Characterization of translation invariant MMD on R d and connections with Wasserstein distances

Thibault Modeste, Clément Dombry

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

Thibault Modeste * Clément Dombry †

November 16, 2022

Abstract

Kernel mean embeddings and maximum mean discrepancies (MMD) associated with positive semi-definite kernels are important tools in machine learning that allow to compare probability measures and sample distributions. Two kernels are said equivalent if their associated MMDs are equal. We characterize the equivalence of kernels in terms of their variogram and deduce that MMDs are in one to one correspondance with negative semi-definite functions. As a consequence, we provide a full characterization of translation invariant MMDs on $\mathbb{R}^d$ that are parametrized by a spectral measure and a semi-definite 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 $\alpha \geq 1$.

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

*Univ. Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France E-mail: modeste@math.univ-lyon1.fr

†Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, F-25000 Besançon, France. E-mail: clement.dombry@univ-fcomte.fr

Contents

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 Minimum 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 semi-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 (4) 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 adressed 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. (2021). 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 has recently been considered in several learning algorithms (Frogner et al., 2015). 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 (2021) establish non-asymptotic inequalities relating MMD and Wasserstein distances.

Main contributions. Our main findings are the following:

- The notion of equivalent kernels is introduced (Definition 2.4) and characterized via the variogram (Proposition 2.6), showing that MMDs are in one-to-one correspondence with negative semi-definite functions.

- The class of translation invariant MMD on $\mathbb{R}^d$ is characterized by a spectral measure and a symmetric positive semi-definite matrix (Corollary 2.10). Extending the results of Sriperumbudur et al. (2010), we provide an explicit formula for the MMD in terms of Fourier transform (Proposition 2.13) and provide a necessary and sufficient condition for the kernel to be characteristic over probability measures (Proposition 2.14).

- Strong connections between Energy kernels and Wasserstein distances are established (Theorem 3.4). 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 3.7).

- We provide non-asymptotic inequalities between W_α and d_α for tight subsets of probability measures (Proposition 3.9).

Potential applications. Although our focus here is 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 has 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). Based on the relationships between Wasserstein distances and Energy Kernel MMDs discussed in this paper, it would for instance be interesting to compare Wasserstein-GAN and MMD-GAN based on the Energy Kernel.

Structure of the paper. Section 2 gathers our main results on translation invariant MMD. We first introduce some background on reproducing kernel Hilbert spaces, kernel mean embeddings and maximum mean discrepancies in Section 2.1. The notion of equivalent kernels and its characterization via variograms are the purpose of Section 2.2. Translation invariant MMDs and their properties are studied in Section 2.3. Section 3 focuses 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 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 4.

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. For $\alpha > 0$, we define

$$\mathcal{M}_\alpha = \left\{ \mu \in \mathcal{M} \colon : \int_{\mathbb{R}^d} \|x\|^\alpha \, |\mu|(\mathrm{d}x) < \infty \right\} \quad \text{and} \quad \mathcal{P}_\alpha = \mathcal{M}_\alpha \cap \mathcal{P} \tag{1}$$

the set of signed measures (resp. probability measures) with finite α-moment.

2 Kernel Mean Embedding and Maximum Mean Discrepancy

2.1 Hilbert space embedding of measures

We present some basic elements of the theory of Reproducing Kernel Hilbert Space (RKHS), Kernel Mean Embedding (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 semi-definite. The latter conditions means that

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

Definition 2.1. *A 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

$$\forall f \in \mathcal{H},\ f(x) = \langle f, k(x,\cdot)\rangle. \tag{2}$$

In particular, we have $K(x) = k(x,\cdot)$. The reproducing kernel characterizes the RKHS and conversely, according to Aronszajn's theorem, any kernel defines a unique RKHS.

Theorem 2.2 (Aronszajn's theorem)**.** *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 (2) 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

$$\forall f \in \mathcal{H}_k,\ \langle f, K(\mu)\rangle = \int_{\mathcal{X}} f(x)\,\mu(\mathrm{d}x). \tag{3}$$

The KME extends this property to the class of measures

$$\mathcal{M}_k = \Big\{\mu \in \mathcal{M}\colon \int_{\mathcal{X}} \sqrt{k(x,x)}\,|\mu|(\mathrm{d}x) < +\infty\Big\}. \tag{4}$$

The following proposition defines the KME on $\mathcal{M}_k$. See, e.g., Steinwart and Christmann (2008, Theorem 4.26) for the proof – note that, the kernel k being measurable, all functions $f \in \mathcal{H}_k$ are measurable (Steinwart and Christmann, 2008, Lemma 4.24).

Proposition 2.3. *For all $\mu \in \mathcal{M}_k$, $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$ and there exists an unique $K(\mu) \in \mathcal{H}_k$ satisfying Equation (3).*

The map $K : \mathcal{M}_k \to \mathcal{H}_k$ is the KME associated with k; the vector $K(\mu)$ represents the measure μ in the same way as the vector $K(x)$ represents the point x (identified with the Dirac measure δ_x). One of the main argument in the proof of Proposition 2.3 is the continuity of the linear form $f \in \mathcal{H}_k \mapsto \int f\,\mathrm{d}\mu$ for all $\mu \in \mathcal{M}_k$. It follows from the inequality

$$|f(x)| = |\langle f, k(x,.)\rangle_{\mathcal{H}_k}| \le \|f\|_{\mathcal{H}_k}\|k(x,\cdot)\|_{\mathcal{H}_k} = \|f\|_{\mathcal{H}_k}\sqrt{k(x,x)}$$

which entails

$$\Big|\int_{\mathcal{X}} f\,\mathrm{d}\mu\Big| \le \|f\|_{\mathcal{H}_k}\int_{\mathcal{X}}\sqrt{k(x,x)}\,|\mu|(\mathrm{d}x).$$

Remark 1. One can find in the literature a different construction of the KME on an extended space of measures in terms of the Pettis integral (Diestel and Uhl, 1977, Section 2.3). Using the Closed Graph Theorem for Banach spaces, Steinwart and Ziegel (2021, Section 2) proves the continuity of the linear form $f \in \mathcal{H}_k \mapsto \int_{\mathcal{X}} f\mathrm{d}\mu$ as soon as $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$. The KME is then defined on the subspace $\mathcal{M}_k' = \{\mu \in \mathcal{M}\colon \mathcal{H}_k \subset \mathcal{L}^1(\mu)\}$. This subspace always contains $\mathcal{M}_k$ and the two constructions of the KME coincide there.

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 (3) - applied twice - implies

$$\begin{aligned} d_k^2(\mu, \nu) &= \langle K(\mu - \nu), K(\mu - \nu) \rangle_{\mathcal{H}_k} \\ &= \int_{\mathcal{X} \times \mathcal{X}} k(x, y) \, (\mu - \nu) \otimes (\mu - \nu)(\mathrm{d}x \mathrm{d}y). \end{aligned} \tag{5}$$

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

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

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

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

$$k(x, y) = \exp(-\|x - y\|_2^2/2), \quad x, y \in \mathbb{R}^d.$$

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

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

where φ denotes the multivariate standard Gaussian density on $\mathbb{R}^d$ and $\hat{\mu}$ (resp. $\hat{\nu}$) the characteristic function of μ (resp. ν). By Theorem 7 of Simon-Gabriel et al. (2021), the MMD metrizes weak convergence on $\mathcal{P}$.

2.2 Equivalent kernels and variograms

Given different measurable kernels on $\mathcal{X} \times \mathcal{X}$, one can wonder in which case the associated MMDs are equal. This gives rise to the following definition.

