Journal of Machine Learning Research 25 (2024) 1-39 Submitted 11/22; Revised 4/24; Published 5/24

Characterization of translation invariant MMD on $\mathbb{R}^d$ and connections with Wasserstein distances

Thibault Modeste MODESTE@MATH.UNIV-LYON1.FR
Institut Camille Jordan
Université Claude Bernard Lyon 1
CNRS UMR 5208, F-69622 Villeurbanne, France

Clément Dombry CLEMENT.DOMBRY@UNIV-FCOMTE.FR
Université de Franche-Comté,
CNRS, LmB (UMR 6623),
F-25000 Besançon, France

Editor: Zaid Harchaoui

Abstract

Kernel mean embeddings and maximum mean discrepancies (MMD) associated with positive definite kernels are important tools in machine learning that allow to compare probability measures and sample distributions. We provide a full characterization of translation invariant MMDs on $\mathbb{R}^d$ that are parametrized by a spectral measure and a semi-definite positive symmetric matrix. Furthermore, we investigate the connections between translation invariant MMDs and Wasserstein distances on $\mathbb{R}^d$. We show in particular that convergence with respect to the MMD associated with the Energy Kernel of order $\alpha \in (0, 1)$ implies convergence with respect to the Wasserstein distance of order $\beta < \alpha$. We also provide examples of kernels metrizing the Wasserstein space of order $\alpha \geq 1$. A short numerical experiment illustrates our findings in the framework of the one-sample-test.

Keywords: Reproducing Kernel Hilbert Space, Kernel Mean Embedding, Maximum Mean Discrepancy, translation invariance, Wasserstein distance.

1 Introduction

Background. Many problems in statistics and machine learning require comparing several probability measures and/or sample distributions: goodness-of-fit testing compares a sample distribution to a reference distribution (Chwialkowski et al., 2016); two-sample testing compares two sample distributions (Gretton et al., 2012); independence testing compares a joint distribution to a product distribution (Gretton et al., 2005); generative model fitting compares the distributions of real and fake data (Dziugaite et al., 2015; Sutherland et al., 2017). The different methods proposed in these references all rely on the important notion of Maximum Mean Discrepancy (MMD).

MMDs are semi-metrics between probability measures and their definition relies on the theory of Reproducing Kernel Hilbert Spaces (RKHS) and Kernel Mean Embeddings (KME). Given a symmetric positive definite kernel k and its associated RKHS $\mathcal{H}_k$, the KME is a map $\mu \mapsto K(\mu)$ that assigns a function $K(\mu) \in \mathcal{H}_k$ to each signed measure μ in a suitable subspace $\mathcal{M}_k$ (defined in Equation (3) below). The corresponding MMD between

two measures μ and ν is defined as the RKHS distance between their embeddings, i.e. $d_k(\mu,\nu) := \|K(\mu) - K(\nu)\|_{\mathcal{H}_k}$. When the KME is injective, in which case the kernel is called characteristic, the MMD defines a proper distance that can be used to compare probability measures and/or sample distributions. Due to their theoretical tractability and computational efficiency, KMEs and MMDs are widely used in many areas of machine learning. We refer to Smola et al. (2007) for an overview on distribution Hilbert space embeddings and their applications in machine learning.

Related works. In the last decade, an important line of research has focused on theoretical properties of KMEs and MMDs. Sriperumbudur et al. (2010) and Sriperumbudur et al. (2011) consider conditions ensuring that a kernel is characteristic, meaning that the associated kernel mean embedding is injective. In the particular case of invariant kernels on $\mathbb{R}^d$, the question can be addressed thanks to Fourier analysis and the kernel is shown to be characteristic if and only if the spectral measure has a full support on $\mathbb{R}^d \setminus \{0\}$ (Sriperumbudur et al., 2010, Theorem 9). Already considered in the latter references, the question of whether MMD can metrize weak convergence of distributions has been fully addressed by Simon-Gabriel and Schölkopf (2018) and Simon-Gabriel et al. (2023). The main result is that, for a continuous kernel with RKHS included in the space of continuous functions vanishing at infinity, the MMD metrizes weak convergence if and only if the kernel is characteristic.
Although weak convergence is an important concept and a minimal requirement, this notion of convergence is very weak, as its name suggests. A stronger notion of convergence, which has turned out to be very useful and successful in machine learning, is the convergence in Wasserstein space. The Wasserstein distance is related to optimal transport (Villani, 2008) and was recently used successfully in statistics and machine learning, as described in the recent monograph by Panaretos and Zemel (2020) or survey by Montesuma et al. (2023) – see also the numerous references therein. To cite only a few, optimal transport is used in learning algorithms (Frogner et al., 2015), signal processing (Kolouri et al., 2017), generative models (Lei et al., 2019)..., algorithmic fairness Si et al. (2021)... One of the main question addressed in the present paper is whether a MMD can metrize the Wasserstein space. We show that the answer is positive and that the use of unbounded kernels is needed. In a slightly different perspective, Auricchio et al. (2020) and Vayer and Gribonval (2023) establish non-asymptotic inequalities relating MMD and Wasserstein distances.

Main contributions. Our main findings are the following:

- The class of translation invariant MMD on $\mathbb{R}^d$ is characterized by a spectral measure and a symmetric positive semi-definite matrix (Corollary 5). Extending the results of Sriperumbudur et al. (2010), we provide an explicit formula for the MMD in terms of Fourier transform (Proposition 7) and provide a necessary and sufficient condition for the kernel to be characteristic over probability measures (Proposition 9).

- Strong connections between Energy kernels and Wasserstein distances are established (Theorem 13) in Section 3.3. More precisely, for $\alpha \in (0,1)$, we denote by d_α the MMD associated with the energy kernel of order α and by W_α the Wasserstein distance of order α; we prove that convergence of probability measures with respect to W_α implies

convergence with respect to d_β for all $0 < \beta \leq \alpha$ and, conversely, that convergence with respect to d_α implies convergence with respect to W_β for all $0 < \beta < \alpha$.

- We exhibit new families of kernels that metrize the Wasserstein spaces of order $\alpha \geq 1$ (Theorem 14) in Section 3.4.

- We provide non-asymptotic inequalities between W_1 and d_α for tight subsets of probability measures (Proposition 17) in Section 3.5.

Structure of the paper. Section 2 gathers some necessary material on reproducing kernel Hilbert spaces, kernel mean embeddings and maximum mean discrepancies in Section 2.1 and some important results about equivalent kernels and their characterization via variograms closely related to Sejdinovic et al. (2013) in Section 2.2. Original results on the characterization of translation invariant MMDs on $\mathbb{R}^d$ are presented in Section 2.3, Corollary 5 being the main new result. Next we focus in Section 3 on the connections between MMDs and Wasserstein distances. Some background on Wasserstein spaces is presented in Section 3.1 and some preliminary results in Section 3.2. The relationships between MMDs associated with the Energy Kernel of order $\alpha < 1$ and Wasserstein distances of order $\alpha < 1$ are investigated in Section 3.3. New families of kernels metrizing the Wasserstein spaces of order $\alpha \geq 1$ are studied in Section 3.4. Finally, some nonasymptotic inequalities relating MMDs and Wasserstein distances are established in Section 3.5. All the proofs are postponed to Section 6.

Notation. In Sections 2.1 and 2.2, $(\mathcal{X}, \mathcal{B})$ denotes a measurable space and $\mathcal{M}$ (resp. $\mathcal{P}$) the sets of signed measures (resp. probability measures) on $(\mathcal{X}, \mathcal{B})$. The total variation measure of a signed measure $\mu \in \mathcal{M}$ is denoted by $|\mu|$. In the rest of the paper, we take $\mathcal{X} = \mathbb{R}^d$ endowed with its Borel sigma-field and $\mathcal{M}$ (resp. $\mathcal{P}$) denotes the space of Borel signed measures (resp. probability measures) on $\mathbb{R}^d$. We equip $\mathbb{R}^d$ with its canonical Euclidean structure and we write $\|x\|$ and $x \cdot y$ respectively for the norm of x and the inner product between x and y. The characteristic function of $\mu \in \mathcal{M}$ is denoted by $\hat{\mu}$ and defined by

$$\hat{\mu}(\xi) = \int_{\mathbb{R}^d} \mathrm{e}^{i(x-y)\cdot\xi}\, \mu(\mathrm{d}\xi), \quad \xi \in \mathbb{R}^d.$$

For $\alpha > 0$, we define $\mathcal{M}^\alpha = \left\{\mu \in \mathcal{M} \colon \int_{\mathbb{R}^d} \|x\|^\alpha\, |\mu|(\mathrm{d}x) < \infty\right\}$ and $\mathcal{P}^\alpha = \mathcal{M}^\alpha \cap \mathcal{P}$ as the set of signed measures (resp. probability measures) with finite moment of order α. The minimum between $a, b \in \mathbb{R}$ is denoted by $a \wedge b$.

2 Kernel Mean Embeddings and Maximum Mean Discrepancy

This section is dedicated to the characterization of translation invariant maximum mean discrepancies (MMD) where we show that the underlying kernel need not be translation invariant in order that the MMD be. In Section 2.3, such kernels are characterized by a semidefinite positive matrix Σ and a possibly infinite measure ν on $\mathbb{R}^d \setminus \{0\}$, see Corollary 5. Necessary material on MMDs and variograms are introduced in Sections 2.1 and 2.2 respectively.

2.1 Preliminary: Hilbert space embedding of measures

We present some basic elements of the theory of Reproducing Kernel Hilbert Spaces (RKHS), Kernel Mean Embeddings (KME) and Maximum Mean Discrepancy (MMD). For more details, the reader could refer to Berlinet and Thomas-Agnan (2004), Smola et al. (2007) or Steinwart and Christmann (2008, Section 4).

Reproducing Kernel Hilbert Space (RKHS). Let $\mathcal{X}$ be an arbitrary space and $\mathcal{F}(\mathcal{X},\mathbb{R})$ denote the space of real valued function on $\mathcal{X}$. A function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is called a kernel if it is symmetric and positive definite. The latter condition means that

$$\sum_{1\leq i,j\leq n} a_i a_j k(x_i,x_j)\geq 0, \quad \text{for all } n\geq 1,\, x_1,\dots,x_n\in\mathcal{X},\, a_1,\dots,a_n\in\mathbb{R}.$$

An Hilbert space $\mathcal{H}\subset\mathcal{F}(\mathcal{X},\mathbb{R})$ is called a RKHS if, for all $x\in\mathcal{X}$, the evaluation map $f\mapsto f(x)$ is continuous. By the Riesz representation theorem, there exists, for all $x\in\mathcal{X}$, a unique representer $K(x)\in\mathcal{H}$ such that

$$\forall f\in\mathcal{H},\ f(x)=\langle f,K(x)\rangle.$$

Then, the function $k(x,y)=\langle K(x),K(y)\rangle$ is a kernel and is called the *reproducing kernel* of $\mathcal{H}$ because of the following *reproducing property*: for all $x\in\mathcal{X}$, $k(x,\cdot)\in\mathcal{H}$ and

$$\forall f\in\mathcal{H},\ f(x)=\langle f,k(x,\cdot)\rangle. \tag{1}$$

In particular, we have $K(x)=k(x,\cdot)$. The reproducing kernel characterizes the RKHS. Conversely, Aronszajn's theorem states that for any kernel k on $\mathcal{X}\times\mathcal{X}$, there exists an unique RKHS, noted $\mathcal{H}_k$, with reproducing kernel k.

Kernel Mean Embedding (KME). We assume that $(\mathcal{X},\mathcal{B})$ is a measurable space and the kernel k is measurable on $\mathcal{X}\times\mathcal{X}$. The space of signed finite measures (resp. probability measures) μ on $(\mathcal{X},\mathcal{B})$ is denoted by $\mathcal{M}$ (resp. $\mathcal{P}$) and the total variation measure of μ by $|\mu|$. The reproducing kernel property (1) readily implies that for any finite discrete measure $\mu=\sum_{i=1}^n a_i\delta_{x_i}$, the function $K(\mu)=\sum_{i=1}^n a_i K(x_i)\in\mathcal{H}_k$ satisfies

$$\forall f\in\mathcal{H}_k,\ \langle f,K(\mu)\rangle=\int_{\mathcal{X}} f(x)\,\mu(\mathrm{d}x). \tag{2}$$

The KME extends this property to the class of measures

$$\mathcal{M}_k=\Big\{\mu\in\mathcal{M}\colon \int_{\mathcal{X}}\sqrt{k(x,x)}\,|\mu|(\mathrm{d}x)<+\infty\Big\}. \tag{3}$$

More precisely, for all $\mu\in\mathcal{M}_k$, the RKHS $\mathcal{H}_k$ is included in $\mathcal{L}^1(\mu)$ and there exists a unique $K(\mu)\in\mathcal{H}_k$ satisfying Equation (2) –see e.g. Steinwart and Christmann (2008, Theorem 4.26). The map $K:\mathcal{M}_k\to\mathcal{H}_k$ is the KME associated with k. The measure $\mu\in\mathcal{M}_k$ is represented in the RKHS by the vector $K(\mu)$ in the same way as the point x (identified with the Dirac measure δ_x) is represented by $K(x)$.

Maximum Mean Discrepancy (MMD). To compare two measures in $\mathcal{M}_k$, we compare their images in $\mathcal{H}_k$ under the KME: the MMD is defined by

$$d_k(\mu,\nu) = \|K(\mu) - K(\nu)\|_{\mathcal{H}_k}, \quad \mu,\nu \in \mathcal{M}_k.$$

The reproducing kernel property (2) - applied twice - implies

$$\begin{aligned} d_k^2(\mu,\nu) &= \langle K(\mu-\nu), K(\mu-\nu)\rangle_{\mathcal{H}_k} \\ &= \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y). \end{aligned} \tag{4}$$

For sample distributions $\mu_n = n^{-1}\sum_{k=1}^n \delta_{x_k}$ and $\nu_m = m^{-1}\sum_{l=1}^m \delta_{y_l}$, the MMD reduces to

$$d_k^2(\mu_n,\nu_m) = n^{-2}\sum_{1\le k,l\le n} k(x_k,x_l) + m^{-2}\sum_{1\le k,l\le m} k(y_k,y_l) - 2n^{-1}m^{-1}\sum_{1\le k\le n}\sum_{1\le l\le m} k(x_k,y_l)$$

and is easily computed (for sample of reasonable size). Furthermore, using the dual representation of the Hilbert norm in $\mathcal{H}_k$, the MMD can also be expressed as

$$d_k(\mu,\nu) = \sup_{\|f\|_{\mathcal{H}_k}\le 1}\left|\int_{\mathcal{X}} f\,\mathrm{d}\mu - \int_{\mathcal{X}} f\,\mathrm{d}\nu\right|. \tag{5}$$

This form corresponds to an Integral Probability Metric (Müller, 1997) with test functions belonging to the unit ball of the RKHS.

2.2 Preliminary: variograms and equivalent kernels

Given different measurable kernels on $\mathcal{X}\times\mathcal{X}$, one can wonder in which case the associated MMDs are equal. This question is investigated in Sejdinovic et al. (2013), where the authors consider the relationships between kernels and distances of negative type (Section 4 in Sejdinovic et al. 2013) and the corresponding MMDs and energy distances (Section 5 and Theorem 22 in Sejdinovic et al. (2013)). We adopt here a slightly different terminology more related to geostatistics (see Remark 2 below).

