Theta\int_\Theta\left|\frac{\partial}{\partial\theta_i}\hat{Y}_\theta(\mu)\right|^{p+1}d\theta \\ &+ A_\Theta\sum_{j=1}^{p}\int_\Theta\left|\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}\hat{Y}_\theta(\mu)\right|^{p+1}d\theta.\end{aligned}$$

Therefore in order to prove the theorem it is sufficient to show that for $w_\theta(\mu)$ of the form $r_\theta(\mu)$ or $\frac{\partial}{\partial\theta_i}r_\theta(\mu)$ or $\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}r_\theta(\mu)$, and for W_θ equal to a product of the matrices R_θ^{-1} or $\frac{\partial}{\partial\theta_i}R_\theta$ or $\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}R_\theta$, we have

$$\int_\Theta\left|w_\theta^t(\mu)W_\theta y\right|^{p+1}d\theta = O_{\mathbb{P}}(1).$$

From Fubini theorem for positive integrands we have

$$\mathbb{E}\left[\int_\Theta\left|w_\theta^t(\mu)W_\theta y\right|^{p+1}d\theta\right] = \int_\Theta\mathbb{E}\left(\left|w_\theta^t(\mu)W_\theta y\right|^{p+1}\right)d\theta.$$

Now there exists a constant c_{p+1} so that for X a centred Gaussian random variable,

$$\mathbb{E}\left(|X|^{p+1}\right) = c_{p+1}\left(\mathrm{Var}(X)\right)^{(p+1)/2},$$

hence

$$\begin{aligned}&\mathbb{E}\left(\int_\Theta\left|w_\theta^t(\mu)W_\theta y\right|^{p+1}d\theta\right)\\ &= c_{p+1}\int_\Theta\left(\mathrm{Var}\left(w_\theta^t(\mu)W_\theta y\right)\right)^{(p+1)/2}d\theta\\ &= c_{p+1}\int_\Theta\left(w_\theta^t(\mu)W_\theta R_{\theta_0}W_\theta^t w_\theta(\mu)\right)^{(p+1)/2}d\theta.\end{aligned}$$

Now from Lemmas A.5 and A.6 there exists $B < \infty$ such that

$$\sup_{\theta\in\Theta}\lambda_{\max}\left(W_\theta R_{\theta_0}W_\theta\right) \le B.$$

Thus

$$\mathbb{E}\left(\int_\Theta\left|w_\theta^t W_\theta y\right|^{p+1}d\theta\right) \le B^{(p+1)/2}c_{p+1}\int_\Theta\left\|w_\theta^t(\mu)\right\|^{(p+1)/2}d\theta.$$

Finally for some $q \in \{0,1,2\}$ and for $i_1, \cdots, i_q \in \{1, \cdots p\}$ we have

$$\begin{aligned}\sup_{\theta\in\Theta}\left\|w_\theta^t(\mu)\right\|^2 &= \sup_{\theta\in\Theta}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta_{i_1}}\cdots\frac{\partial}{\partial\theta_{i_q}}F_\theta(W_2(\mu,\mu_i))\right)^2\\ &\le C\sum_{i=1}^{n}\left|\frac{1}{1+W_2(\mu,\mu_i)^{1+\tau}}\right|,\end{aligned}$$

with $C < \infty$ coming from Condition V.2, V.6, and V.7.

Using the proof of Lemma A.4 we see that this quantity is bounded, which finishes the proof of Theorem V.11. □

C. Technical lemmas for Section V-B

Lemma A.4.

$$\sup_{\mu\in W_2(\mathbb{R})}\sup_{\theta\in\Theta}\sum_{j=1}^{n}|K_\theta(\mu,\mu_j)|$$

is bounded as $n \to \infty$.