Definition 2.4. *The measurable kernels k_1 and k_2 on $\mathcal{X} \times \mathcal{X}$ are said equivalent if*

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

Let us stress that, in this definition, the equality of MMDs is required for probability measures only.
Our characterization of equivalent kernel relies on the notion of variogram that we now define.

Definition 2.5. *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.

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

Furthermore, according to Berg et al. (1984, Lemma 2.1 p.74), the variogram is a conditionnally negative definite function on $\mathcal{X} \times \mathcal{X}$, meaning that

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

for all $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$. See Chapter 3 in Berg et al. (1984) for more details on the strong relationships between positive definite and negative definite functions.
Our main result in this section is the following simple, yet new to our best knowledge, characterization of equivalent kernels.

Proposition 2.6. *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 normalized kernel. Fix an arbitrary origin $o \in \mathcal{X}$. A kernel k is said to be normalized (with origin o) if

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

Proposition 2.7. *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*

$$k_0(x,y) = k(x,y) - k(x,o) - k(o,y) + k(o,o).$$

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

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

Remark 2. The term *variogram* comes from the theory of stochastic processes and geostatistic (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 process. 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

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

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. We study now 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 $\tau_h(x) = x + h$.

Definition 2.8. *The MMD associated with a kernel k on $\mathbb{R}^d \times \mathbb{R}^d$ is said translation invariant if, for all $h \in \mathbb{R}^d$, $\mu \in \mathcal{M}_k$ implies $\mu \circ \tau_h^{-1} \in \mathcal{M}_k$ and*

$$d_k(\mu \circ \tau_h^{-1}, \nu \circ \tau_h^{-1}) = d_k(\mu, \nu) \quad \text{for all } \mu, \nu \in \mathcal{M}_k. \tag{6}$$

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 were studied by Sriperumbudur et al. (2010, Section 3.2). Note that a translation invariant kernel is always bounded since

$$|k(x,y)| \leq \sqrt{k(x,x)}\sqrt{k(y,y)} = \psi(0).$$

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 negative definite if

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

for all $x_1, \ldots, x_n \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$.

Theorem 2.9. *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 $\rho(x, y) = \gamma(y - x)$.

In other words, Theorem 2.9 establishes a one-to-one correspondence between translation invariant MMDs and negative definite functions.

Remark 3. As a continuation of Remark 2 relating kernels, variogram and stochastic processes, one can relate translation invariant MMDs with stationary increment processes. A process $(B(x))_{x\in\mathbb{R}^d}$ is said to have stationary increments if for any $x_0, \dots, x_n$ and $h \in \mathbb{R}^d$, we have

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

where $\overset{d}{=}$ stands for equality in distribution. We can reformulate Theorem 2.9 as follows: let k be a kernel on $\mathbb{R}^d \times \mathbb{R}^d$ and ρ the associated variogram; then the MMD associated with k is translation invariant if and only if the Gaussian process B with origin 0 and variogram ρ has stationary increments.

Using the previous remark and the characterization of stationary increment Gaussian processes by Yaglom and Silverman (1962) (see Formula (3.59) in Section 3.18), we can characterize all normalized kernels associated with a translation invariant MMD.

Corollary 2.10. *Let k be a normalized (with origin 0) and continuous kernel on $\mathbb{R}^d \times \mathbb{R}^d$. If the MMD associated to k is translation invariant, then there exists a symmetric Borel measure Λ on $\mathbb{R}^d \setminus \{0\}$ satisfying*

$$\int_{\mathbb{R}^d} \left(\|\xi\|^2 \wedge 1\right) \Lambda(\mathrm{d}\xi) < \infty \tag{7}$$

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

$$k(x, y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + x^T\Sigma y. \tag{8}$$

Conversely, for any such Λ and Σ, the kernel k defined by (8) is continuous on $\mathbb{R}^d \times \mathbb{R}^d$, normalized, and the associated MMD is translation invariant.

Note that the integrability condition (7) ensures that the integral in Equation (8) is well-defined because

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

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

$$k(x, y) = \int_{\mathbb{R}^d} \left(1 - \cos(x \cdot \xi) - \cos(y \cdot \xi) + \cos((x - y) \cdot \xi)\right)\Lambda(\mathrm{d}\xi) + x^T\Sigma y. \tag{9}$$

Clearly, in the case when $\Sigma = 0$ and Λ is finite, the kernel k is equivalent to

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

This class of bounded translation invariant kernels is studied in Sriperumbudur et al. (2010, Section 3.2). Most of the available literature on KME and MMD focuses on bounded kernels; the following lemma characterizes the boundedness of k in terms of Λ and Σ.

Lemma 2.11. *Let k be the kernel defined by (8). The following statements are equivalent:*

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

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

We next provide examples of translation MMD associated with unbounded kernel, the so-called Energy Kernels, that will be the focus of Section 3.3.

Example 2. Brownian motion is a classical stationary increment process. In dimension 1, its covariance function $k(x, y) = \min(x, y)$ for $x, y \geq 0$ can be rewritten as

$$k(x, y) = \frac{1}{2}(|x| + |y| - |x - y|).$$

The Energy Kernels can be seen as an extension of this formula. Let $H \in (0, 1)$ and define, for $x, y \in \mathbb{R}^d$,

$$k_H(x, y) = \|x\|^{2H} + \|y\|^{2H} - \|x - y\|^{2H}. \tag{10}$$

This kernel corresponds to the covariance of so-called Fractional Brownian Motion, see Herbin and Merzbach (2007) or Cohen and Istas (2013, Section 3). It is a well-studied family of kernels in statistics and is connected with the α-distance correlation for independence test (Székely and Rizzo, 2009, Section 4). Lemma 1 in Székely and Rizzo (2005) gives us the spectral representation of these kernels, for $H \in (0, 1)$, $x \in \mathbb{R}^d$,

$$\|x\|_2^{2H} = \frac{1}{C(d, 2H)} \int_{\mathbb{R}^d} \frac{1 - \cos(\xi \cdot x)}{\|\xi\|^{d+2H}} \, \mathrm{d}\xi,$$

where $C(d, 2H)$ is a constant depending only on d and H. Then by a direct computation with Equation (9),

$$k_H(x, y) = \frac{1}{C(d, 2H)_{\mathbb{R}^d}} \int_{\mathbb{R}^d} \frac{\left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right)}{\|\xi\|^{d+2H}} \, \mathrm{d}\xi. \tag{11}$$

This shows that the Energy Kernel corresponds to the spectral measure

$$\Lambda(\mathrm{d}\xi) = \frac{1}{C(d, 2H)} \|\xi\|^{-d-2H} \, \mathrm{d}\xi.$$

We next discuss the domain of definition $\mathcal{M}_k$ of the KME associated with k and the form of the corresponding MMD d_k. Since the kernel k decomposes into two terms

$$k_\Lambda(x, y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) \tag{12}$$

$$k_\Sigma(x, y) = x^T \Sigma y, \tag{13}$$

the following lemma will be useful.

Lemma 2.12. *Let* k_1, k_2 *be two kernels and* $k = k_1 + k_2$. *Then* $\mathcal{M}_k = \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$ *and*

$$d_k^2(\mu, \nu) = d_{k_1}^2(\mu, \nu) + d_{k_2}^2(\mu, \nu), \quad \textit{for all } \mu, \nu \in \mathcal{M}_k.$$

Lemma 2.12 suggests that one can study k_Λ and k_Σ separately. 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 the definition (1) of the set $\mathcal{M}_\alpha$ of finite signed measures with a finite absolute moment of order $\alpha > 0$.