Variogram. We call *variogram* associated with a kernel k the function

$$\rho(x,y) = \frac{1}{2}k(x,x) + \frac{1}{2}k(y,y) - k(x,y), \quad x,y\in\mathcal{X}.$$

Clearly, the variogram ρ is a symmetric function on $\mathcal{X}\times\mathcal{X}$ and vanishes on the diagonal, i.e. $\rho(x,x) = 0$ for all $x\in\mathcal{X}$. Furthermore, according to Berg et al. (1984, Lemma 2.1 p.74), the variogram is a conditionally negative definite function on $\mathcal{X}\times\mathcal{X}$, meaning that

$$\sum_{1\le i,j\le n} a_i a_j \rho(x_i,x_j) \le 0$$

for all $x_1,\dots,x_n\in\mathcal{X}$ and $a_1,\dots,a_n\in\mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$. See Berg et al. (1984, Chapter 3) for more details on the strong relationships between positive definite and negative

definite functions. Note that Sejdinovic et al. (2013, Section 4) uses the terminology *semi-metric of negative type induced by k* instead of *variogram associated with k* but the objects considered are the same.

Equivalent kernels. Two measurable kernels k_1 and k_2 on $\mathcal{X} \times \mathcal{X}$ are called *equivalent* if

$$\mathcal{M}_{k_1} = \mathcal{M}_{k_2} \quad \text{and} \quad d_{k_1}(\mu, \nu) = d_{k_2}(\mu, \nu) \qquad \text{for all } \mu, \nu \in \mathcal{M}_{k_1} \cap \mathcal{P}. \tag{6}$$

Let us stress that, in this definition, the equality of MMDs is required for probability measures only. The following statement is a reformulation of Sejdinovic et al. (2013, Theorem 22).

Proposition 1. *(Sejdinovic et al., 2013, Theorem 22) Two measurable kernels are equivalent if and only if they have the same variogram.*

In order to have a form of uniqueness, we consider the notion of a normalized kernel. Fix an arbitrary origin $o \in \mathcal{X}$. A kernel k is said to be normalized (with origin o) if

$$k(x, o) = k(o, x) = 0 \quad \text{for all } x \in \mathcal{X}.$$

For any kernel k on $\mathcal{X} \times \mathcal{X}$, there exists a unique kernel k_0 which is equivalent to k and normalized (with origin o); it is given by

$$k_0(x, y) = k(x, y) - k(x, o) - k(o, y) + k(o, o). \tag{7}$$

Denoting by ρ the common variogram of k and k_0, one can easily check that k_0 can be written as

$$k_0(x, y) = \rho(x, o) + \rho(o, y) - \rho(x, y). \tag{8}$$

Remark 2. *The term variogram comes from the theory of stochastic processes and geostatistics (Cressie, 1993). Let $(B(x))_{x \in \mathcal{X}}$ be a square integrable stochastic process on $\mathcal{X}$. The covariance function is a symmetric and positive definite function on $\mathcal{X} \times \mathcal{X}$, that is*

$$k(x, y) = \mathrm{Cov}(B(x), B(y))$$

is a kernel. The associated variogram

$$\begin{aligned} \rho(x, y) &= \frac{1}{2}k(x, x) + \frac{1}{2}k(y, y) - k(x, y) \\ &= \frac{1}{2}\mathrm{Var}(B(y) - B(x)) \end{aligned}$$

corresponds to half the variance of the increment $B(y) - B(x)$. Given an origin $o \in \mathcal{X}$, the process $(B(x) - B(o))_{x \in \mathcal{X}}$ of increments at the origin has covariance

$$\begin{aligned} k_0(x, y) &= \mathrm{Cov}(B(x) - B(o), B(y) - B(o)) \\ &= k(x, y) - k(x, o) - k(o, y) + k(o, o), \end{aligned}$$

which is the unique normalized kernel with variogram ρ. We focus next on the class of Gaussian processes. If the process B is centered and Gaussian, then its distribution is fully characterized by its covariance function. It follows that, given an origin o and a variogram ρ, there exists a (unique in distribution) centered Gaussian process $B = (B(x))_{x \in \mathcal{X}}$ such that

$$\mathrm{Var}(B(y) - B(x)) = 2\rho(x, y) \quad \textit{and} \quad B(o) = 0 \ \textit{a.s.}$$

The process B is called the Gaussian process with variogram ρ and origin o.

2.3 Translation invariant MMD on $\mathbb{R}^d$

In the rest of the paper, we consider $\mathcal{X} = \mathbb{R}^d$ endowed with its Borel sigma-field.

Translation invariant MMD. We study translation invariant MMDs as in the following definition. For $h \in \mathbb{R}^d$, we note $\tau_h : \mathbb{R}^d \to \mathbb{R}^d$ the translation defined by $\tau_h(x) = x + h$ and by $\tau_{h\#}\mu$ the image (pushforward) of a measure μ on $\mathbb{R}^d$.

Definition 3. *The MMD associated with a kernel k on $\mathbb{R}^d \times \mathbb{R}^d$ is said to be translation invariant if, for all $\mu, \nu \in \mathcal{M}_k$ and $h \in \mathbb{R}^d$, $\tau_{h\#}\mu, \tau_{h\#}\nu \in \mathcal{M}_k$ and*

$$d_k(\tau_{h\#}\mu, \tau_{h\#}\nu) = d_k(\mu, \nu). \tag{9}$$

Clearly, if the kernel k is translation invariant, i.e. satisfies

$$k(x + h, y + h) = k(x, y), \quad \text{for all } x, y, h \in \mathbb{R}^d,$$

then the associated MMD is invariant. Such kernels are of the form $k(x, y) = \psi(x - y)$ with ψ a positive definite function and are always bounded since $|k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(y, y)} = \psi(0)$. Under a continuity assumption, the class of translation invariant kernels is studied in Sriperumbudur et al. (2010, Sections 2 and 3.2) where Bochner Theorem is shown to imply the existence of a finite symmetric nonnegative Borel measure Λ on $\mathbb{R}^d$ such that

$$k(x, y) = \int_{\mathbb{R}^d} \mathrm{e}^{i(x-y)\cdot\xi}\, \Lambda(\mathrm{d}\xi). \tag{10}$$

Furthermore the associated MMD is expressed, for $\mu, \nu \in \mathcal{M}$, as

$$d_k^2(\mu, \nu) = \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2 \Lambda(\mathrm{d}\xi) = \|\hat{\mu} - \hat{\nu}\|^2_{L^2(\Lambda)}. \tag{11}$$

Example 1. *When $\mathcal{X} = \mathbb{R}^d$, the Gaussian kernel is the most popular one in machine learning and is defined by*

$$k(x, y) = \exp(-\|x - y\|^2/2), \quad x, y \in \mathbb{R}^d.$$

This kernel being bounded, we have $\mathcal{M}_k = \mathcal{M}$ and, using Fourier theory, the MMD can be rewritten as

$$d_k^2(\mu, \nu) = \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2\ \varphi(\xi)\, \mathrm{d}\xi,$$

where φ denotes the multivariate standard Gaussian density on $\mathbb{R}^d$. Simon-Gabriel et al. (2023, Theorem 7) states that this MMD metrizes weak convergence on $\mathcal{P}$.

Characterization of translation invariant MMDs. Interestingly, the class of translation invariant MMDs is much larger and is fully characterized in the next theorem. A function $\gamma : \mathbb{R}^d \to \mathbb{R}$ is said to be negative definite if

$$\sum_{i=1}^{n} a_i a_j \gamma(x_i - x_j) \leq 0 \tag{12}$$

for all $x_1, \ldots, x_n \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$. The following result is a consequence of Proposition 1 and states a one-to-one correspondence between translation invariant MMDs and negative definite functions.

Corollary 4. *The MMD associated with the kernel k is translation invariant if and only if there exists a negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that the variogram ρ associated with k satisfies $\rho(x, y) = \gamma(y - x)$.*
Conversely, for all negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that $\gamma(0) = 0$, the MMD associated with the normalized kernel $k_0(x, y) = \gamma(x) + \gamma(y) - \gamma(y - x)$ is translation invariant and its variogram is $\rho(x, y) = \gamma(y - x)$.

Using the point of view of geostatistics and random processes discussed in Remark 2, the negative definite function γ can be related to the variogram of a *stationary increment process*. A process $(B(x))_{x\in\mathbb{R}^d}$ is said to have stationary increments if for all $x_0, \dots, x_n$ and $h \in \mathbb{R}^d$, we have

$$(B(x_i) - B(x_0))_{1\le i\le n} \overset{d}{=} (B(x_i + h) - B(x_0 + h))_{1\le i\le n},$$

where $\overset{d}{=}$ stands for equality in distribution. Then Corollary 4 can be reformulated as follows: let k be a kernel on $\mathbb{R}^d \times \mathbb{R}^d$, ρ the associated variogram and $(B(x))_{x\in\mathbb{R}^d}$ the Gaussian process with origin 0 and variogram ρ (Remark 2); then the MMD associated with k is translation invariant if and only if $(B(x))_{x\in\mathbb{R}^d}$ has stationary increments.
Then we can exploit the fact that the structure of stationary increment Gaussian processes is well-known and has been characterized in Yaglom and Silverman (1962, Section 3.18) or Matheron (1973, Theorem 2.1). See also Chilès and Delfiner (2012, Chapter 4) where the different terminology of Intrinsic Random Function of order 0 (IRF-0) is used or the more recent article by Shen et al. (2022). The following result follows from Corollary 4 by exploiting the structure of negative definite function (or equivalently of stationary increment Gaussian processes).

Corollary 5. *Let k be a normalized (with origin 0) and continuous kernel on $\mathbb{R}^d \times \mathbb{R}^d$. If the MMD associated with k is translation invariant, then there exists a symmetric nonnegative Borel measure Λ on $\mathbb{R}^d \setminus \{0\}$ satisfying*

$$\int_{\mathbb{R}^d} \left(\|\xi\|^2 \wedge 1\right) \Lambda(\mathrm{d}\xi) < \infty \tag{13}$$

and a $d \times d$ symmetric positive semi-definite matrix Σ such that

$$k(x, y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + x^T \Sigma y. \tag{14}$$

Conversely, for any such Λ and Σ, the kernel k defined by (14) is continuous on $\mathbb{R}^d \times \mathbb{R}^d$, normalized, and the associated MMD is translation invariant.

Note that the integrability condition (13) ensures that the integral in Equation (14) is well-defined because

$$\left|\left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right)\right| \le 4 \wedge \left(\|x\|\|y\|\|\xi\|^2\right).$$

The symmetry condition implies that the kernel is real-valued and given by

$$k(x, y) = \int_{\mathbb{R}^d} \left(1 - \cos(x \cdot \xi) - \cos(y \cdot \xi) + \cos((x - y) \cdot \xi)\right)\Lambda(\mathrm{d}\xi) + x^T \Sigma y. \tag{15}$$

Properties of translation invariant MMDs. In the light of Corollary 5, we next establish several properties of translation invariant MMDs. First, the following proposition characterizes translation invariant MMDs that are bounded.

Proposition 6. *Let k be the kernel defined by* (14)*. The following statements are equivalent:*

i) k is bounded on $\mathbb{R}^d \times \mathbb{R}^d$;

ii) Λ is a finite measure and $\Sigma = 0$.

In this case, the two kernels defined in Equations (10) and (14) respectively are easily shown to be equivalent and thus associated with the same MMD defined by Equation (11).

Next we discuss the domain of definition $\mathcal{M}_k$ of the KME associated with k and the form of the corresponding MMD d_k. Note that the kernel decomposes into $k = k_\Lambda + k_\Sigma$ with

$$k_\Lambda(x, y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) \tag{16}$$

$$k_\Sigma(x, y) = x^T \Sigma y, . \tag{17}$$

As a consequence, Steinwart and Ziegel (2021, Lemma 3.3) implies $\mathcal{M}_k = \mathcal{M}_{k_\Lambda} \cap \mathcal{M}_{k_\Sigma}$ and

$$d_k^2(\mu, \nu) = d_{k_\Lambda}^2(\mu, \nu) + d_{k_\Sigma}^2(\mu, \nu) \quad \text{for all } \mu, \nu \in \mathcal{M}_k.$$

One can therefore study k_Λ and k_Σ separately and, for the sake of readability, we use the short notation $\mathcal{M}_\Lambda$ and d_Λ (resp. $\mathcal{M}_\Sigma$ and d_Σ) instead of $\mathcal{M}_{k_\Lambda}$ and d_{k_Λ} (resp. $\mathcal{M}_{k_\Sigma}$ and d_{k_Σ}).
Recall that $\mathcal{M}^\alpha$ denotes the set of finite signed measures with a finite absolute moment of order $\alpha > 0$.

Proposition 7. *Let k_Λ be the kernel defined by Equation* (16)*. If $\int_{\mathbb{R}^d} (\|\xi\|^\alpha \wedge 1)\, \Lambda(\mathrm{d}\xi) < \infty$ for some $\alpha > 0$, then $\mathcal{M}^{\alpha/2} \subset \mathcal{M}_\Lambda$. In particular, Equation* (13) *implies that $\mathcal{M}^1 \subset \mathcal{M}_\Lambda$. For $\mu, \nu \in \mathcal{M}_\Lambda$,*

$$d_\Lambda^2(\mu, \nu) = \int_{\mathbb{R}^d} \left|\hat{\mu}(\xi) - \hat{\nu}(\xi) - \mu(\mathbb{R}^d) + \nu(\mathbb{R}^d)\right|^2 \Lambda(\mathrm{d}\xi). \tag{18}$$

Note that for probability measures $\mu, \nu \in \mathcal{M}_\Lambda \cap \mathcal{P}$, Equation (18) yields $d_\Lambda^2 = \|\hat{\mu} - \hat{\nu}\|_{L^2(\Lambda)}$ as in Equation (11), because the difference $\mu(\mathbb{R}^d) - \nu(\mathbb{R}^d)$ vanishes.

Proposition 8. *Let k_Σ be the kernel defined by Equation* (17)*. Then the space $\mathcal{M}_\Sigma$ is characterized by*

$$\mathcal{M}_\Sigma = \Big\{\mu \in \mathcal{M} \colon \int_{\mathbb{R}^d} |e_j \cdot x|\, |\mu|(\mathrm{d}x) < \infty \quad \textit{for all } 1 \le j \le r\Big\},$$

where r denotes the rank of Σ and $(e_1, \ldots, e_r)$ an orthonormal system of eigenvectors associated with the positive eigenvalues $0 < \lambda_1 \le \ldots \le \lambda_r$. For $\mu, \nu \in \mathcal{M}_\Sigma$,

$$d_\Sigma^2(\mu, \nu) = \sum_{j=1}^r \lambda_j \Big| \int_{\mathbb{R}^d} (e_j \cdot x)\, \mu(\mathrm{d}x) - \int_{\mathbb{R}^d} (e_j \cdot x)\, \nu(\mathrm{d}x)\Big|^2.$$

In particular, if Σ is strictly positive definite, then $\mathcal{M}_\Sigma = \mathcal{M}^1$ and, for $\mu, \nu \in \mathcal{M}^1$,

$$d_\Sigma^2(\mu, \nu) = \|e(\mu) - e(\nu)\|_\Sigma^2$$

where $e(\mu) = \int_{\mathbb{R}^d} x\, \mu(\mathrm{d}x)$ is the expectation of μ and $\|x\|_\Sigma^2 = x^T \Sigma x$ the squared norm associated with Σ.

We finally focus on conditions ensuring that the kernel k is characteristic over probability measures, meaning that d_k defines a proper distance (and not only a semi-metric) on $\mathcal{M}_k \cap \mathcal{P}$, which happens exactly when the KME is injective on $\mathcal{M}_k \cap \mathcal{P}$. Note that the kernel k is never characteristic on $\mathcal{M}_k$ because $d_k^2(\mu, \mu + \alpha\delta_0) = 0$ for all $\mu \in \mathcal{M}_k$ and $\alpha \in \mathbb{R}$, showing that the KME is not injective on $\mathcal{M}_k$. The following theorem provides a necessary and sufficient condition and generalizes Theorem 9 in Sriperumbudur et al. (2010) which considers bounded kernels only.