Proof. Let $\mu \in \mathcal{W}_2(\mathbb{R})$ and $i^* \in \mathrm{argmin}_{k\in\{1,\cdots n\}}W_2(\mu_k,\mu)$. For every $j \in \{1, \cdots, n\}$, $W_2(\mu,\mu_j) \ge W_2(\mu,\mu_{i^*})$. Moreover from the triangle inequality we have

$$W_2(\mu,\mu_j) \ge W_2(\mu_j,\mu_{i^*}) - W_2(\mu_{i^*},\mu),$$

hence

$$W_2(\mu,\mu_j) \ge \frac{W_2(\mu_j,\mu_{i^*})}{2}.$$

Let us define

$$r_\mu := \sup_{\theta\in\Theta}\sum_{i=1}^{n}F_\theta(W_2(\mu_i,\mu))$$

From Condition V.2 we have

$$r_\mu \le \sum_{i=1}^{n}\frac{A}{1+W_2(\mu_i,\mu)^{1+\tau}} \le \sum_{i=1}^{n}\frac{A}{1+\left(\frac{W_2(\mu_j,\mu_{i^*})}{2}\right)^{1+\tau}}.$$

Now

$$W_2^2(\mu_j,\mu_{i^*}) = \int_0^1\left|q_{\mu_j}(t) - q_{\mu_{i^*}}(t)\right|^2dt,$$

where for every $t \in [0,1]$

$$q_\mu(t) = \inf\{x \in \mathbb{R}|\ F_\mu(x) \ge t\}.$$

Note that from Condition V.1 for every $t \in [0,1]$,

$$q_{\mu_i}(t) \in [i, i+L].$$

If $|j - i^*| \ge L$ we have

$$\forall t \in \mathbb{R},\ |q_{\mu_{i^*}}(t) - q_{\mu_j}(t)| \ge |j - i^*| - L$$

so that

$$W_2(\mu_{i^*},\mu_j) \ge |j - i^*| - L. \tag{28}$$

Hence

$$\begin{aligned}r_\mu &\le 2AL + \sum_{j,\ |j-i^*|\ge L}\frac{A}{1+\left(\frac{|j-i^*|-L}{2}\right)^{1+\tau}}\\ &\le 2AL + \sum_{j=-\infty}^{+\infty}\frac{A}{1+\left|\frac{j}{2}\right|^{1+\tau}} < \infty.\end{aligned}$$

□

Lemma A.5. *Under Conditions V.1 to V.4,*

$$\sup_{\theta\in\Theta}\lambda_{\max}(R_\theta)$$

and

$$\sup_{\theta\in\Theta}\max_{i=1\cdots p}\lambda_{\max}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)$$

are bounded as $n \to \infty$.

Proof.

$$\sup_{\theta\in\Theta}\lambda_{\max}(R_\theta)\leq\sup_{\theta\in\Theta}\max_{i=1,\ldots,n}\sum_{j=1}^{n}\left|F_\theta(W_2(\mu_i,\mu_j))\right|$$

is bounded as $n\to\infty$ from Lemma A.4. The proof is similar for

$$\sup_{\theta\in\Theta}\max_{i=1\cdots p}\lambda_{\max}\left(\frac{\partial}{\partial\theta_i}R_\theta\right).$$

□

In a similar way we also obtain the following Lemma.

Lemma A.6. $\forall q\in\{2,3\}$, $\forall i_1,\cdots,i_q\in\{1,\cdots p\}$,

$$\sup_{\theta\in\Theta}\lambda_{\max}\left(\frac{\partial}{\partial\theta_{i_1}}\cdots\frac{\partial}{\partial\theta_{i_q}}R_\theta\right)$$

is bounded as $n\to\infty$.

D. Proofs for Section V-C

Proposition A.7. *Under the setting of Proposition V.12, almost surely as $n\to\infty$,*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i,j=1}^{n}\left[K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right]^2\right.$$
$$\left.-\sum_{j=-\infty}^{\infty}\mathbb{E}\left(\left[K_\theta(\mu_0,\mu_j)-K_{\theta_0}(\mu_0,\mu_j)\right]^2\right)\right|\to 0$$

and the sum in the right-hand side of the above display is a continuous function of θ.

