Controlling Wasserstein Distances by Kernel Norms with Application to Compressive Statistical Learning

Titouan Vayer, Rémi Gribonval

Journal of Machine Learning Research 24 (2023) 1-51 Submitted 12/21; Revised 3/23; Published 4/23

Controlling Wasserstein Distances by Kernel Norms with Application to Compressive Statistical Learning

Titouan Vayer TITOUAN.VAYER@INRIA.FR
Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France

Rémi Gribonval REMI.GRIBONVAL@INRIA.FR
Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France

Editor: Marco Cuturi

Abstract

Comparing probability distributions is at the crux of many machine learning algorithms. Maximum Mean Discrepancies (MMD) and Wasserstein distances are two classes of distances between probability distributions that have attracted abundant attention in past years. This paper establishes some conditions under which the Wasserstein distance can be controlled by MMD norms. Our work is motivated by the *compressive statistical learning* (CSL) theory, a general framework for resource-efficient large scale learning in which the training data is summarized in a single vector (called *sketch*) that captures the information relevant to the considered learning task. Inspired by existing results in CSL, we introduce the *Hölder Lower Restricted Isometric Property* and show that this property comes with interesting guarantees for compressive statistical learning. Based on the relations between the MMD and the Wasserstein distances, we provide guarantees for compressive statistical learning by introducing and studying the concept of *Wasserstein regularity* of the learning task, that is when some task-specific metric between probability distributions can be bounded by a Wasserstein distance.

Keywords: optimal transport, maximum mean discrepancy, statistical learning, compressive learning, kernel methods, inverse problems.

1. Introduction

Countless methods in machine learning (ML) and data science rely on comparing probability distributions. Whether it is to measure errors between parametric models and empirical datasets or to produce statistical tests, a recurring problem is to define loss functions that could faithfully quantify the discrepancy between two probability distributions π and π'. Divergences and metrics are frequently used to address this problem and are at the core of numerous works, ranging from signal processing (Kolouri et al., 2017), generative modeling (Arjovsky et al., 2017; Genevay et al., 2018), supervised and semi-supervised learning (Frogner et al., 2015; Solomon et al., 2014), fairness (Gordaliza et al., 2019), two-sample testing (Gretton et al., 2012) or in information theory (Liese and Vajda, 2006). The choice of such a metric is an important issue, as finding a suitable one is delicate and often depends on many criteria such as its associated topology, its computational cost, the type of the problem being considered, the task at hand ... Consequently it is often of great interest to understand the links/relationships between them. *Integral Probability Metrics* (IPMs) introduced by Mueller (1997) (see also Sriperumbudur et al., 2009, 2012) offer an important class of distances that take the form

$$d_{\mathcal{G}}(\pi, \pi') := \sup_{g \in \mathcal{G}} \left| \int g \mathrm{d}\pi - \int g \mathrm{d}\pi' \right|, \tag{1}$$

where π, π' are appropriately integrable distributions and $\mathcal{G}$ is a class of real-valued functions parameterizing the distance. The choice of an adequate function class $\mathcal{G}$ whose generated IPM faithfully describes the "right notion" of discrepancy is not straightforward. One possibility is to choose $\mathcal{G}$ based on the learning task, for example by considering functions $g \in \mathcal{G}$ that depend on the loss and the hypothesis space. This produces *task-specific* pseudo-metrics[1] between probability distributions, abreviated as TaskMetric, that can be used, *inter alia*, to obtain bounds on the generalization error of a learning task (Shalev-Shwartz and Ben-David, 2014; Reid and Williamson, 2011). Another possibility is to rely on *task-agnostic* IPM and to choose $\mathcal{G}$ based on the prior knowledge that this class is appropriate for the task at hand. Notable examples of task-agnostic IPMs include the popular Maximum Mean Discrepancies (MMD) (when $\mathcal{G}$ is the unit ball in a *Reproducible Kernel Hilbert Space* (RKHS), see Berlinet and Thomas-Agnan, 2011) and the 1-Wasserstein distance W_1 (when $\mathcal{G}$ is the class of 1-Lipschitz functions, see Villani, 2008). Both are gaining interest from the machine learning community due to their ability to handle the metric structure of the feature space (see Peyré and Cuturi, 2019; Muandet et al., 2017 and references therein).

Our first contribution is to exhibit some relationships between task-specific metrics between probability distributions, MMD and optimal transport (OT) distances. We first give necessary and sufficient conditions, on the kernel that defines the RKHS, under which the MMD can be bounded by a Wasserstein distance. We study in a second step the other direction, more difficult to obtain, which corresponds to finding the conditions under which the Wasserstein distance W_p can be upper-bounded by an MMD with a "Hölder" exponent, that is when

$$\mathrm{W}_p(\pi, \pi') \lesssim \mathrm{MMD}^{\delta}(\pi, \pi') \text{ for some } \delta \in (0, 1]\,. \tag{2}$$

Especially, we are interested in MMDs associated to RKHSs generated by translation-invariant positive semi-definite kernels that are widely used in many machine learning applications and are at the core of many large-scale learning algorithms (Rahimi and Recht, 2008, 2007). Despite some connections between MMDs and *regularized* OT distances, such as the Sinkhorn divergences (Feydy et al., 2019) or Gaussian smoothed OT (Nietert et al., 2021b; Zhang et al., 2021), little is known regarding the relationships between non-regularized W_p and such MMDs. We show that the bound (2) can not hold in full generality and that one needs to find additional constraints on the distributions π, π'. This will be formalized by the means of a *model set* of distributions $\mathfrak{S}$, so that (2) applies for every $\pi, \pi' \in \mathfrak{S}$. We shed light on several controls of the type (2) depending on the properties of this model set $\mathfrak{S}$ and the TI kernel (see Section 2).

This study is motivated by the compressive statistical learning (CSL) framework whose aim is to provide resource-efficient large-scale learning algorithms (Gribonval et al., 2021a,b; Keriven et al., 2018) and which heavily relies on MMDs with TI kernels. Large-scale ML faces nowadays a number of computational challenges, due to the high dimensionality of data and, often, very large training collections. Compressive statistical learning is one remedy to this situation. Its objective is: 1) to summarize a large dataset $\mathbf{X} \in \mathbb{R}^{d \times n}$, where d is the dimension and n the number of samples, into a single vector $\mathbf{s} \in \mathbb{R}^m$ or $\mathbb{C}^m$ with $m \ll nd$; and 2) to rely *solely* on $\mathbf{s}$ to solve the learning task, such as finding centroids in K-means or learning mixture models (Keriven et al., 2017, 2018; Gribonval et al., 2021b). The generic idea behind compressive learning is that, for many tasks, we only need to have access to informations from a "low-dimensional" subspace, captured by a well-designed sketch vector $\mathbf{s}$.

This framework requires specific statistical tools for establishing learning guarantees compared to standard machine learning approaches. One of the main notion in this context is found in the *Lower Restricted Isometric Property* (LRIP) which is a condition on the sketching operator that maps a dataset to a sketch. However, this property is far from trivial to prove and is usually obtained by: 1) carefully designing a model set of distributions $\mathfrak{S}$; 2) finding a kernel whose MMD dominates

1. A pseudo-metric D satisfies all the axioms of a metric except (possibly) for separation. In other words, D is symmetric $D(x, y) = D(y, x)$, non-negative $D(x, y) \geq 0$, satisfies the triangular inequality $D(x, y) \leq D(x, z) + D(z, y)$ and is such that $D(x, x) = 0$ (but possibly $D(x, y) = 0$ for some $x \neq y$).

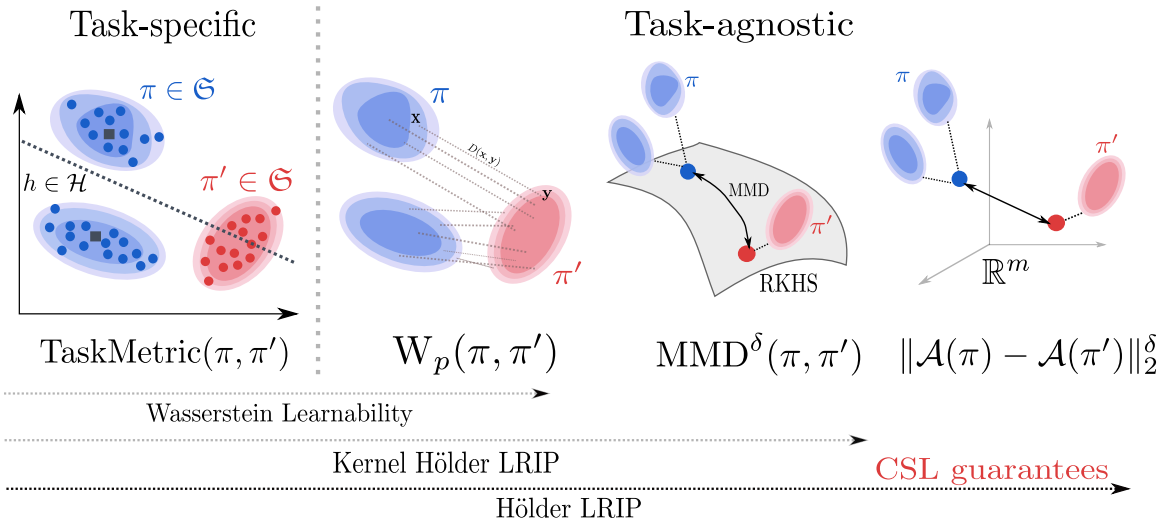

Figure 1: The reasoning used in the paper to obtain compressive statistical learning guarantees. *(left)* Given two distributions π, π' on a model set $\mathfrak{S}$, our goal is to control some task-specific metric $\text{TaskMetric}(\pi, \pi')$ that depends on the learning task. *(middle left)* In Section 3, we use an upper-bound $\text{TaskMetric}(\pi, \pi') \lesssim W_p(\pi, \pi')$ by introducing the notion of Wasserstein regularity of the task. *(middle right)* In Section 2, we first show how to control the MMD by the Wasserstein distance, then we study the other direction that is controling W_p by an MMD with a Hölder exponent $\delta \in (0, 1]$: $W_p(\pi, \pi') \lesssim \text{MMD}^\delta(\pi, \pi')$. *(right)* In Section 4 we discuss how to control the MMD by the distance between the finite dimensional sketches of the distributions $\mathcal{A}(\pi), \mathcal{A}(\pi')$ in $\mathbb{R}^m$. The whole pipeline gives the Hölder LRIP property which allows us to derive CSL guarantees.

TaskMetric, a property being known as the *Kernel LRIP*; and 3) approximating this MMD using random features (Gribonval et al., 2021a).

Based on the relationships between the MMD and the Wasserstein distance discussed above we will show that a slightly different property, namely the *Kernel Hölder LRIP*, can be proved for a wide range of tasks where it is natural to control TaskMetric by a Wasserstein distance (*Wasserstein regularity*). In particular we prove that many unsupervised learning tasks such as *compression-type tasks* (K-means/medians, PCA, see Gribonval et al., 2021a) or supervised learning tasks, such as regression and binary classification with Lipschitz regressors/classifiers, fall into this category. From this study we will propose a property which generalizes the LRIP, namely the *Hölder LRIP*, and we will show that this property also comes with interesting compressive statistical learning guarantees. Figure 1 summarizes the whole reasoning used in this paper to establish these CSL guarantees.

Organization of the paper We start by presenting in Section 2 the relations between the Wasserstein distance and the MMD. We provide conditions so that $W_p \lesssim \text{MMD}^\delta$ holds for some $\delta \in (0, 1]$. In Section 3 we study the relations between task-specific metrics between probability distributions and the Wasserstein distance. For this, we introduce the concept of *Wasserstein regularity* of the learning task. In Section 4 we introduce the compressive statistical learning framework which motivates our study. We study a generalization of the LRIP, namely the Hölder LRIP, and we show that this property has many advantages for CSL.

1.1 Notations and Definitions

We first detail the different usual notations and definitions used in this article.

1.1.1 Metric Spaces

In this article the space $\mathcal{X}$ will always be a complete, separable metric space. The relation $d(\mathbf{x},\mathbf{y}) \lesssim d'(\mathbf{x},\mathbf{y})$ hides a multiplicative constant, *i.e.* $d(\mathbf{x},\mathbf{y}) \leq Cd'(\mathbf{x},\mathbf{y})$ with $C>0$ that *does not depend* on $\mathbf{x},\mathbf{y}$. The class of L-Lipschitz continuous functions from a metric space $(\mathcal{X},d_{\mathcal{X}})$ to $(\mathcal{Y},d_{\mathcal{Y}})$ is denoted by $\mathrm{Lip}_L((\mathcal{X},d_{\mathcal{X}}),(\mathcal{Y},d_{\mathcal{Y}}))$ or simply by $\mathrm{Lip}_L(\mathcal{X},\mathcal{Y})$ when it is clear from context. If $f \in \mathrm{Lip}_L((\mathcal{X},d_{\mathcal{X}}),(\mathcal{Y},d_{\mathcal{Y}}))$ we have $\forall \mathbf{x},\mathbf{x}' \in \mathcal{X}, d_{\mathcal{Y}}(f(\mathbf{x}),f(\mathbf{x}')) \leq Ld_{\mathcal{X}}(\mathbf{x},\mathbf{x}')$. In the following $\|\cdot\|_2$ denotes the ℓ_2 norm, and vectors and matrices are written in bold. On a normed space $(\mathcal{X},\|\cdot\|)$, the ball centered at $\mathbf{x}_0 \in \mathcal{X}$ and with radius $R>0$ is denoted $B_{\|\cdot\|}(\mathbf{x}_0,R)$ or simply by $B(\mathbf{x}_0,R)$ when it is clear from context.