Proposition 9. *The MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

Examples of translation invariant and unbounded MMDs. We next provide examples of translation invariant MMDs associated with unbounded kernels. It is worth emphasizing that these MMDs are translation invariant even if the underlying kernels are not.

Example 2. *The quadratic kernel $k(x, y) = x \cdot y$ on $\mathbb{R}^d \times \mathbb{R}^d$ has been considered in Sriperumbudur et al. (2010, Example 2) and is the simplest unbounded kernel associated with a translation invariant MMD. It corresponds to $\Lambda = 0$ and $\Sigma = \mathrm{Id}_d$ in Equation* (14)*. Clearly, the corresponding RKHS is the finite dimensional space of linear functions on $\mathbb{R}^d$, the corresponding variogram is $\rho(x, y) = \frac{1}{2}\|x - y\|^2$ and the corresponding stationary increment Gaussian process can be represented as $B(x) = G \cdot x$, $x \in \mathbb{R}^d$, with $G \sim \mathcal{N}(0_d, \mathrm{Id}_d)$. According to Proposition 8, the MMD takes the form $d_k(\mu, \nu) = \|e(\mu) - e(\nu)\|$ for $\mu, \nu \in \mathcal{M}^1$. In particular $d_k(\mu, \nu) = 0$ if μ and ν are probabilities with equal moment of order 1 and d_k is not a proper distance on $\mathcal{M}^1$.*

Example 3. *Brownian motion is the most important stationary increment Gaussian process. In dimension $d = 1$, its covariance function is $k(x, y) = \min(x, y)$ for $x, y \geq 0$, and more generally*

$$k(x, y) = \frac{1}{2}(|x| + |y| - |x - y|) \quad \text{for } x, y \in \mathbb{R}.$$

Clearly, $k(x, x) = |x|$ so that $\mathcal{M}_k = \mathcal{M}^{1/2}$. The associated variogram is $\rho(x, y) = \frac{1}{2}|x - y|$. The spectral measure and matrix in the representation (14) *are known to be $\Lambda(\mathrm{d}\xi) = \frac{1}{2\pi}|\xi|^{-2}$ and $\Sigma = 0$ so that the MMD can be rewritten*

$$d_k^2(\mu, \nu) = \frac{1}{2\pi} \int_{\mathbb{R}^d} \frac{|\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2}{|\xi|^2}\, \mathrm{d}\xi, \quad \mu, \nu \in \mathcal{M}^{1/2}.$$

Interestingly, when restricted to probability measures, the MMD coincide with the Cramer defined as the L^2-distance between the cumulative distribution functions. More precisely,

$$d_k^2(\mu, \nu) = \int_{\mathbb{R}^d} |F_\mu(x) - F_\nu(x)|^2\, \mathrm{d}x, \quad \mu, \nu \in \mathcal{M}^{1/2} \cap \mathcal{P}$$

with $F_\mu(x) = \mu(]-\infty, x])$ and similarly for F_ν. In the literature on scoring rule, this kernel is associated with the so-called CRPS (Gneiting and Raftery, 2007).

Example 4. *A well known stationary increment Gaussian process on $\mathbb{R}^d$ is the fractional Brownian random field with Hurst index $H \in (0,1)$ defined by the covariance*

$$k_H(x,y) = \frac{1}{2}\left(\|x\|^{2H} + \|y\|^{2H} - \|x-y\|^{2H}\right), \tag{19}$$

see Herbin and Merzbach (2007) or Cohen and Istas (2013, Section 3). This is a natural extension of the previous example because the particular case $d = 1$ and $H = 1/2$ corresponds to the Brownian motion. The kernel satisfies $k(x,x) = \|x\|^{2H}$ so that $\mathcal{M}_k = \mathcal{P}^H$. The corresponding variogram is $\rho(x,y) = \frac{1}{2}\|x-y\|^{2H}$. The spectral measure and matrix in the representation (14) *are known to be (Cohen and Istas, 2013, Section 3.3.1)*

$$\Lambda(\mathrm{d}\xi) = \frac{1}{c(d,H)}\|\xi\|^{-d-2H}\mathrm{d}\xi \quad \textit{and} \quad \Sigma = 0$$

with constant

$$c(d,H) = \frac{\sqrt{\pi}\Gamma(H+1/2)}{2^{d/2}H\Gamma(2H)\sin(\pi H)\Gamma(H+d/2)}.$$

The MMD can be rewritten

$$d_k^2(\mu,\nu) = \frac{1}{c(d,H)}\int_{\mathbb{R}^d} \frac{|\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2}{\|\xi\|^{d+2H}}\,\mathrm{d}\xi, \quad \mu,\nu \in \mathcal{M}^H. \tag{20}$$

Thus family of kernels is connected with the α-distance correlation for independence tests (Székely and Rizzo, 2009, Section 4). In the literature on scoring rule, this kernel is associated with the so-called Energy Score (Gneiting and Raftery, 2007).

Example 5. *Another extension of Brownian motion to higher dimension is the fractional Brownian sheet (Cohen and Istas, 2013, 3.3.2) defined by the covariance function*

$$k(x,y) = \prod_{l=1}^{d}\frac{1}{2}\left(|x_l|^{2H_l} + |y_l|^{2H_l} - |x_l - y_l|^{2H_l}\right),$$

where $x = (x_l)_{1\le l\le d}, y = (y_l)_{1\le l\le d} \in \mathbb{R}^d$ and $H_1, \ldots, H_d \in (0,1)$. Here the spectral measure takes the product form $\Lambda(\mathrm{d}\xi) = \prod_{l=1}^d c(1,H_l)^{-1}|\xi_l|^{-1-2H_l}\mathrm{d}\xi_l$ and $\Sigma = 0$.

3 Metrizing the Wasserstein space with MMD

The MMD associated with a characteristic kernel defines a distance on the space of probability measures. Understanding the notion of convergence – or equivalently the topology, associated with this distance – is an important question which has been investigated in particular by Sriperumbudur et al. (2010) and Simon-Gabriel and Schölkopf (2018). Most of the results in this line of research consider bounded kernels and the equivalence between weak convergence and convergence in MMD.

In this section, we investigate the case of unbounded kernels and consider whether convergence in Wasserstein spaces can be metrized by an MMD. The intuition behind this is that convergence of probability measures $d_k(\mu_n, \mu) \to 0$ for the MMD implies convergence of integrals $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu_n$ for all test functions $f \in \mathcal{H}_k$ – see Equation (5). When k is bounded and continuous, the RKHS is included in the space of bounded continuous functions and one cannot expect more than weak convergence. On the opposite, when k is unbounded, the RKHS contains unbounded function and one may hope convergence of integrals for power test functions $x \mapsto \|x\|^\beta$, $\beta > 0$, whence the relationship with convergence of moments and Wasserstein spaces. Because the Energy Kernels from Example 4 are naturally related to power functions, the associated MMDs are natural candidates for metrizing the Wasserstein distance.

3.1 Background on Wasserstein spaces

We first provide the necessary background on Wasserstein spaces. For the purpose of this paper, the underlying space will always be $\mathbb{R}^d$ and we therefore restrict our presentation to this case. More general results as well as proofs can be found in Villani (2003, Section 7). Recall that $\mathcal{M}^\alpha$ (resp. $\mathcal{P}^\alpha$) denotes the set of signed measures (resp. probability measures) with a finite absolute moment of order $\alpha > 0$. Given two probability measures μ, ν on $\mathbb{R}^d$, we denote by $\Gamma(\mu, \nu)$ the set of couplings between μ and ν, that is the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that

$$\gamma(B \times \mathbb{R}^d) = \mu(B) \quad \text{and} \quad \gamma(\mathbb{R}^d \times B) = \nu(B),$$

for all Borel set $B \subset \mathbb{R}^d$. The Wasserstein distance of order α is defined, for $\alpha \geq 1$, by

$$W_\alpha(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x - y\|^\alpha \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/\alpha}, \quad \mu, \nu \in \mathcal{P}^\alpha.$$

For $\alpha \in (0, 1)$, it is defined by

$$W_\alpha(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x - y\|^\alpha \, \gamma(\mathrm{d}x, \mathrm{d}y).$$

For all $\alpha > 0$, the Wasserstein space $(\mathcal{P}^\alpha, W_\alpha)$ is a complete and separable metric space. The case $\alpha < 1$ is somewhat less usual and we stress that the Wasserstein distance W_α is then equal to the Wasserstein distance of order 1 on the metric space $(\mathbb{R}^d, \rho_\alpha)$ with the alternative distance $\rho_\alpha(x, y) = \|x - y\|^\alpha$.
An important result in the theory of Wasserstein space is the Kantorovitch-Rubinstein duality which states that

$$W_1(\mu, \nu) = \sup \Big\{ \int_{\mathbb{R}^d} \varphi \, \mathrm{d}(\mu - \nu) \, : \, \varphi : \mathbb{R}^d \to \mathbb{R} \;\; 1\text{-Lipschitz} \Big\}.$$

In the case $\alpha > 1$, a more involved duality theory, called Kantorovitch duality, holds but it will not be needed here. In the case $\alpha < 1$, we have

$$W_\alpha(\mu, \nu) = \sup \Big\{ \int_{\mathbb{R}^d} \varphi \, \mathrm{d}(\mu - \nu) \, : \, \varphi : \mathbb{R}^d \to \mathbb{R} \;\; (\alpha, 1)\text{-Hölder} \Big\}, \tag{21}$$

where a function φ is said to be $(\alpha, 1)$-Hölder if $|\varphi(x) - \varphi(y)| \leq \|x - y\|^\alpha$ for all $x, y \in \mathbb{R}^d$. Note that the set of $(\alpha, 1)$-Hölder functions is equal to the set of 1-Lipschitz functions on $\mathbb{R}^d$ equipped with the distance ρ_α, so that the duality in the case $\alpha < 1$ is a straightforward consequence from the Kantorovitch-Rubinstein duality.
We finally discuss the notion of convergence in Wasserstein spaces. Let $\alpha > 0$ and $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$. According to (Villani, 2003, Theorem 7.12), the following statements are equivalent:

i) $W_\alpha(\mu_n, \mu) \to 0$;

ii) the sequence $(\mu_n)_{n\geq 1}$ converges weakly to μ and

$$\int_{\mathbb{R}^d} \|x\|^\alpha \, \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^\alpha \, \mu(\mathrm{d}x);$$

iii) for all continuous functions $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $|\varphi(x)| = O_{x\to\infty}(\|x\|^\alpha)$, we have

$$\int_{\mathbb{R}^d} \varphi(x) \, \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \varphi(x) \, \mu(\mathrm{d}x).$$

Note that the convergence in $\mathcal{P}^\alpha$ is stronger for larger values of α. More precisely, $\beta < \alpha$ implies $\mathcal{P}^\alpha \subset \mathcal{P}^\beta$, and for all $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$,

$$W_\alpha(\mu_n, \mu) \to 0 \quad \text{implies} \quad W_\beta(\mu_n, \mu) \to 0. \tag{22}$$

3.2 Some negative answers

Our main question is whether an MMD can metrize the Wasserstein distance according to the following definition.

Definition 10. *Let k be a kernel on $\mathbb{R}^d$ and $\alpha > 0$. We say that the MMD d_k associated with the kernels k metrizes the Wasserstein space of order α if $\mathcal{P} \cap \mathcal{M}_k = \mathcal{P}^\alpha$ and, for all $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$,*

$$d_k(\mu_n, \mu) \to 0 \quad \textit{if and only if} \quad W_\alpha(\mu_n, \mu) \to 0.$$

The following proposition is elementary but it emphasizes the need for unbounded kernels.

Proposition 11. *Assume the kernel k metrizes the Wasserstein space of order $\alpha > 0$. Then k is unbounded on $\mathbb{R}^d \times \mathbb{R}^d$.*

Another negative result focuses on translation invariant MMDs associated with kernel of the form (14). According to Proposition 7, such kernels satisfy $\mathcal{P}^1 \subset \mathcal{M}_k$ so that it is natural to ask whether d_k can metrize the Wasserstein space of order 1.

Proposition 12. *There exists no kernel k of the form* (14) *such that d_k metrizes the Wasserstein space of order* 1*.*

The proof relies on a counter-example with measures of the form $\mu_n = (1 - p_n)\delta_0 + p_n\delta_{x_n}$, $n \geq 1$, with sequences $x_n \to \infty$, $p_n \to 0$ chosen so that $(\mu_n)_{n\geq 1}$ converges to δ_0 for the MMD d_k but not for the Wasserstein distance W_1. More generally, as a straightforward adaptation of this construction shows, there exists no translation invariant MMD metrizing the Wasserstein space of order $\alpha \geq 1$.

3.3 Energy kernels and Wasserstein spaces of order $\alpha < 1$

We focus in this section on the special class of Energy Kernels, see Example 3. We recall that, for $\alpha \in (0,1)$, the Energy Kernel is defined by

$$k_\alpha(x,y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x-y\|^{2\alpha}, \quad x,y \in \mathbb{R}^d,$$

and that the associated MMD is defined on $\mathcal{M}^\alpha$ and translation invariant. For clarity of notation, we denote by $d_\alpha = d_{k_\alpha}$ the MMD associated with k_α. The following theorem links Energy Kernels and Wasserstein distances.

Theorem 13. *Let $\alpha \in (0,1)$ and $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$.*

i) $W_\alpha(\mu_n,\mu) \to 0$ *implies* $d_\alpha(\mu_n,\mu) \to 0$.

ii) $d_\alpha(\mu_n,\mu) \to 0$ *implies* $W_\beta(\mu_n,\mu) \to 0$ *for all* $\beta < \alpha$.

The theorem reveals the close relationship between the Wasserstein distance W_α and the MMD d_α. The first point states that W_α is stronger than d_α, while the second point states that d_α is stronger than W_β for all $\beta < \alpha$. Since W_α can be seen as the limit of W_β as $\beta \uparrow \alpha$, this suggests that d_α and W_α are *almost equivalent*. However, we conjecture that the two distances are not equivalent on $\mathcal{P}^\alpha$.

3.4 MMD metrizing the Wasserstein space for $\alpha \geq 1$

In view of the negative result from Proposition 12, we wish to exhibit a MMD that metrizes the Wasserstein space of order 1, or more generally, of order $\alpha \geq 1$. The issue evidenced in the proof of Proposition 12 is that the matrix part d_Σ controls the expectation and not the absolute moment, suggesting the following modification of Equation (14).
Consider the symmetric positive definite kernel

$$k(x,y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + |x|^{\alpha T}\Sigma|y|^\alpha, \tag{23}$$

where Λ is a symmetric measure on $\mathbb{R}^d\setminus\{0\}$ satisfying condition (13), Σ is a $d\times d$ symmetric positive semi-definite matrix, $\alpha \geq 1$ and $|x|^\alpha = (|x_1|^\alpha, \dots, |x_d|^\alpha)$ denotes the componentwise absolute α-power. Note that the introduction of this absolute power breaks the translation invariance of the associated MMD.
Combining Proposition 7 and a straightforward adaptation of Proposition 8, one can prove that $\mathcal{M}_k$ always contains $\mathcal{M}^\alpha$ and that, for $\mu,\nu \in \mathcal{M}^\alpha$,

$$d_k^2(\mu,\nu) = \int_{\mathbb{R}^d} |\hat\mu(\xi) - \hat\nu(\xi) - \mu(\mathbb{R}^d) + \nu(\mathbb{R}^d)|^2 \,\Lambda(\mathrm{d}\xi) + \|m_\alpha(\mu) - m_\alpha(\nu)\|_\Sigma^2, \tag{24}$$

where $m_\alpha(\mu) = \int_{\mathbb{R}^d} |x|^\alpha \,\mu(\mathrm{d}x) \in \mathbb{R}^d$ denotes the absolute α-moment of μ. With similar argument as in the proof of Proposition 8, one can also prove that $\mathcal{M}_k = \mathcal{M}^\alpha$ if and only if $\ker\Sigma \cap \mathbb{R}^d_+ = \{0\}$. The next theorem states two important properties of the MMD.