Proof. Let

$$S_\theta=\frac{1}{n}\sum_{i,j=1}^{n}\left[K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right]^2.$$

Let $(m_n)_{n\in\mathbb{N}}$ be a sequence of integers so that as $n\to\infty$, $m_n\to\infty$ and $n/m_n\to\infty$. Let

$$S_{\theta,m_n}=\frac{1}{n}\sum_{i,j=1}^{n}\mathbf{1}_{\{\lfloor\frac{i-1}{m_n}\rfloor=\lfloor\frac{j-1}{m_n}\rfloor\}}\left[K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right]^2.$$

With the same proof as that of Lemma D.11 in [41], we can show (using (28)) that $|S_\theta-S_{\theta,m_n}|$ goes almost surely to zero as $n\to\infty$. Also

$$S_{\theta,m_n}=\frac{1}{n/m_n}\sum_{k=0}^{\lfloor\frac{n}{m_n}\rfloor-1}\frac{1}{m_n}\sum_{i,j=1}^{m_n}\left[K_\theta(\mu_{km_n+i},\mu_{km_n+j})\right.$$
$$\left.-K_{\theta_0}(\mu_{km_n+i},\mu_{km_n+j})\right]^2$$
$$+\frac{1}{n}\sum_{i,j=m_n\left(\lfloor\frac{n}{m_n}\rfloor-1\right)+1}^{n}$$
$$\mathbf{1}_{\{\lfloor\frac{i-1}{m_n}\rfloor=\lfloor\frac{j-1}{m_n}\rfloor\}}\left[K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right]^2$$
$$=\frac{1}{n/m_n}\sum_{k=0}^{\lfloor\frac{n}{m_n}\rfloor-1}B_k+r,$$

say. From (28), one can show simply that $r\to 0$ almost surely as $n\to\infty$. Also, the B_k are independent random variables with identical distribution, and they are bounded in absolute value by

$$2L+1+\sum_{i=-\infty}^{\infty}2\left(\frac{A}{1+|i|^{1+\tau}}\right)^2<\infty$$

from (28) and Condition V.2. Hence, applying Theorem 2.1 in [60] yields

$$\left(\frac{1}{n/m_n}\sum_{k=0}^{\lfloor\frac{n}{m_n}\rfloor-1}B_k\right)-\mathbb{E}(B_0)\to^{a.s.}_{n\to\infty}0.$$

Hence, finally he have obtained almost surely as $n\to\infty$

$$\left|S_\theta-\frac{1}{m_n}\sum_{i,j=1}^{m_n}\mathbb{E}\left[\left(K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right)^2\right]\right|\to 0.$$

Also, we have, for $|i-j|\geq L$

$$\mathbb{E}\left[\left(K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right)^2\right]$$
$$\leq 2\left(\frac{A}{1+(|i-j|-L)^{1+\tau}}\right)^2$$

from (28). Hence, we can simply show

$$\left|\frac{1}{m_n}\sum_{i,j=1}^{m_n}\mathbb{E}\left[\left(K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right)^2\right]-T_\theta\right|\to^{a.s.}_{n\to\infty}0,$$

with

$$T_\theta=\sum_{j=-\infty}^{\infty}\mathbb{E}\left(\left[K_\theta(\mu_0,\mu_j)-K_{\theta_0}(\mu_0,\mu_j)\right]^2\right).$$

From (28) and Condition V.6, we can show that there exists a deterministic finite constant C so that

$$\sup_{\theta\in\Theta}\max_{i=1,\ldots,p}\left|\frac{\partial}{\partial\theta_i}S_\theta\right|\leq C.$$

Also, by dominated convergence T_θ is a continuously differentiable function of θ and

$$\sup_{\theta\in\Theta}\max_{i=1,\ldots,p}\left|\frac{\partial}{\partial\theta_i}T_\theta\right|\leq C'$$

where C' is also a deterministic finite constant. Hence $\sup_{\theta\in\Theta}|S_\theta-T_\theta|\to 0$ almost surely as $n\to\infty$.

