Gaussian Processes on Distributions based on Regularized Optimal Transport

François Bachoc **Louis Béthune**
IMT, Université Paul sabatier

Alberto Gonzalez-Sanz **Jean-Michel Loubes**
IRIT, Université Paul sabatier

Abstract

We present a novel kernel over the space of probability measures based on the dual formulation of optimal regularized transport. We propose an Hilbertian embedding of the space of probabilities using their Sinkhorn potentials, which are solutions of the dual entropic relaxed optimal transport between the probabilities and a reference measure $\mathcal{U}$. We prove that this construction enables to obtain a valid kernel, by using the Hilbert norms. We prove that the kernel enjoys theoretical properties such as universality and some invariances, while still being computationally feasible. Moreover we provide theoretical guarantees on the behaviour of a Gaussian process based on this kernel. The empirical performances are compared with other traditional choices of kernels for processes indexed on distributions.

1 INTRODUCTION

Context: Gaussian Processes and Kernels Indexed by Distributions. Gaussian process (GP) models are widely used in fields such as geostatistics, computer code experiments and machine learning. We refer to Rasmussen and Williams (2006) for general references. They consist in modeling an unknown function as a realization of a GP, and hence correspond to a functional Bayesian framework. For instance, in computer experiments, the input points of the function are simulation parameters and the output values are quantities of interest obtained from the simulations. GPs rely on the definition of a covariance function that characterises the correlations between the values of the process at different observation points.
In this paper we consider GPs indexed by distributions. Learning functions defined on distributions has gained a special interest over the last decade in the machine learning literature, see for instance Póczos et al. (2013). Distribution-valued inputs are commonly used to describe complex objects such as images, shapes or media as described for instance in Glaunes et al. (2004), Muandet et al. (2012), Ginsbourger et al. (2016) or Szabó et al. (2016). The construction of a kernel for these inputs requires a notion of similitude between the probability distributions. Many methods have been considered to provide kernels for distributions, from the mere extraction of parametric features, such as the mean or higher moments, to the well used Maximum Mean Discrepancy (MMD) method (Gretton et al., 2012).

Context: Optimal Transport for Kernels. On this topic, optimal transport (OT) has imposed itself has a prominent method for comparing or analyzing distributions. Previous works in this direction are, in one dimension, Bachoc et al. (2017) and Thi Thien Trang et al. (2021), where a kernel directly based on the quadratic difference between the quantiles, which yields (on the real line) the quadratic Wasserstein distance, is proposed. In several dimensions, a quite natural generalisation uses the quadratic norm between the multidimensional transport maps between the probabilities and a reference measure, see Bachoc et al. (2020) and Moosmüller and Cloninger (2020). Even though, with a good choice of the reference measure, the generated kernels are translation invariant (see del Barrio et al. (2020); del Barrio et al. (2022b); Hallin et al. (2021)), for machine learning purposes the computation of the transport map (between continuous measures) is rather complicated and depends highly on the dimension, see Peyré et al. (2019). Moreover, even if the transport maps exist almost surely (see McCann (1995)) for a suitable choice of the reference distribution, their continuity—and therefore their approximations from empirical settings–require at least upper-lower bounded densities and the convex supports of the target measures, see Figalli (2017). Regarding GPs based on OT, the high complexity of the transport problem makes it so that their continuity properties are not much studied especially in multi-dimension. As a simplification, Kolouri et al. (2018) proposed the use of a slice-Wasserstein kernel. The idea is to reduce the problem by projecting on the

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

different directions generated by a (uniform) discretization of the unit sphere $\mathbb{S}_{d-1}$, and then integrating w.r.t. the uniform measure on $\mathbb{S}_{d-1}$. This avoids completely the curse of the dimensionality, but this does not discriminate quite well between non convex domains. The corresponding universality properties are studied in the recent work of Meunier et al. (2022).

Contributions. While the initial formulations of OT yield computational challenges, subsequent regularized versions have provided valuable trade-offs between richness and tractability. In this work, we provide a kernel based on regularized OT. We prove that the norm between the potentials derived from entropy relaxation of Wasserstein distances, see Cuturi (2013), provides a natural embedding for the distributions and can be used to construct a valid kernel. Much work on the properties of the potentials has been carried out in particular del Barrio et al. (2022a); Gonzalez-Sanz et al. (2022); Mena and Niles-Weed (2019) but few results exist taking advantage of the natural embedding the potentials provide. **1)** Our contribution is first to propose a novel valid and universal kernel based on Sinkhorn's dual potentials when considering the regularized transport towards a reference measure. **2)** We then propose statistical guarantees for this kernel by studying the properties of its empirical counterpart as well as invariance properties w.r.t. the choice of the reference. **3)** We study the theoretical properties of the corresponding GP, especially the existence of a continuous version. Feasible computations through Sinkhorn's algorithm enable to study the prediction performance of the kernel. **4)** We provide publically available code, together with simulations and real datasets where our kernel competes favorably with state of the art methods: it yields a similar accuracy as them, while being computationally efficient, in particular providing a computational speed-up of order up to 100 compared to the MMD kernel.

Outline. Section 2 is devoted to providing some definitions and notations related to Sinkhorn's transport methods. In Section 3 we define and study the kernel based on the potentials, while Section 4 studies the GP with this kernel as covariance operator. Implementation and experiments are discussed in Section 5. The proofs and complementary content are postponed to the Appendix.

2 DEFINITIONS AND BASIC PROPERTIES OF SINKHORN DISTANCE

2.1 General Definitions and Notations

We let $\mathcal{P}(A)$ be the set of probability measures on a general set $A \subset \mathbb{R}^d$. When A is compact and for $s > 0$, we let $\mathcal{C}^s(A)$ be the space of functions $f : A \to \mathbb{R}$ that are $\lfloor s \rfloor$ times differentiable, with $\lfloor . \rfloor$ the integer part, with $\|f\|_{\mathcal{C}^s(A)} < \infty$ where

$$\|f\|_{\mathcal{C}^s(A)} := \sum_{i=0}^{\lfloor s \rfloor} \sum_{|\boldsymbol{\alpha}|=i} \|D^{\boldsymbol{\alpha}} f\|_\infty. \tag{1}$$

Above $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ with $\sum_{j=1}^d \alpha_j = i$ and $D^{\boldsymbol{\alpha}} = \partial^i / \partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d}$. The space $\mathcal{C}^s(A)$ is endowed with the norm $\|\cdot\|_{\mathcal{C}^s(A)}$. A probability $\mathrm{P} \in \mathcal{P}(A)$ belongs also to the topological dual space of $\mathcal{C}^s(A)$. A distance between two measures $\mathrm{P}, \mathrm{Q} \in \mathcal{P}(A)$ can be defined as

$$\|\mathrm{P} - \mathrm{Q}\|_s := \sup_{f \in \mathcal{C}^s(A),\ \|f\|_{\mathcal{C}^s(A)} \leq 1} \int f(\mathbf{x})(d\mathrm{P}(\mathbf{x}) - d\mathrm{Q}(\mathbf{x})). \tag{2}$$

We let ℓ_d be d-dimensional Lebesgue measure. For $p > 0$ and for $\mathrm{P} \in \mathcal{P}(A)$, we let $L^p(\mathrm{P})$ be the set of functions $f : A \to \mathbb{R}$ such that $\|f\|^p_{L^p(\mathrm{P})} := \int_A |f(\mathbf{x})|^p d\mathrm{P}(\mathbf{x}) < \infty$.

We use the abbreviations "a.s." for "almost surely" and "a.e." for "almost everywhere". For a probability measure P on A, we let $\mathrm{supp}(\mathrm{P})$ be its topological support (the smallest closed set with P-probability one). For two sets A and B, for a probability measure P on A, and for $T : A \to B$, we let $T\sharp A$ be the probability measure of $T(\boldsymbol{X})$ where $\boldsymbol{X}$ is a random vector with law P. For two probability distributions P and Q on A, we write $\mathrm{P} \ll \mathrm{Q}$ when P is absolutely continuous w.r.t. Q and in this case we write $d\mathrm{P}/d\mathrm{Q}$ for the density of P w.r.t. Q. A random vector $\mathbf{V}$ on $A \subset \mathbb{R}^d$ is said to be sub-Gaussian if there is $\sigma^2 < \infty$ such that $\mathbb{E}(\exp(s\mathbf{u}^\top \mathbf{V})) \leq \exp(\sigma^2 s^2/2)$ for any $\mathbf{u} \in \mathbb{R}^d$, $\|\mathbf{u}\| = 1$ and $s \in \mathbb{R}$. We let $\mathcal{P}_{SG}(A)$ be the set of sub-Gaussian probability measures on A. For $\mathbf{x} \in A$ we let $\delta_{\mathbf{x}}$ be the Dirac probability measure at $\mathbf{x}$.

For a set E, a function $k : E \times E \to \mathbb{R}$ is said to be positive definite when for any $x_1, \ldots, x_n \in E$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$. The function is said to be strictly positive definite if in addition the sum is strictly positive when $x_1, \ldots, x_n$ are two-by-two distinct and not all $\alpha_1, \ldots, \alpha_n$ are zero.

For $A \subset \mathbb{R}^d$ we let $\mathrm{diam}(A) = \sup\{\|\mathbf{x} - \mathbf{y}\|; \mathbf{x}, \mathbf{y} \in A\}$. For $t \in \mathbb{R}$, we let $\lceil t \rceil$ be the smallest integer larger or equal to t. For two column vectors $\mathbf{x}, \mathbf{y}$, we let $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ be their scalar product.

2.2 Regularized Optimal Transport

We consider an input space $\Omega \subset \mathbb{R}^d$ that is fixed throughout the paper. For some of the results of the paper, Ω will be assumed to be compact, while for others, we can make the weaker assumption to consider sub-Gaussian measures on Ω (that is not necessarily bounded). Let P, Q be probabilities on Ω and let $\Pi(\mathrm{P}, \mathrm{Q})$ be the set of probability measures

$\pi \in \mathcal{P}(\Omega \times \Omega)$ with marginals P and Q, i.e. for all A, B measurable sets

$$\pi(A \times \Omega) = \mathrm{P}(A), \quad \pi(\Omega \times B) = \mathrm{Q}(B). \tag{3}$$

The OT problem amounts to solve the optimization problem (see Kantorovich (1942))

$$\mathcal{T}_c(\mathrm{P}, \mathrm{Q}) := \min_{\pi \in \Pi(\mathrm{P},\mathrm{Q})} \int c(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}), \tag{4}$$

with a continuous cost $c : \Omega \times \Omega \to [0, \infty)$. It is well known (see eg. Villani (2003)) that $\mathcal{W}_p(\mathrm{P}, \mathrm{Q}) := \left(\mathcal{T}_{\|\cdot\|^p}(\mathrm{P}, \mathrm{Q})\right)^{\frac{1}{p}}$—the value of (4) for a potential cost $(\mathbf{x}, \mathbf{y}) \mapsto \|\mathbf{x} - \mathbf{y}\|^p$, for $p \geq 1$—defines a distance on the space of probabilities with finite moments of order p. This distance is called the Wasserstein distance.

In this paper we will consider the quadratic cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, for simplicity. Note nevertheless that most of our results (except, in particular, those related to the continuity of the GP [Proposition 4.1] and the bounds on the potentials [Statements 3.2 to 3.5] apply as well to general bounded cost functions, see the recent work Rigollet and Stromme (2022)).

When at least one distribution P is absolutely continuous w.r.t. Lebesgue measure, then there exists a P-a.e. unique map $T : \Omega \to \Omega$ such that $T\sharp P = Q$, and $\mathcal{W}_2(P, Q)^2 = \int_\Omega \|T(\mathbf{x}) - \mathbf{x}\|^2 d\mathrm{P}(\mathbf{x})$. Moreover, there exists a lower semi-continuous convex function φ such that $T = \nabla\varphi$ P-a.e., with ∇ the gradient operator, and T is the only map of this type pushing forward P to Q, up to a P-negligible modification. This theorem above is commonly referred to as Brenier's theorem (Brenier, 1991). Note that a similar statement was established earlier independently in a probabilistic framework in Cuesta and Matrán (1989).
This result enables to define a natural Hilbertian embedding of the distributions in $\mathcal{P}(\Omega)$ by considering the distance between the transport maps towards a common reference distribution. This framework has been used in Bachoc et al. (2020) to provide kernels on distributions. Yet such kernels have the drawback of being difficult to compute, preventing their use for large or high-dimensional datasets.
Indeed, computing the OT (4) turns out to be computationally difficult. In the discrete case, different algorithms have been proposed such as the Hungarian algorithm (Kuhn, 1955), the simplex algorithm (Luenberger et al., 1984) or others versions using interior points algorithms (Orlin, 1988). The complexity of these methods is at worst of order $O(n^3 \log(n))$ for two discrete distributions with equal size n. Hence Bachoc et al. (2020) and many statistical methods based on OT suffer from this drawback.

To overcome this issue, regularization methods have been proposed to approximate the OT problem by adding a penalty. The seminal paper by Cuturi (2013) provides the description of the Sinkhorn algorithm to regularize OT by using an entropy penalty.

The relative entropy between two probability measures α, β on Ω, is defined as

$$H(\alpha|\beta) = \int_\Omega \log\left(\frac{d\alpha}{d\beta}(\mathbf{x})\right) d\alpha(\mathbf{x})$$

if $\alpha \ll \beta$ and $|\log(d\alpha/d\beta)| \in L^1(\beta)$, and $+\infty$ otherwise. Set $\epsilon > 0$. Then the entropy regularized version of the OT problem is defined as

$$S_\epsilon(\mathrm{P}, \mathrm{Q}) := \min_{\pi \in \Pi(\mathrm{P},\mathrm{Q})} \int_{\Omega\times\Omega} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 d\pi(\mathbf{x}, \mathbf{y}) + \epsilon H(\pi|\mathrm{P} \times \mathrm{Q}), \tag{5}$$

with $\mathrm{P} \times \mathrm{Q}$ the product measure. The entropy term H modifies the linear term in classical OT (the quadratic transportation cost) to produce a strictly convex functional. The parameter ϵ balances the trade-off between the classical OT problem ($\epsilon = 0$) and the influence of the regularizing penalty.

The minimization of (5) is achieved using the Sinkhorn's algorithm. We refer to Peyré et al. (2019) and references therein for more details. The introduction of the Sinkhorn divergence enables to obtain an ε-approximation of the OT distance which can be computed, as pointed out in Altschuler et al. (2017), with a complexity of algorithm of order $O(\frac{n^2}{\varepsilon^3})$, hence in a much faster way than the original OT problem. Several toolboxes have been developed to compute regularized OT such among others as Flamary and Courty (2017) for Python, Klatt et al. (2017) for R.

Contrary to (unregularized) OT, Sinkhorn OT does not provide transport maps, which would in turn provide a Hilbertian embedding. Hence, we consider the dual formulation of (5) pointed out in Genevay (2019):

$$S_\epsilon(\mathrm{P}, \mathrm{Q}) = \sup_{f \in L^1(\mathrm{P}), g \in L^1(\mathrm{Q})} \int_\Omega f(\mathbf{x}) d\mathrm{P}(\mathbf{x}) + \int_\Omega g(\mathbf{y}) d\mathrm{Q}(\mathbf{y}) - \epsilon \int_{\Omega\times\Omega} e^{\frac{1}{\epsilon}\left(f(\mathbf{x}) + g(\mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right)} d\mathrm{P}(\mathbf{x}) dQ(\mathbf{y}) + \epsilon. \tag{6}$$

Note that this formulation is a convex relaxation of the duality of the usual OT. Both primal and dual problems have solutions if P and Q have finite second moments.