Theorem 14. *Let $\alpha \geq 1$, k be the kernel defined by Equation* (23) *and d_k the corresponding MMD given by Equation* (24).

1. *The MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$.
2. *The MMD d_k metrizes the Wasserstein space $\mathcal{P}^\alpha$ if and only if and only if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$ *and* $\ker \Sigma \cap \mathbb{R}^d_+ = \{0\}$.

Example 6. *Taking Λ the standard Gaussian measure on $\mathbb{R}^d$ as in Example 1 and $\alpha \geq 1$, we obtain the modified Gaussian kernel*

$$k(x,y) = \exp(-\|x-y\|^2/2) + |x|^\alpha \cdot |y|^\alpha, \quad x, y \in \mathbb{R}^d.$$

According to Theorem 14, the corresponding MMD metrizes the Wasserstein space $\mathcal{P}^\alpha$.

Remark 15. *It is interesting to see that the condition* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$ *does not imply that the MMD metrizes weak convergence. It is indeed tempting to think that, in Equation 24, the first term guarantees weak convergence while the second term ensures convergence of moment of order $\alpha \geq 1$, whence the convergence in Wasserstein space. However this heuristic is not valid, as the following example shows. Let $\Sigma = 0$ and $\Lambda = \sum_{j=1}^{+\infty} j^{-2} \delta_{\pi x_j}$, with $(x_j)_{j\geq 1}$ an enumeration of the dyadic rational numbers $\left\{\pm a/2^b \colon a, b \in \mathbb{N}\right\}$. Consider the sequence of probability measures $\mu_n = \delta_{2^n}$, $n \geq 1$. For $j \geq 1$, it holds $\hat{\mu}_n(x_j) = 1$ for large enough n, so that $\hat{\mu}_n(\xi) \to 1$ $\Lambda(\mathrm{d}\xi)$-a.e. as $n \to \infty$. Then, in view of Equation (24), the Dominated Convergence Theorem implies $d_k(\mu_n, \delta_0) \to 0$ as $n \to \infty$. Since $(\mu_n)_{n\geq 1}$ does not converge weakly to δ_0, the MMD does not metrize weak convergence. To put this result in perspective with Theorem 7 in Simon-Gabriel et al. (2023), note that the RKHS $\mathcal{H}_k$ is not included in the subspace of functions vanishing at infinity, showing that this hypothesis is crucial in the theorem mentioned.*

3.5 Non asymptotic inequalities for the control of Wasserstein distances

We have considered in Theorem 13 the topological equivalence between MMDs associated with energy kernels and Wasserstein distances. In the following we consider stronger results establishing non asymptotic upper bounds.

The first result gives a strong control of the Wasserstein distance W_1 on subsets of probability measures with uniformly bounded support. A result similar in spirit is due to Auricchio et al. (2020) where discrete measures on a regular grid of $[0,1]^d$ are considered; their analysis relies on Fourier analysis and the Wasserstein distance between discrete measures is bounded by the L^2-norm between Fourier transform which is closely related to MMDs. For $K > 0$, let

$$\mathcal{T}_{\infty,K}(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \colon \mu(\mathbb{R}^d \setminus B(0,K)) = 0\right\}.$$

Proposition 16. *Let $\alpha \in (0,1)$. For all $\mu, \nu \in \mathcal{T}_{\infty,K}(\mathbb{R}^d)$,*

$$W_1(\mu,\nu) \leq C\, d_\alpha(\mu,\nu)^{1/(d+1+\alpha)}$$

with constant $C = C(d,K,\alpha)$ explicitly given in Equation (36).

Next we consider a similar result where the assumption of uniformly bounded support is relaxed. We focus here on the case of uniformly bounded moment of order $\gamma > 1$ but what

really matters is to consider a relatively compact subspace $\mathcal{T} \subset \mathcal{W}_1(\mathbb{R}^d)$. We recall that the relative compactness of $\mathcal{T}$ in the Wasserstein space $\mathcal{W}_1(\mathbb{R}^d)$ is equivalent to

$$\lim_{K\to\infty} \sup_{\mu\in\mathcal{T}} \int_{\mathbb{R}^d} \|x\| \mathbb{1}_{\{\|x\|>K\}}\, \mu(\mathrm{d}x) = 0,$$

see e.g. Proposition 2.2.3 in Panaretos and Zemel (2020). For $\gamma > 1$ and $S > 0$, define

$$\mathcal{T}_{\gamma,S}(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \colon \int_{\mathbb{R}^d} \|x\|^\gamma \, \mu(\mathrm{d}x) < S \right\}.$$

The upper bound

$$\int_{\mathbb{R}^d} \|x\| \mathbb{1}_{\{\|x\|>K\}}\, \mu(\mathrm{d}x) \le \frac{S}{K^{\gamma-1}}, \quad \mu \in \mathcal{T}_{\gamma,S}(\mathbb{R}^d), \tag{25}$$

implies that $\mathcal{T}_{\gamma,S}(\mathbb{R}^d)$ is relatively compact in $\mathcal{W}_1(\mathbb{R}^d)$.

Proposition 17. *Let $\alpha \in (0,1)$. For all $\mu, \nu \in \mathcal{T}_{\gamma,S}(\mathbb{R}^d)$,*

$$W_1(\mu,\nu) \le C\, d_\alpha(\mu,\nu)^\rho \tag{26}$$

with exponent $\rho = \frac{\gamma-1}{\gamma(d+\alpha+1)-\alpha}$ and constant $C = C(d,\gamma,S,\alpha)$ explicitly given in Equation (37).

Unfortunately, we can see that in Propositions 16 and 17, the exponent in the upper bound depends on the dimension so that the upper bound gets worse as the dimension increases.

4 Application to the One Sample Test

We propose a simple numerical experiment illustrating the behaviour of the various MMDs considered in this paper in the context of the One-Sample-Test. This simulation study is very close to those for the Two-Sample-Test proposed in Sejdinovic et al. (2013, Section 8.1) or Gretton et al. (2012, Section 8.1), except that we provide a comparison with the test based on Wasserstein distance in order to illustrate Theorem 14 and support the idea that suitable MMDs can be used as surrogate for the Wasserstein distance. We consider the One-Sample-Test rather than the Two-Sample-Test merely for simplicity and our point is to illustrate the comparison between MMD and Wasserstein distance in the simplest setting.

The One-Sample-Test problem is to determine whether a sample $(X_1, \ldots, X_n)$ comes from a reference distribution P_0. The null assumption is thus

$$(H_0): \quad (X_1, \ldots, X_n) \text{ is an independent sample from the distribution } P_0,$$

In the following, we always take the standard Gaussian distribution (in various dimensions) as reference P_0. We consider the alternative

$$(H_1): \quad (X_1, \ldots, X_n) \text{ is an independent sample from the distribution } P \neq P_0.$$

The test relies on the empirical distribution $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ that we compare to the empirical distribution $P_{0,m} = m^{-1}\sum_{i=1}^m \delta_{Z_i}$ of a simulated independent sample $(Z_1, \ldots, Z_m)$

from the distribution P_0. For this second sample, a large sample size m can be used to reduce the sample fluctuations. The comparison between the two samples relies on the choice of a distance d between probability measures. The quantity $d(P_n, P_{0,n})$ is an approximation for $d(P, P_0)$ and a small distance supports the null hypothesis (H_0). In order to calibrate the test, we use simulated samples $(X_1^\star, \dots, X_n^\star)$ and $(Z_1^\star, \dots, Z_m^\star)$ both with distribution P_0 and compute the distance between their empirical distributions $P_n^\star$ and $P_{0,m}^\star$. Under the null hypothesis (H_0), $d(P_n^\star, P_{m,0}^\star)$ is an independent copy of $d(P_n, P_{m,0})$. Similarly as in the parametric bootstrap, we use multiples copies $d(P_n^{\star b}, P_{m,0}^{\star b})$, $b = 1, \dots, B$, and compute their empirical quantile $q_{1-\alpha}^\star$ of order $1-\alpha$. Typically, $B = 1000$ and $\alpha = 5\%$. We define a (randomized) test with level α by rejecting (H_0) whenever $d(P_n, P_{n,0}) > q_{1-\alpha}^\star$.

In the following, we are interested by the impact of the choice of the distance d on the power of the test and we consider various MMDs as well as the Wasserstein distance. In our numerical experiment, the reference distribution P_0 is the standard Gaussian distribution (in various dimension d), the sample size $n = 100$ is fixed and we consider two families of alternatives with the following data generating process (DGP):

- **DGP1**: the sample $(X_1, \dots, X_n)$ comes from a standard Student distribution $T_d(\mathrm{df})$ in dimension d with df degrees of freedom; we use the continuous parametrization $\mathrm{df} = 1/\varepsilon$ with $\varepsilon \in [0, 1]$ so that $\mathrm{df} \in [1, +\infty]$ and the convention that the Student distribution with $\mathrm{df} = +\infty$ is the standard normal Gaussian distribution;
- **DGP2**: the sample $(X_1, \dots, X_n)$ comes from the mixture distribution $(1-\varepsilon)\mathcal{N}_d(0, 1) + \varepsilon T_d(2)$; that is we have a contaminated standard Gaussian sample with each observations replaced by a Student $T(2)$ alternative with probability $\varepsilon \in [0, 1]$.

For the two data generating processes, (H_0) corresponds to $\varepsilon = 0$ and H_1 to $\varepsilon > 0$, with larger value of ε corresponding to stronger departure from the null assumption. We consider the tests as described above with $n = 100$, $m = 500$, $B = 1000$ and $\alpha = 0.05$ and the following distances:

- **GK**: the MMD associated with the Gaussian kernel with variance $\sigma^2 = d$, i.e. $k(x, y) = \exp(-\|x - y\|^2/(2d))$ (similar to Example 1);
- **ESK1-ESK3**: the MMD associated with energy score kernel with power $\alpha = 0.25$, 0.5 and 0.75 respectively (see Example 4);
- **MGK**: the MMD associated with the modified Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/(2d)) + d^{-1}x \cdot y$ (see Example 6).
- **W1**: the Wasserstein distance of order 1.

We report in Figure 1 the rejection rates of the tests corresponding to these different distances for DGP1 and DGP2 respectively. We use Monte-Carlo estimation based on $N = 1000$ replications to estimate the probability of rejecting (H_0). Recall that when $\varepsilon = 0$, we expect a rejection rate equal to the nominal level $\alpha = 5\%$. When $\varepsilon > 0$, the rejection rate corresponds to the power of the test under the alternative and a higher rejection rate indicates a better ability of the test to discriminate between the null and alternative hypotheses.

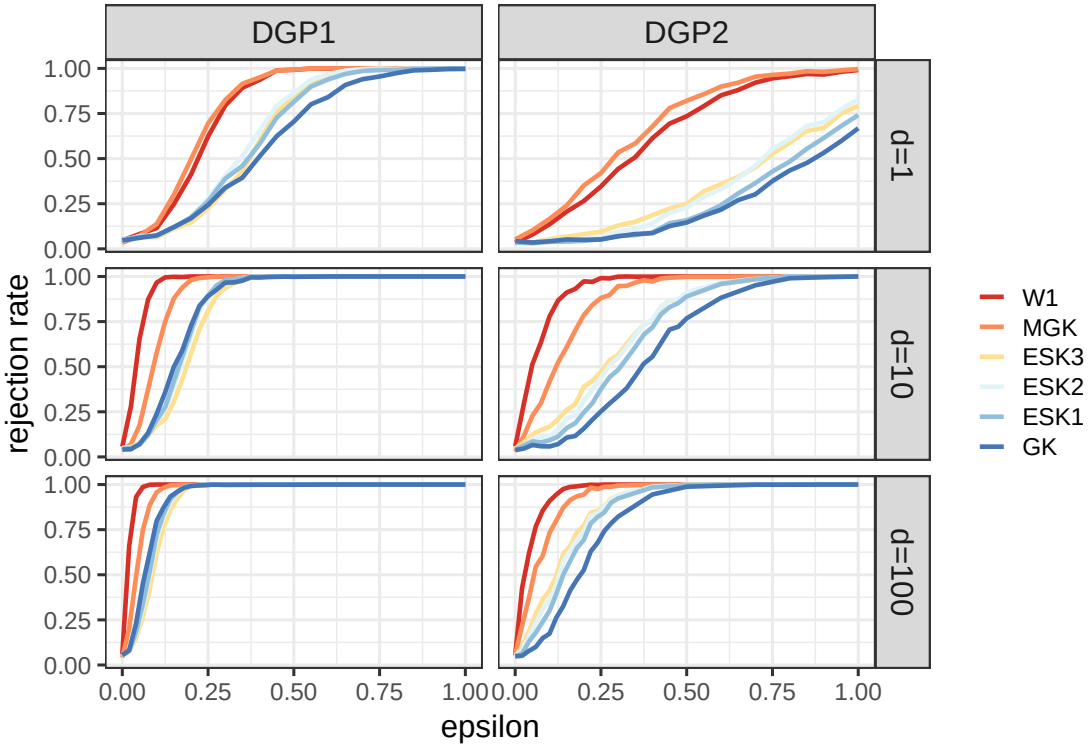

Figure 1: Rejection rates for the one-sample-test based on various distances with data generating process DGP1 (left) and DGP2 (right) in dimension $d = 1$ (top), 10 (middle) and 100 (bottom). The distances considered are the Wasserstein distance (W1) and the MMDs based on the gaussian kernel (GK), various energy score kernels (ESK1-ESK3) and the modified gaussian kernel (MGK).

As should be expected, for the 6 different tests, the rejection rate is roughly equal to 5% when $\varepsilon = 0$ and it increases when ε increases. In both setting, one can see that higher dimension yields higher rejection rate, which is due to an increased effective population size of order nd (note that in our alternatives, all the marginal distributions are deviating from the normal distribution). More importantly, one can compare the different tests. For DGP2, the results are similar in all dimensions: the Gaussian kernel yields the lowest power; Energy Score kernels with increasing α provide tests with increasing power; the Modified Gaussian kernel achieves the best power with a performance very similar to the one of the Wasserstein distance in dimension 1 and slightly lower in higher dimension. For DGP1 in dimension 1, the same comments still hold. However, in higher dimension, the Energy Score Kernels and Gaussian Kernel yield similar performances and only the Modified Gaussian Kernel stands out.

In this simulation framework where deviation from normality arise with heavy tailed Student distribution, our numerical experiments reveals that the Gaussian MMD has a low

expressivity compared to the Wasserstein distance; furthermore, the MMD based on the modified Gaussian kernel almost reaches the same level of expressivity as the Wasserstein distance. Furthermore, these results seem remarkably stable across dimension.

Finally, we compare our findings with similar experiments from the literature.

- Fukumizu et al. (2009, Section 5) study a two sample test and the use of generalized MMD defined as the maximum MMD over a family of kernel; the alternative to the Gaussian distribution is a sinusoidal perturbation interpreted as a high frequency perturbation. The focus is on kernel hyperparameter selection (e.g. bandwidth in the Gaussian kernel). Only bounded kernels are considered with a different alternative distribution from our, making comparison with our result difficult.

- Gretton et al. (2012, Section 8.1) propose a study of the two kernel test with Gaussian distributions with a shift in mean or variance; the Gaussian kernel test, t-test, Kolmogorov-Smirnov test or Hall test are considered among others. Here again only bounded kernels are considered and the emphasis is put on the calibration of the test (based on universal bound, limiting distribution or bootstrap for instance), making comparison with our result difficult.