Proposition 2.13. *Let* k_Λ *and* k_Σ *be the kernels defined by Equations* (12) *and* (13) *respectively.*

- *If $\alpha > 0$ is such that $\int_{\mathbb{R}^d}(\|\xi\|^\alpha \wedge 1)\,\Lambda(\mathrm{d}\xi) < \infty$, then $\mathcal{M}_{\alpha/2} \subset \mathcal{M}_\Lambda$; in particular, Equation (7) implies that we always have $\mathcal{M}_1 \subset \mathcal{M}_\Lambda$. For $\mu, \nu \in \mathcal{M}_\Lambda$,*

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

where $\hat{\mu}(\xi) = \int_{\mathbb{R}^d} e^{i\xi\cdot x}\,\mu(\mathrm{d}x)$ (resp. $\hat{\nu}$) denotes the characteristic function of μ (resp. ν).

- *The space $\mathcal{M}_\Sigma$ is characterized by*

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

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

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

Remark 4. The following special cases are important:

1. If Λ is finite, then $\mathcal{M}_\Lambda = \mathcal{M}$; this corresponds exactly to the case when the kernel k_Λ is bounded and this case has been studied in Sriperumbudur et al. (2010, Section 3.2).

2. For $\mu, \nu \in \mathcal{M}_\Lambda$ with the same total mass, in particular for probability measures,

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

   The MMD is equal to the norm in $L^2(\Lambda)$ distance between characteristic functions and the spectral measure Λ puts more or less weight to the different frequencies in the spectral domain.

3. 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. This quadratic kernel has been considered in Sriperumbudur et al. (2010, Example 2).

Example 3. As a continuation of Example 2, consider the Energy Kernel with index $H \in (0,1)$ defined in Equation (10). We have $\Sigma = 0$ and $\Lambda(\mathrm{d}\xi) = C(d,2H)^{-1}\|\xi\|^{-d-2H}\,\mathrm{d}\xi$. The equality $k(x,x) = 2\|x\|^H$ implies $\mathcal{M}_\Lambda = \mathcal{M}_H$. For probability measures $\mu, \nu \in \mathcal{M}_H \cap \mathcal{P}$,

$$d_H^2(\mu,\nu) = \frac{1}{C(d,2H)} \int_{\mathbb{R}^d} \frac{|\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2}{\|\xi\|^{d+2H}}\,\mathrm{d}\xi. \tag{14}$$

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 $\mathcal{M}_k \cap \mathcal{P}$. This happens exactly when the KME $K : \mathcal{M}_k \cap \mathcal{P} \to \mathcal{H}_k$ is injective. The following Theorem generalizes Theorem 9 in Sriperumbudur et al. (2010) which considers bounded kernels only. Proposition 3.6 states a similar result and we will prove only this latter one.

Proposition 2.14. *The MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

Note that the kernel k is not characteristic on $\mathcal{M}_k$ – i.e. the KME is not injective on $\mathcal{M}_k$ – because $d_k^2(\mu, \mu + \alpha\delta_0) = 0$ for all $\mu \in \mathcal{M}_k$ and $\alpha \in \mathbb{R}$.

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 show the equivalence between weak convergence and convergence in MMD for bounded kernels. In this section, we investigate whether convergence in Wasserstein spaces can be metrized by a MMD.

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 from Equation (1) the notation $\mathcal{M}_\alpha$ (resp. $\mathcal{P}_\alpha$) for 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 coupling between μ and ν, that is the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that

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

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

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

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

$$W_\alpha(\mu, \nu) = \sup \left\{ \int_{\mathbb{R}^d} \varphi \, \mathrm{d}(\mu - \nu) \, : \, \varphi : \mathbb{R}^d \to \mathbb{R} \;\; (\alpha, 1)\text{-Hölder} \right\}, \tag{15}$$

where a function φ is said $(\alpha, 1)$-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 each continuous function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ 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 $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_\alpha$,

$$W_\alpha(\mu_n, \mu) \to 0 \quad \text{implies} \quad W_\beta(\mu_n, \mu) \to 0. \tag{16}$$

Equivalently, the injection $(\mathcal{P}_\alpha, W_\alpha) \to (\mathcal{P}_\beta, W_\beta)$ is continuous.

3.2 Some negative answers

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

Definition 3.1. *Let k be a kernel on $\mathbb{R}^d$ and $\alpha > 0$. We say that the MMD d_k associated with the kernel 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 3.2. *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 MMD associated with kernels of the form (8). According to Proposition 2.13, 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 3.3. *There exists no kernel k of the form (8) such that d_k metrizes the Wasserstein space of order 1.*

More generally, as a straightforward adaptation of the proof of Proposition 3.3 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 2. 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 the 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 3.4. *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 $\mathcal{P}_\alpha$.

Conjecture 1. Let $\alpha \in (0,1)$. There exist $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_\alpha$ such that

$$d_\alpha(\mu_n, \mu) \to 0 \quad \text{and} \quad W_\alpha(\mu_n, \mu) \not\to 0.$$

Remark 5. It is easy to show that, for $\beta < \alpha < 1$, there exist $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_\alpha$ such that

$$W_\beta(\mu_n, \mu) \to 0 \quad \text{and} \quad d_\alpha(\mu_n, \mu) \not\to 0, \tag{17}$$

or, similarly,

$$d_\beta(\mu_n, \mu) \to 0 \quad \text{and} \quad W_\alpha(\mu_n, \mu) \not\to 0. \tag{18}$$

We construct simple examples by considering

$$\mu_n = (1 - p_n)\delta_0 + p_n \delta_{nx} \quad \text{and} \quad \mu = \delta_0,$$

where $x \in \mathbb{R}^d \setminus \{0\}$ and $p_n \in (0,1)$ is suitably chosen. We easily compute

$$d_\alpha(\mu_n, \mu) = \sqrt{2} p_n n^\alpha \|x\|^\alpha \quad \text{and} \quad W_\alpha(\mu_n, \mu) = p_n n^\alpha \|x\|^\alpha.$$

Similar equations hold for d_β and W_β. Taking $p_n = 1/n^\alpha$, we obtain an example for Equation (17). Taking $p_n = 1/(n^\beta \log n)$, we obtain an example for Equation (18).

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

In view of the negative result from Proposition 3.3, 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 3.3 is that the matrix part d_Σ controls the expectation and not the absolute moment, suggesting the following modification of Equation (8).
Consider the symmetric positive definite kernel

$$k(x,y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + |x|^{\alpha T} \Sigma |y|^\alpha, \tag{19}$$

where Λ is a symmetric measure on $\mathbb{R}^d \setminus \{0\}$ satisfying condition (7), Σ is a $d \times d$ symmetric positive semi-definite matrix, $\alpha \geq 1$ and $|x|^\alpha = (|x_1|^\alpha, \ldots, |x_1|^\alpha)$ denotes the componentwise absolute α-power. Note that the introduction of this absolute power breaks the translation invariance of the associated MMD.
We first consider the domain of definition.

Lemma 3.5. *Let k be the kernel defined by Equation (19).*

1. *$\mathcal{M}_k$ contains the set of measures $\mathcal{M}_\alpha$ that have a finite moment of order α.*

2. *If* $\ker \Sigma \cap (\mathbb{R}^+)^d = \{0\}$ *then* $\mathcal{M}_k = \mathcal{M}_\alpha$.