1.1.2 Measures and Probability Distributions

We note $\mathcal{P}(\mathcal{X})$ the set of probability measures on $\mathcal{X}$. $\mathcal{M}(\mathcal{X})$ is the space of finite signed measures on $\mathcal{X}$. For the sake of brevity, for a probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ that admits a density f *w.r.t.* the Lebesgue measure on $\mathbb{R}^d$ we adopt the notation $\pi = f\mathrm{d}\mathbf{x}$. Given a probability distribution $\pi \in \mathcal{P}(\mathcal{X})$ and a measurable function $T : \mathcal{X} \to \mathcal{Y}$ the pushforward operator $\#$ defines a probability distribution $T\#\pi \in \mathcal{P}(\mathcal{Y})$ *via* the relation $T\#\pi(A) = \pi(T^{-1}(A))$ for every measurable set A in $\mathcal{Y}$. In other words, if $X \sim \pi$ is a random variable then $Y = T(X)$ has the law $T\#\pi$. The support of a probability distribution is denoted as $\mathrm{supp}(\pi)$ and it is defined as the smallest closed set S such that $\pi(S) = 1$.

1.1.3 Integrability, Fourier Transform and Sobolev Space

For a measurable space $\mathcal{X}$ and a Borel measure μ on $\mathcal{X}$ we note $L_p(\mu)$ the space of real-valued p-integrable functions *w.r.t* μ, *i.e.* that satisfy $\int_{\mathcal{X}} |f(\mathbf{x})|^p \mathrm{d}\mu(\mathbf{x}) < +\infty$. When $\mathcal{X} = \mathbb{R}^d$ we note $L_p(\mathbb{R}^d)$ the space of p-integrable functions with respect to the Lebesgue measure. For an integrable function $f \in L_1(\mathbb{R}^d)$ we adopt the convention of the Fourier transform $\hat{f}(\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega}) := \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} f(\mathbf{x})\mathrm{d}\mathbf{x}$. The Fourier transform of a non-negative finite measure $\mu \in \mathcal{M}_+(\mathbb{R}^d)$ is defined for $\boldsymbol{\omega} \in \mathbb{R}^d$ by $\widehat{\mu}(\boldsymbol{\omega}) := \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\mu(\mathbf{x})$. For $s \geq 0$, we define the Sobolev space of order s as (Adams and Fournier, 2003):

$$H^s(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \boldsymbol{\omega} \to (1+\|\boldsymbol{\omega}\|_2^2)^{s/2} \mathcal{F}[f](\boldsymbol{\omega}) \in L_2(\mathbb{R}^d) \right\}.$$

It is a Hilbert space whose corresponding norm is $\|f\|_{H^s(\mathbb{R}^d)} := \left(\int_{\mathbb{R}^d} (1+\|\boldsymbol{\omega}\|_2^2)^s |\mathcal{F}[f](\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}\right)^{1/2}$. It corresponds to the space of functions whose weak derivatives up to order s are squared-integrable.

2. Controlling Wasserstein Distances by Kernel Norms

We focus in this section on the first main contributions of this paper, that is the comparison of optimal transport distances and maximum mean discrepancies. We begin by describing the main notions related to these two metrics.

The interest of optimal transport lies in both its ability to provide correspondences between sets of points and its ability to induce a geometric notion of distance between probability distributions thanks to the popular Wasserstein distances (Villani, 2008; Santambrogio, 2015; Peyré and Cuturi, 2019). Considering a complete and separable metric space $(\mathcal{X},D)$ and $p \in [1,+\infty)$, the Wasserstein

distance of order p between two probability distributions $\pi, \pi' \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathrm{W}_p(\pi, \pi') := \left(\inf_{\gamma \in \Pi(\pi, \pi')} \int_{\mathcal{X} \times \mathcal{X}} D(\mathbf{x}, \mathbf{y})^p \mathrm{d}\gamma(\mathbf{x}, \mathbf{y}) \right)^{1/p}, \tag{3}$$

where $\Pi(\pi, \pi')$ is the set of couplings of π and π' *i.e.* the set of joint distributions $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that both marginals of γ are respectively π and π'. More formally $\Pi(\pi, \pi') = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \forall A, B \subseteq \mathcal{X}, \gamma(A \times \mathcal{X}) = \pi(A),\ \gamma(\mathcal{X} \times B) = \pi'(B)\}$. This quantity satisfies all the axioms of a distance and endows the space

$$\mathcal{P}_p(\mathcal{X}) := \{\pi \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} D(\mathbf{x}_0, \mathbf{y})^p \mathrm{d}\pi(\mathbf{y}) < +\infty \text{ for some arbitrary } \mathbf{x}_0 \in \mathcal{X}\},$$

with a metric structure[2] (Villani, 2008). When $(\mathcal{X}, D)$ is a normed space such as $(\mathbb{R}^d, \|\cdot\|_2)$ the space $\mathcal{P}_p(\mathcal{X})$ is the space of probability distributions with finite p-th moment $\int_{\mathcal{X}} \|\mathbf{x}\|_2^p \mathrm{d}\pi(\mathbf{x}) < +\infty$. More generally, we can define OT problems by using a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ instead of a distance D and by minimizing the quantity $\int c(\mathbf{x}, \mathbf{y}) \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$ over $\gamma \in \Pi(\pi, \pi')$. With a slight abuse of terminology we will denote the optimal value of both problems by the term *Wasserstein distance* and we will specify, when necessary, the choice of the cost function. A coupling γ^* minimizing (3) is called *optimal coupling* and it provides a probabilistic matching of the points in the support of the distributions π, π'. As such, computing an OT distance equals to finding the most cost-efficient way to "match" one distribution to the other. An important property of the Wasserstein distance relies on its dual formulation. It allows, among others, to characterize W_1 by considering the maximization problem

$$\mathrm{W}_1(\pi, \pi') = \sup_{f \in \mathrm{Lip}_1(\mathcal{X}, \mathbb{R})} \left| \int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y}) \right|,$$

where $\mathrm{Lip}_1(\mathcal{X}, \mathbb{R})$ is the set of 1-Lipschitz function from $(\mathcal{X}, D)$ to $\mathbb{R}$ (Santambrogio, 2015).

The other important technical ingredient of this section, the theory of kernels, has a long history when it comes to learning problems or more generally to probability and statistics (Aronszajn, 1950; Berlinet and Thomas-Agnan, 2011; Muandet et al., 2017). In the rest of the paper κ will denote a *positive semi-definite* (PSD) kernel[3] on a space $\mathcal{X}$. It defines a Hilbert space of functions from $\mathcal{X}$ to $\mathbb{C}$ denoted by $\mathcal{H}_\kappa$ endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$. This space is called a reproducing kernel Hilbert space and is characterized by the property $\forall \mathbf{x} \in \mathcal{X}, \kappa(\cdot, \mathbf{x}) \in \mathcal{H}_\kappa$ and the reproducing property: each $f \in \mathcal{H}_\kappa$ can be evaluated as $f(\mathbf{x}) = \langle f, \kappa(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_\kappa}$ for any $\mathbf{x} \in \mathcal{X}$. A PSD kernel also defines the so-called *Maximum Mean Discrepancy* (MMD) which can be used to compare two probability distributions $\pi \in \mathcal{P}(\mathcal{X})$ and $\pi' \in \mathcal{P}(\mathcal{X})$ with the formula[4]

$$\mathrm{MMD}_\kappa(\pi, \pi') := \left(\mathop{\mathbb{E}}_{\mathbf{x}, \mathbf{x}' \sim \pi} [\kappa(\mathbf{x}, \mathbf{x}')] + \mathop{\mathbb{E}}_{\mathbf{y}, \mathbf{y}' \sim \pi'} [\kappa(\mathbf{y}, \mathbf{y}')] - 2 \operatorname{Re}(\mathop{\mathbb{E}}_{\mathbf{x} \sim \pi, \mathbf{y} \sim \pi'} [\kappa(\mathbf{x}, \mathbf{y})]) \right)^{1/2}.$$

This quantity defines a pseudo-metric on the space of probability distributions and is a true metric when the kernel is *characteristic*: $\mathrm{MMD}_\kappa(\pi, \pi') = 0 \iff \pi = \pi'$ (Simon-Gabriel et al., 2020; Sriperumbudur et al., 2010). The MMD is also characterized by the relation $\mathrm{MMD}_\kappa(\pi, \pi') = \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} |\int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{x}) \mathrm{d}\pi'(\mathbf{x})|$. Moreover, it can be extended to any finite signed measure $\mu \in \mathcal{M}(\mathcal{X})$ by defining a semi-norm[5] on $\mathcal{M}(\mathcal{X})$ with the formula

$$\|\mu\|_\kappa := \left(\int \int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{y}) \right)^{1/2}. \tag{4}$$

2. The space $\mathcal{P}_p(\mathcal{X})$ is here to formalize that W_p is finite and thus defines a proper distance.
3. A function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a PSD kernel if it is *Hermitian i.e.* $\kappa(\mathbf{x}, \mathbf{y}) = \overline{\kappa(\mathbf{y}, \mathbf{x})}$ and for all $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{X}$ and any $c_1, \cdots, c_n \in \mathbb{C}$ we have $\sum_{i,j=1}^n c_i \overline{c_j} \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
4. When the kernel κ is bounded, the MMD $\|\pi - \pi'\|_\kappa$ is finite for any probability distributions π, π'.
5. A semi-norm $\|\cdot\|$ on a vector space is non-negative, satisfies the triangle inequality, is such that: a) if $\mathbf{x} = 0$ then $\|\mathbf{x}\| = 0$ (but not necessarily the converse); and b) for $\lambda \in \mathbb{R}$, $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$.

When κ is a PSD kernel this quantity is well defined, *i.e.* the integral in (4) is non-negative, and we have $\forall \pi, \pi' \in \mathcal{P}(\mathcal{X}), \mathrm{MMD}_\kappa(\pi, \pi') = \|\pi - \pi'\|_\kappa$. In the rest of the paper we informally denote $\|\cdot\|_\kappa$ by the term *kernel norm* or *MMD norm*. An important family of kernels, namely *translation-invariant (TI), PSD kernels*, are particularly interesting in our context. They are defined for $\mathcal{X} = \mathbb{R}^d$ and when $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ for some *continuous* PSD function[6] $\kappa_0 : \mathbb{R}^d \to \mathbb{C}$. This family encompasses many popular kernels such as Gaussian or Laplacian kernels, or kernels of the Matèrn class (Sriperumbudur et al., 2010). The following characterization of such kernels is due to the celebrated Bochner's theorem (see Theorem 6.6 and Theorem 6.11 in Wendland, 2004):

Theorem 1 (Bochner) *Let $\kappa_0 : \mathbb{R}^d \to \mathbb{C}$. A function κ of the form $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$, where κ_0 is continuous, is a PSD kernel if and only if there exists a probability distribution $\Lambda \in \mathcal{P}(\mathbb{R}^d)$ such that*

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \kappa_0(\mathbf{x}) = \kappa_0(0) \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\Lambda(\boldsymbol{\omega})\,.$$

If κ_0 is continuous and *in $L_1(\mathbb{R}^d)$ then $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ is a PSD kernel if and only if $\forall \boldsymbol{\omega} \in \mathbb{R}^d, \widehat{\kappa_0}(\boldsymbol{\omega}) \geq 0$.*

Bochner's theorem shows that a translation invariant PSD kernel κ (when properly scaled to ensure $\kappa_0(0) = 1$) can be written as an expectation $\kappa(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{\omega} \sim \Lambda}[\phi(\mathbf{x}, \boldsymbol{\omega})\overline{\phi(\mathbf{y}, \boldsymbol{\omega})}]$ where $\Lambda \in \mathcal{P}(\mathbb{R}^d)$ and $\phi(\mathbf{x}, \boldsymbol{\omega}) = e^{-i\boldsymbol{\omega}^\top \mathbf{x}}$. An interesting property of such kernels is that they can be approximated using finite dimensional vectors by sampling from the frequencies $\boldsymbol{\omega} \sim \Lambda$ and approximating $\mathbb{E}_{\boldsymbol{\omega} \sim \Lambda}[\phi(\mathbf{x}, \boldsymbol{\omega})\overline{\phi(\mathbf{y}, \boldsymbol{\omega})}]$ using a Monte-Carlo algorithm (Li et al., 2021; Sutherland and Schneider, 2015; Sriperumbudur and Szabo, 2015). This property is at the core of methods that rely on *random Fourier features* to accelerate kernel learning algorithms (Rahimi and Recht, 2007, 2008).