- To our best knowledge, Sejdinovic et al. (2013, Section 8.1) is the only reference where unbounded kernels are considered for the two-sample-test problem. Power distances (equivalent to our Energy Score Kernel MMD) are compared with the Gaussian MMD in three different settings: shift in mean, shift in variance or sinusoidal perturbation of a Gaussian distribution. The Gaussian MMD exhibits good performance in the first two cases, while the energy score kernel with a small power $\alpha = 1/3$ performs best in the case of a sinusoidal perturbation.

This last reference allows us to compare the use of energy score kernel in two different settings: sinusoidal perturbation (high frequency perturbation) or heavy-tail perturbation (low frequency perturbation). It appears that smaller α in the energy score kernel yield better performance in the former situation, whereas larger α perfom better in the latter situation. This suggests that the expressivity of kernels strongly depends on the alternative considered. As a final remark, let us emphasize that the framework of heavy-tail perturbation has received little attention so far and that we could see that the Wasserstein distance offers the best performance in this setting, followed by the Modified Gaussian Kernel test.

5 Conclusion

Summary. Our main contributions provide new insight into the theory of MMDs associated with unbounded kernels. First, we show that the class of translation invariant MMDs is not restricted to translation invariant kernels (well studied in the literature) but is characterized by translation invariant variograms that can be specified with a spectral measure Λ and a symmetric semidefinite matrix Σ. Second, we consider the relationships between such MMDs and Wasserstein distances: we prove that the Wasserstein distance of order 1 cannot be metrized by a bounded MMD; we prove that the MMDs associated with energy kernels of order $\alpha \in (0, 1)$ *almost* metrizes the Wasserstein distance of order α; finally, for all $\alpha \geq 1$,

we propose a class of kernels metrizing the Wasserstein distance of order $\alpha \geq 1$ (without the translation invariant constraint). A short simulation based on the one-sample-test illustrates the good properties of this last class of MMDs that achieves the closest performances to the Wasserstein distance in the framework proposed with a reduced computational cost.

Potential applications. Although our focus was mostly on theoretical properties, we believe that the present work advocates for further and possibly more applied research to connect MMD- and Wasserstein-based learning. Due to its implicit definition as the minimum of the transport cost, the computation of Wasserstein distances remains challenging, even if efficient algorithms have been designed and surrogate distances have been considered to reduce the computational burden (Kolouri et al., 2019; Bayraktar and Guo, 2021). Interestingly, in the framework of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), both MMD and Wasserstein distances have been studied (Li et al., 2015; Arjovsky et al., 2017; Li et al., 2021). For instance, based on the relationships between Wasserstein distances and MMDs discussed in this paper, it would be interesting to compare the performances and computational costs between Wasserstein-GANs and MMD-GANs.

Acknowledgments and Disclosure of Funding

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (TREX project). They are grateful to the Associate Editor and anonymous referees for their numerous suggestions that have significantly improved the paper.

6 Proofs

6.1 Proofs related to Section 2

Proof of Corollary 4. Assume the MMD associated with k is translation invariant. For $h \in \mathbb{R}^d$, define the translated kernel $k_h(x, y) = k(x + h, y + h)$. Clearly, we have

$$d_k(\tau_{h\#}\mu, \tau_{h\#}\nu) = d_{k_h}(\mu, \nu)$$

and Equation (9) implies that the kernel k and k_h are equivalent (in the sense of Definition 6). Proposition 1 implies that k_h and k have the same variogram, which implies

$$\rho(x, y) = \rho(x + h, y + h), \quad \text{for all } x, y \in \mathbb{R}^d.$$

Since h is arbitrary, we can take $h = y - x$ and define the function $\gamma(h) = \rho(0, h)$ so as to obtain $\rho(x, y) = \rho(0, y - x) = \gamma(y - x)$. The function γ is negative definite because ρ is negative definite. Furthermore, $\gamma(0) = \rho(0, 0) = 0$.

Conversely, given a negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that $\gamma(0) = 0$, the function $\rho(x, y) = \gamma(y - x)$ is negative definite on $\mathbb{R}^d \times \mathbb{R}^d$ and

$$k_0(x, y) = \rho(x, 0) + \rho(0, y) - \rho(x, y) - \rho(0, 0)$$

is positive definite, see Berg et al. (1984, Lemma 2.1 p.74). One can easily check that $k_0(x, y) = \gamma(x) + \gamma(y) - \gamma(y - x)$. Furthermore, the translated kernel

$$k_h(x, y) = k_0(x + h, y + h) = \gamma(x + h) + \gamma(y + h) - \gamma(y - x)$$

has variogram

$$\rho_h(x,y) = \frac{1}{2}k_h(x,x) + \frac{1}{2}k_h(y,y) - k_h(x,y) = \gamma(y-x).$$

The kernels k_h and k have the same variogram and are thus equivalent, which proves that the MMD is translation invariant. ■

The following proof may potentially exist in the literature but we were not able to find a reference with the precise form required. The literature in the field of IRF-n is often quite general and the purpose of this proof is to simplify the form of the characterization of the generalized covariance of IRF-n for the specific case $n = 0$.

Proof of Corollary 5. Let k be a normalized kernel such that its MMD is translation invariant. By Corollary 4, there exists a negative definite function γ such that $\rho(x,y) = \gamma(x-y)$ for all $x, y \in \mathbb{R}^d$. Property (12) corresponds to $-\gamma$ being conditionally positive definite of order 0 (Matheron, 1973, Section 2.1) and therefore according to Theorem 2.1 in the same reference, for $h \in \mathbb{R}^d$,

$$-\gamma(h) = \int_{\mathbb{R}^d} (\cos(h\cdot\xi) - \mathbb{1}_B(\xi)) \frac{\chi(\mathrm{d}\xi)}{\|\xi\|^2} + Q(h),$$

where Q is an even conditionally positive definite of order 0 polynomial of degree ≤ 2 , B is an arbitrary neighborhood of 0 and χ is a positive symmetric measure with no atom at the origin and satisfying

$$\int_{\mathbb{R}^d} \frac{\chi(\mathrm{d}\xi)}{1 + 4\pi^2\|\xi\|^2} < \infty. \tag{27}$$

We define $\Lambda(\mathrm{d}\xi) = \chi(\mathrm{d}\xi)/\|\xi\|^2$ and Equation (27) implies

$$\int_{\mathbb{R}^d} (1 \wedge \|\xi\|^2)\ \Lambda(\mathrm{d}\xi) < \infty.$$

The neighborhood can be chosen $B = \mathbb{R}^d$, which amounts to changing the constant term in the polynomial Q. Moreover, as $\gamma(0) = 0$, the constant term of Q is null and then by parity of Q, for $h \in \mathbb{R}^d$, $Q(h) = h^T M h$ where $M \in \mathcal{M}_d(\mathbb{R})$. We can assume that M is symmetric because for an asymmetric matrix A, $h^T A h = 0$ and any matrix M is the sum of a symmetric and antisymmetric matrix. As Q is conditionally positive definite of order 0 polynomial, for any $h \in \mathbb{R}^d$,

$$Q(h-h) + (-1)^2 Q(0-0) - 2Q(h) \geq 0,$$

then M is a symmetric negative semi-definite matrix.
Equation (8) gives for $x, y \in \mathbb{R}^d$,

$$\begin{aligned} k(x,y) &= \gamma(x) + \gamma(-y) - \gamma(x-y) \\ &= \int_{\mathbb{R}^d} (1 - \cos(x\cdot\xi) - \cos(-y\cdot\xi) + \cos((x-y)\cdot\xi))\ \Lambda(\mathrm{d}\xi) - 2x^T M y \\ &= \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + x^T \Sigma y, \end{aligned}$$

where $\Sigma = -2M$. The last equality comes from the symmetry of Λ. ■

Proof of Proposition 6. If Λ is finite then k_Λ is bounded. Now, assume that Λ is not finite. Let $R > 0$, we denote by B_R the ball with center 0 and radius R in $\mathbb{R}^d$ and by λ_R its volume for the Lebesgue measure λ. By Fubini-Tonelli Theorem

$$\frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x,x)\,\lambda(\mathrm{d}x) = \frac{1}{\lambda_R}\int_{\mathbb{R}^d}\int_{B_R} |1-\mathrm{e}^{ix\cdot\xi}|^2\,\lambda(\mathrm{d}x)\Lambda(\mathrm{d}\xi).$$

We consider

$$f_R(\xi) = \frac{1}{\lambda_R}\int_{B_R} |1-\mathrm{e}^{ix\cdot\xi}|^2\,\lambda(\mathrm{d}x) = \frac{1}{\lambda_R}\int_{B_R} \big(2-2\cos(x\cdot\xi)\big)\,\lambda(\mathrm{d}x).$$

By Fatou's Lemma, as $R\to+\infty$,

$$\liminf \frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x,x)\,\lambda(\mathrm{d}x) = \liminf \int_{\mathbb{R}^d} f_R(\xi)\,\Lambda(\mathrm{d}\xi) \geq \int_{\mathbb{R}^d} \liminf f_R(\xi)\,\Lambda(\mathrm{d}\xi).$$

If $\xi \neq 0$, Riemann-Lebesgue Lemma entails, as $R\to+\infty$,

$$\lim f_R(\xi) = \lim \frac{1}{\lambda_R}\int_{B_R} \big(2-2\cos(x\cdot\xi)\big)\,\lambda(\mathrm{d}x) = 2,$$

whence we deduce

$$\liminf \frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x,x)\,\lambda(\mathrm{d}x) \geq 2\Lambda(\mathbb{R}^d) = +\infty.$$

This shows that k_Λ is not bounded. We have proven that k_Λ is bounded if and only if Λ is bounded. The condition on $k = k_\Lambda + k_\Sigma$ follows easily. ■

The following Lemma gives an upper bound on the growth of the kernel k_Λ and will be useful in the proof of Proposition 7.

Lemma 18. *Let k_Λ be a kernel of the form* (16) *and assume that, for some $0<\alpha\leq 2$, we have $\int_{\mathbb{R}^d}(\|\xi\|^\alpha\wedge 1)\,\Lambda(\mathrm{d}\xi) < +\infty$. Then $k_\Lambda(x,x) = o(\|x\|^\alpha)$, as $\|x\|\to+\infty$, and $\mathcal{M}_{\alpha/2}\subset\mathcal{M}_\Lambda$.*

Proof Assume $\int_{\mathbb{R}^d}(\|\xi\|^\alpha\wedge 1)\Lambda(\mathrm{d}\xi)<\infty$ with $0<\alpha\leq 2$. We show that for all $\varepsilon>0$, there exists $C>0$ such that

$$|k_\Lambda(x,x)| \leq C + \varepsilon\|x\|^\alpha, \quad x\in\mathbb{R}^d. \tag{28}$$

Since ε can be chosen arbitrary small, this shows $k_\Lambda(x,x) = o(\|x\|^\alpha)$ as $\|x\|\to+\infty$.
We compute

$$k_\Lambda(x,x) = \int_{\mathbb{R}^d}\left|1-\mathrm{e}^{ix\cdot\xi}\right|^2\,\Lambda(\mathrm{d}\xi) \leq 4\int_{\mathbb{R}^d}\left(\left(\|x\|\|\xi\|\right)^2\wedge 1\right)\Lambda(\mathrm{d}\xi)$$

and divide the integral into two parts, depending whether $\|\xi\|$ is larger or smaller than some $\eta > 0$ that will be fixed later. The inequality $u^2 \wedge 1 \leq 1$ implies

$$\int_{\{\|\xi\|\geq\eta\}} \left((\|x\|\|\xi\|)^2 \wedge 1 \right) \Lambda(\mathrm{d}\xi) \leq \Lambda(\|\xi\| \geq \eta).$$

For $0 < \alpha \leq 2$, the inequality $u^2 \wedge 1 \leq |u|^\alpha$ implies

$$\int_{\{\|\xi\|<\eta\}} \left((\|x\|\|\xi\|)^2 \wedge 1 \right) \Lambda(\mathrm{d}\xi) \leq \int_{\{\|\xi\|<\eta\}} (\|x\|\|\xi\|)^\alpha \, \Lambda(\mathrm{d}\xi) \leq \|x\|^\alpha \int_{\{\|\xi\|<\eta\}} \|\xi\|^\alpha \, \Lambda(\mathrm{d}\xi).$$

Since $\int_{\mathbb{R}^d} (\|\xi\|^\alpha \wedge 1) \Lambda(\mathrm{d}\xi) < \infty$, for any fixed $\varepsilon > 0$, one can find $\eta > 0$ small enough such that $\int_{\{\|\xi\|<\eta\}} \|\xi\|^\alpha \, \Lambda(\mathrm{d}\xi) < \varepsilon/4$. Setting $C = 4\Lambda(\|\xi\| \geq \eta)$, the upper bounds for the two terms above entail Equation (28).
As a direct consequence of Equation (28), any measure $\mu \in \mathcal{M}$ satisfying $\int_{\mathbb{R}^d} \|x\|^\alpha \, |\mu|(\mathrm{d}x) < \infty$ satisfies also $\int_{\mathbb{R}^d} \sqrt{k_\Lambda(x,x)} \, |\mu|(\mathrm{d}x) < \infty$. In other words, $\mathcal{M}^\alpha \subset \mathcal{M}_\Lambda$ and this concludes the proof of the Lemma. ∎

Proof of Proposition 7. The inclusion $\mathcal{M}^{\alpha/2} \subset \mathcal{M}_\Lambda$ is proven in Lemma 18. Equation (13) implies that $\mathcal{M}^1 \subset \mathcal{M}_\Lambda$. The computation of the MMD in terms of characteristic function follows the lines Sriperumbudur et al. (2010, Corollary 4 and its proof). For $\mu, \nu \in \mathcal{M}_\Lambda$,

$$\begin{aligned}
d_\Lambda^2(\mu,\nu) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} k_\Lambda(x,y) \, (\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi)(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\
&= \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)(\mu-\nu)(\mathrm{d}x) \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{-iy\cdot\xi}\right)(\mu-\nu)(\mathrm{d}y) \right] \Lambda(\mathrm{d}\xi) \\
&= \int_{\mathbb{R}^d} \left(\mu(\mathbb{R}^d) - \nu(\mathbb{R}^d) - \hat{\mu}(\xi) + \hat{\nu}(\xi)\right) \overline{\left(\mu(\mathbb{R}^d) - \nu(\mathbb{R}^d) - \hat{\mu}(\xi) + \hat{\nu}(\xi)\right)} \, \Lambda(\mathrm{d}\xi) \\
&= \int_{\mathbb{R}^d} \left|\mu(\mathbb{R}^d) - \nu(\mathbb{R}^d) - \hat{\mu}(\xi) + \hat{\nu}(\xi)\right|^2 \Lambda(\mathrm{d}\xi).
\end{aligned}$$

In these lines, we have used successively Equations (4) and (16), Fubini's theorem and the definition of the characteristic function. ∎

Proof of Proposition 8. The Spectral Theorem for the symmetric positive semidefinite matrix Σ implies

$$k_\Sigma(x,y) = x^T \Sigma y = \sum_{j=1}^r \lambda_j x^T e_j e_j^T y, \quad x, y \in \mathbb{R}^d,$$

where $\lambda_1 \geq \ldots \geq \lambda_r > 0$ are the positive eigenvalues of Σ associated with the orthonormal eigenvectors $(e_1, \ldots, e_r)$. Together with the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for $a, b \geq 0$, we deduce

$$\sqrt{\lambda_l} \left|e_l^T x\right| \leq \sqrt{k_\Sigma(x,x)} \leq \sum_{j=1}^r \sqrt{\lambda_r} \left|e_j^T x\right|, \quad l = 1, \ldots, r.$$