Lemma 2.12 and similar arguments as in the proof of Proposition 2.13 show that, for $\mu, \nu \in \mathcal{M}_k$,

$$d_k^2(\mu,\nu) = \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi) - \mu(\mathbb{R}^d) + \nu(\mathbb{R}^d)|^2 \, \Lambda(\mathrm{d}\xi) + \|m_\alpha(\mu) - m_\alpha(\nu)\|_\Sigma^2, \tag{20}$$

where $m_\alpha(\mu) = \int_{\mathbb{R}^d} |x|^\alpha \, \mu(\mathrm{d}x) \in \mathbb{R}^d$ is absolute α-moment of μ. Similarly as in Proposition 2.14, one can easily characterize characteristic kernels in this class.

Proposition 3.6. *Let k be the kernel defined by Equation (19). Then the MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

Remark 6. The condition on the support is not sufficient even for metrizing the weak convergence. Indeed, consider

$$\Lambda = \sum_{n=1}^{+\infty} \frac{1}{n^2} \delta_{x_n} \text{ and } \Sigma = 0,$$

where $(x_n)_{n\geq 1}$ is an enumeration of countable set

$$\left\{ \pm \frac{2a+1}{2^b} \pi \mid a, b \in \mathbb{N} \right\}.$$

This set is dense in $\mathbb{R}$ but for $\mu_n := \delta_{2^n}$, one notes that for all $j \in \mathbb{N}$, $\widehat{\mu_n}(x_j) = 1$ for n large enough. Then $\xi \mapsto 1 - \hat{\mu}_n(\xi)$ converges $\Lambda - ae$ to 0. Then by Dominated Convergence Theorem $d_\Lambda(\mu_n, \delta_0) \to 0$, but the sequence $(\mu_n)_{n\geq 1}$ does not converge weakly to δ_0. Note that this kernel verifies all the assumptions of Theorem 7 of Simon-Gabriel et al. (2021), expected $\mathcal{H}_k \subset \mathcal{C}_0$ where $\mathcal{C}_0$ is the subspace of functions that vanish at infinity.

The following theorem is the main result of this section. It provides an example of MMD that metrizes the Wasserstein space of order $\alpha \geq 1$.

Theorem 3.7. *Let k be the kernel defined by Equation (19). Then the MMD d_k metrizes the Wasserstein space of order α if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ *and* $\ker \Sigma \cap (\mathbb{R}^+)^d = \{0\}$.

Example 4. The Gaussian Kernel, Example 1, is generalizable in this way by considering

$$k(x,y) = \exp(-\|x-y\|_2^2/2) + |x| \cdot |y|, \quad x, y \in \mathbb{R}^d.$$

The previous theorem states that this kernel metrizes the convergence in Wasserstein W_1.

3.5 Non asymptotic inequalities for the control of Wasserstein distances

The translation invariant MMD a L^2-distance of Fourier Transform. The link between the Wasserstein distance and this L^2-distance, for a measure Λ, has already been established in Auricchio et al. (2020) for discrete measures on a regular grid of $[0,1]^d$. But the problem of the Wasserstein distance is its computational cost. That's why a strong equivalence with another distance could be more useful than a topological equivalence. The current equivalences, present in the literature, do not allow us to conclude in our case, as we do not want to consider a specific class of probability measures. We pay the cost of the lack of assumption on the form

of our measure by the uniformly integrability assumption. Moreover, we do not prove a strong equivalence, ie an upper bound of a distance by another but only a partial upper bound. This type of inequality has already been introduced and obtained for the MMD in the Section 4 of Vayer and Gribonval (2021). The authors treat the case where the kernel k is bounded and especially the case where the kernel k is translation invariant. Our results concern only the Energy Kernels, Equation (10). The first proposition concerns the Fortet-Mourier distance d_{FM}, ie a distance which metrizes the weak convergence, defined by

$$d_{FM}(\mu,\nu) = \sup\Big\{ \int_{\mathbb{R}^d} \varphi \,\mathrm{d}(\mu-\nu) \;:\; \varphi : \mathbb{R}^d \to \mathbb{R} \;\text{ 1-Lipschitz and } \|f\|_\infty \leq 1 \Big\}.$$

We recall the dual formulation of W_1

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

Proposition 3.8. *Let $\alpha \in (0,1)$ and $\mathcal{T} \subset \mathcal{P}_\alpha(\mathbb{R}^d)$ be a tight subset, i.e.*

$$\forall \varepsilon > 0,\ \exists K \subset \mathbb{R}^d \text{ compact}, \forall \mu \in \mathcal{T},\ \mu(K^c) \leq \varepsilon.$$

Then, for all $\varepsilon > 0$, there exists $C > 0$ such that

$$\forall \mu,\nu \in \mathcal{T},\ d_{FM}(\mu,\nu) \leq C d_\alpha(\mu,\nu) + \varepsilon.$$

With a stronger assumption, we can get a similar result for the Wasserstein distance W_1.

Proposition 3.9. *Let $\alpha \in (0,1)$ and $\mathcal{T} \subset \mathcal{P}_1(\mathbb{R}^d)$ be an uniformly integrable subset, i.e.*

$$\forall \varepsilon > 0,\ \exists K \subset \mathbb{R}^d \text{ compact}, \forall \mu \in \mathcal{T},\ \int_{K^c} \|x\| \ \mu(\mathrm{d}x) \leq \varepsilon.$$

Then, for all $\varepsilon > 0$, there exists $C > 0$ such that

$$\forall \mu,\nu \in \mathcal{T},\ W_1(\mu,\nu) \leq C d_\alpha(\mu,\nu) + \varepsilon.$$

Remark 7. For both propositions, the constant C depends on ε and on the set $\mathcal{T}$. If we assume that the absolute moment of order $\beta > 1$ are bounded by a constant M, we can give an explicit form to the constant C only in terms of ε. Indeed, Markov's and Hölder's inequality allow to quantify the tighness and the uniform integrability of the set $\mathcal{T}$, i.e. one has an explicit expression of the compact K in function of ε.

Acknowledgments : The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (TREX project).

4 Proofs

4.1 Proofs related to Section 2

Proof of Proposition 2.6. For a kernel k, Equation (5) implies that

$$d_k^2(\delta_x,\delta_y) = k(x,x) + k(y,y) - 2k(x,y) = 2\rho(x,y),$$

for all $x, y \in \mathcal{X}$. It follows that if k_1 and k_2 are equivalent kernels, then they have the same variogram.
Conversely, we prove that kernels with the same variograms are equivalent. Let k be a kernel with variogram ρ. We fix an origin $o \in \mathcal{X}$ and consider the kernel

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

which has the same variogram ρ. The application k_0 is indeed a kernel by the Lemma 2.1 of Berg et al. (1984) cause $-k$ is negative definite. We show that $\mathcal{M}_k = \mathcal{M}_{k_0}$ and $d_k = d_{k_0}$ on $\mathcal{M}_k \cap \mathcal{P}$. Since k_0 depends only on the variogram ρ, this implies that two kernels with the same variogram are equivalent.
The inequality

$$k(x,x) - 2|k(x,o)| + k(o,o) \leq k_0(x,x) \leq k(x,x) + 2|k(x,o)| + k(o,o)$$

together with the Cauchy Schwarz inequality entail

$$\left(\sqrt{k(x,x)} - \sqrt{k(o,o)}\right)^2 \leq k_0(x,x) \leq \left(\sqrt{k(x,x)} + \sqrt{k(o,o)}\right)^2.$$