Let π be the solution to (5) which will be denoted as the *optimal entropic plan*. Let (f, g) be the solution to (6), which will be denoted as the *optimal entropic potentials*. For $\mathrm{P}, \mathrm{Q} \in \mathcal{P}_{SG}(\Omega)$, both quantities can be related using the formula

$$\frac{d\pi}{d\mathrm{P}dQ} = \exp\left(-\frac{1}{\varepsilon}\left(f(\mathbf{x}) + g(\mathbf{y}) - \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2\right)\right). \tag{7}$$

A consequence of this relation is that we have the *optimality conditions*

$$\int e^{\frac{1}{\epsilon}(f(\mathbf{x})+g(\mathbf{y})-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2)} d\mathrm{P}(\mathbf{x}) = 1, \quad \forall \mathbf{y} \in \Omega, \quad (8)$$

$$\int e^{\frac{1}{\epsilon}(f(\mathbf{x})+g(\mathbf{y})-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2)} d\mathrm{Q}(\mathbf{y}) = 1, \quad \forall \mathbf{x} \in \Omega. \quad (9)$$

3 A KERNEL BASED ON REGULARIZED OPTIMAL TRANSPORT

3.1 Construction of Positive Definite Kernels

Consider a reference measure $\mathcal{U}$ on Ω. For two distributions P and Q, consider the two regularized OTs respectively between P and $\mathcal{U}$ and between Q and $\mathcal{U}$. Let $\pi^{\mathrm{P}}_{\mathcal{U}}$ and $\pi^{\mathrm{Q}}_{\mathcal{U}}$ be the optimal entropic plans and $(f^{\mathrm{P}}_{\mathcal{U}}, g^{\mathrm{P}}_{\mathcal{U}})$ and $(f^{\mathrm{Q}}_{\mathcal{U}}, g^{\mathrm{Q}}_{\mathcal{U}})$ the optimal entropic potentials:

$$\frac{d\pi^{\mathrm{P}}_{\mathcal{U}}}{dP d\mathcal{U}} = \exp\left(-\frac{1}{\varepsilon}\left(f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x}) + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right)\right) \quad (10)$$

$$\frac{d\pi^{\mathrm{Q}}_{\mathcal{U}}}{dQ d\mathcal{U}} = \exp\left(-\frac{1}{\varepsilon}\left(f^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{x}) + g^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right)\right). \quad (11)$$

Our aim is to use the distance $\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})}$ to build Sinkhorn kernels. Note first that the uniqueness of Sinkhorn potentials holds up to additive constants. To obtain uniqueness, from now on, we will define $g^{\mathrm{P}}_{\mathcal{U}}$ as the unique centered (w.r.t. to $\mathcal{U}$) potential. This implies that $g^{\mathrm{P}}_{\mathcal{U}} = g^{\mathrm{P}}_{\mathcal{U}} - \mathbb{E}(g^{\mathrm{P}}_{\mathcal{U}}(U))$, which yields the following equality

$$\mathrm{Var}_{\mathbf{U}\sim\mathcal{U}}(g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U}) - g^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{U})) = \|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|^2_{L^2(\mathcal{U})}.$$

Then, a function $f : [0,\infty) \to \mathbb{R}$ is said to be completely monotone if it is C^∞ on $(0,\infty)$, continuous at 0 and satisfies $(-1)^\ell f^{(\ell)}(r) \geq 0$ for $r > 0$ and $\ell \in \mathbb{N}$. Let $F : [0,\infty) \to \mathbb{R}$ be continuous.
The following theorem provides the kernel construction and its validity (positive-definiteness).

Theorem 3.1. *Let $K : \mathcal{P}_{SG}(\Omega) \times \mathcal{P}_{SG}(\Omega) \to \mathbb{R}$ be the function defined as*

$$(\mathrm{P},\mathrm{Q}) \mapsto K(\mathrm{P},\mathrm{Q}) = F(\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})}), \quad (12)$$

for some $\mathcal{U} \in \mathcal{P}_{SG}(\Omega)$. Then the two following conditions are sufficient conditions for K to be a positive definite kernel on $\mathcal{P}_{SG}(\Omega)$.

1. *$F(\sqrt{\cdot})$ is completely monotone on $[0,\infty)$.*
2. *There exists a finite nonnegative Borel measure ν on $[0,\infty)$ such that for $t \geq 0$ $F(t) = \int_0^\infty e^{-ut^2} d\nu(u)$.*

Remark that the quantity $\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})}$ in Theorem 3.1 is finite via (Mena and Niles-Weed, 2019, Proposition 1).

Examples of functions F for which the assumptions of Theorem 3.1 are satisfied are the well-known square exponential, power exponential and Matérn covariance functions, see Bachoc et al. (2020) and the references therein.

The following proposition bounds the $L^2(\mathcal{U})$ distance between the potentials as a function of the distance between the distributions.

Proposition 3.2. *Let $s \in \mathbb{N}$. Assume that Ω is compact and let $\mathrm{P},\mathrm{Q} \in \mathcal{P}(\Omega)$. Then there exists a constant c_d, depending on the dimension, such that*

$$\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})} \leq c_d \,\mathrm{diam}(\Omega)^s e^{\frac{19}{2}\,\mathrm{diam}(\Omega)^2} \|P-\mathrm{Q}\|_s.$$

Note that the previous bound is still valid if we replace $\|P-\mathrm{Q}\|_s$ by $\mathcal{W}_1(\mathrm{P},\mathrm{Q})$. This remark follows directly from Kantorovich's duality, see Theorem 1.14 in Villani (2003).

The following proposition guarantees that the entropic potentials $g^{\mathrm{P}}_{\mathcal{U}}$ and $g^{\mathrm{Q}}_{\mathcal{U}}$ can be used to characterize the distributions P and Q. It also guarantees that our suggested kernel is not only positive definite but also strictly positive definite.

Proposition 3.3. *Let $\mathrm{P},\mathrm{Q},\mathcal{U} \in \mathcal{P}_{SG}(\Omega)$. The potentials $g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u})$ and $g^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{u})$ can be extended continuously (see Remark 3.5) with (9) for $\mathbf{u} \in \mathbb{R}^d$, which we call the canonical extension. Then $\mathrm{P} = \mathrm{Q}$ if and only if there exists an open set $\mathcal{D}$ with $\mathrm{supp}(\mathcal{U}) \subset \mathcal{D} \subset \Omega$ such that $g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u}) = g^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{u})$, for ℓ_d-a.e. $\mathbf{u} \in \mathcal{D}$ (after extension). Moreover, if there exists an open set $\mathcal{D}' \subset \Omega$ such that $\ell_d \ll \mathcal{U}$ in $\mathcal{D}'$, then $\mathrm{Var}_{\mathbf{U}\sim\mathcal{U}}(g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U}) - g^{\mathrm{Q}}_{\mathcal{U}}(\mathbf{U})) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$.*

Corollary 3.4. *Let $\mathcal{U} \in \mathcal{P}_{SG}(\Omega)$ and assume that there exists an open set $\mathcal{D}' \subset \Omega$ such that $\ell_d \ll \mathcal{U}$ in $\mathcal{D}'$. Assume also that F in Theorem 3.1 is non-constant. Then the function K in Theorem 3.1 is strictly positive definite on $\mathcal{P}_{SG}(\Omega)$.*

Remark 3.5. *The previous result is stated for the case where $\mathcal{U}$ dominates Lebesgue measure on a ball; in particular $\mathcal{U}$ cannot be discrete. Nevertheless, even when $\mathcal{U}$ does not satisfies this assumption, we can still construct a strictly positive definite kernel as follows. Let $\mathcal{U},\mathrm{P},\mathrm{Q} \in \mathcal{P}_{SG}(\Omega)$. First note that from (8) and (9) we have*

$$g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{y}) = -\epsilon \log \int \exp\left(-\frac{1}{\varepsilon}\left(f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right)\right) d\mathrm{P}(\mathbf{x})$$

which extends (as in Proposition 3.3) $g^{\mathrm{P}}_{\mathcal{U}}$ out of the support of $\mathcal{U}$, on an open ball B of $\mathbb{R}^d$ containing $\mathrm{supp}(\mathcal{U})$. Then we have $g^{\mathrm{P}}_{\mathcal{U}} = g^{\mathrm{Q}}_{\mathcal{U}}$ ℓ_d-a.e. on B implies $P = Q$. Thus, let F be as in Theorem 3.1 and assume that it is non-constant. Define the function $K : \mathcal{P}_{SG}(\Omega) \times \mathcal{P}_{SG}(\Omega) \mapsto \mathbb{R}$ as

$$K(\mathrm{P},\mathrm{Q}) = F(\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\ell_d,B)}), \quad (13)$$

with $\|.\|_{L^2(\ell_d,B)}$ *the square norm w.r.t. the measure* ℓ_d *on* B*. Then* K *is strictly definite positive on* $\mathcal{P}_{SG}(\Omega) \times \mathcal{P}_{SG}(\Omega)$.

As an example for Remark 3.5, set $\epsilon = 1$, suppose $\mathbf{0} \in \Omega$ and consider the discrete measure $\mathcal{U} = \delta_{\mathbf{0}}$. In this case, $g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{y}) = \frac{\|\mathbf{y}\|^2}{2} - \log(\mathrm{M}_{\mathrm{P}}(\mathbf{y}))$, for $\mathbf{y}$ in a neighborhood of $\mathbf{0}$, after extension, where $\mathrm{M}_{\mathrm{P}}(\mathbf{y}) = \int e^{\langle \mathbf{y},\mathbf{x}\rangle} d\mathrm{P}(\mathbf{x})$ is the moment generating function.

A kernel K is said to be universal on $\mathcal{P}(\Omega)$ as soon as the space generated by all possible linear combinations $\mu \mapsto \sum_{i=1}^n \alpha_i K(\mu, \mu_i)$ has good approximation properties, in the sense that it is dense in the set continuous functions on $\mathcal{P}(\Omega)$, endowed with the weak convergence of probabilities. We prove that the squared exponential kernel built with the distance between the potentials $\|g^P_{\mathcal{U}} - g^Q_{\mathcal{U}}\|_{L^2(\mathcal{U})}$ is universal.

Proposition 3.6 (Universality of Sinkhorn based kernel)**.** *Assume that* Ω *is compact and that there exists an open set* $\mathcal{D}' \subset \Omega$ *such that* $\ell_d \ll \mathcal{U}$ *in* $\mathcal{D}'$ *Consider for every distribution* P, Q *in* $\mathcal{P}(\Omega)$*, their potentials* $g^P_{\mathcal{U}}$*,* $g^Q_{\mathcal{U}}$ *as in* (10) *and* (11)*. Then for any* $\sigma > 0$ *the kernel defined by*

$$K_\sigma(P, Q) = \exp(-\sigma \|g^P_{\mathcal{U}} - g^Q_{\mathcal{U}}\|^2_{L^2(\mathcal{U})})$$

is universal.

3.2 Consistency Property of the Empirical Kernel

In practical situations, the distributions may not be known but only random samples may be at hand. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be mutually independent sequences of random vectors with distributions P and Q respectively. Denote as P_n and Q_m their empirical measures: $\mathrm{P}_n = (1/n)\sum_{i=1}^n \delta_{\mathbf{X}_i}$ and $\mathrm{Q}_m = (1/m)\sum_{i=1}^m \delta_{\mathbf{Y}_i}$. Consider the optimal entropic transport potentials of the empirical distributions towards a common fixed measure $\mathcal{U}$ denoted by $(f^{\mathrm{P}_n}, g^{\mathrm{P}_n})$ and $(f^{\mathrm{Q}_m}, g^{\mathrm{Q}_m})$. Finally, define the empirical kernel by $K(\mathrm{P}_n, \mathrm{Q}_m) = F(\|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})})$. The following proposition proves its consistency.

Proposition 3.7 (Consistency of the empirical kernel)**.** *Assume that* Ω *is compact and let* $\mathrm{P}, \mathrm{Q} \in \mathcal{P}(\Omega)$*. When* F *is continuous, the empirical kernel* $K(\mathrm{P}_n, \mathrm{Q}_m)$ *converges almost-surely when both* $n, m \to \infty$ *to the true kernel* $K(\mathrm{P}, \mathrm{Q})$*. Moreover if we assume that* F *satisfies* $|F(t) - F(s)| \leq A|t - s|^a$ *for constants* $0 < A < \infty$ *and* $0 < a \leq 1$ *and for* $t \geq 0$*, then we have the following bound, with a constant* C_d*,*

$$\mathbb{E}|K(\mathrm{P}_n, \mathrm{Q}_m) - K(\mathrm{P}, \mathrm{Q})| \leq$$
$$C_d \left(\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right) \mathrm{diam}(\Omega)^{2^{d+1}} e^{\frac{19}{2}\mathrm{diam}(\Omega)^4} \right)^a. \quad (14)$$

3.3 Influence of $\mathcal{U}$ and Invariance Properties.

In this section we investigate the impact of the reference distribution $\mathcal{U}$. Consider two distributions $\mathcal{U}$ and $\mathcal{U}'$ that will be used to build two different kernels. Consider the Sinkhorn OT towards respectively $\mathcal{U}$ and $\mathcal{U}'$ for both distribution P and Q. We will write the corresponding entropic potentials as $(f^P_{\mathcal{U}}, g^{\mathrm{Q}}_{\mathcal{U}})$ and $(f^{\mathrm{P}}_{\mathcal{U}'}, g^{\mathrm{Q}}_{\mathcal{U}'})$ (defined as in Section 3.1). We thus have the two kernels $K_{\mathcal{U}}(\mathrm{P}, \mathrm{Q}) = F(\|g^P_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})})$ and $K_{\mathcal{U}'}(\mathrm{P}, \mathrm{Q}) = F(\|g^P_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U}')})$. One desirable property is translation invariance, which means that the kernel does not change whenever the reference distribution is changed by translation. This is shown next for our kernel construction.

Lemma 3.8. *Let* $\mathcal{U}, \mathrm{P}, \mathrm{Q} \in \mathcal{P}_{SG}(\Omega)$ *and* $T_0 : \Omega \to \Omega$ *be a translation* ($T_0(\mathbf{u}) = \mathbf{u} + \mathbf{u}_0$ *for a fixed* $\mathbf{u}_0 \in \mathbb{R}^d$)*, then* $K_{T_0\sharp\mathcal{U}}(\mathrm{P}, \mathrm{Q}) = K_{\mathcal{U}}(\mathrm{P}, \mathrm{Q})$.

The choice of the reference measure merits an important discussion in this work. Indeed, many possible probabilities with density could be used. For the sake of applications, uniformly distributed measures (on the square, on the ball, spherical uniform) are beneficial, as they are easily approximated on a discrete set. Also, the uniform distribution allows us to compare the Sinkhorn potentials by factorising into independent lower dimensional marginals. Note that an issue of using a unit-squared reference is the high influence of the coordinate system, which can be arbitrary in some applications.

A benefit of using a spherical (invariant to linear isometries) reference distributions is rotation invariance, as shown next.