2.1 Controlling MMDs by Wasserstein distances

When it comes to comparing MMD and W_p, one direction is easier: controlling MMD by W_p. More precisely we have the following result (the proof can be found in Appendix A.1):

Proposition 2 *Let $(\mathcal{X}, D)$ be a complete separable metric space, $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PSD kernel, $\mathcal{H}_\kappa$ the associated RKHS and $B_\kappa := \{f \in \mathcal{H}_\kappa : \|f\|_{\mathcal{H}_\kappa} \leq 1\}$ the unit ball in $\mathcal{H}_\kappa$. Consider the Wasserstein distances computed with the metric D. For any $C > 0$ the following statements are equivalent:*

(i)
$$B_\kappa \subseteq \mathrm{Lip}_C((\mathcal{X}, D), \mathbb{R}) \tag{5}$$

(ii)
$$\forall p \in [1, +\infty), \forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_\kappa \leq C\, \mathrm{W}_p(\pi, \pi') \tag{6}$$

(iii)
$$\exists p \in [1, +\infty), \forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_\kappa \leq C\, \mathrm{W}_p(\pi, \pi') \tag{7}$$

(iv)
$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X},\ \kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{y}, \mathbf{y}) - 2\kappa(\mathbf{x}, \mathbf{y}) \leq C^2 D^2(\mathbf{x}, \mathbf{y}) \tag{8}$$

For the sake of clarity, we restrict ourselves to the case where D is a proper metric but extensions of this result are possible by considering an OT problem with a more general cost. In particular, this type of bound has already been considered in Arbel et al. (2018); Sriperumbudur et al. (2010) with the pseudo-metric $D(\mathbf{x}, \mathbf{y}) = \|\kappa(\mathbf{x}, \cdot) - \kappa(\mathbf{y}, \cdot)\|_{\mathcal{H}_\kappa}$ which gives $C = 1$ and an equality in (8). As a corollary of this proposition we have the following result (see Appendix A.1 for a proof):

6. A function $\kappa_0 : \mathbb{R}^d \to \mathbb{C}$ is PSD if for all $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and $c_1, \cdots, c_n \in \mathbb{C}$ we have $\sum_{i,j=1}^n c_i \overline{c_j} \kappa_0(\mathbf{x}_i - \mathbf{x}_j) \geq 0$. Such function is bounded $|\kappa_0(\mathbf{x})| \leq \kappa_0(0)$ and satisfies $\kappa_0(-\mathbf{x}) = \overline{\kappa_0(\mathbf{x})}$ (Wendland, 2004, Theorem 6.2). When κ_0 is even ($\kappa_0(-\mathbf{x}) = \kappa_0(\mathbf{x})$) then κ_0 and thus κ are real-valued.

Corollary 3 *Consider $\mathcal{X} = \mathbb{R}^d$ equipped with the Euclidean distance $D(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$ and a PSD kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that is normalized, i.e. $\kappa(\mathbf{x}, \mathbf{x}) = 1$ for every $\mathbf{x} \in \mathcal{X}$. Assume that for each $\mathbf{x} \in \mathcal{X}$ the function $\phi_{\mathbf{x}} : \mathbf{y} \mapsto \kappa(\mathbf{x}, \mathbf{y})$ is C^2 in a neighborhood of $\mathbf{x}$, and denote $\mathbf{H}_{\mathbf{x}} = -\nabla^2[\phi_{\mathbf{x}}](\mathbf{x})$ its negative Hessian matrix evaluated at $\mathbf{x}$. Then the following holds:*

(i) Any of the four equivalent properties of Proposition 2 implies

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \lambda_{\max}(\mathbf{H}_{\mathbf{x}}) \leq C^2 \,, \tag{9}$$

where $\lambda_{\max}(\mathbf{H}_{\mathbf{x}})$ denotes the largest eigenvalue of $\mathbf{H}_{\mathbf{x}}$.

(ii) If κ is translation invariant, i.e. $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, then conversely, (6) holds with $C := \sqrt{\sup_{\mathbf{x}} \lambda_{\max}(\mathbf{H}_{\mathbf{x}})} = \sqrt{\lambda_{\max}(-\nabla^2[\kappa_0](0))}$.

The second point of the previous result shows that under mild assumptions on a TI kernel the MMD is bounded by a constant times a Wasserstein distance, for *any* distributions π, π' for which these quantities are well-defined. In particular it holds for popular kernels such as the Gaussian kernel, or kernels of the Matérn class with parameter[7] $\nu > 1$:

Example 4 *An important family of TI kernels is the Matérn class (Rasmussen and Williams, 2005, Section 4.2.1), given in any dimension by the relation $\kappa(\mathbf{x}, \mathbf{y}) := \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2}{\sigma})^\nu K_\nu(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{y}\|_2}{\sigma})$ for $\nu > 0, \sigma > 0$ where Γ is the gamma function, and K_ν is the modified Bessel function of the second kind of order ν. This family of kernel admits the following Fourier transform[8] :*

$$\widehat{\kappa_0}(\boldsymbol{\omega}) = \frac{2^{d+\nu}\pi^{d/2}\Gamma(\nu + d/2)\nu^\nu}{\Gamma(\nu)\sigma^{2\nu}} \left(\frac{2\nu}{\sigma^2} + \|\boldsymbol{\omega}\|_2^2\right)^{-(\nu+d/2)} . \tag{10}$$

Interestingly, $\nu = \frac{1}{2}$ corresponds to the Laplacian kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2/\sigma)$ whose Fourier transform is $\frac{2^d \pi^{\frac{d-1}{2}} \Gamma(\frac{d+1}{2})}{\sigma}(\frac{1}{\sigma^2} + \|\boldsymbol{\omega}\|_2^2)^{-\frac{d+1}{2}}$ while $\nu \to +\infty$ recovers the RBF kernel see Rasmussen and Williams (2005, Section 4.2.1)[9].

Note that when the kernel is TI but is not normalized the second point of Corollary 3 holds also with $C = \kappa_0(0)\sqrt{\lambda_{\max}(-\nabla^2[\kappa_0](0))}$. For other types of *normalized* kernels, condition (8) is a necessary and sufficient condition that amounts to checking if there is a constant $C > 0$ such that $1 - \kappa(\mathbf{x}, \mathbf{y}) \leq \frac{C^2}{2} D^2(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Interestingly, it echoes the "C-strongly locally characteristic" property of the kernel as in Gribonval et al. (2021b, Definition 5.14) but with the reverse inequality. When the kernel is C^2 a necessary condition is given by the maximum eigenvalue of the negative Hessian as in (9).

Overall Proposition 2 shows that it is not too difficult to find necessary and sufficient conditions under which the MMD can be controlled by a Wasserstein distance. What is more difficult to characterize is the inequality in the other direction.

2.2 Controlling Wasserstein distances by MMDs ?

Thereafter, the objective is thus to find reasonable conditions on a subset of probability distributions $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ and on a PSD kernel κ such that the Wasserstein distance can be controlled with the MMD with kernel κ uniformly on $\mathfrak{S}$. We adopt the following definition:

7. In this case κ_0 is C^2 in a neighbourhood of 0 since $\kappa_0 \in L_1(\mathbb{R}^d)$ and $\boldsymbol{\omega} \to \|\boldsymbol{\omega}\|_2^2 \widehat{\kappa_0}(\boldsymbol{\omega}) \in L_1(\mathbb{R}^d)$ when $\nu > 1$
8. See Rasmussen and Williams (2005, Section 4.2.1) with slightly modified conventions on Fourier transforms.
9. Likewise, with adapted conventions on Fourier transforms.

Definition 5 *Let $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ be a subset of probability distributions, $p \in [1, +\infty)$, κ a real-valued PSD kernel on $\mathcal{X}$ and $\delta \in (0, 1]$. We say that the space $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable with error $\eta \geq 0$ if*

$$\exists C > 0, \forall \pi, \pi' \in \mathfrak{S}, \mathrm{W}_p(\pi, \pi') \leq C \|\pi - \pi'\|_\kappa^\delta + \eta \,. \tag{11}$$

When $\eta = 0$ we simply say that $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable.

Note that the constants C, η, δ in (11) *do not depend* on the probability distributions π, π': we want to bound uniformly on the whole subset $\mathfrak{S}$. In the following, we will call *model set* this subset $\mathfrak{S}$. As discussed later in Section 4, introducing $\mathfrak{S}$ will also be crucial in order to obtain compressive statistical learning guarantees. Moreover, we are particularly interested in establishing such an inequality for translation-invariant PSD kernels that at the core of the CSL theory since they admit a random Fourier feature expansion useful to find a sketching operator based on random Fourier features (Gribonval et al., 2021a).

Remark 6 *An immediate consequence of Definition 5 is that when $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable (i.e. with no error) then the kernel κ is necessarily characteristic to $\mathfrak{S}$ (Simon-Gabriel et al., 2020, Section 1.2), in other words $\|\pi - \pi'\|_\kappa = 0 \iff \pi = \pi'$ for all $\pi, \pi' \in \mathfrak{S}$ (indeed when the MMD vanishes then the Wasserstein distance also vanishes which implies equality of the distributions). Moreover, if $(\mathfrak{S}, \mathrm{W}_p)$ is $(\kappa, \delta = 1)$-embeddable and if the condition (8) is also fulfilled, then W_p and $\|\cdot\|_\kappa$ induce the same topology on $\mathfrak{S}$ and define equivalent metrics on $\mathfrak{S}$.*

Remark 7 *If $\mathfrak{S} \subseteq \mathfrak{S}'$ where $(\mathfrak{S}', \mathrm{W}_p)$ is (κ, δ)-embeddable then $(\mathfrak{S}, \mathrm{W}_p)$ is also (κ, δ)-embeddable. In other words, if $\mathfrak{S}$ is contained in a space that is (κ, δ)-embeddable it is also (κ, δ)-embeddable. On the other hand, if $\mathfrak{S}'$ contains a subspace $\mathfrak{S}$ for which there is a necessary condition to the (κ, δ)-embeddability property then the same condition applies to $\mathfrak{S}'$.*

In the following we focus on property (11) with no error $\eta = 0$. First we consider necessary conditions, that is, we argue that property (11) with no error can only be expected to hold for a kernel κ and a model set $\mathfrak{S}$ if certain appropriate assumptions are made. Conversely, we then derive some sufficient conditions on $\mathfrak{S}$ and κ such that $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable.

2.3 Necessary Conditions

Let us first review some necessary conditions for property (11) with no errror.

2.3.1 Boundedness of the Model Set is Necessary.

Consider a model set $\mathfrak{S} \subseteq \mathcal{P}_1(\mathbb{R}^d)$ and denote by

$$\mathrm{m}(\pi) := \int \mathbf{x} \mathrm{d}\pi(\mathbf{x})$$

the mean of $\pi \in \mathcal{P}_1(\mathbb{R}^d)$. On the one hand, simple calculus (Lemma 42 in Appendix A.3) shows that for any $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ and $p \in [1, +\infty)$, if W_p is defined based on some norm $\|\cdot\|$ and $\|\cdot\|_\star$ denotes the dual norm defined by $\|\mathbf{z}\|_\star = \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{z} \rangle$, then

$$\mathrm{W}_p(\pi, \pi') \geq \|\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_\star.$$

On the other hand, if κ is a bounded PSD kernel (*i.e.*, $\sup_{\mathbf{x}} \kappa(\mathbf{x}, \mathbf{x}) \leq K < +\infty$) then, by the Cauchy-Schwarz inequality for kernels we have $\forall \mathbf{x}, \mathbf{y},\ |\kappa(\mathbf{x}, \mathbf{y})| \leq \sqrt{\kappa(\mathbf{x}, \mathbf{x})}\sqrt{\kappa(\mathbf{y}, \mathbf{y})} \leq K$. Hence, for any $(\pi, \pi') \in \mathfrak{S},\ \|\pi - \pi'\|_\kappa \leq \sqrt{2K}$. As a result, if $\mathfrak{S}$ is unbounded in the sense that $\sup_{\pi, \pi' \in \mathfrak{S}} \|\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_\star = +\infty$, then for each $\delta > 0$,

$$\sup_{(\pi, \pi') \in \mathfrak{S}} \frac{\mathrm{W}_p(\pi, \pi')}{\|\pi - \pi'\|_\kappa^\delta} = +\infty \,. \tag{12}$$

Consequently, we can not have (11) for any $\delta > 0$. Since all norms are equivalent in finite dimension the following lemma holds:

Lemma 8 *Consider $\mathcal{X} = \mathbb{R}^d, p \in [1, +\infty)$ and assume that W_p is based on a norm on $\mathbb{R}^d$. If κ is bounded and $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable for some $\delta > 0$ then $\mathfrak{S}$ is bounded:*

$$\text{m-diam}(\mathfrak{S}) := \sup_{\pi, \pi' \in \mathfrak{S}} \| \mathrm{m}(\pi) - \mathrm{m}(\pi') \|_2 < +\infty \,.$$

2.3.2 Bounds on δ due to the Convergence Rate of Empirical Measures.

Another obstacle to (11) concerns the samples rate of convergence of both terms with empirical measures : it is known that the Wasserstein distance suffers from the curse of dimensionality while the MMD does not. More precisely if $\pi \in \mathcal{P}_1(\mathbb{R}^d)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ then it is known that $\mathbb{E}[\mathrm{W}_1(\pi, \pi_n)] \gtrsim n^{-1/d}$ where $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, $\mathbf{x}_i \sim \pi$ and the expectation is taken *w.r.t.* the draws of $\mathbf{x}_i$ (Dudley, 1969; Weed and Bach, 2019). By monotonicity of W_p in p this is also true for W_p with $p \geq 1$ (since for $p \leq q, \mathrm{W}_p(\pi, \pi) \leq \mathrm{W}_q(\pi, \pi')$ for any[10] π, π'). On the contrary, it is not difficult to see that if the PSD kernel κ is bounded by K then $\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta] \leq (2K)^{\delta/2} n^{-\delta/2}$ (see Lemma 41 in Appendix A.2). Consequently, even when the model set $\mathfrak{S} \subseteq \mathcal{P}_1(\mathbb{R}^d)$ satisfies $\text{m-diam}(\mathfrak{S}) < +\infty$ (to avoid the obstacles to (11) already identified in Lemma 8), if $\mathfrak{S}$ is rich enough to contain a distribution π that is absolutely continuous *w.r.t.* the Lebesgue measure, as well as its empirical distributions π_n for every n, then (11) implies $n^{-1/d} \lesssim n^{-\delta/2}$, so necessarily $\delta \leq 2/d$. An example of such a model set is the set of all probability distributions producing almost surely vectors in a prescribed ball, leading to the following result:

Lemma 9 *Consider $R > 0$, $\Omega = B(0, R) \subseteq \mathcal{X} = \mathbb{R}^d$, $\mathfrak{S} := \{\pi \in \mathcal{P}(\mathcal{X}) : \pi(\Omega) = 1\}$, κ a bounded PSD kernel, and W_p based on a norm in $\mathbb{R}^d$ with $p \in [1, +\infty)$. If $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable then $\delta \leq 2/d$.*

In the context of CSL, as described in Section 4, such $\delta \leq 2/d$ would imply a very slow convergence rate of the order of $O(n^{-\frac{1}{d}})$. In other words, if the strategy described in Section 4 is followed we would require an exponential amount of samples in order to have reasonable CSL guarantees which is problematic for a large scale scenario where d is usually large. This discussion suggests that we must find suitable constraints on p, δ, κ and $\mathfrak{S}$ to avoid such a curse of dimensionality. Sufficient conditions to achieve this goal will be discussed later, but first we continue with some additional necessary conditions.