It follows that $\int_{\mathcal{X}} \sqrt{k(x,x)}\,|\mu|(\mathrm{d}x) < \infty$ if and only if $\int_{\mathcal{X}} \sqrt{k_0(x,x)}\,|\mu|(\mathrm{d}x) < \infty$ so that $\mathcal{M}_{k_0} = \mathcal{M}_k$. Let μ and ν be probability measures in $\mathcal{M}_k$. By Equation (5),

$$\begin{aligned} d_{k_0}^2(\mu,\nu) &= \int_{\mathcal{X}\times\mathcal{X}} \Big(k(x,y) - k(x,o) - k(o,y) + k(o,o)\Big)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= d_k^2(\mu,\nu). \end{aligned}$$

The second equality uses that $\mu - \nu$ has total mass 0 (since μ and ν are probability measures) so that only $k(x,y)$ yields a non null integral. Interestingly, a similar computation shows that the MMD can be directly written in terms of the variogram: for $\mu, \nu \in \mathcal{M}_k$ with the same mass,

$$\begin{aligned} d_k^2(\mu,\nu) &= d_{k_0}^2(\mu,\nu) \\ &= \int_{\mathcal{X}\times\mathcal{X}} \Big(\rho(x,o) + \rho(o,y) - \rho(x,y)\Big)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= -\int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y). \end{aligned}$$

□

Proof of Proposition 2.7. Let k be a kernel and $o \in \mathcal{X}$ an arbitrary origin. The kernel k_0 is naturally normalized with origin o. It is easy to check that k_0 has the same variogram than k, then these two kernels are equivalent by Proposition 2.6. Conversely, let K_0 be another kernel equivalent to k normalized, then K_0 and k_0 have the same variogram then for $x \in \mathcal{X}$,

$$k_0(x,x) = 2\rho(o,x) = K_0(x,x).$$

So for any $x, y \in \mathcal{X}$, the equality of the variogram implies $K_0(x,y) = k_0(x,y)$, then this kernel is unique. □

Proof of Theorem 2.9. 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

$$d_k(\mu \circ \tau_h^{-1}, \nu \circ \tau_h^{-1}) = d_{k_h}(\mu, \nu)$$

and Equation (6) implies that the kernel k and k_h are equivalent (in the sense of Definition 2.4). Proposition 2.6 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. □

Proof of Lemma 2.11. 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. □

Proof of Lemma 2.12. The proof of $\mathcal{M}_k = \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$ relies on the inequality

$$\max\left(\sqrt{k_1(x,x)}, \sqrt{k_2(x,x)}\right) \leq \sqrt{k_1(x,x) + k_2(x,x)} \leq \sqrt{k_1(x,x)} + \sqrt{k_2(x,x)},$$

which implies that $\sqrt{k_1(x,x) + k_2(x,x)}$ is $|\mu|(\mathrm{d}x)$-integrable if and only if both $\sqrt{k_1(x,x)}$ and $\sqrt{k_2(x,x)}$ are. Then, for $\mu, \nu \in \mathcal{M}_k$, we can compute $d_k^2(\mu,\nu)$ according to Equation (5) with k replaced by k_1 and k_2; since $\mu, \nu \in \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$, the integral can be slit into two integrals, one for k_1 and one for k_2, and we obtain $d_k^2(\mu,\nu) = d_{k_1}^2(\mu,\nu) + d_{k_2}^2(\mu,\nu)$. □

The following Lemma gives an upper bound on the growth of the kernel k_Λ and will be useful in the proof of Proposition 2.13.

Lemma 4.1. *Let k_Λ be a kernel of the form (12) 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 of Lemma 4.1. 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

$$|k_\Lambda(x,x)| \leq C + \varepsilon\|x\|^\alpha, \quad x \in \mathbb{R}^d. \tag{21}$$

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

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

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 (21).
As a direct consequence of Equation (21), 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 2.13. • The inclusion $\mathcal{M}_{\alpha/2} \subset \mathcal{M}_\Lambda$ is proven in Lemma 4.1. Assumption ?? 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 (5) and (12), Fubini's theorem and the definition of the characteristic function.

- 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 $j=1,\ldots,r$. This proves the characterization of M_Σ. On the other hand, a direct computation gives, for $\mu,\nu\in\mathcal{M}_\Sigma$,

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

□

4.2 Proofs related to Section 3

4.2.1 Proofs of Subection 3.2

Proof of Proposition 3.2. 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 3.3. 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 2.13. 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 4.1, $|k_\Lambda(nx, nx)| = o(n^2)$, $|k_\Lambda(-nx, -nx)| = o(n^2)$ 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 (8) can metrize the Wasserstein space of order 1. □

4.2.2 Proof of Theorem 3.4

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 3.4 into two parts. The next two lemma will be useful for the first part.

Lemma 4.2. *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 of Lemma 4.2. 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 4.3. *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 of Lemma 4.3. We recall that, for $\alpha \in (0,1)$, the Kantorovitch-Rubinstein duality implies that

$$W_\alpha(\mu,\nu) = \sup \Big| \int_{\mathbb{R}^d} \varphi(x)\,(\mu - \nu)(\mathrm{d}x)\Big| \tag{22}$$

with the supremum taken over the set of Hölder-continuous function with exponent α and constant 1.
Starting from Equation (5) 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 4.2, the function $K_\alpha(\mu-\nu)$ is Hölder continuous with exponent α and constant $c_\alpha(\mu-\nu)$. Then, Equation (22) 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 3.4 (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 4.3,

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

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 3.4. The following lemma is the key of the proof.

Lemma 4.4. *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 of Lemma 4.4. 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 rotationnal 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

$$\int_{\mathbb{R}^d} \left(\frac{u\|z\|}{1 + u\|z\|} \right)^{d+r} \frac{1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}}{\|z\|^{d+r}} \, \mathrm{d}z \to d(\alpha, r) > 0, \quad \text{as } u \to +\infty. \tag{23}$$

We observe that, for all $z \in \mathbb{R}^d \setminus \{0\}$,

$$\left(\frac{u\|z\|}{1+u\|z\|}\right)^{d+r} \longrightarrow 1, \quad \text{as } u \to \infty,$$

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

$$\left(\frac{u\|z\|}{1+u\|z\|}\right)^{d+r} \leq 1$$

and

$$g(z) = \frac{1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}}{\|z\|^{d+r}} \text{ is integrable.}$$

This last claim holds because:

- 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 (23) 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

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

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 $\alpha \in (0, 1)$. □

In the proof of this second point, one will also need this technical lemma. It is a generalization of the classical characterization of the Wasserstein convergence (Theorem 7.12, Villani (2003)).

Lemma 4.5. *Let $f \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ and $\beta \in (0,1)$ such that*

$$f(x) \sim C\|x\|^\beta, \quad \text{as } \|x\| \to +\infty,$$

with $C > 0$. Let $(\mu_n)_{n\geq 1}$ be a sequence of probability measures and $\mu \in \mathcal{P}_\beta$. If the sequence $(\mu_n)_{n\geq 1}$ converges weakly to μ and $\int_{\mathbb{R}^d} f(x)\ \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} f(x)\ \mu(\mathrm{d}x)$ then $W_\beta(\mu_n, \mu) \to 0$.