We deduce that $\int_{\mathbb{R}^d} \sqrt{k_\Sigma(x,x)}\,|\mu|(\mathrm{d}x)$ is finite if and only if $\int_{\mathbb{R}^d} |e_j^T x|\,|\mu|(\mathrm{d}x)$ is finite for all $j = 1,\dots,r$. This proves the characterization of M_Σ. On the other hand, a direct computation gives, for $\mu,\nu \in \mathcal{M}_\Sigma$,

$$\begin{aligned} d_\Sigma^2(\mu,\nu) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} k_\Lambda(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\ &= \sum_{j=1}^r \lambda_j \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(x^T e_j e_j^T y\right)(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\ &= \sum_{j=1}^r \lambda_j \Big| \int_{\mathbb{R}^d} (e_j^T x)\,\mu(\mathrm{d}x) - \int_{\mathbb{R}^d} (e_j^T x)\,\mu(\mathrm{d}x)\Big|^2. \end{aligned}$$

■

6.2 Proofs related to Section 3

6.2.1 Proofs of Subection 3.2

Proof of Proposition 11. The proof is done by contraposition. Assume that the kernel k is bounded and let $\alpha > 0$. We prove that d_k does not metrize the Wasserstein space of order α. The assumption that k is bounded implies $\mathcal{M}_k = \mathcal{M}$. For $x \in \mathbb{R}^d \setminus \{0\}$ and $n \geq 1$, we consider the probability measures

$$\mu_n = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{n^{1/\alpha}x} \quad \text{and} \quad \mu = \delta_0.$$

Then, since k is bounded,

$$d_k^2(\mu_n,\mu) = \frac{1}{n^2}\Big(k(0,0) + k(n^{1/\alpha}x, nn^{1/\alpha}x) - 2k(n^{1/\alpha}x,0)\Big) \to 0.$$

On the other hand,

$$W_\alpha(\mu_n,\mu) = \int_{\mathbb{R}^d} \|y\|^\alpha\,\mu_n(\mathrm{d}y) = \|x\| \nrightarrow 0.$$

This shows that d_k does not metrize the Wasserstein space of order α. ■

Proof of Proposition 12. For $x \in \mathbb{R}^d \setminus \{0\}$ and $n \geq 2$, we consider the probability measures

$$\mu_n = \frac{n-2}{n}\delta_0 + \frac{1}{n}\delta_{-nx} + \frac{1}{n}\delta_{nx} \quad \text{and} \quad \mu = \delta_0.$$

On the one hand, the measures μ_n and μ are symmetric and thus have expectation 0. It follows that $e(\mu) = e(\mu_n) = 0$ and $d_\Sigma(\mu_n,\delta_0) = 0$ according to Proposition 7. Furthermore, we compute

$$d_\Lambda^2(\mu_n,\mu) = \frac{1}{n^2}\big(k_\Lambda(nx,nx) + k_\Lambda(-nx,-nx) + 2k_\Lambda(nx,-nx)\big)$$

and, according to Lemma 18, $|k_\Lambda(nx, nx)| = o(n^2)$, $|k_\Lambda(-nx, -nx)| = o(n^2)$ and

$$|k_\Lambda(-nx, nx)| \leq \sqrt{k_\Lambda(nx, nx)}\sqrt{k_\Lambda(-nx, -nx)} = o(n^2).$$

We deduce $d_k(\mu_n, \mu) = d_\Lambda(\mu_n, \mu) \to 0$. On the other hand,

$$W_1(\mu_n, \mu) = \int_{\mathbb{R}^d} \|y\| \, \mu_n(\mathrm{d}y) = \|x\| \nrightarrow 0.$$

This proves that no kernel of the form (14) can metrize the Wasserstein space of order 1. ■

6.2.2 Proof of Theorem 13

For $\alpha \in (0, 1)$, we recall that the Energy Kernel is defined by

$$k_\alpha(x, y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x - y\|^{2\alpha}$$

and we denote by $\mathcal{H}_\alpha = \mathcal{H}_{k_\alpha}$ and $d_\alpha = d_{k_\alpha}$ the associated RKHS and the MMD. We recall that $\mathcal{M}_{k_\alpha} = \mathcal{M}^\alpha$. The kernel mean embedding is denoted by $K_\alpha : \mathcal{M}^\alpha \to \mathcal{H}_\alpha$ and is defined by

$$K_\alpha(\mu)(x) = \int_{\mathbb{R}^d} k_\alpha(x, y) \, \mu(\mathrm{d}y), \quad x \in \mathbb{R}^d.$$

For the sake of clarity, we divide the proof of Theorem 13 into two parts. The next two lemma will be useful for the first part.

Lemma 19. *For all $\mu \in \mathcal{M}^\alpha$, the kernel mean embedding $K_\alpha(\mu)$ is α-Hölder continuous with constant $c_\alpha(\mu) = 2\int_{\mathbb{R}^d} \|y\|^\alpha \, |\mu|(\mathrm{d}y)$, i.e.*

$$|K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| \leq c_\alpha(\mu)\|x - x'\|^\alpha, \quad x, x' \in \mathbb{R}^d.$$

Proof We have, for $x, x' \in \mathbb{R}^d$,

$$\begin{aligned} |K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| &= \Big| \int_{\mathbb{R}^d} k_\alpha(x, y) \, \mu(\mathrm{d}y) - \int_{\mathbb{R}^d} k_\alpha(x', y) \, \mu(\mathrm{d}y) \Big| \\ &\leq \int_{\mathbb{R}^d} \big| k_\alpha(x, y) - k_\alpha(x', y) \big| \, |\mu|(\mathrm{d}y). \end{aligned}$$

Using the reproducing kernel property and Cauchy-Schwartz inequality, the integrand satisfies

$$\begin{aligned} |k_\alpha(x, y) - k_\alpha(x', y)| &= |\langle K_\alpha(x), K_\alpha(y)\rangle - \langle K_\alpha(x'), K_\alpha(y)\rangle| \\ &= |\langle K_\alpha(x) - K_\alpha(x'), K_\alpha(y)| \\ &\leq \|K_\alpha(x) - K_\alpha(x')\| \|K_\alpha(y)\| \\ &= \sqrt{k_\alpha(x, x) + k_\alpha(x', x') - 2k_\alpha(x, x')}\sqrt{k_\alpha(y, y)} \\ &= 2\|x - x'\|^\alpha \|y\|^\alpha. \end{aligned}$$

Integrating with respect to $|\mu|(\mathrm{d}y)$, we deduce

$$|K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| \leq 2\|x - x'\|^\alpha \int_{\mathbb{R}^d} \|y\|^\alpha \, |\mu|(\mathrm{d}y),$$

whence the function $K_\alpha(\mu)$ is Hölder-continuous with exponent α. ■

Lemma 20. *For all $\mu, \nu \in \mathcal{P}^\alpha$, we have*

$$d_\alpha^2(\mu, \nu) \leq \big(c_\alpha(\mu) + c_\alpha(\nu)\big) W_\alpha(\mu, \nu).$$

Proof We recall that, for $\alpha \in (0, 1)$, the Kantorovitch-Rubinstein duality implies that

$$W_\alpha(\mu, \nu) = \sup \Big| \int_{\mathbb{R}^d} \varphi(x) \, (\mu - \nu)(\mathrm{d}x) \Big| \tag{29}$$

with the supremum taken over the set of Hölder-continuous function with exponent α and constant 1.
Starting from Equation (4) and integrating with respect to y, we get

$$\begin{aligned} d_\alpha^2(\mu, \nu) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} k_\alpha(x, y) \, (\mu - \nu) \otimes (\mu - \nu)(\mathrm{d}x \mathrm{d}y) \\ &= \int_{\mathbb{R}^d} K_\alpha(\mu - \nu)(x) \, (\mu - \nu)(\mathrm{d}x). \end{aligned}$$

According to Lemma 19, the function $K_\alpha(\mu - \nu)$ is Hölder continuous with exponent α and constant $c_\alpha(\mu - \nu)$. Then, Equation (29) implies

$$\begin{aligned} d_\alpha^2(\mu, \nu) &= \int_{\mathbb{R}^d} K_\alpha(\mu - \nu)(x) \, (\mu - \nu)(\mathrm{d}x) \\ &\leq c_\alpha(\mu - \nu) W_\alpha(\mu, \nu). \end{aligned}$$

We conclude by using the fact that

$$\begin{aligned} c_\alpha(\mu - \nu) &= 2 \int_{\mathbb{R}^d} \|y\|^\alpha |\mu - \nu|(\mathrm{d}y) \\ &\leq 2 \int_{\mathbb{R}^d} \|y\|^\alpha \mu(\mathrm{d}y) + 2 \int_{\mathbb{R}^d} \|y\|^\alpha \nu(\mathrm{d}y) \\ &= c_\alpha(\mu) + c_\alpha(\nu). \end{aligned}$$

■

Proof of Theorem 13 (first point). Let $(\mu_n)_{n \geq 1}, \mu \in \mathcal{P}^\alpha$ be such that $W_\alpha(\mu_n, \mu) \to 0$. By Lemma 20,

$$d_\alpha^2(\mu_n, \mu) \leq \big(c_\alpha(\mu_n) + c_\alpha(\mu)\big) W_\alpha(\mu_n, \mu).$$

It is enough to prove that the sequence $(c_\alpha(\mu_n))_{n\geq 1}$ remains bounded in order to conclude $d_\alpha(\mu_n, \mu) \to 0$. This is indeed the case since the convergence $\mu_n \to \mu$ in Wasserstein space of order α implies the convergence of absolute moments

$$\int_{\mathbb{R}^d} \|x\|^\alpha \, \mu_n(\mathrm{d}x) \longrightarrow \int_{\mathbb{R}^d} \|x\|^\alpha \, \mu(\mathrm{d}x),$$

which yields $c_\alpha(\mu_n) \to c_\alpha(\mu)$. Being convergent, the sequence $(c_\alpha(\mu_n))_{n\geq 1}$ is bounded. ■

We next consider the proof of the second point in Theorem 13. The following lemma is the key of the proof.

Lemma 21. *For $r > 0$, we define the measure $\mu_r(\mathrm{d}s) = (1 + \|s\|)^{-d-r}\mathrm{d}s$. Then, for $r > \alpha$, $\mu_r \in \mathcal{M}^\alpha$. Furthermore, for $\alpha < r < 1 \wedge 2\alpha$, the kernel mean embedding satisfies*

$$K_\alpha(\mu_r)(x) \sim d(\alpha, r)\|x\|^{2\alpha - r}, \quad \text{as } \|x\| \to +\infty,$$

with $d(\alpha, r) > 0$.

Proof As $r > \alpha$, the function $\sqrt{k_\alpha(x,x)} = \sqrt{2}\|x\|^\alpha$ is μ_r-integrable and hence $\mu_r \in \mathcal{M}^\alpha$. The KME $K_\alpha(\mu_r) \in \mathcal{H}_\alpha$ is defined by

$$\begin{aligned} K(\mu_r)(x) &= \int_{\mathbb{R}^d} k_\alpha(x, y) \, \mu_r(\mathrm{d}y) \\ &= \int_{\mathbb{R}^d} \Big(\|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x - y\|^{2\alpha} \Big) \big(1 + \|y\|\big)^{-(d+r)} \, \mathrm{d}y. \end{aligned}$$

The change of variable $z = y/\|x\|$ yields

$$K(\mu_r)(x) = \|x\|^{2\alpha + d} \int_{\mathbb{R}^d} \Big(1 + \|z\|^{2\alpha} - \|x/\|x\| - z\|^{2\alpha} \Big) \big(1 + \|x\|\|z\|\big)^{-(d+r)} \, \mathrm{d}z.$$

By the rotational invariance of the Euclidean norm and the Lebesgue measure, the integral does not change if we replace the unit vector $x/\|x\|$ by $e_1 = (1, 0, \ldots, 0)$. This yields

$$K(\mu_r)(x) = \|x\|^{2\alpha + d} \int_{\mathbb{R}^d} \Big(1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha} \Big) \big(1 + \|x\|\|z\|\big)^{-(d+r)} \, \mathrm{d}z.$$

Note that $K_\alpha(\mu_r)(x)$ is rotation invariant and depends only on $\|x\|$. We next consider the asymptotic as $\|x\| \to +\infty$. In order to ease the analysis, we use the following form

$$K(\mu_r)(x) = \|x\|^{2\alpha - r} \int_{\mathbb{R}^d} \left(\frac{\|x\|\|z\|}{1 + \|x\|\|z\|} \right)^{d+r} \frac{1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z.$$

Using this expression, the proof of the Lemma is reduced to the proof of the convergence

$$\int_{\mathbb{R}^d} \left(\frac{u\|z\|}{1 + u\|z\|} \right)^{d+r} \frac{1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z \to d(\alpha, r) > 0, \quad \text{as } u \to +\infty. \tag{30}$$

We observe that, for all $z \in \mathbb{R}^d \setminus \{0\}$, $(u\|z\|/(1+u\|z\|))^{d+r} \to 1$, as $u \to \infty$, suggesting the convergence with limit

$$d(\alpha, r) = \int_{\mathbb{R}^d} \frac{1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z.$$

This is justified by Lebesgue dominated convergence Theorem, since $(u\|z\|/(1+u\|z\|))^{d+r} \leq 1$ and $g(z) = (1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha})/\|z\|^{d+r}$ is integrable. Indeed:

- for $\|z\| > 1/2$, the upper bound

$$|g(z)| = \|z\|^{-(d+r)} |k_\alpha(e_1, z)| \leq \|z\|^{-(d+r)} \sqrt{k_\alpha(e_1, e_1)} \sqrt{k_\alpha(z, z)} = 2\|z\|^{\alpha - d - r},$$

  implies integrability on $\{z : \|z\| > 1/2\}$ since $r > \alpha$;

- for $\|z\| \leq 1/2$, the function $z \mapsto 1 - \|e_1 - u\|^{2\alpha}$ is continuously differentiable on the compact ball $\{z : \|z\| \leq 1/2\}$ and vanishes at 0 so that $\left|1 - \|e_1 - z\|^{2\alpha}\right| \leq C\|z\|$ for some $C > 0$; we deduce

$$|g(z)| \leq \|z\|^{2\alpha - d - r} + C\|z\|^{1 - d - r}$$

  which implies integrability on $\{z : \|z\| \leq 1/2\}$ since $r < 1 \wedge 2\alpha$.

The convergence (30) is proved and it remains to show that the limit is positive. By rotation invariance,

$$d(\alpha, r) = \int_{\mathbb{R}^d} \frac{1 + \|z\|^{2\alpha} - \|z - e_1\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z = \int_{\mathbb{R}^d} \frac{1 + \|z\|^{2\alpha} - \|z + e_1\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z.$$

Then, taking the mean of the two expressions,we get

$$\begin{aligned}
d(\alpha, r) &= \int_{\mathbb{R}^d} \|z\|^{-(d+r)} \Big(1 + \|z\|^{2\alpha} - \frac{\|z - e_1\|^{2\alpha} + \|z + e_1\|^{2\alpha}}{2}\Big) \, \mathrm{d}z \\
&\geq \int_{\mathbb{R}^d} \|z\|^{-(d+r)} \Big(1 + \|z\|^{2\alpha} - \Big[\frac{\|z - e_1\|^2 + \|z + e_1\|^2}{2}\Big]^\alpha\Big) \, \mathrm{d}z \\
&= \int_{\mathbb{R}^d} \|z\|^{-(d+r)} \Big(1 + \|z\|^{2\alpha} - \big(1 + \|z\|^2\big)^\alpha\Big) \, \mathrm{d}z \\
&> \int_{\mathbb{R}^d} \|z\|^{-(d+r)} \Big(1 + \|z\|^{2\alpha} - 1 - \|z\|^{2\alpha}\Big) \, \mathrm{d}z \\
&= 0.
\end{aligned}$$

The first inequality uses the concavity of the function $u \mapsto u^\alpha$ on $(0, +\infty)$ and the second inequality uses $(1+u)^\alpha < 1 + u^\alpha$ for $u > 0$. Both properties hold because $\alpha \in (0, 1)$. ■

The following lemma is a generalization of the classical characterization of the Wasserstein convergence (Theorem 7.12, Villani 2003). The proof is easily adapted and omitted for the sake of brevity.