2.3.3 Another Bound on δ for Certain Model Sets

Another restriction comes from the type of distributions in the model set. We will prove that, as soon as $\mathfrak{S}$ contains two distributions whose supports are disjoint, as well as the convex segment between these distributions, we cannot hope to have (11) with error $\eta = 0$ when $p \cdot \delta > 1$.

Proposition 10 *Let $(\mathcal{X}, D)$ be a complete and separable metric space and consider the Wasserstein distances computed with the distance D. Let κ be any PSD kernel. Consider two arbitrary probability distributions $\pi_0, \pi_1 \in \mathcal{P}(\mathcal{X})$ such that $\|\pi_0 - \pi_1\|_\kappa < +\infty$ and $\mathrm{supp}(\pi_0)$ and $\mathrm{supp}(\pi_1)$ are disjoint*[11]*. Consider $\mathfrak{S} := \{(1-t)\pi_0 + t\pi_1 : t \in [0, 1]\}$. If $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable then $\delta \leq 1/p$.*

The result is mostly based on Niles-Weed and Berthet (2022). Its proof in Appendix A.4 essentially amounts to showing (12) as soon as $p \cdot \delta > 1$. Following Remark 7, the same conclusion holds if $\mathfrak{S}$ only *contains* the convex combinations of distributions π_0, π_1 as in the above proposition. For

10. This is a consequence of Jensen inequality (Santambrogio, 2015, Section 5.1).
11. We recall that the support $\mathrm{supp}(\pi)$ of a probability distribution $\pi \in \mathcal{P}(\mathcal{X})$ is the smallest closed set S such that $\pi(S) = 1$.

a bounded kernel, since $\|\pi_0 - \pi_1\|_\kappa$ is always finite, the same result is thus valid in particular when the model set $\mathfrak{S}$ contains a segment whose extreme points have disjoint supports. This is notably the case when $\mathfrak{S}$ is convex and contains two distributions with disjoint supports. As a consequence, given any PSD kernel κ, $(\mathfrak{S}, \mathrm{W}_p)$ is *not* (κ, δ)-embeddable for $\delta > 1/p$ when $\mathfrak{S}$ contains for example mixtures of two Diracs or more generally mixtures of two compactly supported distributions. We emphasize that this result does not depend on the dimension of the ambient space and is true for any PSD kernel.

2.3.4 Bound on δ for Mixture Models and Smooth TI Kernels

In most concrete applications, one often has to compare *discrete* distributions. We show in this section that the regularity of the kernel plays an important role when trying to control the Wasserstein distance with an MMD for model sets made of discrete distributions. In the following we define, for $K \in \mathbb{N}^*$ and $\Omega \subseteq \mathcal{X} = \mathbb{R}^d$, the space of mixtures of K diracs located in Ω:

$$\mathfrak{S}_K(\Omega) := \Big\{ \sum_{i=1}^K a_i \delta_{\mathbf{x}_i} : a_i \in \mathbb{R}_+, \sum_{i=1}^K a_i = 1, \forall i \in [\![K]\!], \mathbf{x}_i \in \Omega \Big\}.$$

This type of model with $\Omega = B(0, R)$ for some $R > 0$ plays a central role in compressive learning theory and is used to show that the LRIP (Section 4) does not hold for tasks such as K-means without separability assumptions on the diracs (Gribonval et al., 2021b). We show in the next theorem (proof in Appendix A.5) that there is a trade-off between the exponent δ and the regularity of the kernel provided that the model set is rich enough to contain discrete distributions with enough diracs.

Theorem 11 *Consider a TI, PSD kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ such that κ_0 is k times differentiable at 0 with $k \in \mathbb{N}^*$. Consider $p \in [1, +\infty)$, a Wasserstein distance W_p based on a norm in $\mathbb{R}^d$, a vector $\mathbf{x}_0 \in \mathbb{R}^d$, $R > 0$ and $\Omega = B(\mathbf{x}_0, R)$. If $(\mathfrak{S}_{\lfloor \frac{k}{2} \rfloor + 1}(\Omega), \mathrm{W}_p)$ is (κ, δ)-embeddable then $\delta \leq 2/k$.*

Following Remark 7, the same conclusion holds if $\mathfrak{S}$ only *contains* all mixtures of Dirac supported in some arbitrary Euclidean ball. Theorem 11 proves that if the kernel is k times differentiable and if $\mathfrak{S}$ is rich enough to contain $\lfloor \frac{k}{2} \rfloor + 1$ diracs then we can not control the Wasserstein distance with MMD^δ *uniformly* over $\mathfrak{S}$ when $\delta > 2/k$. As an immediate consequence we have the following corollary when the kernel is smooth:

Corollary 12 *Consider a TI, PSD kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ such that $\kappa_0 \in C^\infty(\mathbb{R}^d, \mathbb{R})$ and a model set $\mathfrak{S} \subseteq \mathcal{P}(\mathbb{R}^d)$. Assume that $\mathfrak{S}_K(\Omega) \subseteq \mathfrak{S}$ with $K \geq 2$ where $\Omega \subseteq \mathbb{R}^d$ is an open set. If $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable, where W_p is based on a norm in $\mathbb{R}^d$ and $p \in [1, +\infty)$, then $\delta \leq 2/K$.*

These results have many consequences. First it shows that when κ is smooth and $\mathfrak{S}$ *contains* mixtures of arbitrarily many diracs located in some open set, $(\mathfrak{S}, \mathrm{W}_p)$ is *not* (κ, δ)-embeddable for any $\delta > 0$. In other words, it proves that finding a absolute constant $C > 0$ such that $\mathrm{W}_p(\pi, \pi') \leq C \, \mathrm{MMD}_\kappa^\delta(\pi, \pi')$ for all discrete distributions π, π' is hopeless when the kernel κ is smooth *even if* these distributions lie also in some fixed ball of $\mathbb{R}^d$ (to take care of the necessary condition associated to Lemma 8). It suggest that finding suitable constraints on the model set $\mathfrak{S}$ *and* on the kernel κ is required in order to have the control (11). We will show in the next sections how to obtain these types of control with additional hypotheses on the regularity of the distributions in $\mathfrak{S}$. The Figure 2 summarizes the necessary conditions established in the previous sections.

2.4 Sufficient Conditions: Regular Distributions

We are now interested in sufficient conditions allowing to uniformly control the Wasserstein distance by MMD^δ on a subset of distributions $\mathfrak{S} \subset \mathcal{P}(\mathbb{R}^d)$. In the following we consider Wasserstein distances

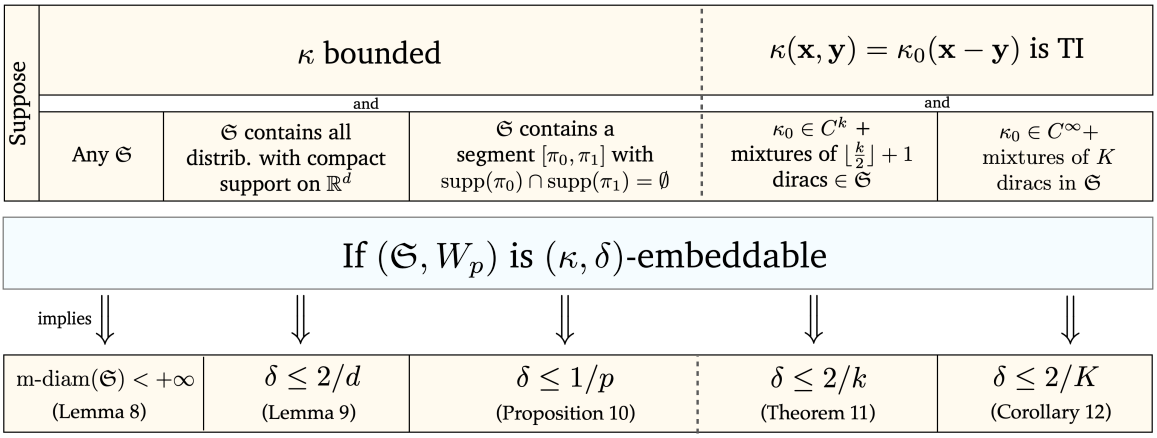

Figure 2: Summary of the established necessary conditions to the (κ,δ)-embedabbility property.

defined with respect to the Euclidean norm $\|\cdot\|_2$, and denote

$$\mathrm{M}_r[\pi] := (\mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^r])^{1/r}$$

the moment of order r of $\pi \in \mathcal{P}(\mathbb{R}^d)$. At first we restrict to the case of "regular" distributions, in the sense that probability distributions in $\mathfrak{S}$ are assumed to admit densities with respect to the Lebesgue measure (non-regular distributions will be studied in the next section). We recall that the shorthand $\pi = f\mathrm{d}\mathbf{x}$ indicates that π has density f with respect to the Lebesgue measure.

Our first Lemma (proved in Appendix A.6) controls W_p by a distance L_2 between densities, under the assumption that distributions in the model set $\mathfrak{S}$ have a certain number of bounded moments:

Proposition 13 *Consider $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ with densities f, g with respect to the Lebesgue measure, i.e. $\pi = f\mathrm{d}\mathbf{x}, \pi' = g\mathrm{d}\mathbf{x}$. If $\max\{\mathrm{M}_r[\pi], \mathrm{M}_r[\pi']\} \leq M$, where $r > 1$, then for each $1 \leq p < r$ we have*

$$\mathrm{W}_p(\pi,\pi') \leq C \left(\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \right)^{\frac{r-p}{(d+2r)p}} , \tag{13}$$

with $C = 2(\max\{V_d, 1\})^{\frac{1}{2p}} M^{\frac{(d+2p)r}{(d+2r)p}}$ with $V_d = \pi^{d/2}/\Gamma(d/2+1)$ the volume of the d-dimensional unit sphere.

The L_2 distance between densities that appears in the right hand side of (13) can be further bounded by an MMD with an appropriate kernel. Indeed, using Plancherel's formula and introducing the Fourier transform $\widehat{\kappa_0}$ of a TI, PSD kernel, Cauchy-Schwarz inequality yields

$$\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \propto \int_{\mathbb{R}^d} |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \leq \Big(\int_{\mathbb{R}^d} \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega} \Big)^{\frac{1}{2}} \Big(\int_{\mathbb{R}^d} \widehat{\kappa_0}(\boldsymbol{\omega}) |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \Big)^{\frac{1}{2}} .$$

where $\hat{f}, \hat{g}$ denote the Fourier transforms of f, g. The second integral of the right hand side of this expression being proportional to the MMD (Lemma 48) one can transform the bound (13) into a bound involving an MMD if we can control the integral $\int_{\mathbb{R}^d} \widehat{\kappa_0}(\boldsymbol{\omega})^{-1} |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\mathbf{x})|^2 \mathrm{d}\boldsymbol{\omega}$ by a constant. Moreover, we also have the following relation (see[12] Wendland 2004, Theorem 10.12):

$$(2\pi)^{-d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega} = \|f - g\|_{\mathcal{H}_\kappa}^2 ,$$

12. With adapted conventions on Fourier transforms.

where $\mathcal{H}_\kappa$ is the RKHS associated to the kernel κ and $\|\cdot\|_{\mathcal{H}_\kappa}$ is the corresponding RKHS norm. Consequently, when the distributions in $\mathfrak{S}$ have densities in some RKHS ball, we can bound $\int_{\mathbb{R}^d} \widehat{\kappa_0}(\boldsymbol{\omega})^{-1}|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\mathbf{x})|^2 \mathrm{d}\boldsymbol{\omega}$ by a constant:

Theorem 14 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ such that $\kappa_0 \in L_1(\mathbb{R}^d)$, $\widehat{\kappa_0}(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$. For $B, M, r \geq 0$, denote*

$$\mathfrak{S}_{B,M,r,\kappa} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \ \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{\mathcal{H}_\kappa} \leq B \text{ and } \ \mathrm{M}_r[\pi] \leq M\right\} \subset \mathcal{P}_r(\mathbb{R}^d)\,. \tag{14}$$

If $r > 1$ then for each $1 \leq p < r$ we have

$$\forall \pi, \pi' \in \mathfrak{S}_{B,M,r,\kappa},\ \mathrm{W}_p(\pi,\pi') \leq C' \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,,$$

where $C' = 8(\max\{V_d, 1\})^{\frac{1}{2p}} B^{\frac{r-p}{(d+2r)p}} M^{\frac{(d+2p)r}{(d+2r)p}}$.