Proof of Lemma 4.5. The purpose of this proof is to show a kind of Wasserstein tigntness as stated in point (ii) of (Theorem 7.12, Villani (2003)),

$$\lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) = 0.$$

By this theorem, this condition will imply the Wasserstein convergence. Let $R > 1$ such that $\|x\|^\beta \leq 2Cf(x)$ for all $\|x\| \geq R-1$. Let $\chi_R : \mathbb{R}^d \to \mathbb{R}$ be the continuous function defined by

$$\chi_R(x) = \mathbb{1}_{\|x\|_2\leq R-1} + (R - \|x\|_2)\mathbb{1}_{R-1<\|x\|_2<R}, \quad \text{for } x \in \mathbb{R}^d.$$

Let $n \geq 1$, noting that $1 - \chi_R(x) = 1$ for $\|x\| \geq R$ and $f(x) \geq 0$ for $\|x\| \geq R-1$,

$$\begin{aligned}\int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) &\leq 2C \int_{\mathbb{R}^d} (1-\chi_R(x))f(x)\ \mu_n(\mathrm{d}x) \\ &= 2C \int_{\mathbb{R}^d} f(x)\ \mu_n(\mathrm{d}x) - 2C \int_{\mathbb{R}^d} \chi_R(x)f(x)\ \mu_n(\mathrm{d}x).\end{aligned}$$

The function $\chi_R f$ is continuous and bounded then by the weak convergence

$$\limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) \leq 2C \int_{\mathbb{R}^d} f(x)\ \mu(\mathrm{d}x) - 2C \int_{\mathbb{R}^d} \chi_R(x)f(x)\ \mu(\mathrm{d}x).$$

As f is integrable, the Dominated Convergence Theorem gives

$$\lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) = 0.$$

This tightness condition implies $W_\beta(\mu_n, \mu) \to 0$. □

Proof of Theorem 3.4 (second point). Let $(\mu_n)_{n\geq 1}$ and μ be probability measures such that $d_\alpha(\mu_n, \mu) \to 0$. Then the sequence of KME $(K(\mu_n))_{n\geq 1}$ converges weakly (in Hilbert sense) to $K(\mu)$, ie

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

in particular, for any fonctions $K(\mu_r)$ of Lemma 4.4 with $\alpha < r < 1 \wedge 2\alpha$. For $\beta \in (2\alpha - 1 \vee 0, \alpha)$, let's consider $r := 2\alpha - \beta \in (\alpha, 1 \wedge 2\alpha)$. Again by the Lemma 4.4,

$$K(\mu_r)(x) \sim d(\alpha, r)\|x\|^\beta, \quad \text{as } \|x\| \to +\infty.$$

Hence there exists a constant $C > 0$ such that $\|x\|^\beta \leq C(K(\mu_r)(x)+1)$ for all $x \in \mathbb{R}^d$. Then the sequence $(m_\beta(\mu_n))_{n\geq 1}$ of β-moment is bounded. The Markov Inequality ensures the tightness of the sequence $(\mu_n)_{n\geq 1}$. Let us recall the Equation (14) which gives the form of d_α^2

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

As the convergence L^1 implies the converges almost everywhere to a sub-sequence and the characteristic function is continuous, the probability measure μ is the unique adherent point of the tight sequence $(\mu_n)_{n\geq 1}$, then by the Prokorhov Theorem, the sequence converges weakly to the measure μ.

The kernel k_α is continuous in its 2 variables, so it is separately continuous and locally bounded. Thus by the Corollary 3 of Simon-Gabriel and Schölkopf (2018), all functions $f \in \mathcal{H}_\alpha$ are continuous, including the function $K(\mu_r)$.

The assumptions of Lemma 4.5 are therefore satisfied, so $W_\beta(\mu_n, \mu) \to 0$. The continuity of the injection, recalled in formula (16), generalizes this convergence for any $\beta \in (0, \alpha)$. □

4.2.3 Proof of Subsection 3.4

Proof of Lemma 3.5. We first state a simple property that will be useful for the proof : there exists $M \geq 0$ such that

$$x^T \Sigma x \leq M\|x\|^2 \quad \text{for all } x \in (\mathbb{R}_+)^d, \tag{24}$$

and, if $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d) = \{0\}$, there exists also $m > 0$ such that

$$x^T \Sigma x \geq m\|x\|^2 \quad \text{for all } x \in (\mathbb{R}_+)^d. \tag{25}$$

To prove this, we consider $K = \{x \in (\mathbb{R}^+)^d \colon \|x\| = 1\}$ and we set

$$m = \min_{x\in K} x^T \Sigma x \quad \text{and} \quad M = \max_{x\in K} x^T \Sigma x.$$

The min and max are well defined because $x \mapsto x^T \Sigma x$ is continuous on K compact. Inequalities (24) and (25) are clearly satisfied for all $x \in K$, and, by a standard homogeneity argument, they also holds for all $x \in (\mathbb{R}_+)^d$. Finally, m and M are non negative because Σ is positive semi-definite and the conditions $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d = \emptyset$ implies that m and M are positive.
We now prove Lemma 3.5. The kernel k defined by Equation (19) is the sum of two kernels

$$k(x, y) = k_\Lambda(x, y) + k_{\Sigma,\alpha}(x, y) \tag{26}$$

with k_Λ defined in Equation (12) and $k_{\Sigma,\alpha}(x, y) = |x|^{\alpha T} \Sigma |y|^\alpha$. Therefore Lemma 2.12 implies - with straightforward notation - $\mathcal{M}_k = \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha}$. According to Proposition 2.13, $\mathcal{M}_1 \subset \mathcal{M}_\Lambda$. According to Equation (24),

$$0 \leq k_{\Sigma,\alpha}(x, x) = |x|^{\alpha T} \Sigma |x|^\alpha \leq M\|x\|^{2\alpha},$$

which implies $\mathcal{M}_\alpha \subset \mathcal{M}_{\Sigma,\alpha}$. Then, for $\alpha \geq 1$, the inclusion $\mathcal{M}_\alpha \subset \mathcal{M}_1$ implies

$$\mathcal{M}_\alpha = \mathcal{M}_1 \cap \mathcal{M}_\alpha \subset \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_k.$$

When $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d = \{0\}$, Equations (24) and (25) together imply

$$m\|x\|^{2\alpha} \leq k_{\Sigma,\alpha}(x, x) \leq M\|x\|^{2\alpha},$$

and $\mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha$. Then, for $\alpha \geq 1$, the inclusions $\mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha \subset \mathcal{M}_1 \subset \mathcal{M}_\Lambda$ imply

$$\mathcal{M}_k = \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha.$$

□

The key ingredient of the Proposition 3.6 is this following lemma. Our proof is largely inspired by the proof of Theorem 9 of Sriperumbudur et al. (2010).

Lemma 4.6. *Let $U \subset \mathbb{R}^d \setminus \{0\}$ 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 $\theta_1, \ldots, \theta_{d+1}$ are linearly independent because their Fourier transforms $\widehat{\theta}_1, \ldots, \widehat{\theta}_{d+1}$ 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 $u_1, \ldots, u_{d+1} \in \mathbb{R}$, 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 $\theta_1, \ldots, \theta_{d+1}$ are linearly independent. □

Proof of Proposition 3.6. Recall the decompostion $k = k_\Lambda + k_{\Sigma,\alpha}$ in Equation (26).
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 2.13, as $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d) = 1$

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

Since Λ has a full support and the integrand is continuous, we must have

$$\hat{\mu}(\xi) - \hat{\nu}(\xi) = 0 \quad \text{for all } \xi \in \mathbb{R}^d.$$

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 4.6, 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 (20)). 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. □

The following Lemma is the sequential version of the Equations (24) and (25).