Lemma 3.9. *Let* $\mathcal{U}, \mathrm{P}, \mathrm{Q} \in \mathcal{P}_{SG}(\Omega)$*, with* $\mathcal{U}$ *spherical and* $T_0 : \Omega \to \Omega$ *be a rigid transformation (i.e.* $T_0(\mathbf{x}) = \boldsymbol{R}\mathbf{x} + \boldsymbol{t}$ *with* $\boldsymbol{R}^T = \boldsymbol{R}^{-1}$ *and* $\boldsymbol{t} \in \mathbb{R}^d$*), then* $K_{T_0\sharp\mathcal{U}}(\mathrm{P}, \mathrm{Q}) = K_{\mathcal{U}}(\mathrm{P}, \mathrm{Q})$.

Again for spherical distributions, the following result shows that a dilatation of factor δ of the reference measure is equivalent to a change of order $\epsilon = 1/\delta^2$ on the Sinkhorn problem.

Proposition 3.10. *Let* $\mathcal{U}, \mathrm{P}, \mathrm{Q} \in \mathcal{P}_{SG}(\Omega)$*, with* $\mathcal{U}$ *spherical and* $T_\delta(\mathbf{u}) = \delta\,\mathbf{u}$*, with* $\delta > 0$*, then*

$$\mathrm{Var}_{\mathbf{U}_\delta \sim T_\delta\sharp\mathcal{U}}(g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta) - g^{\mathrm{Q}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta))$$
$$= \delta^4 \mathrm{Var}_{\mathbf{U}\sim\mathcal{U}}(g^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}(\mathbf{U}) - g^{T_{\frac{1}{\delta}}\sharp\mathrm{Q}}_{\mathcal{U},\delta}(\mathbf{U})),$$

where $g^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}$ *and* $g^{T_{\frac{1}{\delta}}\sharp\mathrm{Q}}_{\mathcal{U},\delta}$ *solve the dual formulation* (6) *of* $S_\epsilon(T_{\frac{1}{\delta}}\sharp\mathrm{P}, \mathcal{U})$ *and* $S_\epsilon(T_{\frac{1}{\delta}}\sharp\mathrm{Q}, \mathcal{U})$*, for* $\epsilon = \frac{1}{\delta^2}$*. Above,* $g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}$ *and* $g^{\mathrm{Q}}_{T_\delta\sharp\mathcal{U}}$ *correspond to* $\epsilon = 1$.

For generic changes of reference distribution, the following proposition quantifies the corresponding kernel changes.

Proposition 3.11. *Assume that* Ω *is compact. Let* $s \in \mathbb{N}$*. There exists a constant* $c(\Omega, d, \epsilon, s)$ *such that for*

$\mathcal{U}, \mathcal{U}', \mathrm{P}, \mathrm{Q} \in \mathcal{P}(\Omega)$,

$$|K_{\mathcal{U}}(\mathrm{P},\mathrm{Q}) - K_{\mathcal{U}'}(\mathrm{P},\mathrm{Q})| \leq 2\mathrm{diam}(\Omega)\|\mathcal{U}-\mathcal{U}'\|_s + c(\Omega,d,\epsilon,s)\left(\|\mathcal{U}-\mathcal{U}'\|_s\|\mathrm{P}-\mathrm{Q}\|_s\right)^{1/2}.$$

4 GAUSSIAN PROCESSES USING SINKHORN'S POTENTIAL KERNEL

Let us recall that a GP $(Z(x))_{x\in E}$ indexed by a set E is entirely characterised by its mean and covariance functions. Its covariance function is defined by $(x,y) \in E^2 \mapsto \mathrm{Cov}(Z(x),Z(y))$. In this section we consider the GP on distributions defined by the Sinkhorn's potential Kernel K with $\mathrm{Cov}(Z(P),Z(Q)) = K(P,Q)$ with K as in Theorem 3.1. We study its properties in this section.

4.1 Continuity of the Gaussian Process

For any positive definite kernel, a GP is guaranteed to exist having this kernel as covariance function. Nevertheless, this GP is defined only as a collection of Gaussian variables, and not necessarily as a random continuous function. Being able to define a GP as a random continuous function is at the same time satisfying from a functional Bayesian point of view, and also technically useful to tackle advanced convergence results, see for instance Bect et al. (2019). Next, we establish the existence of a continuous GP with our kernel construction, under mild regularity assumptions on the space of input probability measures.

For a set $S \subset \mathbb{R}^d$, we let ∂S be its boundary and for $\mathbf{t} \in \mathbb{R}^d$, we let $d(\mathbf{t}, S)$ be the smallest distance between $\mathbf{t}$ and an element of S.

Proposition 4.1. *Let Ω be compact with non-empty interior. Let F in (12) satisfy $|F(t) - F(0)| \leq A|t|^a$ for constants $0 < A < \infty$ and $0 < a \leq 1$ and for $t \geq 0$. Let $b > 0$ be fixed. Let $\mathcal{P}_\delta$ be the set of distributions P on Ω that have a continuous density p w.r.t. Lebesgue measure, such that p is zero on $\{\mathbf{x} \in \Omega, d(\mathbf{x}, \partial\Omega) \leq b\}$. Consider $\mathcal{P}_\delta$ as a metric space with the 1-Wasserstein distance $\mathcal{W}_1$. Then there exists a GP Z on $\mathcal{P}_\delta$ with covariance function as in (12) that is almost surely continuous on $\mathcal{P}_\delta$.*

The proof of Proposition 4.1 is based on a control of the covering numbers of the canonical distance defined through the covariance function in (12). A multi-dimensional integration by part allows us to upper bound this quantity by the covering numbers of $\mathcal{C}^{\lceil \frac{d}{a}+1 \rceil}(\Omega)$, which is enough for the continuity of the process (see (van der Vaart and Wellner, 2013, Theorem 2.7.1) and (Adler, 1990, Theorem 1.1)).

4.2 Estimation of the Parameters and Prediction

Parametrization of the Kernel. The kernel is

$$K_{\boldsymbol{\theta},\mathbf{u}}(\mathrm{P},\mathrm{Q}) = F_{\boldsymbol{\theta}}(\|g^{\mathrm{P}}_{\mathbf{u}} - g^{\mathrm{Q}}_{\mathbf{u}}\|_{L^2(\mathcal{U})}),$$

where $F_{\boldsymbol{\theta}}$ is the function F in Theorem 3.1, depending on the vector of covariance parameters $\boldsymbol{\theta}$. For instance for the square exponential covariance function, $\boldsymbol{\theta}$ consists of a scalar variance and length scale. Furthermore, the Hilbertian embedding yielding $g^{\mathrm{P}}_{\mathbf{u}}$ and $g^{\mathrm{Q}}_{\mathbf{u}}$ depends on the choice of the reference measure $\mathcal{U}$ (see Section 3.1). This choice is indexed by a vector $\mathbf{u}$. For instance, in our numerical experiments, $\mathcal{U}$ will be a discrete measure and $\mathbf{u}$ gathers the support points and weights. The presentation of (standard) likelihood methods for selecting $\boldsymbol{\theta}, \mathbf{u}$ in regression and classification, together with a discussion on microergodicity, are given in the Appendix, for the sake of brevity.

Prediction. The GP framework enables to predict the outputs corresponding to new input probability measures, by using conditional distributions given observed outputs. This is reviewed in the Appendix for regression and classification.

5 IMPLEMENTATION AND EXPERIMENTS

All the experiments of the paper were run on the publicly available GPU Colab hardware.

The code can be found on an anonymous repository: https://anonymous.4open.science/r/SinkhornMuGP-D37E/README.md.

Parametrization of the Reference Measure $\mathcal{U}$. We choose a suitable machine representation for $\mathcal{U}$ (see Section 4.2) as a weighted sum of Diracs:

$$\mathcal{U} = \sum_{i=1}^{q} w_i \delta(\mathbf{x}_i) \text{ with } \sum_{i=1}^{q} w_i = 1, w_i \geq 0, \mathbf{x}_i \in \mathbb{R}^d.$$

In this form $\mathcal{U}$ is not absolutely continuous w.r.t. Lebesgue measure, but see Remark 3.5. The parameters $\mathbf{u}$ for $\mathcal{U}$ gather $w_1, \ldots, w_q, \mathbf{x}_1, \ldots, \mathbf{x}_q$. The procedure for the estimation of $\mathbf{u}, \boldsymbol{\theta}$ is sketched in Algorithm 1.

Gradient Computations. We will use the L-BFGS method for optimization (Liu and Nocedal, 1989). This requires the gradients of the likelihood function in regression and classification w.r.t. $\boldsymbol{\theta}$ and $\mathbf{u}$. The derivatives of relevant quantities w.r.t. $\boldsymbol{\theta}$ can be found in the literature, see for instance Rasmussen and Williams (2006). A specificity of $\mathbf{u}$ is that for some measures P, Q, we need to differentiate $\|g^{\mathrm{P}}_{\mathbf{u}} - g^{\mathrm{Q}}_{\mathbf{u}}\|_{L^2(\mathcal{U})}$ w.r.t. $\mathbf{u}$, that is we need to differentiate regularized OT plans. This is possible either by backpropagating through unrolled Sinkhorn iterations (Genevay et al., 2018), or by using implicit differentiation (Eisenberger et al., 2022). In practice we noticed that, while being slower, unrolling of Sinkhorn iterates was actually more stable numerically.

Software Framework Used. For automatic support of autodifferentiation, we use the Jax framework (Brad-

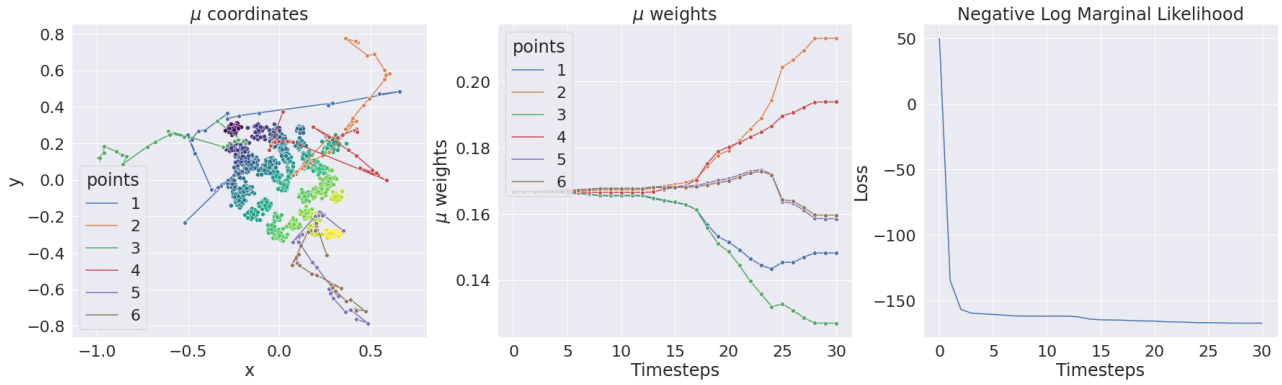

Figure 1: Toy example. **Left**: 50 point clouds of the train set, with color scale depending on the random field Z. The trajectories of the points $\mathbf{x}_i$ of $\mathbf{u}$ are depicted in different colors. **Center**: evolution of the weights $\mathbf{w}$ of $\mathbf{u}$ during training. **Right**: evolution of the negative log marginal likelihood during training.

bury et al., 2018) with the libraries GPJax (Pinder and Dodd, 2022) to implement GP regression, OTT-Jax (Cuturi et al., 2022) for differentiable Sinkhorn algorithm, and Jaxopt (Blondel et al., 2021) for optimization with L-BFGS. The computation of inverse covariance matrices is done efficiently using Cholesky decomposition (Press et al., 2007), which allows efficient computation of *matrix inverse-vector products* without materializing the inverse in memory. The computations are performed in *float32* arithmetic and take advantage of GPU for matrix operations, that are the bottleneck of the algorithm.

Other Numerical Aspects. For $\mathbf{u}$, the point coordinates are parameterized as $\mathbf{x} = S\tanh(\tilde{\mathbf{x}})$ with $S \in \mathbb{R}$ to ensure they remain bounded, the weights are parametrized as $\mathbf{w} = \text{softmax}(\tilde{\mathbf{w}})$ to ensure they represent a valid probability distribution. The dual variables $g^{\mathrm{P}}_{\mathbf{u}}$ computed at each time step during the optimization of $\mathbf{u}$ are cached to speed-up Sinkhorn iterations: this strategy is reasonable since when $\mathbf{u}$ and $\mathbf{u}'$ are close then the dual variables $g^{\mathrm{P}}_{\mathbf{u}}$ and $g^{\mathrm{P}}_{\mathbf{u}'}$ are close too.

Computational Cost of $\mathbf{u}$-Sinkhorn Kernels. We denote by $|\mathbf{u}|$ the size of the support of $\mathbf{u}$ (written q above). For another point cloud of size n, according to Altschuler et al. (2017); Dvurechensky et al. (2018) the time complexity of the Sinkhorn algorithm is $\mathcal{O}(\frac{n|\mathbf{u}|\log(n|\mathbf{u}|)}{\epsilon^2})$ to reach precision ϵ, while the complexity of the MMD kernel is $\mathcal{O}(n^2)$. It follows that for a reference measure with $|\mathbf{u}| \ll n$ the runtime cost of Sinkhorn $\mathbf{u}$-kernel becomes competitive. Runtimes against MMD are reported in Table 3 (in the Appendix), with a speed-up of up to 100 for our method.

Once $\mathbf{u}$ is chosen, the embeddings $g^{\mathrm{P}}_{\mathbf{u}}$ can be pre-computed once for all for each point cloud $\mathrm{P}_1, \ldots, \mathrm{P}_n$ and used as a low dimension embedding of $\mathcal{P}(\Omega)$ into $\mathbb{R}^{|\mathbf{u}|}$. The distribution support $|\mathbf{u}|$ needs to be big enough to capture the similarities between the P_is up to the precision required by the task, but does not need to be bigger (see Section 5.1).

Task	$\lvert\mathbf{u}\rvert$	m	Ours	Bachoc et al. (2020)
Toy example	6	30	0.997	0.81

Table 1: Explained Variance Score (EVS) on the test set for regression tasks, with train set of size $n = 50$ in dimension d=2. $|\mathbf{u}|$: dimension of the embedding. m: cloud size.

Task	$\lvert\mathbf{u}\rvert$	Ours	RBF
"4" vs "6"	4	94.2 ± 1.2	✗
"4" vs "6"	5	95.5 ± 1.0	✗
"4" vs "6"	6	95.0 ± 0.6	98.8 ± 0.2
"shirt" vs "sandals"	12	99.5 ± 0.2	99.7 ± 0.2
"sneakers" vs "sandals"	12	88.6 ± 1.8	91.9 ± 1.2

Table 2: Test Accuracy for **classification** tasks, with train set of size $n = 200$ in dimension d=2 with clouds of size $m = 24 \times 24 = 576$. $|\mathbf{u}|$: embedding dimension. We compare against RBF. Average over 25 runs.