The proof is given in Appendix A.7. With the model set $\mathfrak{S} = \mathfrak{S}_{B,M,r,\kappa}$, this theorem implies that $(\mathfrak{S}, \mathrm{W}_p)$ is $(\kappa, \delta = \frac{r-p}{p(d+2r)})$-embeddable for every $1 \leq p < r$ as soon as κ is a TI, PSD kernel with very few assumptions. A limitation of this result is that the model set $\mathfrak{S}$ depends on the kernel κ so that it is not clear which family of distributions belongs to $\mathfrak{S}$. In the next theorem we decouple the assumptions on the kernel from those on the model set. Assuming that the distributions have densities that are sufficiently regular (Sobolev), a certain number of bounded moments and with some assumptions on the kernel κ the following holds:

Theorem 15 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ such that $\kappa_0 \in L_1(\mathbb{R}^d)$, $\widehat{\kappa_0}(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$, and assume there is $s_\kappa > 0$ such that*

$$\frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})} = O(\|\boldsymbol{\omega}\|_2^{s_\kappa}) \text{ as } \|\boldsymbol{\omega}\|_2 \to +\infty\,. \tag{15}$$

For $r, B, M, s \geq 0$, denote

$$\mathfrak{S}_{B,M,r,s} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \ \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{H^s(\mathbb{R}^d)} \leq B \text{ and } \ \mathrm{M}_r[\pi] \leq M\right\} \subset \mathcal{P}_r(\mathbb{R}^d)\,. \tag{16}$$

If $s \geq s_\kappa/2$ and $r > 1$ then for each $1 \leq p < r$ there exists $C = C(B, M, r, s, d, \kappa, p) > 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}_{B,M,r,s},\ \mathrm{W}_p(\pi,\pi') \leq C \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,.$$

The proof is given in Appendix A.7. With the model set $\mathfrak{S} = \mathfrak{S}_{B,M,r,s}$, this theorem implies that $(\mathfrak{S}, \mathrm{W}_p)$ is $(\kappa, \delta = \frac{r-p}{p(d+2r)})$-embeddable for every $1 \leq p < r$ as soon as κ is a TI, PSD kernel with some regularity, and the distributions in $\mathfrak{S}$ are sufficiently regular with bounded r-moments. This latter hypothesis is not very limiting in practice since it is also required in order to have finite Wasserstein distances. The Sobolev condition on the densities requires that densities are in L_2 and have at least $s \geq s_\kappa/2$ (weak)derivatives in L_2. In particular this is the case for the classical model sets considered in compressive statistical learning literature such as Gaussian mixtures (Gribonval et al., 2021b).

Remark 16 *Since the distributions in $\mathfrak{S}$ admit a density, the constraints of Theorem 11 (mixtures of Diracs) do not apply here and, as such, the kernel is allowed to be smooth.*

An important family of TI kernels satisfying the hypothesis of Theorem 15 is the Matérn class (Rasmussen and Williams, 2005, Section 4.2.1), with parameter ν, as detailed in Example 4. The limit of a Matèrn kernel when the parameter $\nu \to \infty$ is the RBF kernel, which is too regular: its Fourier transform decays too fast to satisfy the assumption (15) of Theorem 15. In the context of compressive learning, translation invariant kernels are most useful if they can be approximated with

random Fourier features with good concentration properties (see Section 4). An interesting question for future work is thus whether the "slow decay" of the Fourier transform needed to apply Theorem 15 appears as a strong constraint in such a context.

Observe that for fixed p and large r the exponent $\delta = \frac{r-p}{p(d+2r)}$ tends to $\frac{1}{2p}$. Another consequence of Theorem 15 is for distributions that have infinitely many bounded moments. In this case the exponent δ can be *independent of the dimension*, as shown in the following two examples:

Example 17 (Uniformly bounded moments) *Consider a kernel κ and an exponent s with the same assumptions as in Theorem 15 and a function $m : \mathbb{R} \to \mathbb{R}_+^*$ along with the following model set:*

$$\mathfrak{S}_{B,m,s} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{H^s(\mathbb{R}^d)} \leq B \text{ and } \forall r > 1,\ \mathrm{M}_r[\pi] \leq m(r)\right\}. \tag{17}$$

i.e., the intersection of the model sets $\mathfrak{S}_{B,m(r),r,s}$, $r > 1$. For any $p \in [1, +\infty)$ and $0 < \delta < \frac{1}{2p}$ we can find a constant[13] *$C = C(B, m(\cdot), \delta, s, d, \kappa, p) > 0$ such that $\forall \pi, \pi' \in \mathfrak{S}_{B,m,s}, \mathrm{W}_p(\pi, \pi') \leq C\|\pi - \pi'\|_\kappa^\delta$. In other words $(\mathfrak{S}_{B,m,s}, \mathrm{W}_p)$ is (κ, δ)-embeddable for an exponent that is as close as we want to $\delta^* = \frac{1}{2p}$.*

A notable example where such a model is relevant is in compressive statistical learning, where the model set associated to Gaussian mixtures with bounded parameters fits into this framework (Gribonval et al., 2021a). More generally one can also consider a model set made of sub-Gaussian variables with smooth densities and bounded sub-Gaussiannity parameter σ. In this case $m(r) = c\sigma_{\max}\sqrt{r}$ for some constant $c > 0$ since, by the sub-Gaussian property, we have $\forall r \geq 1, \mathrm{M}_r[\pi] \leq c\sigma\sqrt{r} \leq c\sigma_{\max}\sqrt{r}$ (see e.g. Foucart and Rauhut 2013, Section 7.4).

Example 18 (Compactly supported distributions) *With the same assumptions of κ and s, when all the distributions in $\mathfrak{S}$ are smooth and have the same compact support, they can be shown to belong to $\mathfrak{S}_{B,m,s}$ where the function $m : \mathbb{R} \to \mathbb{R}_+^*$ can be chosen as constant. Indeed if $\mathrm{supp}(\pi) \subseteq B(0, M)$ for some ball of radius M then $\forall r > 1, \mathrm{M}_r[\pi] \leq M$. In this case the exponent $\delta = \frac{1}{2p}$ is* exactly *attainable as shown in Appendix A.8.*

Remark 19 *We recall that, due to the constraints of Proposition 10, the best possible rate achievable is $\delta = 1/p$ since the model set $\mathfrak{S}_{B,M,r,s}$ in (16) contains a convex combination of two probability distributions whose support are disjoint. Indeed, it is not difficult to construct two measures in the model set $\pi_1 = f_1\mathrm{d}\mathbf{x}$ and $\pi_2 = f_2\mathrm{d}\mathbf{x}$ with $\|f_1\|_{H^s(\mathbb{R}^d)}, \|f_2\|_{H^s(\mathbb{R}^d)} \leq B$ and such that $\mathrm{supp}(\pi_1) \cap \mathrm{supp}(\pi_2) = \emptyset$. Then for any $t \in [0, 1], (1-t)\pi_1 + t\pi_2 \in \mathfrak{S}_{s,B,M,r}$ since it has density $(1-t)f_1 + tf_2$ such that $\|(1-t)f_1 + tf_2\|_{H^s(\mathbb{R}^d)} \leq B$ and $\mathrm{M}_r^r[(1-t)\pi_1 + t\pi_2] = (1-t)\,\mathrm{M}_r^r[\pi_1] + t\,\mathrm{M}_r^r[\pi_2]$ by linearity (with respect to the distribution) thus $\mathrm{M}_r^r[(1-t)\pi_1 + t\pi_2] \leq M^r$ which implies $\mathrm{M}_r[(1-t)\pi_1 + t\pi_2] \leq M$. It remains open whether exponents $\delta \in (1/2p, 1/p)$ are actually achievable on $\mathfrak{S}_{B,M,r,s}$.*

2.5 Sufficient Conditions: Non-Regular Distributions

The case of measures on $\mathbb{R}^d$ and that do not admit a density is more delicate to study. We will however prove that, at the price of an arbitrary small additive term $\eta > 0$, we have the control (11) under mild assumptions on the model set $\mathfrak{S}$. The core idea is to regularize the probability distributions π, π' and to obtain bounds between the true Wasserstein and the "smoothed" Wasserstein distance which is easier to relate to an MMD. We adopt the following definition:

Definition 20 (Regularizer) *We say that a function $\alpha : \mathbb{R}^d \to \mathbb{R}_+$ is a* regularizer *if it is a* non-negative, continuous, even *and* bounded *function such that $\int_{\mathbb{R}^d} \alpha(\mathbf{z})\mathrm{d}\mathbf{z} = 1$ and $\alpha \in L_2(\mathbb{R}^d)$. We say that the regularizer has r-finite moments if $\int \|\mathbf{z}\|_2^r \alpha(\mathbf{z})\mathrm{d}\mathbf{z} < +\infty$ for some $r \geq 1$.*

13. It suffices to apply Theorem 15 with $\mathfrak{S}_{B,m(r),r,s}$ where $r = \frac{(1+\delta d)p}{1-2\delta p} > p$ since $\delta = \frac{r-p}{p(d+2r)}$.

When considering a regularizer α and a probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ (not necessarily regular) the convolution $\alpha * \pi$ defines a probability density function[14] on $\mathbb{R}^d$ *via* $\alpha*\pi(\mathbf{x}) = \int_{\mathbb{R}^d} \alpha(\mathbf{x}-\mathbf{y})\mathrm{d}\pi(\mathbf{y})$. In the following we will note π_α the probability distribution associated to the density $\alpha*\pi$. Note that π_α is usually regular by imposing that α is (such as when α is the Gaussian density). The interpretation behind π_α is the following: if $X \sim \pi$ and Y_α is a random variable independant of X and whose distribution has density α then the random variable $X + Y_\alpha$ has distribution π_α. The idea of regularizing the measure to derive properties on the Wasserstein distance is not new and was used in various contexts (Dedecker and Michel, 2013; Niles-Weed and Berthet, 2022; Goldfeld and Greenewald, 2020; Nguyen, 2013). We have the following lemma which relates the Wasserstein distance W_p to its regularized counterpart:

Lemma 21 *Consider a regularizer α with p-finite moments where $p \geq 1$. Then*

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d),\ \mathrm{W}_p(\pi,\pi') \leq \mathrm{W}_p(\pi_\alpha,\pi'_\alpha) + 2\left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z})\mathrm{d}\mathbf{z}\right)^{1/p}.$$

Proof Using the triangle inequality we have $\mathrm{W}_p(\pi,\pi') \leq \mathrm{W}_p(\pi,\pi_\alpha) + \mathrm{W}_p(\pi_\alpha,\pi'_\alpha) + \mathrm{W}_p(\pi',\pi'_\alpha)$. Let $X \sim \pi$ and Y_α be a random variable independent of X and whose distribution has density α so that $X + Y_\alpha \sim \pi_\alpha$. By definition of W_p we have $\mathrm{W}_p^p(\pi,\pi_\alpha) = \inf_{\gamma\in\Pi(\pi,\pi_\alpha)} \mathbb{E}_{(Z_1,Z_2)\sim\gamma}[\|Z_1 - Z_2\|_2^p]$ hence taking $(Z_1,Z_2) = (X, X+Y_\alpha)$ we obtain $\mathrm{W}_p^p(\pi,\pi_\alpha) \leq \mathbb{E}[\|X - (X+Y_\alpha)\|_2^p] = \mathbb{E}[\|Y_\alpha\|_2^p]$. Consequently $\mathrm{W}_p^p(\pi,\pi_\alpha) \leq \int \|\mathbf{y}\|_2^p \alpha(\mathbf{y})\mathrm{d}\mathbf{y}$. The same applies for the term $\mathrm{W}_p(\pi',\pi'_\alpha)$. ■

When α is the density of the Gaussian $\mathcal{N}(0,\sigma^2\mathbf{I})$ the distance $\mathrm{W}_p(\pi_\alpha,\pi'_\alpha)$ is usually called the Gaussian-smoothed OT and enjoys good properties in terms of sample-complexity and topological properties (Goldfeld and Greenewald, 2020; Nietert et al., 2021a). Our formalism is more general as it considers any type of regularizers. The main idea now is to show that, given the regularizer, $\mathrm{W}_p(\pi_\alpha,\pi'_\alpha)$ can be controlled by the MMD associated to a TI kernel. Since π_α, π'_α admit a density we will use the same idea as in the Proposition 13 to control $\mathrm{W}_p(\pi_\alpha,\pi'_\alpha)$ by $\|\alpha*\pi - \alpha*\pi'\|^\delta_{L_2(\mathbb{R}^d)}$ for some $\delta \in (0,1)$. To connect with the MMD we will rely on the following result whose proof is given in Appendix A.9:

Lemma 22 *Let α be a regularizer and $\kappa_0 := \alpha*\alpha$. Then $\kappa_0 \in L_1(\mathbb{R}^d)$ is even, bounded, continuous and has non-negative Fourier transform. Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) := \kappa_0(\mathbf{x}-\mathbf{y})$. Then κ defines a TI, PSD kernel. Moreover, for $\pi,\pi' \in \mathcal{P}(\mathbb{R}^d)$,*

$$\|\pi - \pi'\|_\kappa = \|\alpha*\pi - \alpha*\pi'\|_{L_2(\mathbb{R}^d)}.$$

Based on these results we have the following upper-bound on $\mathrm{W}_p(\pi_\alpha,\pi'_\alpha)$ using the MMD associated to a TI, PSD kernel (the proof can be found in Appendix A.9):

Proposition 23 *Let $r > 1$. Consider a regularizer α with r-finite moments and the kernel $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ where $\kappa_0 := \alpha*\alpha$. It defines a TI, PSD kernel by Lemma 22. Moreover, for any $\pi,\pi' \in \mathcal{P}_r(\mathbb{R}^d)$ and $1 \leq p < r$, W_p defined with the Euclidean norm on $\mathbb{R}^d$ satisfies*

$$\mathrm{W}_p(\pi_\alpha,\pi'_\alpha) \leq C_{d,r,p}\left(\mathbb{E}_{\mathbf{x}\sim\pi_\alpha}[\|\mathbf{x}\|_2^r] + \mathbb{E}_{\mathbf{y}\sim\pi'_\alpha}[\|\mathbf{y}\|_2^r]\right)^{\frac{2p+d}{(d+2r)p}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}},$$

for some constant $C_{d,r,p} > 0$.