Lemma 22. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous function satisfying $f(x) \sim C\|x\|^\beta$ as $\|x\| \to +\infty$ for some $C > 0$ and $\beta > 0$. For measures $\mu, (\mu_n)_{n\geq 1} \in \mathcal{P}^\beta$, the weak convergence $\mu_n \to \mu$ together with the convergence of integrals $\int_{\mathbb{R}^d} f(x)\ \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} f(x)\ \mu(\mathrm{d}x)$ implies the Wasserstein convergence $W_\beta(\mu_n, \mu) \to 0$.*

Proof of Theorem 13 (second point). Let $\mu, (\mu_n)_{n\geq 1} \in \mathcal{P}^\alpha$ such that $d_\alpha(\mu_n, \mu) \to 0$. Then $K(\mu_n) \to K(\mu)$ in $\mathcal{H}_\alpha$ and it follows

$$\forall f \in \mathcal{H}_\alpha,\ \langle f, K(\mu_n)\rangle = \int_{\mathbb{R}^d} f\ \mathrm{d}\mu_n \longrightarrow \langle f, K(\mu)\rangle = \int_{\mathbb{R}^d} f\ \mathrm{d}\mu.$$

In particular, the result holds for the functions from Lemma 21: for $\beta \in (2\alpha - 1 \vee 0, \alpha)$, $r = 2\alpha - \beta \in (\alpha, 1 \wedge 2\alpha)$ and $f = K(\mu_r) \in \mathcal{H}_\alpha$, we have $f(x) \sim d(\alpha, r)\|x\|^\beta$ as $\|x\| \to +\infty$. The function $f \in \mathcal{H}_{k_\alpha}$ is continuous because the kernel k_α is continuous in its two variables so that all functions in the RKHS are continuous (Simon-Gabriel and Schölkopf, 2018, Corollary 3).
In order to apply Lemma 22 and conclude to the convergence $W_\beta(\mu_n, \mu) \to 0$, it remains to prove the weak convergence $\mu_n \to \mu$. By the discussion above, the moments of order β of the measures (μ_n) are uniformly bounded (note that $\|x\|^\beta \leq C(f(x) + 1)$ for some $C > 0$) and hence the sequence (μ_n) is tight. By Equation (20),

$$d_\alpha^2(\mu_n, \mu) = \frac{1}{c(d, 2\alpha)} \int_{\mathbb{R}^d} \frac{|\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2}{\|\xi\|^{d+2\alpha}}\ \mathrm{d}\xi \longrightarrow 0.$$

This implies that μ is the only possible adherent point of the sequence (μ_n). Tightness and uniqueness of adherent point implies the weak convergence $\mu_n \to \mu$. ■

6.2.3 Proof of Subsection 3.4

The key ingredient of the first point of Theorem 14 is this following lemma. Our proof is largely inspired by the proof of Theorem 9 of Sriperumbudur et al. (2010).

Lemma 23. *Let $U \subset \mathbb{R}^d \backslash \{0\}$ be a symmetric open set and $\alpha \geq 1$. There exists a real-valued Schwartz function $\theta \neq 0$ which has a non null Fourier transform outside U and satisfies*

$$\int_{\mathbb{R}^d} \theta(x)\ \mathrm{d}x = 0 \quad \textit{and} \quad \int_{\mathbb{R}^d} |x_i|^\alpha \theta(x)\ \mathrm{d}x = 0, \quad 1 \leq i \leq d.$$

Proof For $w \in \mathbb{R}^d$ and $\varepsilon \in (0, +\infty)^d$, we define the function

$$f_{w,\varepsilon}(\xi) = \prod_{i=1}^d \mathrm{e}^{-\frac{\varepsilon_i^2}{\varepsilon_i^2 - (\xi_i - w_i)^2}} \mathbb{1}_{[-\varepsilon_i, \varepsilon_i]}(\xi_i - w_i), \quad \xi \in \mathbb{R}^d.$$

Clearly, $f_{w,\varepsilon}$ is a Schwartz function with support equal to the hypercube $[w - \varepsilon, w + \varepsilon]$. Because U is open and symmetric, there exist $w_1, \ldots, w_{d+1} \in U$ and $\varepsilon \in (0, +\infty)^d$ such that the symmetric sets $[w_j - \varepsilon, w_j + \varepsilon] \cup [-w_j - \varepsilon, -w_j + \varepsilon]$, $1 \leq j \leq d+1$, are all included in U and pairwise disjoint. Then the Schwartz functions

$$\widehat{\theta}_j = f_{w_j,\varepsilon} + f_{-w_j,\varepsilon}, \quad 1 \leq j \leq d+1,$$

are symmetric with disjoint support included in U. As the Fourier Transform is a bijection on the Schwartz class, there is a unique Schwartz function θ_j with Fourier transform $\widehat{\theta}_j$, $1 \leq j \leq d+1$. Note that the functions $\theta_1, \dots, \theta_{d+1}$ are linearly independent because their Fourier transforms $\widehat{\theta}_1, \dots, \widehat{\theta}_{d+1}$ have disjoint support and thus are linearly independent. Furthermore, θ_j is real-valued because $\widehat{\theta}_j$ is symmetric and its integral vanishes because the condition $0 \notin U$ implies

$$\int_{\mathbb{R}^d} \theta_i(x) \, \mathrm{d}x = \widehat{\theta}_i(0) = 0.$$

The $d+1$ vectors in dimension d

$$\left(\int_{\mathbb{R}^d} |x_i|^\alpha \theta_j(x) \, \mathrm{d}x \right)_{1 \leq i \leq d} \in \mathbb{R}^d, \quad 1 \leq j \leq d+1,$$

are not linearly independent so that there exist $u_1, \dots, u_{d+1} \in \mathbb{R}$, non all zero, such that

$$\sum_{j=1}^{d+1} u_j \int_{\mathbb{R}^d} |x_i|^\alpha \theta_j(x) \, \mathrm{d}x = 0 \quad \text{for all } 1 \leq i \leq d.$$

Then the function $\theta = \sum_{j=1}^d u_j \theta_j$ satisfies the required properties. It is non null because the functions $\theta_1, \dots, \theta_{d+1}$ are linearly independent. ■

Proof of Theorem 14 (first point). Consider the decomposition

$$k(x,y) = k_\Lambda(x,y) + k_{\Sigma,\alpha}(x,y) \tag{31}$$

with k_Λ defined in Equation (16) and $k_{\Sigma,\alpha}(x,y) = |x|^{\alpha T} \Sigma |y|^\alpha$.
If $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, we prove that the kernel k_Λ is characteristic over probability measures and hence k is also characteristic. The proof is similar to the proof of Theorem 9 in Sriperumbudur et al. (2010) and we recall only the key arguments. By Proposition 7, as $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d) = 1$

$$d_\Lambda^2(\mu,\nu) = 0 \quad \text{if and only if} \quad \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2 \, \Lambda(\mathrm{d}\xi) = 0.$$

Since Λ has a full support and the integrand is continuous, we must have $\hat{\mu}(\xi) = \hat{\nu}(\xi)$ for all $\xi \in \mathbb{R}^d$. We deduce $\mu = \nu$, showing that k_Λ is characteristic over probability measures.
Conversely, we now suppose that $\mathrm{supp}(\Lambda) \neq \mathbb{R}^d$ and show that $k = k_\Lambda + k_{\Sigma,\alpha}$ is not characteristic. Let $U \subset \mathbb{R}^d \setminus \{0\}$ be a symmetric open set such that $\Lambda(U) = 0$. By Lemma 23, there exists a Schwartz function $\theta \neq 0$ such that

$$\int_{\mathbb{R}^d} \theta(x) \, \mathrm{d}x = 0, \quad \int_{\mathbb{R}^d} |x_i|^\alpha \theta(x) \, \mathrm{d}x = 0, \quad 1 \leq i \leq d,$$

and $\widehat{\theta}(x) = 0$ for $x \notin U$. Let $n \geq 1$ and $C > 0$, such that the measure

$$\mu(\mathrm{d}x) = \frac{C}{1 + \|x\|^n} \, \mathrm{d}x$$

is a probability measure with a finite absolute moment of order p. As θ is continuous and with a fast decay at infinity, there exists $u > 0$, such that the function $C(1+\|x\|)^{-n} + u\theta(x)$ remains positive on $\mathbb{R}^d$. Then the measure

$$\nu(\mathrm{d}x) = \left(\frac{C}{1+\|x\|^n} + u\theta(x)\right) \mathrm{d}x$$

is probability measure (recall that θ has a vanishing integral on $\mathbb{R}^d$). By the properties of θ, the measures μ and ν have the same absolute moment of order p:

$$\int_{\mathbb{R}^d} |x_i|^\alpha \, \mu(\mathrm{d}x) = \int_{\mathbb{R}^d} |x_i|^\alpha \, \nu(\mathrm{d}x), \quad 1 \leq i \leq d,$$

so that $m_\alpha(\mu) = m_\alpha(\nu)$ and $d^2_{\Sigma,\alpha}(\mu,\nu) = 0$, see Equation (24). Furthermore, they have the same Fourier transforms outside U, and together with $\Lambda(U) = 0$, this entails

$$d^2_\Lambda(\mu,\nu) = \int_{\mathbb{R}^d} |\widehat{\mu}(\xi) - \widehat{\nu}(\xi)|^2 \, \Lambda(\mathrm{d}\xi) = 0.$$

We conclude that $d^2_k(\mu,\nu) = d^2_\Lambda(\mu,\nu) + d^2_{\Sigma,\alpha}(\mu,\nu) = 0$, so that the MMD is not a distance on $\mathcal{M}_k \cap \mathcal{P}$ and k is not characteristic. ■

Proof of Theorem 14 (second point).
Assume that d_k metrizes the Wasserstein space $\mathcal{P}^\alpha$. Then d_k is a distance, and, by the first point of the theorem, $\mathrm{supp}(\Lambda) = \mathbb{R}^d$. We next prove that $\ker\Sigma \cap \mathbb{R}^d_+ \neq \{0\}$ leads to a contradiction. If $x \in \mathbb{R}^d_+$ is non zero and such that $|x|^\alpha \in \mathrm{Ker}\Sigma$, we consider the sequence $\mu_n = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{nx}$. Clearly $W_\alpha(\mu_n,\delta_0) \not\to 0$ because the α-moment of μ_n does not converge to 0. On the other hand, $d^2_k(\mu_n,\delta_0) = \frac{1}{n^2}k_\Lambda(nx,nx)$ because $|x|^p \in \ker\Sigma$. Then Lemma 18, implies $d^2_k(\mu_n,\delta_0) \to 0$. This shows that d_k does not metrize $\mathcal{P}^\alpha$ and leads to a contradiction, whence $\ker\Sigma \cap \mathbb{R}^d_+ = \{0\}$.

We now assume that $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ and $\mathrm{ker}\Sigma \cap \mathbb{R}^d_+ = \{0\}$. It must be shown that, for $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$, $W_\alpha(\mu_n,\mu) \to 0$ if and only if $d_k(\mu_n,\mu) \to 0$.

- If $W_\alpha(\mu_n,\mu) \to 0$, then $m_\alpha(\mu_n) \to m_\alpha(\mu)$ and $d_\Sigma(\mu_n,\mu) = \|m_\alpha(\mu_n) - m_\alpha(\mu)\|_\Sigma \to 0$. Moreover, as $\alpha \geq 1$, $m_1(\mu_n) \to m_1(\mu)$ and hence these moments are uniformly bounded by some constant C. This implies that the the Fourier transforms $(\hat{\mu}_n)_{n\geq 1}$, $\hat{\mu}$ are all C-Lipschitz continuous and hence

  $$|\hat{\mu}_n(\xi) - \hat{\mu}(\xi)|^2 \leq 4(1 \wedge C^2\|\xi\|^2) \in \mathrm{L}^1(\Lambda).$$

  Also, Wasserstein convergence implying weak convergence, $\hat{\mu}_n \to \hat{\mu}$ pointwise. The Dominated Convergence Theorem then implies,

  $$d^2_\Lambda(\mu_n,\mu) = \int_{\mathbb{R}^d} |\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \, \Lambda(\mathrm{d}\xi) \to 0,$$

  and we deduce $d^2_k(\mu_n,\mu) = d^2_\Lambda(\mu_n,\mu) + d^2_\Sigma(\mu_n,\mu) \to 0$.

- Conversely, if $d_k(\mu_n, \mu) \to 0$, then $d_\Sigma(\mu_n, \mu) = \|m_\alpha(\mu_n) - m_\alpha(\mu)\|_\Sigma \to 0$. This implies $m_\alpha(\mu_n) \to m_\alpha(\mu)$. Indeed, the condition $\mathrm{Ker}\Sigma \cap \mathbb{R}^d_+ = \{0\}$ implies the existence of $c > 0$ such that $\|x\|_\Sigma \geq c\|x\|$ for all $x \in \mathbb{R}^d_+$ (take c has the minimum of the positive continuous function $x \mapsto x^T \Sigma x$ on the compact $\{x \in \mathbb{R}^d_+ : \|x\| = 1\}$).

  Since the moment $m_\alpha(\mu_n)$ converge, the sequence μ_n is tight. Moreover, the convergence

$$d^2_\Lambda(\mu_n, \mu) = \int_{\mathbb{R}^d} |\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \ \Lambda(\mathrm{d}\xi) \to 0$$

  implies that the measure μ is the unique adherent point of the sequence $(\mu_n)_{n\geq 1}$ and hence $(\mu_n)_{n\geq 1}$ converges weakly to μ. Together with the convergence of the absolute moment of order α, this implies the convergence in Wasserstein space $\mathcal{P}^\alpha$.

■

6.2.4 Proof of Subsection 3.5

The proof of Proposition 16 is based on the following lemma, where $*$ denotes the convolution product and h_σ the Gaussian density defined by

$$h_\sigma(x) = (\sigma\sqrt{2\pi})^{-d} \exp(-\|x\|_2^2/2\sigma^2), \quad x \in \mathbb{R}^d. \tag{32}$$

Lemma 24. *For $\varphi \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ and F a probability measure on $\mathbb{R}^d$, we have*

$$\int_{\mathbb{R}^d} \varphi * h_\sigma \ \mathrm{d}F = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \hat{f}(t) h_1(\sigma t) \exp(-iy \cdot t) \ \mathrm{d}t\mathrm{d}y,$$

where $\hat{f}$ is the characteristic function of F.

Proof The proof of this lemma can be found in Ouvrard (2004). By definition of the convolution product and Fubini Theorem, we have

$$\int_{\mathbb{R}^d} \varphi * h_\sigma \ \mathrm{d}F = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi(y) h_\sigma(t - y) \ \mathrm{d}yF(\mathrm{d}t) = \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} h_\sigma(t - y) \ F(\mathrm{d}t)\mathrm{d}y.$$

By a standard result of Fourier theory (see also Ouvrard (2004) lemma 12.5),

$$\int_{\mathbb{R}^d} h_\sigma(t - y) \ F(\mathrm{d}t) = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \hat{f}(t) h_1(\sigma t) \exp(-iy \cdot t) \ \mathrm{d}t,$$

whence the Lemma follows. ■

Lemma 25. *For all $a, b > 0$ and $p, q > 0$,*

$$\inf_{\sigma > 0} \left(a\sigma^p + b\sigma^{-q}\right) = C_{p,q} a^{\frac{q}{p+q}} b^{\frac{p}{p+q}},$$

with $C_{p,q} = \left(\frac{p}{q}\right)^{\frac{q}{p+q}} + \left(\frac{q}{p}\right)^{\frac{p}{p+q}}$.