□

Proof of Proposition V.12. Assume that

$$\liminf_{n\to\infty}\inf_{\|\theta-\theta_0\|\geq\alpha}\frac{1}{n}\sum_{i,j=1}^{n}\left[K_\theta(\mu_i,\mu_j)-K_{\theta_0}(\mu_i,\mu_j)\right]^2=0.$$

Then from Proposition A.7 and by compacity, there exists $\theta_1\neq\theta_0$ so that

$$\sum_{j=-\infty}^{\infty}\mathbb{E}\left(\left[K_{\theta_1}(\mu_0,\mu_j)-K_{\theta_0}(\mu_0,\mu_j)\right]^2\right)=0.$$

From the conditions on $\{F_\theta\}$, there exists $\beta>0$, $\delta>0$, $a\geq 0$ so that for $u\in[a-\delta,a+\delta]$ we have $|F_{\theta_1}(u)-F_{\theta_0}(u)|\geq\beta$. Hence, we have

$$\beta^2P(W_2(\mu_0,\mu_{k-1})\in[a-\delta,a+\delta])=0,$$

for k so that $a \in (k-1, k]$.

Let now $g_0 : [0, L] \to \mathbb{R}^+$ be defined by $g_0(u) = D_0 \exp(-1/(1-u^2)) \mathbf{1}_{\{u \in [-1,1]\}}$ where $0 < D_0 < \infty$ is so that $\int_{\mathbb{R}} g_0(u) du = 1$. Then, g_0 is infinitely differentiable. Let $h_0(u) = (1/\sigma) g_0((u - \delta/4)/\sigma)$ and $h_{k-1}(u) = (1/\sigma) g_0((u - a)/\sigma)$, where $\sigma > 0$ is chosen small enough so that, with ν_0 and ν_{k-1} the distributions with probability density functions h_0 and h_{k-1} we have $W_2(\nu_0, \nu_{k-1}) \in [a - \delta/2, a + \delta/2]$ and ν_0, ν_{k-1} have supports in $[0, L], [k-1, k-1+L]$.

Let now P_1, P_2 be two distributions with support in $[0, L]$, with quantile functions q_1, q_2, with cumulative distribution functions F_1, F_2 and with probability density functions f_1, f_2. Then we have

$$
\begin{aligned}
& W_2(P_1, P_2) \qquad (29) \\
& = \sqrt{\int_0^1 (q_1 - q_2)^2} \\
& \leq \sqrt{L} \sqrt{\int_0^1 |q_1 - q_2|} \\
& = \sqrt{L} \sqrt{\int_0^L |F_1 - F_2|} \\
& \leq L \sqrt{\sup_{u \in [0,L]} |F_1(u) - F_2(u)|} \\
& \leq L \sqrt{\int_0^L |f_1 - f_2|} \\
& \leq L^{3/2} \sqrt{\sup_{u \in [0,L]} |f_1(u) - f_2(u)|}.
\end{aligned}
$$

Let $\tau > 0$ be so that $L^{3/2} \tau^{1/2} \leq \delta/5$. Then, for any $f : [0, L] \to \mathbb{R}$ and $g : [k-1, k-1+L] \to \mathbb{R}$, we have that $|f/(\int_0^L f) - h_0|_\infty \leq \tau$ and $|g/(\int_0^L g) - h_{k-1}|_\infty \leq \tau$ imply $W_2(\nu_f, \nu_g) \in [a - \delta, a + \delta]$, where ν_f, ν_g are the measures with probability density functions f and g. Since h_0 and h_{k-1} are infinitely differentiable, have integral one, and have respective supports included in $[0, L]$ and $[k-1, k-1+L]$, it is easy to see that there exists $\epsilon > 0$ so that $|f - h_0|_\infty \leq \epsilon$ implies $|f/(\int_0^L f) - h_0|_\infty \leq \tau$, and similarly for g and h_{k-1}. Hence, if we can show that

$$
P(\sup_{u \in [0,L]} |h_0(u) - \exp(Z_0(u))| \leq \epsilon) > 0
$$

and

$$
P(\sup_{u \in [0,L]} |h_{k-1}(u + (k-1)) - \exp(Z_{k-1}(u))| \leq \epsilon) > 0,
$$

we obtain a contradiction. The two probabilities above are shown to be non-zero similarly and we will address the first one only. It is sufficient to show that

$$
P(\sup_{u \in [0,L]} |h_0(u) + \epsilon/2 - \exp(Z_0(u))| \leq \epsilon/2) > 0.
$$