Lemma 4.7. *Let F be a non empty closed linear cone and $\Sigma \in \mathcal{S}_d(\mathbb{R})$ be a non negative matrix,*

$$\ker \Sigma \cap F = \{0\} \iff \left[\forall (x_n)_n \in F^{\mathbf{N}}, \ (x_n^T \Sigma x_n \to 0 \implies x_n \to 0) \right]$$

Proof. $\boxed{\Leftarrow}$ This implication is proved by contraposition. If $\ker\Sigma \cap F \neq \{0\}$ then let $y \neq 0$ in this intersection. Let $(x_n)_n$ be te constant sequence equal to y. This sequence checks $x_n^T\Sigma x_n \to 0$ but $x_n \not\to 0$.
$\boxed{\Rightarrow}$ Let $(x_n)_n \in F^{\mathbf{N}}$ such that $x_n^T\Sigma x_n \to 0$ then by the Equation (25),

$$0 \leq m\|x_n\|_2^2 \leq x_n^T\Sigma x_n \to 0,$$

where $m > 0$ then $x_n \to 0$. □

Proof of Theorem 3.7. $\boxed{\Rightarrow}$ This implication is proved by contraposition. If $\mathrm{supp}(\Lambda) \neq \mathbb{R}^d$, then by the Proposition 3.6, the MMD d_k is not a distance So the MMD cannot metrize the Wasserstein space.

If $\ker\Sigma \cap (\mathbb{R}^+)^d \neq \{0\}$, let $x \in (\mathbb{R}^+)^d$ be a non null vector such that $(|x|^p)^T\Sigma|x|^p = 0$. Let's define the sequence of probabilty measures $\mu_n = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{nx}$. It is easy to see that $W_p(\mu_n, \delta_0) \not\to 0$ since the moment of order p does not coverge. But $d_k^2(\mu_n, \delta_0) = \frac{1}{n^2}k_\Lambda(nx, nx)$ cause $|x|^p \in \ker\Sigma$ and $|x| = x$. Then by Lemma 4.1,

$$d_k^2(\mu_n, \delta_0) = o_n(1),$$

then it vanishes. So the MMD does not metrize the Wasserstein space of order p.

$\boxed{\Leftarrow}$ First of all, by the Lemma 3.5, $\mathcal{M}_k \cap \mathcal{P} = \mathcal{P}_p$. Let $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_p$, it must be shown that $W_p(\mu_n, \mu) \to 0$ if and only if $d_k(\mu_n, \mu) \to 0$.

- if $W_p(\mu_n, \mu) \to 0$, then $m_p(\mu_n) \to m_p(\mu)$. Then by the Equation (24),

  $$d_\Sigma(\mu_n, \mu) = \|m_p(\mu_n) - m_p(\mu)\|_\Sigma \to 0.$$

  Moreover, as $(\mu_n)_{n\geq 1}$ (resp. μ) have a first moment, their Fourier Transforms are $\|m_1(\mu_n)\|_2$ (resp. $\|m_1(\mu)\|_2$)-Lipschitz continuous and as the convergence for W_p implies the convergence of W_1,

  $$\|m_1(\mu_n)\|_2 \to \|m_1(\mu)\|_2.$$

  Then these Fourier Transforms are C-Lipschitz continuous with $C := \sup(\|m_1(\mu_n)\|_2)$. Then

  $$|\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \leq 4(1 \wedge C^2\|\xi\|^2) \in \mathrm{L}^1(\Lambda). \tag{27}$$

  As $(\mu_n)_{n\geq 1}$ converges weakly to μ, their Fourier transforms converge to $\hat{\mu}$. By (27) and Dominated Convergence Theorem,

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

  Then $d_k^2(\mu_n, \mu) = d_\Lambda^2(\mu_n, \mu) + d_\Sigma^2(\mu_n, \mu) \to 0$.

- if $d_k(\mu_n, \mu) \to 0$, then $d_\Sigma(\mu_n, \mu) = \|m_p(\mu_n) - m_p(\mu)\|_\Sigma \to 0$ so by the Lemma 4.7

  $$m_p(\mu_n) \to m_p(\mu).$$

  Then the sequence $(\mu_n)_{n\geq 1}$ is tight by the Markov Inequality. Moreover as

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

  the measure μ is the unique adherent value of the sequence $(\mu_n)_{n\geq 1}$ then by the Prokhorov's theorem, $(\mu_n)_{n\geq 1}$ converges weakly to μ. However, the weak convergence and the convergence of the absolute moment of order p implies the W_p convergence, then $W_p(\mu_n, \mu) \to 0$.

□

4.2.4 Proof of Subsection 3.5

The proof of the Proposition 3.8 is based on this lemma. We denote by $*$ the convolution product.

Lemma 4.8. *For $\varphi \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ with $\|\varphi\|_{\text{Lip}}$ and let F be a probability 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 and $h_\sigma(x) = (\sigma\sqrt{2\pi})^{-d} \exp(-\|x\|_2^2/2\sigma^2)$.

Proof. The proof of this lemma is present in Ouvrard (2004). This equality is not directly written. So we will quickly prove the equality. One has

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi * h_\sigma \, \mathrm{d}F &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi(y) h_\sigma(t-y) \, \mathrm{d}y F(\mathrm{d}t) \\ &= \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} h_\sigma(t-y) \, F(\mathrm{d}t)\mathrm{d}y,\end{aligned}$$

by the Fubini Theorem. The lemma 12.5 of this reference states

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

And so using this last equality, we get the desired result. □

Proof of Proposition 3.8. Let $\varphi \colon \mathbb{R}^d \to [-1, 1]$ be a 1-Lipschitz continuous function bounded by the constant 1. Let $\mu, \nu \in \mathcal{T}$, we define

$$\mu(\varphi) = \int_{\mathbb{R}^d} \varphi(x) \, \mu(\mathrm{d}x) \quad \text{and} \quad \nu(\varphi) = \int_{\mathbb{R}^d} \varphi(x) \, \nu(\mathrm{d}x).$$

Let $h_\sigma(x) = (\sigma\sqrt{2\pi})^{-d} \exp(-\|x\|_2^2/2\sigma^2)$ denote the multivariate Gaussian density function with standard deviation $\sigma > 0$. We use an approximation argument and consider, for a sequence $\sigma_n \to 0$, the approximations

$$\mu_n(\varphi) = \int_{\mathbb{R}^d} \varphi * h_{\sigma_n} \, \mathrm{d}\mu \quad \text{and} \quad \nu_n(\varphi) = \int_{\mathbb{R}^d} \varphi * h_{\sigma_n} \, \mathrm{d}\nu.$$

Note that the convolution is well-defined because φ is bounded and h_{σ_n} is integrable. Since the function φ is 1-Lipschitz continuous, we have

$$\|\varphi - \varphi * h_{\sigma_n}\|_\infty \leq \int_{\mathbb{R}^d} \|y\|_2 h_{\sigma_n}(y) \, \mathrm{d}y = \sigma_n \times m_d \to 0, \tag{28}$$

where m_d is the absolute moment of a d dimensional standard gaussian. Let $\varepsilon > 0$ and $N \in \mathbb{N}$ be such that $\|\varphi - \varphi * h_{\sigma_N}\|_\infty < \varepsilon$ then

$$\left|\mu(\varphi) - \mu_N(\varphi)\right| \leq \varepsilon \quad \text{and} \quad \left|\nu(\varphi) - \nu_N(\varphi)\right| \leq \varepsilon,$$

whence we deduce

$$\left|\mu(\varphi) - \nu(\varphi)\right| \leq \left|\mu_N(\varphi) - \nu_N(\varphi)\right| + 2\varepsilon. \tag{29}$$