5.1 Regression on Toy Example of Bachoc et al. (2020)

In this section we re-use the example introduced in Section 5.3 of Bachoc et al. (2020). We simulate 100 random two-dimensional isotropic Gaussian distributions. The means are sampled uniformly from $[-0.3, 0.3]^2$, and the variance uniformly from $[0.01^2, 0.02^2]$. The value of the random field induced by a Gaussian of means (m_1, m_2) and variance σ^2 is $Z = \frac{(m_1+0.5-(m_2+0.5)^2)}{1+\sigma}$. Gaussians are approximated by point clouds of size 30 sampled from the distribution. The dataset is splitted into train (50 clouds) and test (50 clouds). The $\mathbf{u}$-measure consists of 6 points on the ball of radius 0.5. Their position $\mathbf{x}_i$ and weight w_i are trained for 30 iterations jointly with kernel parameters. The results are highlighted in Figure 1 and Table 1. The role of $\mathbf{u}$ is investigated in Figure 2 with $|\mathbf{u}| = 2$: the position of the $\mathbf{x}_i$'s makes the embedding more or less suitable for the downstream task, as illustrated by the Explained Variance

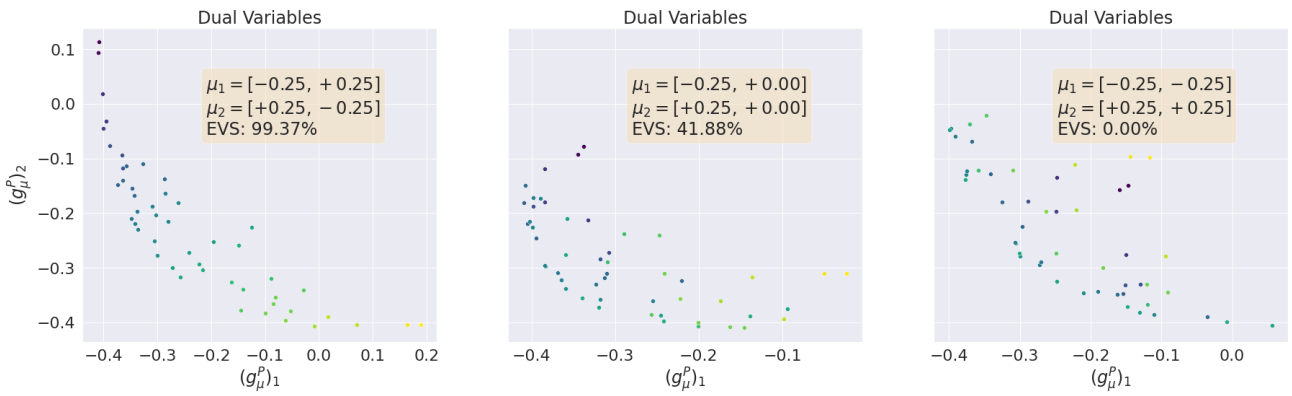

Figure 2: Role of $\mathbf{u}$ in quality of embeddings when $|\mathbf{u}| = 2$ for the example of Section 5.1. Each dot is the 2D embedding of a Gaussian where the color depends on the random field Z. **Left**: optimal choice for $\mathbf{u}$ that ensures the task can be solved. **Center**: sub-optimal choice for $\mathbf{u}$. **Right**: bad choice of $\mathbf{u}$ that prevents learning.

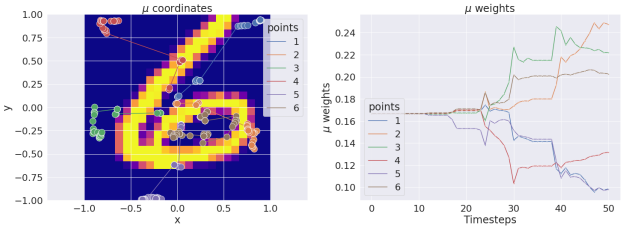

Figure 3: Optimization of $\mathbf{u}$ on Mnist "4" versus "6" task with $|\mathbf{u}| = 6$. An image from the train set is displayed on the background to better grasp the scale of $\mathbf{u}$.

Score (EVS). The EVS of our method is higher than that of Bachoc et al. (2020).

5.2 Binary Classification on Mnist and Fashion-Mnist

We perform binary classification on Mnist by learning to separate digits "4" and "6". The dataset consists of 200 train images, and 1000 test images. Each 28×28 images is centered crop to 24×24 to generate a cloud of size 576 matching pixel coordinates. The normalized pixel intensity is used as a weight in OT. The likelihood is modeled with Bernoulli distributions (not Gaussian, see the Appendix on GP classification), and the log marginal likelihood is maximized using maximum a posteriori (MAP) estimates. We tested different sizes for $|\mathbf{u}| \in [4, 5, 6]$. The training is depicted in Figure 3. The experiment is repeated 10 time with random splits. It shows that Mnist images can be embedded in a space of small dimension that preserves most information about labels, achieving a compression rate of $R = \frac{|\mathbf{u}|}{584} \in [0.006, 0.013]$ tailored for the learning task. We also perform binary classification on Fashion-Mnist.

For both Mnist and Fashion-Mnist, the accuracy is quantified in Table 2. It is approximately similar to the accuracy of a Radial Basis Function (RBF) kernel (also called squared exponential) applied to the "vectorized" images (see Section D.1). Remark that the RBF kernel cannot be applied to general point clouds, while our Sinkhorn kernel is designed for this. On Mnist and Fashion-Mnist, the MMD kernel could not provide comparable accuracy as Sinkhorn and RBF, due to its higher computational cost, see also Table 3.

Algorithm 1: Learn Kernel parameters.

input $(P_i, y_i)_{1 \leq i \leq N}$: dataset of distributions.
input $\boldsymbol{\theta}_0 = (\mathbf{u}_0, \sigma_0, l_0)$: initial parameters.
repeat
  for all P_i **do**
    Solve regularized OT problem between $P_i, \mathbf{u}$.
    Compute Sinkhorn dual potential $g_{\mathbf{u}}^{P_i}$.
  end for
  Build Kernel $K_{ij} := \sigma^2 \exp - \frac{\|g_{\mathbf{u}}^{P_i} - g_{\mathbf{u}}^{P_j}\|^2}{2l^2}$.
  Compute log marginal likelihood $\mathcal{L}(\mathbf{u}, \sigma, l, K, y)$ (Sections B.1 and B.2).
  Compute gradients $\nabla_{(\mathbf{u},\sigma,l)}\mathcal{L}$ with Auto-Diff.
  Perform one step of L-BFGS on $(\mathbf{u}, \sigma, l)$.
until convergence of $(\mathbf{u}, \sigma, l)$.
Return optimal parameters (u_*, σ_*, l_*).

5.3 Texture Classification with C-SVM

We follow the experimental procedure of Kolouri et al. (2016) on the University of Illinois Urbana Champaign (UIUC) texture dataset (Lazebnik et al., 2005). We transform the images into two dimensional probability distributions by computing the gray-level co-occurence matrices (GLCM) (Haralick et al., 1973). C-SVM yields a quadratic programming problem. Convexity guarantees that the algorithm will converge to a global minimum. Our kernel matches the performances of Kolouri et al. (2016) on the same experimental protocol, in Table 4 (see Appendix).

6 CONCLUSION

In this paper we proposed a new positive definite kernel for distributions. It is universal, allows to embed distributions in a space of smaller dimension controlled by $\mathbf{u}$, and is consistent so it scales with the number of points available to approximate the distributions. Empirically, we showed that the reference measure $\mathbf{u}$ was of crucial importance and could be optimized directly with maximum likelihood. Our numerical experiments also highlight that our kernel yields a similar accuracy as other methods, while providing an important computational speed-up compared to MMD.

Acknowledgments

This work was supported by the Project GAP (ANR-21-CE40-0007) of the French National Research Agency (ANR).

References

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes*. IMS.

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in neural information processing systems*, pages 1964–1974.

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification. *Computational Statistics and Data Analysis*, 66:55–69.

Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35.

Bachoc, F., Gamboa, F., Loubes, J.-M., and Venet, N. (2017). A Gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, 64(10):6620–6637.

Bachoc, F., Suvorikova, A., Ginsbourger, D., Loubes, J.-M., and Spokoiny, V. (2020). Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding. *Electronic journal of statistics*, 14(2):2742–2772.

Bect, J., Bachoc, F., and Ginsbourger, D. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919.

Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, second edition.

Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., and Vert, J.-P. (2021). Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Cuesta, J. A. and Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *Annals of Probability*, 17(3):1264–1276.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal transport tools (ott): A JAX toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*.

del Barrio, E., González-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022a). An improved central limit theorem and fast convergence rates for entropic transportation costs. *ArXiv*, abs/2204.09105.

del Barrio, E., González-Sanz, A., and Hallin, M. (2020). A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *Journal of Multivariate Analysis*, page 104671.

del Barrio, E., Sanz, A. G., and Hallin, M. (2022b). Nonparametric multiple-output center-outward quantile regression. *arXiv preprint arXiv:2204.11756*.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR.

Eisenberger, M., Toker, A., Leal-Taixé, L., Bernard, F., and Cremers, D. (2022). A unified framework for implicit Sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 509–518.

Figalli, A. (2017). *The Monge-Ampère Equation and Its Applications*. Zurich Lectures in Advanced Mathematics, European Mathematical Society (EMS), Zurich.

Flamary, R. and Courty, N. (2017). POT Python optimal transport library.

Genevay, A. (2019). *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE).

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Ginsbourger, D., Baccou, J., Chevalier, C., and Perales, F. (2016). Design of computer experiments using competing distances between set-valued inputs. In *mODa 11-Advances in Model-Oriented Design and Analysis*, pages 123–131. Springer.

Glaunes, J., Trouvé, A., and Younes, L. (2004). Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE.

Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. *arXiv preprint arXiv:2207.07427*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Klatt, M., Klatt, M. M., LazyData, T., and Rcpp, L. (2017). Package 'barycenter'. *R package*.

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2018). Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436.

Kolouri, S., Zou, Y., and Rohde, G. K. (2016). Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Lazebnik, S., Schmid, C., and Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1265–1278.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Luenberger, D. G., Ye, Y., et al. (1984). *Linear and nonlinear programming*, volume 2. Springer.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80:309–323.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Meunier, D., Pontil, M., and Ciliberto, C. (2022). Distribution regression with sliced Wasserstein kernels. *arXiv preprint arXiv:2202.03926*.

Moosmüller, C. and Cloninger, A. (2020). Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. *arXiv preprint arXiv:2008.09165*.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from distributions via support measure machines. *Advances in neural information processing systems*, 25.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

Orlin, J. (1988). A faster strongly polynomial minimum cost flow algorithm. STOC'88: Proceedings of the twentieth annual ACM symposium on theory of computing (pp. 377–387).

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pinder, T. and Dodd, D. (2022). Gpjax: A Gaussian process framework in JAX. *Journal of Open Source Software*, 7(75):4455.

Póczos, B., Singh, A., Rinaldo, A., and Wasserman, L. (2013). Distribution-free distribution regression. In *Artificial Intelligence and Statistics*, pages 507–515. PMLR.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Rigollet, P. and Stromme, A. (2022). On the sample complexity of entropic optimal transport. *arXiv:2206.13472*.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for Kriging*. Springer Science & Business Media.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311.

Thi Thien Trang, B., Loubes, J.-M., Risser, L., and Balaresque, P. (2021). Distribution regression model with a reproducing kernel Hilbert space approach. *Communications in Statistics-Theory and Methods*, 50(9):1955–1977.

van der Vaart, A. and Wellner, J. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). Scikit-image: image processing in Python. *PeerJ*, 2:e453.

Villani, C. (2003). *Topics in Optimal Transportation*. American mathematical society, Providence, Rhode Island.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Zhang, H. and Wang, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):290–304.

A PROOFS AND SOME ADDITIONAL RESULTS FOR SECTION 3

Proof of Theorem 3.1. The result follows from Proposition 4 and Remark 5 in Bachoc et al. (2020). □

Proof of Proposition 3.2. The proof can be obtained *mutatis mutandis* from that of the empirical case (del Barrio et al., 2022a, Theorem 4.5). □

Proof of Proposition 3.3. For ease of notation we suppose that $\epsilon = 1$. We prove both equivalences at the same time. In any of the assertions, $\mathrm{P} = \mathrm{Q}$ implies the equality of the potentials in $\mathbb{R}^d$–defined via the canonical extension. On the other hand, let us suppose that $g^{\mathrm{P}}(\mathbf{u}) = g^{\mathrm{Q}}(\mathbf{u})$, for ℓ_d-a.e. $\mathbf{u} \in \mathcal{D}$, for some open set $\mathcal{D} \subset \mathbb{R}^d$. Then $g^{\mathrm{P}} = g^{\mathrm{Q}}$, $\mathcal{U}$-a.e. by continuity when $\mathrm{supp}(\mathcal{U}) \subset \mathcal{D}$. In consequence, the other potentials, obtained by the relations

$$f^{\mathrm{P}}(\mathbf{x}) = -\log\left(\int e^{g^{\mathrm{P}}(\mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} d\mathcal{U}(\mathbf{y})\right), \quad f^{\mathrm{Q}}(\mathbf{x}) = -\log\left(\int e^{g^{\mathrm{Q}}(\mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} d\mathcal{U}(\mathbf{y})\right),$$

are also equal. Moreover, since $g^{\mathrm{P}} = g^{\mathrm{Q}}$ $\mathcal{U}$-a.s. then $e^{g^{\mathrm{P}}} = e^{g^{\mathrm{Q}}}$ too, and, using the optimally conditions, we have

$$\int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} dP(\mathbf{x}) = e^{g^{\mathrm{P}}(\mathbf{y})} = \int e^{f^{\mathrm{Q}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} dQ(\mathbf{x}), \quad \text{for } \mathcal{U}\text{-a.e. } \mathbf{y} \in \Omega.$$

Moreover, extending $e^{g^{\mathrm{P}}(\mathbf{y})}$ as in Remark 3.5, we obtain

$$\int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} dP(\mathbf{x}) = e^{g^{\mathrm{P}}(\mathbf{y})} = \int e^{f^{\mathrm{Q}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2} dQ(\mathbf{x}), \quad \text{for all } \mathbf{y} \in \mathcal{D},$$

so, due to the equality $f^{\mathrm{P}} = f^{\mathrm{Q}}$ in Ω, we have the equality

$$\int e^{\langle \mathbf{x},\mathbf{y}\rangle} e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dP(\mathbf{x}) = \int e^{\langle \mathbf{x},\mathbf{y}\rangle} e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dQ(\mathbf{x}), \quad \text{for all } \mathbf{y} \in \mathcal{D}.$$

By hypothesis, and without loosing generality, there exists a ball centered in 0, such that $\mathbb{B}_\epsilon(0) \subset \mathcal{D}$. A *fortiori*

$$\int e^{\langle \mathbf{x},\mathbf{y}\rangle} e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dP(\mathbf{x}) = \int e^{\langle \mathbf{x},\mathbf{y}\rangle} e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dQ(\mathbf{x}),$$

for all $\mathbf{y} \in \mathbb{B}_\epsilon(0)$. In particular, its evaluation in $\mathbf{y} = 0$ yields the inequality $\int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dP(\mathbf{x}) = \int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dQ(\mathbf{x}) > 0$. The uniqueness of the moment generating function, (see eg. (Billingsley, 1986, Theorem 22.)) proves that the probabilities $\frac{e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2}}{\int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dP(\mathbf{x})} dP(\mathbf{x})$ and $\frac{e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2}}{\int e^{f^{\mathrm{P}}(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2} dQ(\mathbf{x})} dQ(\mathbf{x})$ are equal, so that $\mathrm{P} = \mathrm{Q}$ too. □