As a corollary of Proposition 23 and Lemma 21 we are now able to prove the main theorem of this section (the proof is in Appendix A.9):

14. Since α is a regularizer we have $\int \alpha = 1$ and consequently $\int(\int \alpha(\mathbf{x}-\mathbf{y})\mathrm{d}\pi(\mathbf{y}))\mathrm{d}\mathbf{x} = \int(\int \alpha(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{x})\mathrm{d}\pi(\mathbf{y}) = 1$ by using Fubini's theorem (α is non-negative) and the fact that the Lebesgue measure is invariant by translation.

Theorem 24 *Let $r > 1$. Consider a regularizer α with r-bounded moments. Consider the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 := \alpha * \alpha$. It defines a TI, PSD kernel by Lemma 22. We consider the model set*

$$\mathfrak{S}_M := \{\pi \in \mathcal{P}(\mathbb{R}^d) : \mathrm{M}_r[\pi] \leq M\} \subset \mathcal{P}_r(\mathbb{R}^d)\,.$$

Then for any $1 \leq p < r$ there exists a constant $C' = C'_{d,r,p} > 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{W}_p(\pi, \pi') \leq C' \left(M^r + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{\frac{2p+d}{p(d+2r)}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}} + 2 \left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{1/p}.$$

This theorem has multiple implications. First it shows that, for a wide range of TI, PSD kernels, and under mild assumptions, $(\mathfrak{S}, \mathrm{W}_p)$ is $(\kappa, \delta = \frac{2(r-p)}{p(d+2r)})$-embeddable with error $\eta > 0$. Note that the exponent δ is twice the exponent found in Section 2.4 for regular distributions, which is due to the fact that we directly regularize the distributions using the kernel associated to the MMD. Consequently, it leads to a slightly better better exponent (closer to 1) than the one of the regular case, but at a price of an additive error term. We will also see in Example 25 how this error term $\eta > 0$ can be controlled. We emphasize that few assumptions on $\mathfrak{S}$ are required: the distributions in the model set must have uniformly bounded r-moment, *i.e.* $\sup_{\pi \in \mathfrak{S}} \mathbb{E}_{\mathbf{x} \sim \pi}[\|\mathbf{x}\|_2^r] < +\infty$. This assumption is verified when, for example, $\mathfrak{S}$ is the space of Gaussian mixtures whose parameters are in a compact subspace as considered in compressive statistical learning (Gribonval et al., 2021b). Interestingly, if r is big compared to d, p then we have $\delta \approx \frac{1}{p}$.

Example 25 (RBF kernel) *As an example of use of Theorem 24 consider the Gaussian density function $\varphi(\mathbf{x}) := (2\pi)^{-d/2} \exp(-\|\mathbf{x}\|_2^2/2)$. Define for $\sigma > 0$ the regularizer $\alpha(\mathbf{x}) := \sigma^{-d} \varphi(\frac{\mathbf{x}}{\sigma})$. The function α is continuous, even, bounded, all r-moments are finite, $\int_{\mathbb{R}^d} \alpha = 1$. The associated kernel is then defined by $\widehat{\kappa_0}(\boldsymbol{\omega}) = (\hat{\varphi}(\sigma\boldsymbol{\omega}))^2 = (e^{-\frac{1}{2}\sigma^2\|\boldsymbol{\omega}\|_2^2})^2 = e^{-\sigma^2\|\boldsymbol{\omega}\|_2^2}$, hence $\kappa(\mathbf{x}, \mathbf{y}) = \pi^{d/2}\sigma^{-d} \exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{4\sigma^2})$. Consider the case $p = 1$ and $r > 1$ of Theorem 24. The error term $2\int \|\mathbf{z}\|_2 \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} = 2\sigma \int \|\mathbf{z}\|_2 \varphi(\mathbf{z}) \mathrm{d}\mathbf{z}$ can be controlled as*

$$2\sigma \int \|\mathbf{x}\|_2 (2\pi)^{-d/2} \exp(-\|\mathbf{x}\|_2^2/2) \mathrm{d}\mathbf{x} \leq 2\sigma \Big(\int \|\mathbf{x}\|_2^2 (2\pi)^{-d/2} \exp(-\|\mathbf{x}\|_2^2/2) \mathrm{d}\mathbf{x} \Big)^{1/2}$$

by Jensen since $\mathbf{x} \to (2\pi)^{-d/2} \exp(-\|\mathbf{x}\|_2^2/2)$ is a probability density function. Thus, we can bound the error therm by $2\sigma(\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0,\mathbf{I})}[\|\mathbf{x}\|_2^2])^{1/2} = 2\sigma\sqrt{d}$. Moreover, $\int \|\mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} = \sigma^r \int \|\mathbf{z}\|_2^r \varphi(\mathbf{z}) \mathrm{d}\mathbf{z} = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0,\mathbf{I})}[\|\mathbf{x}\|_2^r] = 2^{r/2} \frac{\Gamma(\frac{r+d}{2})}{\Gamma(\frac{r}{2})}$ (it is the r-th moment of a χ_2 distribution). Then, using Theorem 24 we have

$$\forall \pi, \pi' \in \mathfrak{S},\ \mathrm{W}_1(\pi, \pi') \leq C' \left(M^r + 2^{r/2} \sigma^r \frac{\Gamma(\frac{r+d}{2})}{\Gamma(\frac{r}{2})} \right)^{\frac{d+2}{d+2r}} \|\pi - \pi'\|_\kappa^{\frac{2(r-1)}{d+2r}} + 2\sigma\sqrt{d}\,.$$

Interestingly enough, the error term behaves as $O(\sigma)$ and can me made as small as possible at a price of a "sharper" kernel (the bound is true for any $\sigma > 0$). Implications of this result wil be discussed in the context of CSL in Section 4.

Remark 26 *The condition $\kappa_0 = \alpha * \alpha$ in Theorem 24 can be met in two ways. First, as done in Example 25, fixing a regularizer α with r-bounded moments gives a TI, PSD kernel so that Theorem 24 holds. This can be achieved for example by considering a PSD function $\alpha \in L_1(\mathbb{R}^d)$ with a sufficient number of bounded moments and that is even, continuous and positive (continuous, integrable and PSD functions are bounded Wendland, 2004). A simple normalization $\alpha \leftarrow \alpha / \int \alpha$ will then produce a suitable α. The second way is to fix the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ and to check that it can be decomposed as $\kappa_0 = \alpha * \alpha$ with α a regularizer with r-bounded moments and $\widehat{\alpha} \geq 0$. This problem is related to the one of finding a so-called* convolution root*, or* Boas–Kac root *of a positive definite function which can be shown to exist under certain assumptions on the function (Ehm et al., 2004; Akopyan and Efimov, 2017; R. P. Boas and Kac, 1945).*

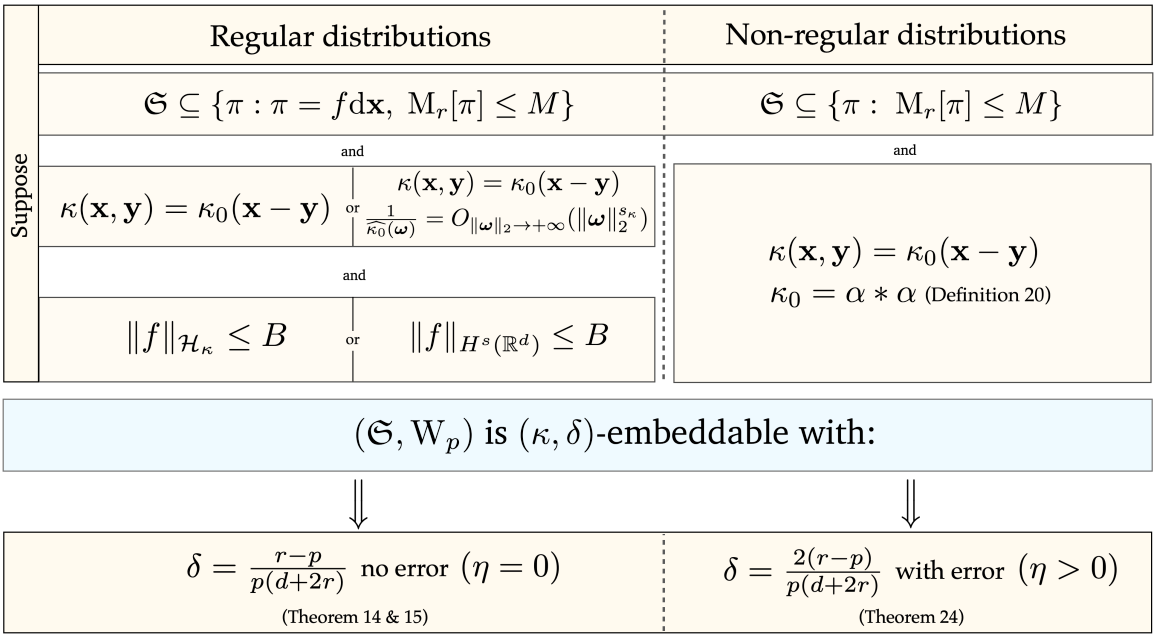

Figure 3: Summary of the different results of Section 2. The mention "with error" means that the relation holds when adding an error $\eta > 0$ that does not depends on $\mathfrak{S}$. $\pi = f\mathrm{d}\mathbf{x}$ means that the measure has density f with respect to the Lebesgue measure.

2.6 Conclusion and Related Works

We established in this section various controls of the form $\mathrm{W}_p \lesssim \mathrm{MMD}_\kappa^\delta$ that depend on $\delta \in (0,1]$, the properties of the model set and the kernel κ. All these results are summarized in Figure 3. Some other connections between MMDs and Wasserstein distances have been explored in the literature. The most simple one is when the metric D used to define the Wasserstein distance is the metric in the RKHS corresponding the the kernel κ, *i.e.* $D(\mathbf{x},\mathbf{y}) = \|\kappa(\cdot,\mathbf{x}) - \kappa(\cdot,\mathbf{y})\|_{\mathcal{H}_\kappa}$. In this case it is known that we can control the Wasserstein distance W_1 by $\sqrt{\mathrm{MMD}_\kappa^2 + K}$ when κ is bounded by K (Sriperumbudur et al., 2010).

2.6.1 Relaxing the Translation-Invariance Property

Other interesting connections are based on the Gaussian-smoothed Wasserstein distance (Goldfeld and Greenewald, 2020) where authors consider α the probability density function of the Gaussian $\mathcal{N}(0,\sigma^2\mathbf{I})$ and the Wasserstein distance between the regularized distributions $\pi_\alpha = \alpha * \pi$. In Zhang et al. (2021) authors show that we can control the Gaussian-smoothed Wasserstein distance with the MMD, by considering a PSD kernel that is *not* translation-invariant and *not* bounded but defined as $\kappa(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{4\sigma^2}\right) I_f\left(\frac{\|\mathbf{x}+\mathbf{y}\|_2}{\sqrt{2}\sigma}\right)$ where I_f is a function parametrized by some probability density function f such as generalized beta-prime distributions. More precisely they prove

$$\forall \pi, \pi' \in \mathfrak{S}_\kappa, \mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq 2\sigma \|\pi - \pi'\|_\kappa^{1/p},$$

where $\mathfrak{S}_\kappa := \{\pi \in \mathcal{P}(\mathbb{R}^d) : \int \sqrt{\kappa(\mathbf{x},\mathbf{x})}\mathrm{d}\pi(\mathbf{x}) < +\infty\}$ (Zhang et al., 2021, Theorem 2). With the same type of arguments as those presented in Lemma 21 we can prove that for any $\pi, \pi' \in \mathfrak{S}_\kappa$ we have $\mathrm{W}_p(\pi,\pi') \leq 2\sigma\|\pi - \pi'\|_\kappa^{1/p} + \eta$ where $\eta = 2\left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z})\mathrm{d}\mathbf{z}\right)^{1/p}$ and W_p is computed with $\|\cdot\|_2$.