Next, we introduce the characteristic function $\hat{\mu}$ (resp. $\hat{\nu}$) of μ (resp. ν). By Lemma 4.8

$$\mu_N(\varphi) = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \hat{\mu}(t) h_1(\sigma_N t) e^{-iy\cdot t} \, \mathrm{d}t\mathrm{d}y,$$

and the same equality holds for $\nu_N(\varphi)$ with $\hat{\mu}$ replaced by $\hat{\nu}$. Taking the difference, we get

$$\big|\mu_N(\varphi) - \nu_N(\varphi)\big| = \Big| \Big(\sqrt{2\pi}\Big)^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \big(\hat{\mu}(t) - \hat{\nu}(t)\big) h_1(\sigma_N t) e^{-iy\cdot t} \, \mathrm{d}t\mathrm{d}y \Big|.$$

Assuming that φ has compact support included in the ball $B(0,K)$ with center 0 and radius K, noted shortly $\mathrm{supp}(\varphi) \subset B(0,K)$, we deduce

$$\begin{aligned}
\big|\mu_N(\varphi) - \nu_N(\varphi)\big|^2 &\leq (2\pi)^{-d} \lambda_d(B(0,K))^2 \Big[\int_{\mathbb{R}^d} \big|\hat{\mu}(t) - \hat{\nu}(t)\big| h_1(\sigma_N t) \, \mathrm{d}t \Big]^2 \\
&\leq (2\pi)^{-d} \lambda_d(B(0,K))^2 \times \int_{\mathbb{R}^d} \|t\|^{d+2\alpha} h_1^2(\sigma_N t) \, \mathrm{d}t \times \int_{\mathbb{R}^d} \frac{|\hat{\mu}(t) - \hat{\nu}(t)|^2}{\|t\|^{d+2\alpha}} \, \mathrm{d}t \\
&= C^2 \times d_\alpha^2(\mu,\nu), \qquad (30)
\end{aligned}$$

where C does not depend to μ and ν. In order to prove the result for the Fortet-Mourier distance, we have to remove the support constraint. The tightness of the subset $\mathcal{T}$ will be useful for this purpose. For any $\varepsilon > 0$, we can choose $K > 1$ such that $\mu(B(0,K)^c) < \varepsilon$ for all $\mu \in \mathcal{T}$. Let $\chi : \mathbb{R}^d \to \mathbb{R}$ be the 1-Lipschitz continuous function defined as in Lemma 4.5,

$$\chi(x) = \mathbb{1}_{\|x\|_2 \leq K-1} + (K - \|x\|_2) \mathbb{1}_{K-1 < \|x\|_2 < K}.$$

The decomposition $\varphi = \chi\varphi + (1 - \chi\varphi)$ implies

$$\begin{aligned}
|\mu(\varphi) - \nu(\varphi)| &\leq |\mu(\chi\varphi) - \nu(\chi\varphi)| + |\mu((1-\chi)\varphi) - \nu((1-\chi)\varphi)| \\
&\leq 2|\mu(\chi\varphi/2) - \nu(\chi\varphi/2)| + 2\varepsilon.
\end{aligned}$$

where $\chi\varphi/2$ is 1-Lipschitz continuous with values in $[-1,1]$ and support included in $B(0,K)$. Taking the supremum over the 1-Lipschitz continuous function $\varphi : \mathbb{R}^d \to [-1,1]$, we get

$$d_{FM}(\mu,\nu) \leq 2 \sup_{\mathrm{supp}(\varphi) \subset B(0,K)} |\mu(\varphi) - \nu(\varphi)| + 2\varepsilon. \qquad (31)$$

By combining, the Equation (29) to (31),

$$d_{FM}(\mu,\nu) \leq 2C d_\alpha(\mu,\nu) + 4\varepsilon$$

□

Proof of the Proposition 3.9. Note that by the dual representation of the Wasserstein distance W_1, we can consider the supremum over 1-Lipschitz function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ and $\varphi(0) = 0$. Such functions satisfy $|\varphi(x)| \leq \|x\|$. Let $\varepsilon > 0$. By the definition of uniformly integrability, there is $K > 1$ such that

$$\int_{B(0,K)^c} \|x\| \, \mu(\mathrm{d}x) < \varepsilon \quad \text{for all } \mu \in \mathcal{T}.$$

Recall the function χ from the proof of Proposition 3.8 which is 1-Lipschitz continuous and such that

$$\mathrm{supp}(\chi) \subset B(0,K+1), \ 0 \leq \chi \leq 1 \text{ and } \chi \equiv 1 \text{ on B(0,K)}.$$

Then, in the decomposition $\varphi = \chi\varphi + (1-\chi)\varphi$, the function $\chi\varphi$ is uniformly bounded and $(K+2)-$Lipschitz with $\|\chi\varphi\|_\infty \leq K+1$. We deduce

$$\Big| \int_{\mathbb{R}^d} \chi\varphi \, \mathrm{d}\mu - \int_{\mathbb{R}^d} \chi\varphi \, \mathrm{d}\nu \Big| \leq (K+2) d_{FM}(\mu,\nu).$$

On the other hand, since $(1-\chi)\varphi$ vanishes on $B(0,K)$ and is bounded from above by the norm of x,

$$\Big| \int_{\mathbb{R}^d} (1-\chi)\varphi \, \mathrm{d}\mu - \int_{\mathbb{R}^d} (1-\chi)\varphi \, \mathrm{d}\nu \Big| \leq \varepsilon.$$

We deduce

$$W_1(\mu,\nu) \leq (K+2) d_{FM}(\mu,\nu) + \varepsilon.$$

Finally, Proposition 3.8 implies the desired inequality. □

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Auricchio, G., Codegoni, A., Gualandi, S., Toscani, G., and Veneroni, M. (2020). The equivalence of fourier-based and wasserstein metrics on imaging problems. *Rendiconti Lincei - Matematica e Applicazioni*, 31:627–649.

Bayraktar, E. and Guo, G. (2021). Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA. With a preface by Persi Diaconis.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In Balcan, M. F. and Weinberger, K. Q., 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. PMLR.

Cohen, S. and Istas, J. (2013). *Fractional Fields and Applications*. Mathématiques et Applications. Springer. volume 76.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

Diestel, J. and Uhl, Jr., J. J. (1977). *Vector measures*. American Mathematical Society, Providence, R.I. With a foreword by B. J. Pettis, Mathematical Surveys, No. 15.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). 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. AUAI Press.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a wasserstein loss. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129.

Herbin, E. and Merzbach, E. (2007). *The Multiparameter Fractional Brownian Motion*, pages 93–101. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced wasserstein distances. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Li, Y., Swersky, K., and Zemel, R. (2015). 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.

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

Ouvrard, J.-Y. (2004). *Probabilité 2*. Cassini.

Simon-Gabriel, C.-J., Barp, A., Schölkopf, B., and Mackey, L. (2021). Metrizing weak convergence with maximum mean discrepancies.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E., editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

Steinwart, I. and Ziegel, J. F. (2021). Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*.

Székely, G. and Rizzo, M. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22:151–183.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236 – 1265.

Vayer, T. and Gribonval, R. (2021). Controlling Wasserstein distances by Kernel norms with application to Compressive Statistical Learning. working paper or preprint.

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Yaglom, A. and Silverman, R. (1962). *An Introduction to the Theory of Stationary Random Functions*. Selected Russian publications in the mathematical sciences. Prentice-Hall.