Proof A straightforward analysis of the function $\sigma \mapsto a\sigma^p + b\sigma^{-q}$ shows that its derivative vanishes at $\sigma = \left(\frac{bq}{ap}\right)^{\frac{1}{p+q}}$ where the minimum is reached. ■

In the following, we note $\mu(f) = \int_{\mathbb{R}^d} \varphi(x)\ \mu(\mathrm{d}x)$ the integral of a function φ with respect to a measure μ.

Lemma 26. *Consider a function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ that is Lipschitz continuous with Lipschitz constant L, bounded by a constant M and with support included in the ball $B(0,K)$. Then, for all probability measures μ, ν on $\mathbb{R}^d$, and $\alpha \in (0,1)$, we have*

$$|\mu(\varphi) - \nu(\varphi)| \leq C d_\alpha(\mu,\nu)^{1/(d+\alpha+1)}$$

with constant C depending only on d, α, K, L, M and given explicitly by Equation (34)*.*

Proof [of Lemma 26] The proof relies on Fourier theory and on an approximation argument using the Gaussian kernel h_σ defined by (32). Since φ is L-Lipschitz continuous, the convolution $\varphi * h_\sigma$ satisfies

$$\|\varphi - \varphi * h_\sigma\|_\infty \leq L \int_{\mathbb{R}^d} \|y\|_2 h_\sigma(y)\ \mathrm{d}y = L m_d \sigma, \tag{33}$$

with $m_d = \int_{\mathbb{R}^d} \|y\|_2 h_1(y)\,\mathrm{d}y$ the absolute moment the d-dimensional standard Gaussian distribution. By the triangle inequality, we have

$$\begin{aligned}|\mu(\varphi) - \nu(\varphi)| &\leq |\mu(\varphi) - \mu(\varphi * h_\sigma)| + |\mu(\varphi * h_\sigma) - \nu(\varphi * h_\sigma)| + |\nu(\varphi * h_\sigma) - \nu(\varphi)| \\ &\leq 2Lm_d\sigma + |\mu(\varphi * h_\sigma) - \nu(\varphi * h_\sigma)|.\end{aligned}$$

The last term is controlled thanks to Fourier analysis and Lemma 24 which implies

$$\begin{aligned}\big|\mu(\varphi * h_\sigma) - \nu(\varphi * h_\sigma)\big| &= (2\pi)^{-d/2}\Big| \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \big(\hat{\mu}(t) - \hat{\nu}(t)\big) h_1(\sigma t) e^{-iy\cdot t}\ \mathrm{d}t\mathrm{d}y\Big| \\ &\leq (2\pi)^{-d/2}\, M\lambda(B(0,K)) \int_{\mathbb{R}^d} \big|\hat{\mu}(t) - \hat{\nu}(t)\big| h_1(\sigma t)\ \mathrm{d}t,\end{aligned}$$

where $\hat{\mu}$ and $\hat{\nu}$ denote the characteristic functions of μ and ν respectively and the last line uses the fact that φ is supported by $B(0,K)$ and bounded by M. Note that the volume of the ball is equal to $\lambda(B(0,K)) = K^d v_d$ with $v_d = \lambda(B(0,1))$ the volume of the unit ball in dimension d. Furthermore, Equation (20) together with the Cauchy-Schwarz inequality implies

$$\begin{aligned}\int_{\mathbb{R}^d} \big|\hat{\mu}(t) - \hat{\nu}(t)\big| h_1(\sigma t)\ \mathrm{d}t &\leq \left(\int_{\mathbb{R}^d} \|t\|^{d+2\alpha} h_1^2(\sigma t)\ \mathrm{d}t \times \int_{\mathbb{R}^d} \frac{|\hat{\mu}(t) - \hat{\nu}(t)|^2}{\|t\|^{d+2\alpha}}\ \mathrm{d}t\right)^{1/2} \\ &= \Big(I(\alpha,d)\sigma^{-2d-2\alpha} \times c(d,2\alpha) d_\alpha^2(\mu,\nu)\Big)^{1/2}\end{aligned}$$

with $I(\alpha,d) = \int_{\mathbb{R}^d} \|t\|^{d+2\alpha} h_1^2(t)\,\mathrm{d}t$. Collecting the different terms, we get

$$|\mu(\varphi) - \nu(\varphi)| \leq 2Lm_d\sigma + (2\pi)^{-d/2}\, MK^d v_d \sqrt{I(\alpha,d)}\sqrt{c(d,2\alpha)} d_\alpha(\mu,\nu)\sigma^{-d-\alpha}.$$

This inequality holds for all $\sigma > 0$ and, minimizing with respect to $\sigma > 0$ according to Lemma 25, we get

$$|\mu(\varphi) - \nu(\varphi)| \leq C d_\alpha(\mu,\nu)^{1/(d+\alpha+1)},$$

where the constant is given by

$$C = DL^{(d+\alpha)/(d+\alpha+1)}(MK^d)^{1/(d+\alpha+1)} \tag{34}$$

with D depending only on d and α and given by

$$D = \frac{d+\alpha+1}{d+\alpha}\,(2m_d)^{(d+\alpha)/(d+\alpha+1)}\left(\frac{(d+\alpha)v_d\sqrt{I(\alpha,d)}\sqrt{c(d,2\alpha)}}{(2\pi)^{d/2}}\right)^{1/(d+\alpha+1)}. \tag{35}$$

■

Proof [of Proposition 16] Let φ be a Lipschitz continuous function with Lipschitz constant $L = 1$ and μ, ν probability measures with support included in $B(0,K)$. Because the quantity $\mu(\varphi) - \nu(\varphi)$ does not change if φ is replaced by $\varphi - \varphi(0)$, we can assume without loss of generality that $\varphi(0) = 0$. Then, because of the Lipschitz property, φ is bounded by K on $B(0,K)$. Therefore, one can easily construct a function $\tilde{\varphi}$ which is 1-Lipschitz on $\mathbb{R}^d$, equal to φ on $B(0,K)$ and equal to 0 on $\mathbb{R}^d \setminus B(0,2K)$. Since μ, ν have their support included in $B(0,K)$, it holds $\mu(\varphi) - \nu(\varphi) = \mu(\tilde{\varphi}) - \nu(\tilde{\varphi})$ and one can apply Lemma 26 to the function $\tilde{\varphi}$ (with $L = 1$, $M = K$ and K replaced by $2K$) and deduce

$$|\mu(\varphi) - \nu(\varphi)| \leq C d_\alpha(\mu,\nu)^{1/(d+\alpha+1)}$$

with constant

$$C = 2^{d/(d+\alpha+1)} D K^{(d+1)/(d+\alpha+1)} \tag{36}$$

and D given in Equation (35). ■

Proof [of Proposition 17] We now remove the support condition and replace it by a weaker moment assumption. For $\gamma > 1$ and $S > 0$, we consider the set $\mathcal{T}_{\gamma,S}$ of measures μ satisfying

$$\int_{\mathbb{R}^d} \|x\|^\gamma \mu(\mathrm{d}x) \leq S.$$

This moment condition implies that, for all $K > 0$,

$$\int_{\mathbb{R}^d} \|x\| \mathbb{1}_{\{\|x\|>K\}}\, \mu(\mathrm{d}x) \leq S K^{1-\gamma}.$$

Consider now probability measures $\mu, \nu \in \mathcal{T}_{\gamma,S}$ and a Lipschitz continuous function φ with Lipschitz constant $L = 1$ and such that $\varphi(0) = 0$. Note that $|\varphi(x)| \leq \|x\|$. For $K > 0$, consider the function $\chi : \mathbb{R}^d \to [0,1]$ defined by

$$\chi(x) = \mathbb{1}_{\|x\|\leq K} + \frac{2K - \|x\|}{K}\mathbb{1}_{K<\|x\|<2K}.$$

Clearly, χ is Lipschtiz continuous with constant $1/K$, is equal to 1 on $B(0,K)$ and to 0 on $\mathbb{R}^d \setminus B(0,2K)$. We introduce the decomposition $\varphi = \chi\varphi + (1-\chi)\varphi$ and the bound

$$|\mu(\varphi) - \nu(\varphi)| \leq |\mu(\chi\varphi) - \nu(\chi\varphi)| + |\mu((1-\chi)\varphi) - \nu((1-\chi)\varphi)|.$$

For the first term, we note that $\chi\varphi$ is supported by $B(0,2K)$, bounded by $2K$ and with Lipschitz constant 2 so that Lemma 26 implies

$$|\mu(\chi\varphi) - \nu(\chi\varphi)| \leq 2^{1+d/(d+\alpha+1)} D K^{(d+1)/(d+\alpha+1)} d_\alpha(\mu,\nu)^{1/(d+\alpha+1)}$$

with D given by (35). For the second term, we note that $(1-\chi)\varphi$ vanishes on $B(0,K)$ and is bounded by $\|x\|$, so that Equation (25) implies

$$\begin{aligned} |\mu((1-\chi)\varphi) - \nu((1-\chi)\varphi)| &\leq \int_{\mathbb{R}^d} \|x\| \mathbb{1}_{\{\|x\|>K\}} \, \mu(\mathrm{d}x) + \int_{\mathbb{R}^d} \|x\| \mathbb{1}_{\{\|x\|>K\}} \, \nu(\mathrm{d}x) \\ &\leq 2SK^{1-\gamma}. \end{aligned}$$

Collecting the two terms, we get

$$|\mu(\varphi) - \nu(\varphi)| \leq 2^{1+d/(d+\alpha+1)} D d_\alpha(\mu,\nu)^{1/(d+\alpha+1)} K^{(d+1)/(d+\alpha+1)} + 2SK^{1-\gamma}.$$

Minimizing the right hand side with respect to $K > 0$ according to Lemma 25, we deduce

$$|\mu(\varphi) - \nu(\varphi)| \leq C d_\alpha(\mu,\nu)^\rho$$

with exponent

$$\rho = \frac{\gamma - 1}{\gamma(d+1) + (\gamma-1)\alpha}$$

and constant

$$C = 2^{\gamma(d+1)/(\gamma(d+1)+\alpha(\gamma-1))} C_{\frac{d+1}{d+\alpha+1}, \gamma-1} D^{(d+\alpha+1)\rho} S^{(d+1)/(\gamma(d+1)+\alpha(\gamma-1))} \tag{37}$$

with D given in Equation (35). Because the 1-Lipschitz function φ in the left hand side is arbitrary, this yields an upper bound for the Wasserstein distance $W_1(\mu,\nu)$. ■

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arjovsky17a.html`.

Gennaro Auricchio, Andrea Codegoni, Stefano Gualandi, Giuseppe Toscani, and Marco Veneroni. The equivalence of fourier-based and wasserstein metrics on imaging problems. *Rendiconti Lincei - Matematica e Applicazioni*, 31:627–649, 11 2020. doi: 10.4171/RLM/908.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13, 2021. doi: 10.1214/21-ECP383. URL https://doi.org/10.1214/21-ECP383.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100 of *Graduate Texts in Mathematics*. Springer, New York, 1984.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7679-7. doi: 10.1007/978-1-4419-9096-9. URL https://doi.org/10.1007/978-1-4419-9096-9. With a preface by Persi Diaconis.

Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty, Second edition*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2012. URL https://minesparis-psl.hal.science/hal-00795336.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/chwialkowski16.html.

Serge Cohen and Jacques Istas. *Fractional Fields and Applications*. Mathématiques et Applications. Springer, 2013. doi: 10.1007/978-3-642-36739-7. URL https://hal.archives-ouvertes.fr/hal-00871783. volume 76.

Noel A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1993. ISBN 0-471-00255-0. doi: 10.1002/9781119115151. URL https://doi.org/10.1002/9781119115151. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 258–267, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/a9eb812238f753132652ae09963a05e9-Paper.pdf.

Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/685ac8cadc1be5ac98da9556bc1c8d9e-Paper.pdf.

Tilmann Gneiting and Adrian Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 03 2007. doi: 10.1198/016214506000001437.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005. URL http://jmlr.org/papers/v6/gretton05a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Erick Herbin and Ely Merzbach. *The Multiparameter Fractional Brownian Motion*, pages 93–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. doi: 10.1109/MSP.2017.2695801.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f0935e4cd5920aa6c7c996a5ee53a70f-Paper.pdf.

Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019. ISSN 0167-8396. doi: https://doi.org/10.1016/j.cagd.2018.10.005. URL https://www.sciencedirect.com/science/article/pii/S0167839618301249.

Yibin Li, Yan Song, Lei Jia, Shengyao Gao, Qiqiang Li, and Meikang Qiu. Intelligent fault diagnosis by fusing domain adversarial training and maximum mean discrepancy via ensemble learning. *IEEE Transactions on Industrial Informatics*, 17(4):2833–2841, 2021. doi: 10.1109/TII.2020.3008010.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1718–1727. JMLR.org, 2015.

G. Matheron. The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468, 1973. doi: 10.2307/1425829.

Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning, 2023.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi: 10.2307/1428011.

Jean-Yves Ouvrard. *Probabilité 2*. Éditions Cassini, 2004.

Victor M. Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham, 2020. ISBN 978-3-030-38437-1; 978-3-030-38438-8. doi: 10.1007/978-3-030-38438-8. URL https://doi.org/10.1007/978-3-030-38438-8.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), Oct 2013. ISSN 0090-5364. doi: 10.1214/13-aos1140. URL http://dx.doi.org/10.1214/13-AOS1140.

Jinqi Shen, Stilian Stoev, and Tailen Hsing. Tangent fields, intrinsic stationarity, and self similarity. *Electronic Journal of Probability*, 27(none):1 – 56, 2022. doi: 10.1214/22-EJP754. URL https://doi.org/10.1214/22-EJP754.

Nian Si, Karthyek Murthy, Jose Blanchet, and Viet Anh Nguyen. Testing group fairness via optimal transport projections. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9649–9659. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/si21a.html.

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29, 2018. URL http://jmlr.org/papers/v19/16-291.html.

Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023. URL http://jmlr.org/papers/v24/21-0599.html.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75225-7.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010. URL http://jmlr.org/papers/v11/sriperumbudur10a.html.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011. URL http://jmlr.org/papers/v11/sriperumbudur10a.html.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008. ISBN 978-0-387-77241-7.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2019.11.005. URL https://www.sciencedirect.com/science/article/pii/S1063520317301483.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HJWHIKqgl.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236 – 1265, 2009. doi: 10.1214/09-AOAS312. URL https://doi.org/10.1214/09-AOAS312.

Titouan Vayer and Rémi Gribonval. Controlling wasserstein distances by kernel norms with application to compressive statistical learning. *Journal of Machine Learning Research*, 24(149):1–51, 2023. URL http://jmlr.org/papers/v24/21-1516.html.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 978-1-4704-1804-5. URL https://books.google.fr/books?id=MyPjjgEACAAJ.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-71050-9. URL https://books.google.fr/books?id=hV8o5R7_5tkC.

A.M. Yaglom and R.A. Silverman. *An Introduction to the Theory of Stationary Random Functions*. Selected Russian publications in the mathematical sciences. Prentice-Hall, 1962. ISBN 9780486605791. URL https://books.google.fr/books?id=cyG9zgEACAAJ.