Proof of Corollary 3.4. Consider two-by-two distinct measures $P_1, \ldots, P_n$. Then from Proposition 3.3, the functions $g_{\mathcal{U}}^{P_1}, \ldots, g_{\mathcal{U}}^{P_n}$ are two-by-two distinct in $L^2(\mathcal{U})$. Then the matrix

$$\left[F(\|g_{\mathcal{U}}^{\mathrm{P_i}} - g_{\mathcal{U}}^{\mathrm{P_j}}\|_{L^2(\mathcal{U})})\right]_{1\leq i,j\leq n}$$

is strictly positive definite from Proposition 4 in Bachoc et al. (2020). □

Proof of Remark 3.5. The fact that $g_{\mathcal{U}}^{\mathrm{P}} = g_{\mathcal{U}}^{\mathrm{Q}}$ ℓ_d-a.e. on B implies $P = Q$ holds from Proposition 3.3. Then, strict positive definiteness is shown as in the proof of Corollary 3.4. □

Proof of Proposition 3.6. First note that $\mathcal{P}(\Omega)$ is a compact metric space endowed with the Wasserstein distance. Consider the map Φ from $\mathcal{P}(\Omega)$ to the separable Hilbert space $L^2(\mathcal{U})$ such that for any $P \in \mathcal{P}(\Omega)$, $\Phi(P) = g_{\mathcal{U}}^P$. This map is continuous w.r.t. the Wasserstein distance $\mathcal{W}_1$ from Proposition 3.2 and the comment after it. Moreover Proposition 3.3 implies that Φ is injective. Hence using Theorem 2.2 in Christmann and Steinwart (2010), we obtain the universality of the kernel. □

Proof of Proposition 3.7. First, using Proposition 3.2, for some constant $C_{d,\Omega}$, we obtain the following bounds

$$\|g^{\mathrm{P}_n} - g^{\mathrm{P}}\|_{L^2(\mathcal{U})} \leq C_{d,\Omega}\|\mathrm{P} - \mathrm{P}_n\|_s, \text{ and } \|g^{\mathrm{Q}_m} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})} \leq C_{d,\Omega}\|\mathrm{Q} - \mathrm{Q}_m\|_s, \tag{15}$$

where, in this case, $s = \lceil \frac{2}{d} \rceil + 1$. Moreover, the triangle inequality

$$\left| \|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})} \right| \leq \|g^{\mathrm{P}_n} - g^{\mathrm{P}}\|_{L^2(\mathcal{U})} + \|g^{\mathrm{Q}} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})}$$

and (15) yield the upper bound

$$\left| \|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})} \right| \leq C_{d,\Omega} \left(\|\mathrm{P} - \mathrm{P}_n\|_s + \|\mathrm{Q} - \mathrm{Q}_m\|_s \right). \tag{16}$$

Hence when $n, m \to +\infty$, consistency of the empirical distributions and continuity of the function F lead to the consistency of the empirical kernel almost surely.
To obtain the upper bound, note that using the Assumption on F we have

$$\begin{aligned} \mathbb{E}|K(\mathrm{P}_n, \mathrm{Q}_m) - K(\mathrm{P}, \mathrm{Q})| &= \mathbb{E}|F(\|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})}) - F(\|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})})| \\ &\leq A\mathbb{E}|\|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})}|^a. \end{aligned}$$

Since $a \in (0, 1]$, Jensen's inequality allows us to say that

$$\mathbb{E}|K(\mathrm{P}_n, \mathrm{Q}_m) - K(\mathrm{P}, \mathrm{Q})| \leq A \left(\mathbb{E}|\|g^{\mathrm{P}_n} - g^{\mathrm{Q}_m}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})}| \right)^a. \tag{17}$$

Therefore, (16) and (17) enable to obtain that

$$\mathbb{E}|K(\mathrm{P}_n, \mathrm{Q}_m) - K(\mathrm{P}, \mathrm{Q})| \leq A \left(\mathbb{E}|\|\mathrm{P} - \mathrm{P}_n\|_s + \|\mathrm{Q} - \mathrm{Q}_m\|_s| \right)^a. \tag{18}$$

The rest of the proof follows by classical empirical processes arguments. □

Proof of Lemma 3.8. Let $\mathbf{U} \sim \mathcal{U} \in \mathcal{P}_{SG}(\Omega)$ and $\mathbf{X} \sim \mathrm{P} \in \mathcal{P}_{SG}(\Omega)$. The potentials are (up to additive constants) characterized by the optimally conditions

$$\begin{aligned} \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{X}) + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u}) - \frac{1}{2}\|\mathbf{X}-\mathbf{u}\|^2} \right) = 1 \quad \mathcal{U} - a.s. \\ \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x}) + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U}) - \frac{1}{2}\|\mathbf{x}-\mathbf{U}\|^2} \right) = 1 \quad \mathrm{P} - a.s. \end{aligned}$$

Let T_0 be a translation—defined as $\mathbf{u} \mapsto \mathbf{u} + \mathbf{u}_0$—and $\mathbf{U}_0 = T_0(\mathbf{U}) \sim \mathcal{U}_0$, then we claim that

$$f^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{x}) = f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x}) + \langle \mathbf{u}_0, \mathbf{x} \rangle + \frac{3}{4}\|\mathbf{u}_0\|^2 \quad \text{and} \quad g^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{u}) = g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u} - \mathbf{u}_0) - \langle \mathbf{u}_0, \mathbf{u} \rangle + \frac{3}{4}\|\mathbf{u}_0\|^2$$

is a pair of OT potentials for $\mathcal{U}_0$. The verification of the optimallity conditions is enough to prove the claim. On the one hand, note that

$$\mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{X}) + g^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{u}') - \frac{1}{2}\|\mathbf{X}-\mathbf{u}'\|^2} \right) = \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{X}) + \langle \mathbf{u}_0, \mathbf{X} \rangle + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u}' - \mathbf{u}_0) - \langle \mathbf{u}_0, \mathbf{u}' \rangle + \frac{3}{2}\|\mathbf{u}_0\|^2 - \frac{1}{2}\|\mathbf{X}-\mathbf{u}'\|^2} \right)$$

and the (evident) change of variables $\mathbf{u} = \mathbf{u}' - \mathbf{u}_0$ yields

$$\begin{aligned} \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{X}) + g^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{u}+\mathbf{u}_0) - \frac{1}{2}\|\mathbf{X}-\mathbf{u}+\mathbf{u}_0\|^2} \right) &= \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{X}) + \langle \mathbf{u}_0, \mathbf{X} \rangle + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u}) - \langle \mathbf{u}_0, \mathbf{u}+\mathbf{u}_0 \rangle + \frac{3}{2}\|\mathbf{u}_0\|^2 - \frac{1}{2}\|\mathbf{X}-(\mathbf{u}+\mathbf{u}_0)\|^2} \right) \\ &= \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{X}) + g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{u}) - \frac{1}{2}\|\mathbf{X}-\mathbf{u}\|^2} \right) \\ &= 1 \quad \mathcal{U} - a.s. \end{aligned}$$

Therefore we obtain the first optimally condition

$$\mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{X}) + g^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{u}') - \frac{1}{2}\|\mathbf{X}-\mathbf{u}'\|^2} \right) = 1 \quad \mathcal{U}_0 - a.s.$$

On the other hand, note that

$$\begin{aligned}\mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{x})+g^{\mathrm{P}}_{\mathcal{U}_0}(\mathbf{U}_0)-\frac{1}{2}\|\mathbf{x}-\mathbf{U}_0\|^2}\right) &= \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x})+\langle \mathbf{u}_0,\mathbf{x}\rangle+g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U}_0-\mathbf{u}_0)-\langle \mathbf{u}_0,\mathbf{U}_0\rangle+\frac{3}{2}\|\mathbf{u}_0\|^2-\frac{1}{2}\|\mathbf{x}-\mathbf{U}_0\|^2}\right)\\ &= \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x})+\langle \mathbf{u}_0,\mathbf{x}\rangle+g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U})-\langle \mathbf{u}_0,\mathbf{U}+\mathbf{u}_0\rangle+\frac{3}{2}\|\mathbf{u}_0\|^2-\frac{1}{2}\|\mathbf{x}-\mathbf{U}+\mathbf{u}_0\|^2}\right)\\ &= \mathbb{E}\left(e^{f^{\mathrm{P}}_{\mathcal{U}}(\mathbf{x})+g^{\mathrm{P}}_{\mathcal{U}}(\mathbf{U})-\frac{1}{2}\|\mathbf{x}-\mathbf{U}\|^2}\right)\\ &= 1 \quad \mathrm{P}-a.s.\end{aligned}$$

which implies the second optimally condition. □

Proof of Lemma 3.9. Via Lemma 3.8, we only need to prove the invariance w.r.t. $T_0(\mathbf{u}) = \boldsymbol{R}\mathbf{u}$ with $\boldsymbol{R}$ a linear isommetry. By definition of spherical measure, $T_0(\mathbf{U}) \sim \mathcal{U}$, for any $\mathbf{U} \sim \mathcal{U}$, so the solutions of (6) are the same. □

Proof of Proposition 3.10. Set $\mathcal{U}_\delta = T_\delta \sharp \mathcal{U}$ and a pair $(g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}, f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}})$ solving the dual formulation (6) of $S_1(\mathrm{P}, \mathrm{T}_\delta\sharp\mathcal{U})$. The optimality conditions yield

$$\mathbb{E}\left(e^{f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{X})+g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{u}_\delta)-\frac{1}{2}\|\mathbf{X}-\mathbf{u}_\delta\|^2}\right) = 1 \quad \mathcal{U}_\delta - a.s.$$

where we can do a change of variables $\mathbf{u} = \frac{1}{\delta}\mathbf{u}_\delta$ to have

$$1 \overset{\mathcal{U}-a.s.}{=} \mathbb{E}\left(e^{f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{X})+g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{u})-\frac{1}{2}\|\mathbf{X}-\delta\mathbf{u}\|^2}\right) \overset{\mathcal{U}-a.s.}{=} \mathbb{E}\left(e^{f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{X})+g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{u})-\frac{\delta^2}{2}\|\frac{1}{\delta}\mathbf{X}-\mathbf{u}\|^2}\right).$$

Set $\mathbf{X}_{\frac{1}{\delta}} = T_{\frac{1}{\delta}}(\mathbf{X}) = \frac{1}{\delta}\mathbf{X}$, then

$$1 \overset{\mathcal{U}-a.s.}{=} \mathbb{E}\left(e^{\delta^2\left(\frac{1}{\delta^2}f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{X}_{\frac{1}{\delta}})+\frac{1}{\delta^2}g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{u})-\frac{1}{2}\|\mathbf{X}_{\frac{1}{\delta}}-\mathbf{u}\|^2\right)}\right).$$

The same argument also shows (with the obvious notation) that

$$1 \overset{\mathrm{P}_{\frac{1}{\delta}}-\mathrm{a.s.}}{=} \mathbb{E}\left(e^{\delta^2\left(\frac{1}{\delta^2}f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{x}_{\frac{1}{\delta}})+\frac{1}{\delta^2}g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\mathbf{U})-\frac{1}{2}\|\mathbf{x}_{\frac{1}{\delta}}-\mathbf{U}\|^2\right)}\right),$$

which means that the pair $\left(\frac{1}{\delta^2}f^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\,\cdot), \frac{1}{\delta^2}g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\,\cdot)\right) = \left(f^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}, \frac{1}{\delta^2}g^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}\right)$ solves the dual formulation (6) of $S_\epsilon(T_{\frac{1}{\delta}}\sharp\mathrm{P}, \mathcal{U})$, for $\epsilon = \frac{1}{\delta^2}$. The same, verbatim, can be done for Q. Finally, we note that

$$\mathbb{E}\left(g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta) - g^{\mathrm{Q}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta)\right)^2 = \mathbb{E}\left(g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\delta\,\mathbf{U}) - g^{\mathrm{Q}}_{T_\delta\sharp\mathcal{U}}(\delta\,\mathbf{U})\right)^2 = \delta^4\mathbb{E}\left(g^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}(\mathbf{U}) - g^{T_{\frac{1}{\delta}}\sharp\mathrm{Q}}_{\mathcal{U},\delta}(\mathbf{U})\right)^2,$$

and

$$0 = \mathbb{E}\left(g^{\mathrm{P}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta)\right) = \delta^2\mathbb{E}\left(g^{T_{\frac{1}{\delta}}\sharp\mathrm{P}}_{\mathcal{U},\delta}(\mathbf{U})\right),\ 0 = \mathbb{E}\left(g^{\mathrm{Q}}_{T_\delta\sharp\mathcal{U}}(\mathbf{U}_\delta)\right) = \delta^2\mathbb{E}\left(g^{T_{\frac{1}{\delta}}\sharp\mathrm{Q}}_{\mathcal{U},\delta}(\mathbf{U})\right),$$

which allows to conclude. □

Proof of Proposition 3.11. For ease of notation we suppose that $\epsilon = 1$. First note that

$$\begin{aligned}\|g^P_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})} &\le \|g^P_{\mathcal{U}} - g^{\mathrm{P}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})} + \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})} + +\|g^{\mathrm{Q}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})}\\ &\le \mathrm{diam}(\Omega)\left(\|g^P_{\mathcal{U}} - g^{\mathrm{P}}_{\mathcal{U}'}\|_\infty + \|g^{\mathrm{Q}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_\infty\right) + \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})}.\end{aligned}$$

Using (del Barrio et al., 2022a, Theorem 4.5), we obtain that

$$\|g^P_{\mathcal{U}} - g^{\mathrm{P}}_{\mathcal{U}'}\|_\infty \le \|\mathcal{U} - \mathcal{U}'\|_s$$

$$\|g^{Q}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_\infty \leq \|\mathcal{U} - \mathcal{U}'\|_s.$$

The last term of the bound can be written as

$$\begin{aligned}
\|g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})} &= \left(\int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d(\mathcal{U}-\mathcal{U}')(\mathbf{x}) + \int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d\mathcal{U}'(\mathbf{x})\right)^{\frac{1}{2}} \\
&\leq \left(\int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d(\mathcal{U}-\mathcal{U}')(\mathbf{x}) + \int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d\mathcal{U}'(\mathbf{x})\right)^{\frac{1}{2}} \\
&\leq \left(\left|\int \frac{(g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2}{\|(g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'})^2\|_{\mathcal{C}^s(\Omega)}} \|(g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'})^2\|_{\mathcal{C}^s(\Omega)} d(\mathcal{U}-\mathcal{U}')(\mathbf{x})\right| + \int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d\mathcal{U}'(\mathbf{x})\right)^{\frac{1}{2}} \\
&\leq \left(\|(g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'})^2\|_{\mathcal{C}^s(\Omega)} \sup_{f\in\mathcal{C}^s(\Omega)} \left|\int f(\mathbf{x}) d(\mathcal{U}-\mathcal{U}')(\mathbf{x})\right| + \int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d\mathcal{U}'(\mathbf{x})\right)^{\frac{1}{2}}.
\end{aligned}$$

Now note on the first hand that

$$\sup_{f\in\mathcal{C}^s(\Omega)} \left|\int f(\mathbf{x}) d(\mathcal{U}-\mathcal{U}')(\mathbf{x})\right| \leq \|\mathcal{U}-\mathcal{U}'\|_s.$$

On the other hand recall that

$$\|(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'})^2\|_{\mathcal{C}^s(\Omega)} = \sum_{i=0}^{s} \sum_{|\boldsymbol{\alpha}|=i} \|D^{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'})^2\|_\infty,$$

with the same notation $D^{\boldsymbol{\alpha}}$ as in Section 2.1.