Since $h_0 + \epsilon/2$ is continuous and bounded away from 0 and infinity on $[0, L]$, it is sufficient to show that for all $\kappa > 0$,

$$
P(\sup_{u \in [0,L]} |\log(h_0(u) + \epsilon/2) - Z_0(u)| \leq \kappa) > 0.
$$

From e.g. Theorem 1.1 in [61], since z_0 has mean function zero, we have

$$
P(\sup_{u \in [0,L]} |Z_0(u)| \leq \kappa) > 0.
$$

Consider now the Gaussian measures $\mathcal{G}_1$ and $\mathcal{G}_2$, on the space of continuous functions from $[0, L] \to \mathbb{R}$, so that $\mathcal{G}_1$ is the measure of the Gaussian process Z_0 and $\mathcal{G}_2$ is that of $Z_0 - \log(h_0 + \epsilon/2)$. Then, from e.g. the discussion in (22) in Chapter 4.2 of [33], since $\log(h_0 + \epsilon/2)$ is infinitely differentiable, and from the assumptions on the covariance function of Z_0, the Gaussian measures $\mathcal{G}_1$ and $\mathcal{G}_2$ are equivalent. Hence, since

$$
\mathcal{G}_1 \left(\{ f \text{ continuous} : [0, L] \to \mathbb{R}; |f|_\infty \leq \kappa \} \right) > 0,
$$

we also have

$$
\mathcal{G}_2 \left(\{ f \text{ continuous} : [0, L] \to \mathbb{R}; |f|_\infty \leq \kappa \} \right) > 0,
$$

which is exactly

$$
P(\sup_{u \in [0,L]} |\log(h_0(u) + \epsilon/2) - Z_0(u)| \leq \kappa) > 0.
$$

This concludes the proof that Condition V.5 holds.

The proof that Condition V.8 holds can be obtained in the same way. In particular, an analog of Proposition A.7 can be obtained. We skip the details. □

Acknowledgements

We thank the anonymous reviewers, whose suggestions have greatly contributed to improve the manuscript.

We thank Yann Richet, from the French Radioprotection and Nuclear Safety Institute (IRSN), for introducing us to the problem of axial burn up analysis of fuel pins, which motivated the present work.

References

[1] N. A. C. Cressie, *Statistics for spatial data*, ser. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991, a Wiley-Interscience Publication.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

[3] V. Vapnik, S. E. Golowich, A. Smola *et al.*, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in neural information processing systems*, pp. 281–287, 1997.

[4] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[5] N. Cristianini and J. Shawe-Taylor, "Support vector machines," 2000.

[6] S. Cohen and M. A. Lifshits, "Stationary Gaussian random fields on hyperbolic spaces and on Euclidean spheres," *ESAIM Probab. Stat.*, vol. 16, pp. 165–221, 2012. [Online]. Available: http://dx.doi.org/10.1051/ps/2011105

[7] J. Istas, "Manifold indexed fractional fields," *ESAIM Probab. Stat.*, vol. 16, pp. 222–276, 2012. [Online]. Available: http://dx.doi.org/10.1051/ps/2011106

[8] A. Feragen, F. Lauze, and S. Hauberg, "Geodesic exponential kernels: When curvature and linearity conflict," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3032–3042.

[9] S. R. Flaxman, Y.-X. Wang, and A. J. Smola, "Who supported obama in 2012?: Ecological inference through distribution regression," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 289–298.

[10] D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin, "Towards a learning theory of cause-effect inference," in *International Conference on Machine Learning*, 2015, pp. 1452–1461.

[11] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf *et al.*, “Kernel mean embedding of distributions: A review and beyond,” *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[12] B. Póczos, L. Xiong, and J. Schneider, “Nonparametric divergence estimation with applications to machine learning on distributions,” *arXiv preprint arXiv:1202.3758*, 2012.

[13] B. Póczos, A. Singh, A. Rinaldo, and L. A. Wasserman, “Distribution-free distribution regression.” in *AISTATS*, 2013, pp. 507–515.