As a corollary, for this kernel that is not TI we can use the result of Zhang et al. (2021) to prove that $(\mathfrak{S}_\kappa, \mathrm{W}_p)$ is $(\kappa, \frac{1}{p})$-embeddable with error $\eta = 2\left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z})\mathrm{d}\mathbf{z}\right)^{1/p}$ that will behave as $O(\sigma)$ as shown in Example 25. We can mention another line of works which draws connections between the Wasserstein distance and some specific dual Sobolev norms which can be related to the MMD. In Nietert et al. (2021b) authors control the Wasserstein distance with an MMD whose kernel, which is not TI, is defined by $\kappa(\mathbf{x}, \mathbf{y}) = -\sigma^2 \operatorname{Ein}(-\langle \mathbf{x}, \mathbf{y}\rangle/\sigma^2)$ where $\operatorname{Ein}(z) = \int_0^z \frac{(1-e^{-t})}{t}\mathrm{d}t$. Despite the fact that our two approaches are related our work differs from the Gaussian-smoothed OT in the sense that we do not want to estimate precisely the smoothed Wasserstein distance $\mathrm{W}_p(\pi_\alpha, \pi'_\alpha)$ by controlling it with an MMD based on a *specific* kernel but instead to control $\mathrm{W}_p(\pi, \pi')$ by kernel norms for *many* types of TI kernels.

2.6.2 Relaxing the PSD Assumption on the Kernel

Beyond PSD kernels other types of kernels can be used to define interesting divergences between probability distributions that can be linked with the Wasserstein distance. These divergences are not *stricly speaking* MMD norms as defined in (4) with PSD kernels but share similar topological properties. For example, by considering the *conditionally* PSD[15] kernel $\kappa(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x}-\mathbf{y}\|_2^\beta$ for $\beta \in (0, 2]$, and $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$, the integral in (4) is non-negative for $\mu = \pi - \pi'$ so that the term $\|\pi - \pi'\|_\kappa$ is well defined (Sejdinovic et al., 2013, Example 15). It is called the energy, or *Cramér*, distance (Székely and Rizzo, 2017; Szekely and Rizzo, 2004; Sejdinovic et al., 2013) and it connects with OT distances in the sense that the Sinkhorn divergence (regularized OT) was shown to interpolate between this MMD and the Wasserstein distance (Feydy et al., 2019). Another notable example is when one considers the so called d-dimensional *Coulomb* kernel defined by $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where

$$\kappa_0(\mathbf{x}) := \begin{cases} -\log \|\mathbf{x}\|_2 & \text{if } d = 2 \\ \|\mathbf{x}\|_2^{2-d} & \text{if } d \geq 3 \end{cases}$$

In this case, for compactly supported $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ with $\int\int \kappa(\mathbf{x}, \mathbf{x}')\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi(\mathbf{x}') < +\infty$ and $\int\int \kappa(\mathbf{y}, \mathbf{y}')\mathrm{d}\pi'(\mathbf{y})\mathrm{d}\pi'(\mathbf{y}') < +\infty$, the quantity $\|\pi - \pi'\|_\kappa$ is well defined, finite, and vanishes if and only if $\pi = \pi'$ (Chafaï et al., 2016; Saff and Totik, 2013). Consequently it defines a valid MMD that remarkably controls the W_1 distance associated to an arbitrary norm in $\mathbb{R}^d$, as described in Chafaï et al. (2016). More precisely consider, for $\Omega \subseteq \mathbb{R}^d$ *compact*, the model set

$$\mathfrak{S} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \operatorname{supp}(\pi) \subseteq \Omega, \int\int \kappa(\mathbf{x}, \mathbf{x}')\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi(\mathbf{x}') < +\infty\right\}.$$

Then Chafaï et al. (2016, Theorem 1) proves that there exists $C = C(\Omega) > 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{W}_1(\pi, \pi') \leq C\|\pi - \pi'\|_\kappa .$$

In particular, with the above $\mathfrak{S}$, $(\mathfrak{S}, \mathrm{W}_1)$ is $(\kappa, \delta = 1)$-embeddable with no error. It is remarkable in the sense that few assumptions on the model set are required (the distributions can be even discrete). An important remark is that the kernel is TI *but not PSD* and, consequently, this result is not in contradiction with Theorem 11. Finally, other connections between W_p and the Cramér distance regarding asymptotic convergence in law can be found in (Modeste and Dombry, 2022).

3. Statistical Learning and Wasserstein Regularity

The bounds obtained previously allow us to control the Wasserstein distance by an MMD under certain conditions. These results will be at the heart of the theoretical guarantees of compressive

15. A conditionally PSD kernel on $\mathcal{X}$ satisfies $\sum_{i,j=1}^n c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ for any $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathcal{X}$ and $c_1, \cdots c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$ (Berg et al., 1984)

learning (Section 4). These guarantees require, in addition, to control metrics related to the learning task (see the reasoning described in Figure 1). In this section we recall the statistical learning framework and introduce more formally these task metrics (referred as TaskMetric in the introduction). We then show how to control them by a Wasserstein distance for various learning tasks.

3.1 Statistical Learning & Task Metrics

Statistical learning is a formalism that offers many tools to study the guarantees of learning algorithms. The problem is usually expressed as follows: given a collection of data $(\mathbf{x}_i)_{i\in[\![n]\!]}$, where $\mathbf{x}_i$ is a *sample* in the data space $\mathcal{X}$, how do we select a hypothesis $h \in \mathcal{H}$ (where $\mathcal{H}$ is called the *hypothesis space*) that best performs the task at hand ? The ideal hypothesis minimizes a certain *risk* which provides a performance measure and is derived from a certain loss function $\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$.

For example, in the context of linear regression the loss is defined as $\ell(\mathbf{x} = (\mathbf{z}, y), h = \boldsymbol{\theta}) = (y - \boldsymbol{\theta}^\top \mathbf{z})^2$ where $y \in \mathbb{R}$ is the value to predict, $h = \boldsymbol{\theta} \in \mathbb{R}^d$ is the parameters to choose and $\mathbf{z} \in \mathbb{R}^d$ is the vector of input features. Given a data-generating distribution $\pi \in \mathcal{P}(\mathcal{X})$, *i.e.* the law under which our samples are produced, most of the machine learning algorithms attempt to minimize the so-called *expected risk* (or generalization error):

$$\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)]\,.$$

This quantity reflects how effective is h on average on the data-generating distribution. The optimal hypothesis $h^* \in \mathcal{H}$, known as the *Bayes prediction function* (Steinwart and Christmann, 2008), is such that $h^* \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h)$. The major difficulty is that the generating distribution π is unknown and that we only have access to finitely many samples $(\mathbf{x}_i)_{i\in[\![n]\!]}$. Methods such as *empirical risk minimization* (ERM) produce an estimated hypothesis $\hat{h}$ from the training dataset by minimizing the risk $\mathcal{R}(\pi_n, \cdot)$ associated to the empirical probability distribution $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$. One aims at guaranteeing, with high probability, the following bound on the *excess risk*:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n\,, \tag{18}$$

where η_n decays as $1/\sqrt{n}$ or better. This simply reflects that we may expect a hypothesis that is close to the best one as the training set grows, *i.e.* when we have access to enough data. To obtain a control of the excess risk by η_n one often relies on the following bound[16]:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq 2\sup_{h\in\mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi_n, h)|\,.$$

Consequently, being able to control the right term in the previous equation is a central problem in statistical learning and for example arguments involving Rademacher complexities can lead to the desired bound in (18) (see Shalev-Shwartz and Ben-David, 2014). The term $\sup_{h\in\mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi_n, h)|$, that was reffered as TaskMetric(π, π') in the introduction, defines a central quantity for our analysis and we introduce the following notation for $\pi, \pi' \in \mathcal{P}(\mathcal{X})$:

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} := \sup_{h\in\mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi', h)|\,. \tag{19}$$

The quantity $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ defines a semi-norm on the space of finite signed measures $\mathcal{M}(\mathcal{X})$ and an integral probability metric (1) with $\mathcal{G} = \mathcal{L}(\mathcal{H}) := \{\mathbf{x} \to \ell(\mathbf{x}, h); h \in \mathcal{H}\}$. It is important to note that this semi-norm is *task-specific i.e.* that it depends on the learning task *via* the family $\mathcal{L}(\mathcal{H})$. In the rest of the paper we will denote, as a language shortcut, $\mathcal{L}(\mathcal{H})$ as "the learning task". As

16. This can be proved by noting that $\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) = \left\{\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi_n, \hat{h})\right\} + \left\{\mathcal{R}(\pi_n, \hat{h}) - \mathcal{R}(\pi_n, h^*)\right\} + \{\mathcal{R}(\pi_n, h^*) - \mathcal{R}(\pi, h^*)\}$. Since $\mathcal{R}(\pi_n, h^*) - \mathcal{R}(\pi_n, \hat{h}) \leq 0$ by definition of $\hat{h}$ we have $\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq 2\sup_{h\in\mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi_n, h)|$.

When do we have $\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \lesssim \mathrm{W}_p(\pi, \pi')$ for some $p \geq 1$ and task $\mathcal{L}(\mathcal{H})$?	
Condition on the task	**Examples**
Compression type-tasks. Loss: $\ell(\mathbf{x}, h) = D(\mathbf{x}, P_h(\mathbf{x}))^p$, P_h projection function	PCA, K-means, K-medians, NMF, dictionary learning (Section 3.3)
Regression tasks. Hypothesis: h Lipschitz, loss: $\ell(\mathbf{x} = (\mathbf{z}, \mathbf{y}), h) = \|\mathbf{y} - h(\mathbf{z})\|^p$	Linear regression, regression using MLP with bounded parameters (Section 3.4.1)
Binary classification. Hypothesis: h Lipschitz, loss: convex surrogate $\ell(\mathbf{x} = (\mathbf{z}, y), h) = \varphi^p(yh(\mathbf{z}))$	MLP classifier with bounded parameters + Lipschitz ouput layer (Section 3.4.2)

Table 1: Summary of the differents results of Section 3.

just described, when $\|\pi - \pi_n\|_{\mathcal{L}(\mathcal{H})} \leq \eta_n$ one can control the excess risk as in (18). Consequently, controlling $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ with other metrics that are more easily computable is of certain interest. When the loss function is non-negative, $\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}_+$, we introduce for $p \geq 1$ the semi-norm

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} := \sup_{h \in \mathcal{H}} |\mathcal{R}^{1/p}(\pi, h) - \mathcal{R}^{1/p}(\pi', h)| \,. \tag{20}$$

A control of this semi-norm implies a slighlty different control of the excess risk as $\|\pi - \pi_n\|_{\mathcal{L}(\mathcal{H}),p} \leq \eta_n$ implies that $\mathcal{R}(\pi, \hat{h})^{1/p} - \mathcal{R}(\pi, h^*)^{1/p} \leq \eta_n$. In the following we often write $\|\pi - \pi_n\|_{\mathcal{L}(\mathcal{H}),p}$ without specifying that the loss function is non-negative and that $p \geq 1$ (this will be implicitly assumed).

Remark 27 *Controlling the quantity $\|\pi - \pi_n\|_{\mathcal{L}(\mathcal{H})}$ sometimes leads to pessimistic bounds on the excess risk. A sharper bound can be produced by considering the following semi-norm $\|\pi - \pi'\|_{\Delta\mathcal{L}(\mathcal{H})} := \sup_{h,h_0 \in \mathcal{H}} [\{\mathcal{R}(\pi, h) - \mathcal{R}(\pi, h_0)\} - \{\mathcal{R}(\pi', h) - \mathcal{R}(\pi', h_0)\}]$ which is related to $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ via the inequality $\|\pi - \pi'\|_{\Delta\mathcal{L}(\mathcal{H})} \leq 2\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ (Gribonval et al., 2021a). However in this work we focus on the quantities defined in* (19) *and* (20) *and leave the analysis of $\|\cdot\|_{\Delta\mathcal{L}(\mathcal{H})}$ for further works.*

3.2 Wasserstein Regularity

The main question investigated in this section, which will find applications to compressive statistical learning in Section 4, is to understand when the task-specific norm $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p}$ can be bounded by the Wasserstein distance between π and π'. We formalize this in the following definition:

Definition 28 (Wasserstein regularity) *Given $p \in [1, +\infty)$, we say that a task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein regular if there exists $C > 0$, such that*

$$\forall \pi, \pi \in \mathcal{P}_p(\mathcal{X}),\ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} = \sup_{h \in \mathcal{H}} |\mathcal{R}^{1/p}(\pi, h) - \mathcal{R}^{1/p}(\pi', h)| \leq C\, \mathrm{W}_p(\pi, \pi') \,.$$

At first sight the Wasserstein regularity seems a bit unexpected since the Wasserstein distance does not take into account the underlying learning task $\mathcal{L}(\mathcal{H})$. However we will show below that this property is quite natural for several learning tasks. We provide a summary of the different results of this section in Table 1.