But for $|\boldsymbol{\alpha}| \geq 1$, $D^{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'})^2$ is a linear combination of product of derivatives of $g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}$, which enables to write that

$$\sum_{i=0}^{s} \sum_{|\boldsymbol{\alpha}|=i} \|D^{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'})^2\|_\infty \leq \sum_{i=0}^{s} \sum_{|\boldsymbol{\alpha}|=i} \|P_{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}, \partial_1(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}), \ldots, D^{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}))\|_\infty$$

for $P_{\boldsymbol{\alpha}}$ polynomial functions. Since all functions are continuous and evaluated on a compact set Ω, their supremum norm is bounded, which enables to write that

$$\|D^{\boldsymbol{\alpha}}(g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'})^2\|_{\mathcal{C}^s(\Omega)} \leq C_{\boldsymbol{\alpha}}(\Omega) \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}\|_{\mathcal{C}^s(\Omega)}$$

for a constant $C_{\boldsymbol{\alpha}}(\Omega)$ which depends on P, Q, $\boldsymbol{\alpha}$ and Ω. Since

$$\|(g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'})^2\|_{\mathcal{C}_s(\Omega)} \leq C_{\boldsymbol{\alpha}}(\Omega) \|\mathrm{P} - \mathrm{Q}\|_s$$

we obtain that

$$\begin{aligned}
\|g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})} &\leq \left(\|\mathcal{U}-\mathcal{U}'\|_s \|\mathrm{P}-\mathrm{Q}\|_s + \int (g^{\mathrm{P}}_{\mathcal{U}'}(\mathbf{x}) - g^{\mathrm{Q}}_{\mathcal{U}'}(\mathbf{x}))^2 d\mathcal{U}'(\mathbf{x})\right)^{\frac{1}{2}} \\
&\leq \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}\|_{L^2(\mathcal{U}')} + \left(\|\mathcal{U}-\mathcal{U}'\|_s \|\mathrm{P}-\mathrm{Q}\|_s\right)^{1/2},
\end{aligned}$$

which gives the inequality

$$\begin{aligned}
&\|g^{\mathrm{P}}_{\mathcal{U}'} - g^{\mathrm{Q}}_{\mathcal{U}'}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}\|_{L^2(\mathcal{U}')} \\
&\leq \left(\|\mathcal{U}-\mathcal{U}'\|_s \|\mathrm{P}-\mathrm{Q}\|_s\right)^{1/2} + 2\mathrm{diam}(\Omega)\|\mathcal{U}-\mathcal{U}'\|_s.
\end{aligned}$$

Finally by symmetry, we obtain that

$$\left|\|g^{\mathrm{P}}_{\mathcal{U}} - g^{\mathrm{Q}}_{\mathcal{U}}\|_{L^2(\mathcal{U})} - \|g^{\mathrm{P}}_{\mathcal{U}'} - g^{Q}_{\mathcal{U}'}\|_{L^2(\mathcal{U}')}\right| \leq \left(\|\mathcal{U}-\mathcal{U}'\|_s \|\mathrm{P}-\mathrm{Q}\|_s\right)^{1/2} + 2\mathrm{diam}(\Omega)\|\mathcal{U}-\mathcal{U}'\|_s,$$

which proves the result. □

B ADDITIONAL CONTENT FOR SECTION 4

B.1 Likelihood in Regression

In regression, we consider a dataset $\mu_1, y_1, \ldots, \mu_n, y_n$, with $y_i = Z(\mu_i)$ where Z is a centered GP with covariance function in $\{K_{\boldsymbol{\theta},\mathbf{u}}\}$. We let U_n be the list $(\mu_1, \ldots, \mu_n)$ and write $\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n, U_n)$ for the $n \times n$ matrix with component i, j equal to $K_{\boldsymbol{\theta}}(\mu_i, \mu_j)$. We also write $\mathbf{Y}_n$ for the $n \times 1$ vector $(y_1, \ldots, y_n)^\top$. Then the likelihood function is $g_{\mathcal{N}}(\mathbf{Y}_n, \mathbf{0}, \mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n, U_n))$, where for any vectors $\mathbf{m}$ and $\mathbf{x}$ and matrix $\boldsymbol{\Sigma}$, in dimension n,

$$g_{\mathcal{N}}(\mathbf{x}, \mathbf{m}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})} \tag{19}$$

is the Gaussian density at $\mathbf{x}$ with mean $\mathbf{m}$ and covariance $\boldsymbol{\Sigma}$. Then $\boldsymbol{\theta}, \mathbf{u}$ can be selected by maximizing the likelihood function. Note that in regression, one can also use cross validation to estimate $\boldsymbol{\theta}$ and $\mathbf{u}$ (Bachoc, 2013; Rasmussen and Williams, 2006; Zhang and Wang, 2010).

B.2 Likelihood in Classification

In classification, Z is as before and we consider a dataset $\mu_1, y_1, \ldots, \mu_n, y_n$, where, conditionally to Z, $y_1, \ldots, y_n$ are independent with, for $i = 1, \ldots, n$,

$$\mathbb{P}(y_i = 1) = 1 - \mathbb{P}(y_i = 0) = \frac{e^{Z(\mu_i)}}{1 + e^{Z(\mu_i)}}.$$

Then, from for instance Equation 3.30 in Rasmussen and Williams (2006), the likelihood function is

$$\int_{\mathbb{R}^n} g_{\mathcal{N}}\left(\mathbf{v}, \mathbf{0}, \mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n, U_n)\right) g(\mathbf{Y}_n|\mathbf{v}) d\mathbf{v},$$

with the density of $\mathbf{Y}_n$ given $(Z(\mu_1), \ldots, Z(\mu_n)) = \mathbf{v}$:

$$g(\mathbf{Y}_n|\mathbf{v}) = \prod_{i=1}^{n} \left(\left(\frac{e^{v_i}}{1 + e^{v_i}} \right)^{y_i} + \left(\frac{1}{1 + e^{v_i}} \right)^{1-y_i} \right). \tag{20}$$

Above, $g_{\mathcal{N}}$ is as in (19).

B.3 Discussion of Microergodicity

For both regression and classification, a natural theoretical question is the consistency of estimators for $\boldsymbol{\theta}$ and $\mathbf{u}$ as $n \to \infty$. This question is essentially open for distributional inputs, as only a few results exist (Bachoc et al., 2017). In contrast, most existing results address standard vector inputs (Bachoc, 2014; Stein, 1999; Zhang, 2004). A necessary condition for this consistency is that $\boldsymbol{\theta}$ and $\mathbf{u}$ are microergodic, which means that changing them always changes the Gaussian measure of Z on the set of functions from the input space to $\mathbb{R}$. We refer to Bachoc et al. (2020); Stein (1999) for more formal details. Related to our setting, Bachoc et al. (2020) shows that microergodicity typically holds when the input space is a Hilbert ball. This result provides positive indications that $\boldsymbol{\theta}$ and $\mathbf{u}$ may be microergodic in fairly general frameworks.

B.4 Prediction

We now aim at predicting a new output, associated to a new measure μ_0, based on $y_1, \ldots, y_n$, that is to compute the conditional distribution of the new output given $y_1, \ldots, y_n$.

First, consider regression, where the output is $Z(\mu_0)$ and $y_1, \ldots, y_n$ are as in Section B.1. The conditional mean of $Z(\mu_0)$ given $y_1, \ldots, y_n$ is

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{u}}\left(Z(\mu_0)|\, Z(\mu_1), \ldots, Z(\mu_n)\right) = \mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(\mu_0, U_n)\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n, U_n)^{-1}\mathbf{Y}_n, \tag{21}$$

where $\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(\mu_0, U_n)$ is the $1 \times n$ vector with component i equal to $K_{\boldsymbol{\theta},\mathbf{u}}(\mu_0, \mu_i)$, $i = 1, \ldots, n$. Thus, classically, GP prediction in regression consists in the conditional mean (also the L^2 projection). We also have the well-known error

indicator (conditional variance)

$$\begin{aligned}&\mathrm{var}_{\boldsymbol{\theta},\mathbf{u}}\left(Z(\mu_0)|\, Z(\mu_1),\ldots,Z(\mu_n)\right) = \\ &\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(\mu_0,\mu_0) - \mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(\mu_0,U_n)\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n,U_n)^{-1}\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n,\mu_0),\end{aligned} \tag{22}$$

where we let $\mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n,\mu_0) = K_{\boldsymbol{\theta},\mathbf{u}}(\mu_0,U_n)^\top$.

Second, consider classification, where the output is y_0, such that conditionally to Z, y_0 is independent from $y_1,\ldots,y_n$ (defined as in Section B.2) and $\mathbb{P}(y_0=1) = 1-\mathbb{P}(y_0=0) = e^{Z(\mu_0)}/(1+e^{Z(\mu_0)})$. Then, as follows from instance from Equations 3.9 and 3.10 in Rasmussen and Williams (2006), the conditional probability that $y_0=1$ given $y_1,\ldots,y_n$ is given by

$$\frac{1}{\kappa}\int_{\mathbb{R}^{n+1}} g_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{v}|\mathbf{Y}_n)\, g_{\boldsymbol{\theta},\mathbf{u}}(z|\mathbf{v})\frac{e^z}{1+e^z}d\mathbf{v}dz.$$

Above, $g_{\boldsymbol{\theta},\mathbf{u}}(z|\mathbf{v})$ is the Gaussian density of $Z(\mu_0)$ at z given $Z(\mu_i)=v_i$, $i=1,\ldots,n$, as obtained from (21) and (22), $\kappa = \int_{\mathbb{R}^n} g_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{v}|\mathbf{Y}_n)d\mathbf{v}$ and

$$g_{\boldsymbol{\theta},\mathbf{u}}(\mathbf{v}|\mathbf{Y}_n) = g_{\mathcal{N}}\left(\mathbf{v}, 0, \mathbf{K}_{\boldsymbol{\theta},\mathbf{u}}(U_n,U_n)\right) g(\mathbf{Y}_n|\mathbf{v}),$$

with $g_{\mathcal{N}}$ as in (19).

C PROOFS FOR SECTION 4

Proof of Proposition 4.1. Let d_Z be the canonical distance on $\mathcal{P}_\delta$ of the covariance function in (12), given by, for $\mathrm{P},\mathrm{Q}\in\mathcal{P}_\delta$,

$$d_Z(\mathrm{P},\mathrm{Q}) = \sqrt{2F(0) - 2F\left(\|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})}\right)}.$$

For $\epsilon>0$, let $\mathcal{N}(\epsilon,\mathcal{P}_\delta,d_Z)$ be the minimum number of d_Z-balls in $\mathcal{P}_\delta$ of radius ϵ needed to cover $\mathcal{P}_\delta$. From for instance Theorem 1.1 in Adler (1990), in order to conclude the proof, it is sufficient to show that

$$\int_0^\infty \sqrt{\log(\mathcal{N}(\epsilon,\mathcal{P}_\delta,d_Z))}d\epsilon < \infty. \tag{23}$$

Let $\alpha_d = \lceil d/a\rceil + 2$ and $s_d = d\alpha_d$. Proposition 3.2 yields that for $\mathrm{P},\mathrm{Q}\in\mathcal{P}_\delta$,

$$\|g^{\mathrm{P}} - g^{\mathrm{Q}}\|_{L^2(\mathcal{U})} \le B_d\|\mathrm{P}-\mathrm{Q}\|_{s_d},$$

where $\|\cdot\|_{s_d}$ is defined in (2) and B_d is a constant not depending on ϵ. For any P and Q in $\mathcal{P}_\delta$, with densities p and q, we have, with $\|\cdot\|_{\mathcal{C}^{s_d}(\Omega)}$ defined in (1),

$$\begin{aligned}d_Z(\mathrm{P},\mathrm{Q}) &\le \sqrt{2A\|g^{\mathrm{P}} - g^{\mathrm{Q}}\|^a_{L^2(\mathcal{U})}} \\ &\le \sqrt{2AB_d^a\|\mathrm{P}-\mathrm{Q}\|^a_{s_d}} \\ &= \sqrt{2AB_d^a}\sup_{f\in\mathcal{C}^{s_d}(\Omega),\|f\|_{\mathcal{C}^{s_d}(\Omega)}\le 1}\left(\int_\Omega f(\mathbf{x})(p(\mathbf{x})-q(\mathbf{x}))d\mathbf{x}\right)^{a/2}.\end{aligned} \tag{24}$$

For $f\in\mathcal{C}^{s_d}(\Omega), \|f\|_{\mathcal{C}^{s_d}(\Omega)}\le 1$, we can multiply f by an infinitely differentiable function that is zero on $\{\mathbf{t}\in\Omega, d(\mathbf{t},\partial\Omega)\le b/2\}$, and one on $\{\mathbf{t}\in\Omega, d(\mathbf{t},\partial\Omega)\ge b\}$ (that exists by Lemma C.1). Let us write $\tilde{f}$ the result of this multiplication. Since P and Q above are in $\mathcal{P}_\delta$, we have

$$\int_\Omega f(\mathbf{x})(p(\mathbf{x})-q(\mathbf{x}))d\mathbf{x} = \int_\Omega \tilde{f}(\mathbf{x})(p(\mathbf{x})-q(\mathbf{x}))d\mathbf{x}. \tag{25}$$

By taking the infinitely differentiable function the same for each f, we obtain $\|\tilde{f}\|_{\mathcal{C}^{s_d}(\Omega)}\le D_d$, where D_d is a constant.

Now we consider a bounded compact hyper-rectangle R such that Ω belongs to the interior of R. Above, p and q are summable and continuous on Ω and are zero on $\{\mathbf{x} \in \Omega, d(\mathbf{x}, \partial\Omega) \leq b\}$, so we can extend them to summable continuous functions on R, that take the value 0 on $R\backslash\Omega$. Let us also extend $\tilde{f}$ on R by taking values zero on $R\backslash\Omega$. We then have $\|\tilde{f}\|_{\mathcal{C}^{s_d}(R)} \leq D_d$, by defining $\|\cdot\|_{\mathcal{C}^{s_d}(R)}$ as in (1) (replacing Ω by R).