[14] D. J. S. J. S. Barnabas and L. X. Poczos, “Nonparametric kernel estimators for image classification,” in *CVPR*, vol. 2012, 2012, p. 1.

[15] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur, “Two-stage sampled learning theory on distributions,” in *Artificial Intelligence and Statistics*, 2015, pp. 948–957.

[16] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, “Kernel Mean Embedding of Distributions: A Review and Beyonds,” *ArXiv e-prints*, May 2016.

[17] S. Kolouri, Y. Zou, and G. K. Rohde, “Sliced wasserstein kernels for probability distributions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5258–5267.

[18] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2009, vol. 338.

[19] A. Munk and C. Czado, “Nonparametric validation of similar distributions and assessment of goodness of fit,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 60, no. 1, pp. 223–241, 1998. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00121

[20] E. Boissard, T. Le Gouic, and J.-M. Loubes, “Distribution’s template estimate with Wasserstein metrics,” *Bernoulli*, vol. 21, no. 2, pp. 740–759, 2015. [Online]. Available: http://dx.doi.org/10.3150/13-BEJ585

[21] T. Le Gouic and J.-M. Loubes, “Existence and consistency of Wasserstein barycenters,” *Probability Theory and Related Fields*, pp. 1–17, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00440-016-0727-z

[22] G. Peyré, M. Cuturi, and J. Solomon, “Gromov-Wasserstein averaging of kernel and distance matrices,” in *ICML 2016*, 2016.

[23] B. Póczos, A. Singh, A. Rinaldo, and L. Wasserman, “Distribution-free distribution regression,” in *In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, volume 31 of JMLR Proceedings*, 2013, pp. 507–515.

[24] N. Venet, F. Bachoc, F. Gamboa, and J.-M. Loubes, “Modèles de régression gaussienne pour des distributions en entrée,” *49è Journées de statistique*, 2016. [Online]. Available: http://www.math.univ-toulouse.fr/~nvenet/pdf/Venet_Bachoc_Gamboa_Loubes_Communication_SFDS2017.pdf

[25] G. Radulescu, D. E. Mueller, and J. C. Wagner, “Sensitivity and uncertainty analysis of commercial reactor criticals for burnup credit,” *Nuclear Technology*, vol. 167, no. 2, pp. 268–287, 2009.

[26] R. Cacciapouti, “Axial burnup profile database for pressurized water reactors.” Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), Tech. Rep., 2000.

[27] S. M. Bowman, D. F. Hollenbach, M. D. DeHART, B. T. Rearden, I. C. Gauld, and S. Goluoglu, “Scale 5: Powerful new criticality safety analysis tools,” *Nuclear Science and Technology*, 2003.

[28] S. T. Rachev, “Monge-kantorovich problem on mass transfer and its applications in stochastics,” *Teoriya Veroyatnostei i ee Primeneniya*, vol. 29, no. 4, pp. 625–653, 1984.

[29] M. Lifshits, “Lectures on gaussian processes,” in *Lectures on Gaussian Processes*. Springer, 2012, pp. 1–117.

[30] B. Kloeckner, “A geometric study of wasserstein spaces: Euclidean spaces,” *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze-Serie V*, vol. 9, no. 2, p. 297, 2010.

[31] B. B. Mandelbrot and J. W. Van Ness, “Fractional brownian motions, fractional noises and applications,” *SIAM review*, vol. 10, no. 4, pp. 422–437, 1968.

[32] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups*. Springer-Verlag, 1984.

[33] M. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

[34] F. Bachoc, “Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification,” *Computational Statistics and Data Analysis*, vol. 66, pp. 55–69, 2013.

[35] ——, “Asymptotic analysis of covariance parameter estimation for gaussian processes in the misspecified case,” *Bernoulli, forthcoming*, 2016.

[36] H. Zhang and Y. Wang, “Kriging and cross validation for massive spatial data,” *Environmetrics*, vol. 21, pp. 290–304, 2010.

[37] K. Mardia and R. Marshall, “Maximum likelihood estimation of models for residual covariance in spatial regression,” *Biometrika*, vol. 71, pp. 135–146, 1984.

[38] N. Cressie and S. Lahiri, “The asymptotic distribution of REML estimators,” *Journal of Multivariate Analysis*, vol. 45, pp. 217–233, 1993.

[39] ——, “Asymptotics for REML estimation of spatial covariance parameters,” *Journal of Statistical Planning and Inference*, vol. 50, pp. 327–341, 1996.

[40] B. A. Shaby and D. Ruppert, “Tapered covariance: Bayesian estimation and asymptotics,” *Journal of Computational and Graphical Statistics*, vol. 21, no. 2, pp. 433–452, 2012.

[41] F. Bachoc, “Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes,” *Journal of Multivariate Analysis*, vol. 125, pp. 1–35, 2014.

[42] R. Furrer, F. Bachoc, and J. Du, “Asymptotic properties of multivariate tapering for estimation and prediction,” *Journal of Multivariate Analysis*, vol. 149, pp. 177–191, 2016.

[43] H. Zhang, “Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics,” *Journal of the American Statistical Association*, vol. 99, pp. 250–261, 2004.

[44] M. Stein, “Asymptotically efficient prediction of a random field with a misspecified covariance function,” *The Annals of Statistics*, vol. 16, pp. 55–63, 1988.

[45] ——, “Bounds on the efficiency of linear predictions using an incorrect covariance function,” *The Annals of Statistics*, vol. 18, pp. 1116–1138, 1990.

[46] ——, “Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure,” *The Annals of Statistics*, vol. 18, pp. 850–872, 1990.

[47] H. Putter and G. A. Young, “On the effect of covariance function estimation on the accuracy of Kriging predictors,” *Bernoulli*, vol. 7, no. 3, pp. 421–438, 2001.

[48] Z. Ying, “Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process,” *Journal of Multivariate Analysis*, vol. 36, pp. 280–296, 1991.

[49] ——, “Maximum likelihood estimation of parameters under a spatial sampling scheme,” *The Annals of Statistics*, vol. 21, pp. 1567–1590, 1993.

[50] H.-S. Chen, D. Simpson, and Z. Ying, “Infill asymptotics for a stochastic process model with measurement error,” *Statistica Sinica*, vol. 10, pp. 141–156, 2000.

[51] W. Loh and T. Lam, “Estimating structured correlation matrices in smooth Gaussian random field models,” *The Annals of Statistics*, vol. 28, pp. 880–904, 2000.

[52] W.-L. Loh, “Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields,” *The Annals of Statistics*, vol. 33, pp. 2344–2394, 2005.

[53] F. Bachoc and R. Furrer, “On the smallest eigenvalues of covariance matrices of multivariate spatial processes,” *Stat*, 2016.

[54] O. Roustant, D. Ginsbourger, and Y. Deville, “DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization,” *Journal of Statistical Software*, vol. 51, no. 1, pp. 1–55, 2012.

[55] T. Muehlenstaedt, J. Fruth, and O. Roustant, “Computer experiments with functional inputs and scalar outputs by a norm-based approach,” *Statistics and Computing*, pp. 1–15, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11222-016-9672-z

[56] S. Nanty, C. Helbert, A. Marrel, N. Pérot, and C. Prieur, “Sampling, metamodeling, and sensitivity analysis of numerical simulators with functional stochastic inputs,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 636–659, 2016.

[57] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer, 2005, vol. 338.

[58] R. Furrer, M. G. Genton, and D. Nychka, “Covariance tapering for interpolation of large spatial datasets,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 502–523, 2006.

[59] R. Adams and J. Fournier, *Sobolev spaces*. Academic Press, Amsterdam, 2003.

[60] T.-C. Hu and R. Taylor, “On the strong law for arrays and for the bootstrap mean and variance.” *International Journal of Mathematics and Mathematical Sciences*, vol. 20, no. 2, pp. 375–382, 1997. [Online]. Available: http://eudml.org/doc/47832

[61] W. V. Li, W. Linde *et al.*, “Approximation, metric entropy and small ball estimates for gaussian measures,” *The Annals of Probability*, vol. 27, no. 3, pp. 1556–1578, 1999.