We can thus write

$$\int_\Omega \tilde{f}(\mathbf{x})(p(\mathbf{x}) - q(\mathbf{x}))d\mathbf{x} = \int_R \tilde{f}(\mathbf{x})(p(\mathbf{x}) - q(\mathbf{x}))d\mathbf{x}, \tag{26}$$

where we use the same notation $p, q, \tilde{f}$ both for the original functions on Ω and their extensions on R. The function $\tilde{f}$ is s_d times differentiable on R, with all the derivatives of order s_d or less that cancel out on the boundary of R. Write the hyper-rectangle R as $\prod_{j=1}^d [\ell_j, u_j]$. Let for $i = 0, \ldots, d$, $I^{(1,i)}q$ be the function defined on R by, for $i = 0$, $I^{(1,0)}q = q$ and for $i \geq 1$, $(x_1, \ldots, x_d) \in R$,

$$(I^{(1,i)}q)(x_1, \ldots, x_d) = \int_{\ell_i}^{x_i} (I^{(1,i-1)}q)(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_d)dt.$$

Let for $i = 0, \ldots, d$, $I^{(2,i)}q$ be the function defined on R by, for $i = 0$, $I^{(2,0)}q = I^{(1,d)}q$ and for $i \geq 1$, $(x_1, \ldots, x_d) \in R$,

$$(I^{(2,i)}q)(x_1, \ldots, x_d) = \int_{\ell_i}^{x_i} (I^{(2,i-1)}q)(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_d)dt.$$

We iterate like that until defining $I^{(\alpha_d,d)}q$ from R to $\mathbb{R}$ that satisfy $D^{(\alpha_1,\ldots,\alpha_d)}I^{(\alpha_d,d)}q = q$. We define $I^{(\alpha_d,d)}p$ similarly.

Hence, we can apply multi-dimensional integration by part on R to obtain

$$\begin{aligned}\int_R \tilde{f}(\mathbf{x})(p(\mathbf{x}) - q(\mathbf{x}))d\mathbf{x} &= (-1)^{d\alpha_d} \int_R (D^{(\alpha_1,\ldots,\alpha_d)}\tilde{f})(\mathbf{x}) \left((I^{(\alpha_d,d)}p)(\mathbf{x}) - (I^{(\alpha_d,d)}q)(\mathbf{x})\right) d\mathbf{x} \\ &= (-1)^{d\alpha_d} \int_\Omega (D^{(\alpha_1,\ldots,\alpha_d)}\tilde{f})(\mathbf{x}) \left((I^{(\alpha_d,d)}p)(\mathbf{x}) - (I^{(\alpha_d,d)}q)(\mathbf{x})\right) d\mathbf{x}.\end{aligned}$$

Hence going back to (24), (25) and (26), with ℓ_d denoting Lebesgue measure, we have

$$\begin{aligned}d_Z(\mathrm{P}, \mathrm{Q}) &\leq \sqrt{2AB_d^a} \sup_{f \in \mathcal{C}^{s_d}(\Omega), \|f\|_{\mathcal{C}^{s_d}(\Omega)} \leq 1} \left(\int_\Omega (D^{(\alpha_1,\ldots,\alpha_d)}\tilde{f})(\mathbf{x}) \left((I^{(\alpha_d,d)}p)(\mathbf{x}) - (I^{(\alpha_d,d)}q)(\mathbf{x})\right) d\mathbf{x} \right)^{a/2} \\ &\leq \sqrt{2AB_d^a} D_d^{a/2} \ell_d(\Omega)^{a/2} \left|\left| I^{(\alpha_d,d)}p - I^{(\alpha_d,d)}q \right|\right|_\infty^{a/2}. \end{aligned} \tag{27}$$

Since p is a density function, we can show by induction that we have, for any $\beta_1, \ldots, \beta_d \in \mathbb{N}$ with $\beta_1 \leq \alpha_1 - 1, \ldots, \beta_d \leq \alpha_d - 1$, $\|D^{(\beta_1,\ldots,\beta_d)}I^{(\alpha_d,d)}p\|_\infty \leq \max(1, \max_{j=1,\ldots,d}(u_j - \ell_j))^{d(\alpha_d-1)}$. Let $E_d = \max(1, \max_{j=1,\ldots,d}(u_j - \ell_j))^{d(\alpha_d-1)}$.

Define the space $\mathcal{C}_{E_d}^{\alpha_d-1}(\Omega)$ as the ball with the norm $\|\cdot\|_{\mathcal{C}^{\alpha_d-1}(\Omega)}$ given by (1), with center 0 and radius E_d. For $\epsilon > 0$, consider a ϵ-covering of this ball, with norm $\|\cdot\|_\infty$, with cardinality N. From Theorem 2.7.1 in van der Vaart and Wellner (2013), we can select N such that

$$\log(N) \leq F_d \epsilon^{-d/(\alpha_d-1)},$$

with a constant F_d that does not depend on ϵ. For each of the N balls that contains one function of the form $I^{(\alpha_d,d)}q$ where q is the density of some $\mathrm{Q} \in \mathcal{P}_\delta$, we consider such a function $I^{(\alpha_d,d)}q$. There are N' such functions that we write $I^{(\alpha_d,d)}q_1, \ldots, I^{(\alpha_d,d)}q_{N'}$. For each $\mathrm{P} \in \mathcal{P}_\delta$ with density p, $I^{(\alpha_d,d)}p$ belongs to $\mathcal{C}_{E_d}^{\alpha_d-1}$ and thus belongs to the same ball as some $I^{(\alpha_d,d)}q_i$ with $i \in \{1, \ldots, N'\}$ and thus $\|I^{(\alpha_d,d)}p - I^{(\alpha_d,d)}q_i\|_\infty \leq 2\epsilon$. Hence from (27) we have, with $Q_i \in \mathcal{P}_\delta$ having density q_i,

$$d_Z(\mathrm{P}, Q_i) \leq \sqrt{2AB_d^a} D_d^{a/2} \ell_d(\Omega)^{a/2} (2\epsilon)^{a/2}.$$

Hence, there are constants G_d, H_d such that for $0 < t \leq 1$,

$$\mathcal{N}(t, \mathcal{P}_\delta, d_Z) \leq G_d e^{H_d t^{-2d/a(\alpha_d-1)}}.$$

Since $d/a(\alpha_d - 1) < 1$, we thus obtain that (23) holds. □

Lemma C.1. *Let Ω be compact and let $b > 0$. There exists an infinitely differentiable function that is zero on $\{\mathbf{t} \in \Omega, d(\mathbf{t}, \delta\Omega) \leq b/2\}$, and one on $\{\mathbf{t} \in \Omega, d(\mathbf{t}, \delta\Omega) \geq b\}$.*

Proof of Lemma C.1. Let g be an infinitely differentiable function with integral one and which support is included in the Euclidean ball of $\mathbb{R}^d$ with center $\mathbf{0}$ and radius $b/4$. Let for $r \geq 0$, $\Omega_r = \{\mathbf{t} \in \Omega, d(\mathbf{t}, \delta\Omega) \geq r\}$. Consider the function h on $\mathbb{R}^d$ defined by, for $\mathbf{t} \in \mathbb{R}^d$,

$$h(\mathbf{t}) = \int_{\mathbb{R}^d} \mathbf{1}_{\{\mathbf{x} \in \Omega_{3b/4}\}} g(\mathbf{x} - \mathbf{t}) d\mathbf{x}.$$

Then h is infinitely differentiable by dominated convergence. For $\mathbf{t} \in \Omega_b$ and $\mathbf{x}$ such that $\|\mathbf{t} - \mathbf{x}\| \leq b/4$, then $\mathbf{x} \in \Omega_{3b/4}$. Hence

$$h(\mathbf{t}) = \int_{\mathbb{R}^d} g(\mathbf{x} - \mathbf{t}) d\mathbf{x} = 1.$$

For $\mathbf{t} \in \Omega$ with $d(\mathbf{t}, \delta\Omega) \leq b/2$, and $\mathbf{x}$ such that $\|\mathbf{t} - \mathbf{x}\| < b/4$, then $d(\mathbf{x}, \delta\Omega) < 3b/4$. Hence

$$h(\mathbf{t}) = \int_{\mathbb{R}^d} 0 d\mathbf{x} = 0.$$

This concludes the proof. □

D ALGORITHMIC DETAILS FOR SECTION 5

D.1 Kernel

We use the kernel:

$$K(\mathrm{P}, \mathrm{Q}) = \sigma^2 \exp\left(-\frac{\|g_{\mathbf{u}}^{\mathrm{P}} - g_{\mathbf{u}}^{\mathrm{Q}}\|^2}{2l^2}\right). \tag{28}$$

Here the parameters are the tuple $\boldsymbol{\theta} = (l, \sigma)$ where $l > 0$ is the length scale and $\sigma > 0$ the scalar variance.

For simplicity we only train a Gaussian process with zero mean function. This does not prevent the GP to reach satisfying levels of EVS/accuracy as illustrated in the experiments.

The RBF kernel uses a similar form:

$$K_{\mathrm{RBF}}(\mathbf{x}, \mathbf{y}) = \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2l^2}\right). \tag{29}$$

Hence the RBF kernel can be applied to vectors (or matrices representing images), but cannot handle general point clouds.

D.2 Sinkhorn's Algorithm

Sinkhorn's algorithm is an iterative algorithm that takes advantage of *approximately good* solutions. Hence, the dual variables are re-used from one optimization step to the other. Using small steps guarantees that the initialization is not far away from the optimum. It allows the algorithm to benefit from a significant speed-up.

D.3 L-BFGS

We apply the Limited Memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS), which is an order 2 method, to enjoy faster convergence than order-1 methods such as Gradient Descent. The dominant cost of the algorithm is induced by the size of the support $\mathbf{u}$ and by the dimension of the points $\mathbf{x}_i \in \mathbb{R}^d$ since $(\sigma, l) \in \mathbb{R}^2$. The total dimension of search space is hence $nd + 2$.

We select the optimal stepsize at each iteration with a *zoom line search* (Algorithm 3.6 of Nocedal and Wright (1999), pg. 59-61. Tries cubic, quadratic, and bisection methods of zooming).

Number of clouds	Cloud size	Sinkhorn with $\lvert\mathbf{u}\rvert = 6$	Sinkhorn with $\lvert\mathbf{u}\rvert = 12$	MMD
$n = 50$	$m = 100$	0.009s	0.001s	0.001s
$n = 100$	$m = 100$	0.013s	0.011s	0.005s
$n = 100$	$m = 400$	0.007s	0.021s	0.055s
$n = 400$	$m = 400$	0.018s	0.059s	0.683s
$n = 400$	$m = 625$	0.026s	0.088s	1.681s
$n = 1000$	$m = 625$	0.064s	0.147s	10.834s
$n = 1000$	$m = 1000$	0.090s	0.158s	14.207s

Table 3: Runtime cost of Sinkhorn $\mathbf{u}$-Kernel (ours) against MMD. The cost reported corresponds to the overall process: computation of regularized OT plan and of the kernel for Sinkhorn $\mathbf{u}$, and computation of MMD distance for MMD. Clouds are in dimension $d = 2$.

D.4 Runtime Cost Against MMD

We choose the MMD distance with RBF as inner kernel:

$$\mathrm{MMD}^2(\mathrm{P},\mathrm{Q}) = \mathbb{E}_{\mathrm{P}}(\mathrm{K}_{\mathrm{RBF}}(\mathbf{X},\mathbf{X}')) + \mathbb{E}_{\mathrm{Q}}(\mathrm{K}_{\mathrm{RBF}}(\mathbf{Y},\mathbf{Y}')) - 2\mathbb{E}_{\mathrm{P},\mathrm{Q}}(\mathrm{K}_{\mathrm{RBF}}(\mathbf{X},\mathbf{Y})), \tag{30}$$

with $\mathbf{X}, \mathbf{X}' \sim \mathrm{P}$, $\mathbf{Y}, \mathbf{Y}' \sim \mathrm{Q}$, with $\mathbf{X}$, $\mathbf{X}'$ $\mathbf{Y}$, $\mathbf{Y}'$ independent.

The MMD distance is turned into a kernel with an additional parameter σ:

$$K_{\mathrm{MMD}}(\mathrm{P},\mathrm{Q}) = \sigma^2 \exp\left(-\mathrm{MMD}^2(\mathrm{P},\mathrm{Q})\right). \tag{31}$$

The kernel in (31) is universal (see Theorem 2.2 of Christmann and Steinwart (2010) for example).

For a fair comparison the Sinkhorn $\mathbf{u}$-Kernel and MMD kernel are benchmarked on the same hardware under '@jax.jit' compiled code to benefit from GPU acceleration.

We report runtime results in Table 3. The clouds all share the same coordinates (but not the same weights). The pairwise distances between points of the clouds are pre-computed to speed-up both MMD and Sinkhorn iterations. We notice that Sinkhorn takes advantage of pre-computing the low dimension embeddings in dimension $|\mathbf{u}| = 6$, independent of the cloud size. We chose $\epsilon = 10^{-2}$ as regularization parameter. The points $\mathbf{u}$ are sampled uniformly in the square $[0,1]^2$, while points from the clouds P_i are a discretization of the square $[0,1]^2$ with equally spaced coordinates. Our Sinkhorn $\mathbf{u}$-kernel shows a speed-up of up to a factor 100 compared to the MMD one.

E RESULT DETAILS FOR SECTION 5

E.1 Visualizing Dual Variables

In Figure 4 we introduce an example with two distributions P and Q obtained by taking finite samples from isotropic Gaussians. For P we sample 30 points from $\mathcal{N}([-2,-2], 0.4)$ and for Q we sample 50 points from $\mathcal{N}([-1,1], 0.3)$. We choose for $\mathbf{u}$ a finite sample of size 120 from the unit ball $\mathbb{B}(\mathbf{0}, 1)$.

We plot both the distributions and the values taken by g^{P} and g^{Q} respectively, by sorting dual variables arbitrarily by increasing error of $|g_i^{\mathrm{P}} - g_i^{\mathrm{Q}}|$.

E.2 Dependence on Dimension

In figure 5 we illustrate the dependence of the dimension of ambient space d on the convergence speed. The reference measure $\mathbf{u}$ is chosen to be 128 points sampled uniformly in unit ball. The task consists of $m \sim \mathbb{U}([100; 200])$ sampled at random from a Gaussian whose center is also sampled uniformly at random in range $\mu_i \sim \mathbb{U}([-10, 10])$. The regression task is the prediction of the mean $Y_i = \mu_i$ of each Gaussian from the finite sample. We use a Support Vector Regression machine (SVR) to perform the task. We report the Normalized mean Square Error by dividing by the dimension d to allow fair comparison on the same scale. We see that convergence speed is similar.

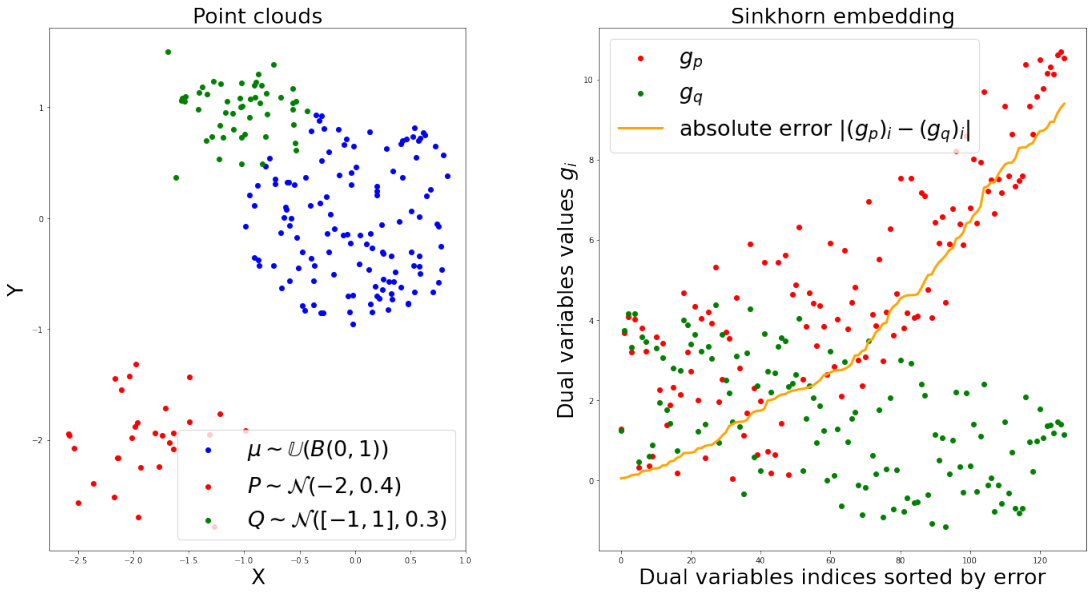

Figure 4: Vizualization of dual variables g^{P} and g^{Q}. For P we sample 30 points from $\mathcal{N}([-2,-2], 0.4)$ and for Q we sample 50 points from $\mathcal{N}([-1,1], 0.3)$. We chose for $\mathbf{u}$ a finite sample of size 120 from the unit ball $\mathbb{B}(\mathbf{0}, \mathbf{1})$.

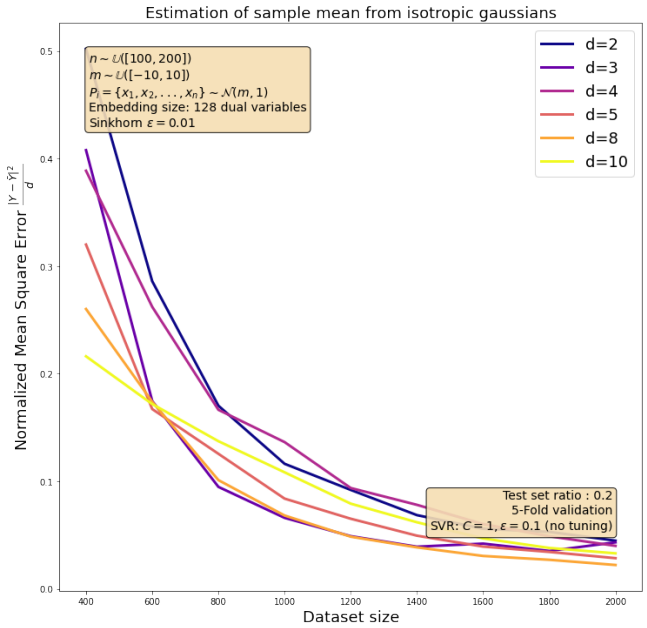

Figure 5: Normalized Mean Square Error as function of dimension d and train set size n for the synthetic task of predicting the mean $\mu_i \sim \mathbb{U}([-10, 10])$ of a Gaussian from finite sample of size $m \in [100; 200]$.

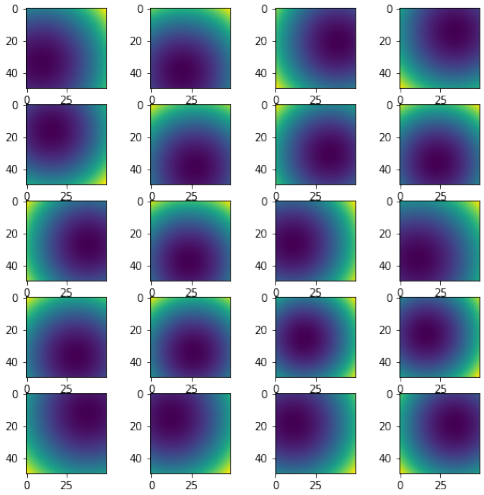

Figure 6: Plot of the dual variables $g_{\mathbf{u}}^{\mathrm{P}}$ for $5 \times 4 = 20$ distributions P_i from the toy example of Section 5.1. The position of the center of each "blob" (i.e. mean of the Gaussian) can be clearly seen by looking at the dual variables.

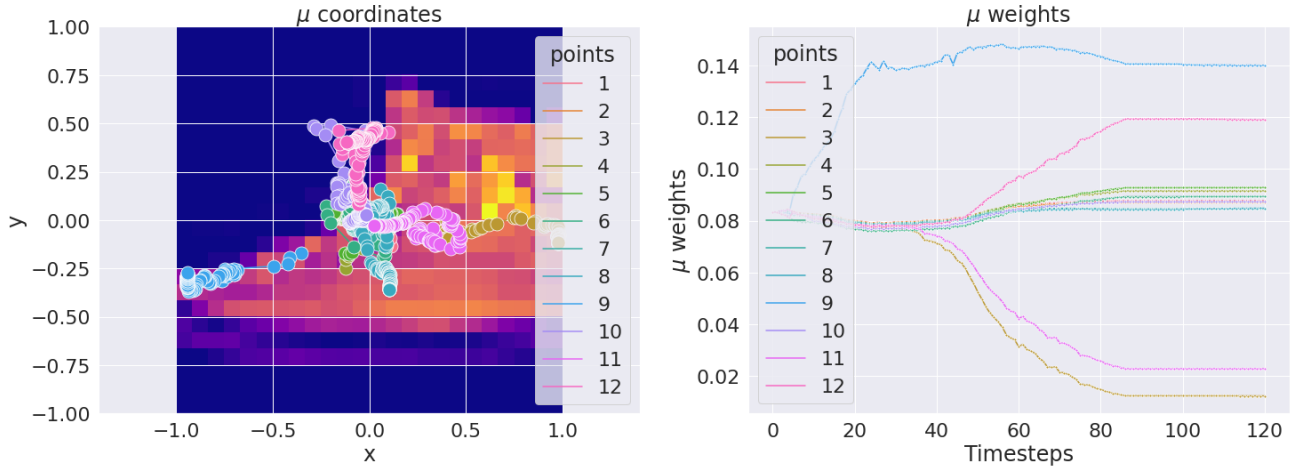

Figure 7: Evolution of $\mathbf{x}_i$'s and w_i's for $\mathbf{u}$ in the "sneakers" versus "sandals" task.

E.3 Toy Dataset

All clouds are centered and rescaled so the overall dataset (obtained by merging all clouds) has zero mean and unit variance across all dimensions.

We study a discretization of $\mathbf{u}$ in the experiment of Section 5.1. We choose $\mathbf{u}$ to be a discretization of the input space $[0, 1]^2$ as a 50×50 grid. The density is chosen uniform over this discretization of 2500 points. Hence each regularized optimal transportation plan is between a Gaussian and an uniform measure over the square $[0, 1]^2$. In this case the dual variable $g_{\mathbf{u}}^{\mathrm{P}}$ can be visualized as an image in definition 50×50.

For 20 train examples, we plot the images $g_{\mathbf{u}}^{\mathrm{P}}$ in Figure 6. We see that all those images appear "blurry", which shows the role of regularization in OT. Moreover those images seem to correspond to a "blob" whose coordinates correspond to the ones of the clouds P_i. This figure helps to understand what the dual variables exactly look like in toy examples.

E.4 Mnist and Fashion-Mnist Datasets

For RBF kernel, the images are normalized so that the pixel intensity lies in $[0, 1]$ range.

Figure 7 illustrates the evolution of $\mathbf{x}_i$'s and w_i's for $\mathbf{u}$ in the case of an image of shoe from Fashion-Mnist.

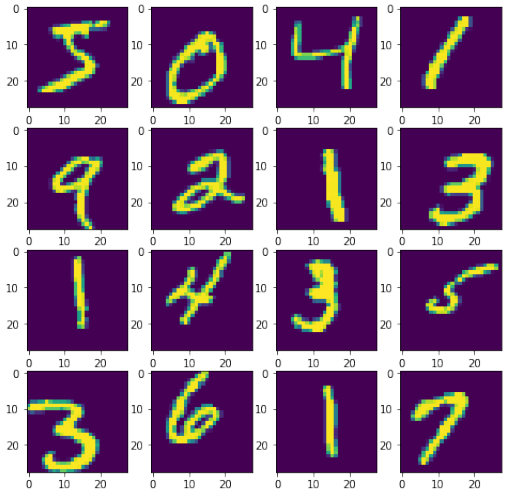

Figure 8: Mnist images with random affine transformations: translation uniformly at random in range $[-6, 6]$ pixels.

Dataset	$\mathbf{u}$-Sinkhorn (ours)	RBF	Sliced Wasserstein (as reported by Kolouri et al. (2016))
UIUC Textures	87.2	87.3	88 ± 1
Mnist (1300 examples)	92.50	92.46	N/A

Table 4: Validation accuracy of C-SVM with different kernels on GLCM embeddings of UIUC texture dataset (Lazebnik et al., 2005), and 1300 example of Mnist (10 classes), with 5 folds cross-validation.

E.4.1 Sensivity to Random Affine Transformations

In Figure 8 we plot a set of Mnist images on which random affine transformations have been applied. We follow the protocol of Meunier et al. (2022) and we sample a translation uniformly at random in range $[-6, 6]$ pixels, and a rotation uniformly at random in range $[-\frac{\pi}{3}, \frac{\pi}{3}]$ rads.

In Figure 9 we study the influence of random affine transformations in dual variable space $g^{\mathrm{P}}_{\mathbf{u}}$, versus pixel space. In this experiment the reference measure $\mathbf{u}$ is chosen to have full support in dimension $28 \times 28 = 784$. The reference measure is chosen uniform on the pixel space. The images are processed as clouds of $28 \times 28 = 784$ pixels in dimension 2. The regularization factor is chosen to be $\epsilon = 10^{-2}$.

We see that dual variables are less sensitive to translations than pixels. In the third row of Figure 9 the dual variables are modified in a way that hints the direction and amplitude of translation, whereas in the fourth row the translation in pixel space has major consequences on the image and exhibits a huge Euclidean norm. This shows that μ Sinkhorn dual variables are better tailored to handle translates than conventional Euclidean metrics, thanks to the properties of OT in translations.

E.5 C-SVM Results

E.5.1 UIUC Dataset

We report here the results of classification with C-SVM on the University of Illinois Urbana Champaign (UIUC) texture dataset (Lazebnik et al., 2005), using the same protocol as Kolouri et al. (2016). Samples are shown in Figure 10. The dataset contains 25 different classes of texture on a total of 1000 images (only 40 images per class).

The Gray Level Co-occurrences Matrices (GLCM) is computed with the Scikit-image library (Van der Walt et al., 2014). The images are illustrated in Figure 11. The parameter γ of the SVM is obtained by following the "scale" policy of Scikit-learn library, applied on normalized features. We apply a grid search in logspace on the parameter C of SVM, ranging from 10^{-1} to 10^{3}. The optimal parameter is selected by the highest average accuracy in 5-fold cross validation.

We compare the result against the RBF kernel applied to the raw (unprocessed) pixels. The results are reported in Table 4. We see that the RBF kernel and our kernel have similar accuracies. We note that our implementation of the RBF kernel provides an higher accuracy for it than the one reported in Kolouri et al. (2016).

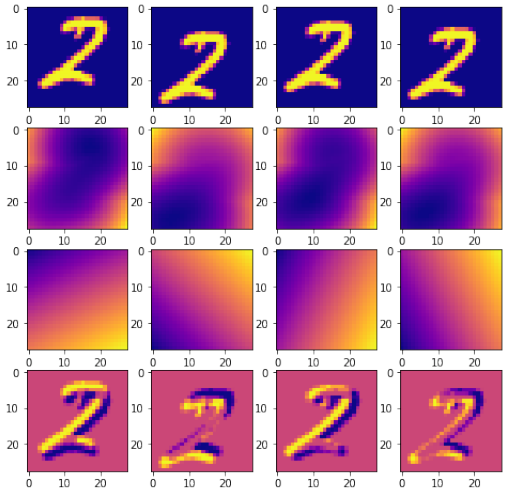

Figure 9: Visualization of translations on dual variables for an Mnist image for $|\mathbf{u}| = 28 \times 28 = 784$. **Top row**: original image $\mathbf{x}$ with different affine transformations $\tilde{\mathbf{x}}$. **Second row**: Dual variables of translated images $g(\tilde{\mathbf{x}})$. **Third row**: pixel-wise difference between the dual variables of original (non modified) image and translated images $g(\mathbf{x}) - g(\tilde{\mathbf{x}})$. **Fourth row**: pixel-wise difference between original image and translated image $\mathbf{x} - \tilde{\mathbf{x}}$, in pixel space. We see that any translation has major impact in the pixel space, but only mild consequences in the dual variables space. Moreover the map $g(\mathbf{x}) - g(\tilde{\mathbf{x}})$ hints the nature of the translation, whereas $\mathbf{x} - \tilde{\mathbf{x}}$ is harder to interpret.

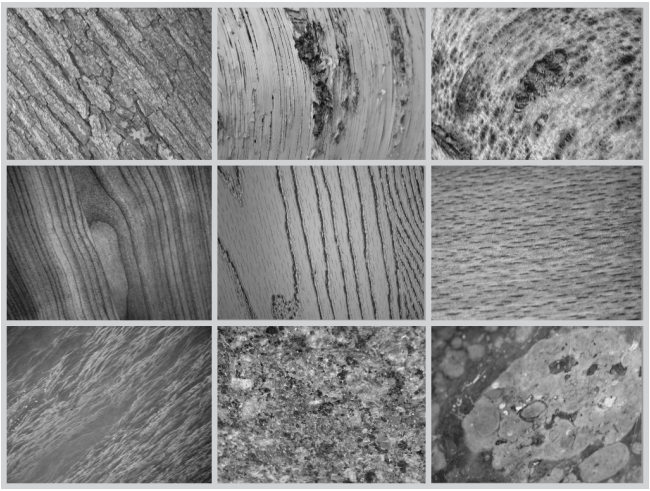

Figure 10: Random samples from the UIUC texture dataset.

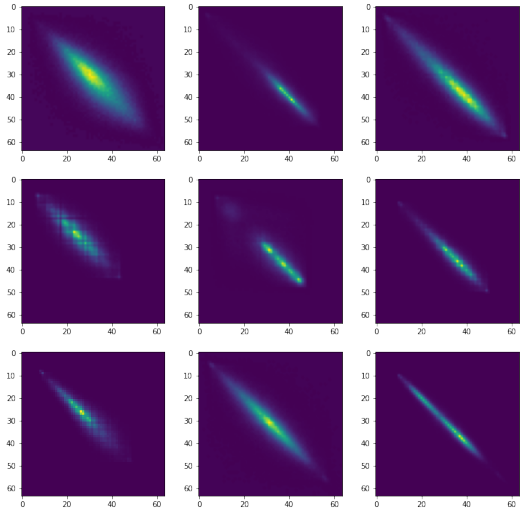

Figure 11: Gray Level Co-occurrences Matrix (GLCM) of random samples of UIUC texture dataset, which can be re-normalized into a discrete 2D distribution.

E.5.2 Mnist C-SVM

We choose a measure $\mathbf{u}$ with full support as in Section E.4. We select 1300 examples at random in the Mnist train set from all 10 classes, and we apply the protocol of Section E.5.1. The results of the best estimator found with 5 fold cross-validation are reported on an independent test set of size 1000 in Table 4. Again, we have similar results as the RBF kernel.

E.6 Hardware and Code

All the experiments were run on the publicly available GPU Colab hardware.

The code can be found on an anonymous repository: https://anonymous.4open.science/r/SinkhornMuGP-D37E/README.md.