Remark 29 *When the task is Wasserstein regular, we can show that the excess-risk is always bounded by a Wasserstein distance, i.e. if $\pi \in \mathcal{P}_p(\mathcal{X})$ is any data generating distribution, and π_n the empirical distribution, then*

$$\mathcal{R}^{1/p}(\pi, \hat{h}) - \mathcal{R}^{1/p}(\pi, h^*) \leq 2C\, \mathrm{W}_p(\pi, \pi_n) \,,$$

where $h^* \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h)$ *is an optimal hypothesis and* $\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi_n, h)$ *the hypothesis found by empirical risk minimization. Therefore, the smaller the Wasserstein distance between* π_n *and* π*, the better* $\hat{h}$ *is.*

We start by showing that many unsupervised tasks, called *compression-type tasks*, are Wasserstein regular. Then we focus on supervised tasks and demonstrate, under certain Lipschitz assumptions on the hypothesis class $\mathcal{H}$, that these tasks are also Wasserstein regular. Unless stated otherwise, until the end of Section 3, Wasserstein distances are defined with respect to the metric D associated to the ambient metric space $(\mathcal{X}, D)$.

3.3 Compression-type Tasks are Wasserstein Regular

The most straightforward case of Wasserstein regularity is when the risk *itself* can be rewritten as a Wasserstein distance. Interestingly, a wide range of unsupervised learning tasks can be recast in this setting. For example, problems such as K-means or PCA can be shown to be performing exactly the task of estimating the data-generating distribution π in the sense of a Wasserstein distance (Canas and Rosasco, 2012). Such problems will be very connected with *compression-type tasks* as defined below :

Definition 30 (Gribonval et al., 2021a) *Consider a metric space* $(\mathcal{X}, D)$ *and a hypothesis space* $\mathcal{H}$*. A task* $\mathcal{L}(\mathcal{H})$ *is called a* compression-type *task if the loss can be written as* $\ell(\mathbf{x}, h) = D(\mathbf{x}, P_h(\mathbf{x}))^p$ *where* $p \geq 1$ *and* $P_h : \mathcal{X} \to \mathcal{X}$ *is a* measurable *projection function that satisfies* $P_h \circ P_h = P_h$ *and* $D(\mathbf{x}, P_h(\mathbf{x})) \leq D(\mathbf{x}, P_h(\mathbf{x}'))$ *for all* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*.*

Notable examples of such tasks are K-means and PCA. In the former, ℓ is defined by $\ell(\mathbf{x}, h = (\mathbf{c}_1, \cdots, \mathbf{c}_K)) = \min_{i\in[\![k]\!]} \|\mathbf{x} - \mathbf{c}_i\|_2^2 = \|\mathbf{x} - P_h(\mathbf{x})\|_2^2$ where $P_h(\mathbf{x})$ is the projection of $\mathbf{x}$ on its closest centroid. In the latter, $P_h(\mathbf{x})$ is the projection of $\mathbf{x}$ on the linear subspace spanned by h. These two problems are actually related to a wider class of problems, namely k-dimensional coding schemes which are particular types of compression-type tasks. As described in Maurer and Pontil (2010), one encounters these problems when $\mathcal{X}$ is a Hilbert space (with some norm $\|\cdot\|$) and when the loss can be written as $\ell(\mathbf{x}, h) = \min_{\mathbf{y}\in Y} \|\mathbf{x} - h\mathbf{y}\|^2$ for $Y \subseteq \mathbb{R}^k$ a prescribed set of *codes* (or *codebook*) and $h : \mathbb{R}^k \to \mathcal{X}$ is a linear map. In particular, non-negative matrix factorization (NMF) (Lee and Seung, 1999; Udell et al., 2016) and dictionary learning (also known as *sparse coding* Lee et al., 2007; Mairal et al., 2009b,a) are other well known unsupervised learning methods which correspond to projection-type tasks. As described in Canas and Rosasco (2012) there are interesting connections between these problems and the Wasserstein distance. More precisely, we have the following lemma (see a proof in Appendix B.1 adapted to our notational context):

Lemma 31 (Canas and Rosasco, 2012) *Let* $S \subseteq \mathcal{X}$*,* $p \in [1, +\infty)$ *and* $\pi \in \mathcal{P}_p(\mathcal{X})$*. Consider* $P_S : \mathcal{X} \to S$*, measurable, such that* $D(\mathbf{x}, P_S(\mathbf{x})) \leq D(\mathbf{x}, \mathbf{y})$ *for all* $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{y} \in S$*. Then*

$$\mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = \mathrm{W}_p^p(\pi, P_S\#\pi)\,.$$

Moreover for any $\nu \in \mathcal{P}_p(\mathcal{X})$ *such that* $\mathrm{supp}(\nu) \subseteq S$ *we have* $\mathrm{W}_p(\pi, P_S\#\pi) \leq \mathrm{W}_p(\pi, \nu)$*.*

We recall that $P_S\#\pi$ is the probability measure defined by $P_S\#\pi(A) := \pi(P_S^{-1}(A))$ for every measurable set A. Based on this lemma we now prove that compression-type tasks are Wasserstein regular, *i.e.* that the task-specific norm $\|\cdot\|_{\mathcal{L}(\mathcal{H}),p}$ can be bounded by a Wasserstein distance.

Proposition 32 (Compression-type tasks are Wasserstein regular) *Consider a metric space* $(\mathcal{X}, D)$*, a hypothesis space* $\mathcal{H}$*,* $p \in [1, +\infty[$*, and a compression-type task* $\mathcal{L}(\mathcal{H})$ *as in Definition 30. Then*

$$\forall h \in \mathcal{H}, \pi \in \mathcal{P}_p(\mathcal{X}),\ \mathcal{R}(\pi, h) = \mathrm{W}_p^p(\pi, P_h\#\pi) \text{ and}$$
$$\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \mathrm{W}_p(\pi, \pi')\,.$$

Proof Let $h \in \mathcal{H}$ and P_h be the projection function. We denote $S = \{P_h(\mathbf{x}); \mathbf{x} \in \mathcal{X}\}$ the image of P_h. Using Lemma 31 we have, for $\pi \in \mathcal{P}_p(\mathcal{X})$,

$$\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)] = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_h(\mathbf{x}))^p] = \mathrm{W}_p^p(\pi, P_h\#\pi)\,.$$

Hence, for $\pi, \pi' \in \mathcal{P}_p(\mathcal{X})$ and $h \in \mathcal{H}$,

$$\begin{aligned}\mathcal{R}(\pi, h)^{1/p} - \mathcal{R}(\pi', h)^{1/p} &= \mathrm{W}_p(\pi, P_h\#\pi) - \mathrm{W}_p(\pi', P_h\#\pi') \leq \mathrm{W}_p(\pi, P_h\#\pi') - \mathrm{W}_p(\pi', P_h\#\pi') \\ &\leq \mathrm{W}_p(\pi, \pi')\,,\end{aligned}$$

where we used $\mathrm{W}_p(\pi, P_h\#\pi) \leq \mathrm{W}_p(\pi, \nu)$ if $\mathrm{supp}(\nu) \subseteq S$ (Lemma 31) and applied it to $\nu = P_h\#\pi'$ (since $\mathrm{supp}(P_h\#\pi') \subseteq S$ by definition of S). The last inequality is due the the triangle inequality. By symmetry $|\mathcal{R}(\pi, h)^{1/p} - \mathcal{R}(\pi', h)^{1/p}| \leq \mathrm{W}_p(\pi, \pi')$. Taking the supremum over $h \in \mathcal{H}$ concludes. ■

Remark 33 *As described in Proposition 32, compression-type tasks can be interpreted as finding a "simple" distribution $\pi_h = P_h\#\pi$ that bests describe the data distribution π in the sense of the Wasserstein distance. In PCA this distribution π_h is given by the best low dimensional projection of π, and in K-means π_h by the best discrete distribution of K centroids. This idea is also related to the problem of fitting densities, i.e. estimating the parameters $h \in \mathcal{H} \subseteq \mathbb{R}^M$ of a parametrized distribution π_h that best fits π. Two notable examples of such a learning task are* Gaussian Mixture Modeling *(GMM) (Dasgupta, 1999) and generative adversarial netwoks (Goodfellow et al., 2020). In order to find $h \in \mathbb{R}^M$ a principled way is to consider the negative likehood loss function $\ell(\mathbf{x}, h) = -\log(\pi_h(\mathbf{x}))$ that corresponds to minimizing the risk $\mathrm{KL}(\pi||\pi_h)$ where* KL *is the Kullback-Leibler divergence. However, this approach is sometimes flawed, e.g. when the data distribution is supported on a low-dimensional space or does not admit a density so that $\mathrm{KL}(\pi||\pi_h)$ is undefined or infinite (Arjovsky and Bottou, 2017). As described in many contexts such as generative modeling (Genevay et al., 2018; Arjovsky et al., 2017) or deconvolution problems (Rigollet and Weed, 2018; Dedecker and Michel, 2013) the Wasserstein distance, or its entropic regularized counterpart, is an interesting alternative fitting criterion to* KL. *It boils down to minimizing a different risk $\widetilde{\mathcal{R}}(\pi, h) := \mathrm{W}_p(\pi, \pi_h)$ which is not based on a loss function but can also be written as a Wasserstein distance. In this context, we directly have the bound $\sup_{h\in\mathcal{H}} |\widetilde{\mathcal{R}}(\pi, h) - \widetilde{\mathcal{R}}(\pi', h)| \leq \mathrm{W}_p(\pi, \pi')$ using the triangle inequality.*

3.4 Loss Functions that are p-th Power of a Lipschitz Function

Compression-type tasks are special cases of loss functions that can be written as the p-th power of a Lipchitz continuous function. Indeed, if P_h is a projection function then $D(\mathbf{x}, P_h(\mathbf{x})) - D(\mathbf{y}, P_h(\mathbf{y})) \leq D(\mathbf{x}, P_h(\mathbf{y})) - D(\mathbf{y}, P_h(\mathbf{y})) \leq D(\mathbf{x}, \mathbf{y})$, thus, by a symmetrical argument, $|D(\mathbf{x}, P_h(\mathbf{x})) - D(\mathbf{y}, P_h(\mathbf{y}))| \leq D(\mathbf{x}, \mathbf{y})$. Interestingly, these more general tasks are also Wasserstein regular:

Proposition 34 *Let $(\mathcal{X}, D)$ be a complete separable metric space. Consider a loss function that can be written for $h \in \mathcal{H}$ as $\ell(\cdot, h) = \phi_h^p$, where $p \geq 1$ and $\phi_h \in \mathrm{Lip}_L(\mathcal{X}, \mathbb{R}_+)$, then*

$$\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq L\,\mathrm{W}_p(\pi, \pi')\,.$$

In other words, the task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein regular with constant L.

Proof Using Villani (2008, Proposition 7.29) we have

$$\left| \left(\int \phi_h(\mathbf{x})^p \mathrm{d}\pi(\mathbf{x}) \right)^{1/p} - \left(\int \phi_h(\mathbf{y})^p \mathrm{d}\pi'(\mathbf{y}) \right)^{1/p} \right| \leq L\,\mathrm{W}_p(\pi, \pi')\,,$$

since $\phi_h \in \mathrm{Lip}_L(\mathcal{X}, \mathbb{R}_+)$. The conclusion follows by taking the supremum over $h \in \mathcal{H}$. ■

As described previously, this argument can be used to recover Wasserstein regularity of compression-type tasks as $\ell(\mathbf{x}, h) = D(\mathbf{x}, P_h(\mathbf{x}))^p$ is the p-th power of a 1-Lipschitz function. More importantly, the previous property allow us to prove that many *supervised* learning tasks are also Wasserstein regular as described in the next example sections.

3.4.1 Regression Tasks

The first example we consider is that of the regression tasks where $\mathcal{X} = \mathbb{R}^{d+K}$ is endowed with the metric $D(\mathbf{x} = (\mathbf{z}, \mathbf{y}), \mathbf{x}' = (\mathbf{z}', \mathbf{y}')) = \|\mathbf{z} - \mathbf{z}'\|_{\mathbb{R}^d} + \|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}^K}$ for some norm $\|\cdot\|_{\mathbb{R}^d}$ (*resp.* $\|\cdot\|_{\mathbb{R}^K}$) on $\mathbb{R}^d$ (*resp.* $\mathbb{R}^K$). The loss function is given by $\ell(\mathbf{x} = (\mathbf{z}, \mathbf{y}), h) = \|\mathbf{y} - h(\mathbf{z})\|_{\mathbb{R}^K}^p$ for some $p \geq 1$ and a regressor h that belongs to the hypothesis space $\mathcal{H} \subseteq \mathrm{Lip}_{\mathrm{L}}(\mathbb{R}^d, \mathbb{R}^K)$. In particular when $p = 2, \|\cdot\|_{\mathbb{R}^K} = \|\cdot\|_2$, the setting corresponds to a standard regression problem with the squared loss, and when $p = 1, \|\cdot\|_{\mathbb{R}^K} = \|\cdot\|_1$, to the least absolute deviation regression problem. Then, for $\mathbf{x} = (\mathbf{z}, \mathbf{y}), \mathbf{x}' = (\mathbf{z}', \mathbf{y}')$, we have

$$\begin{aligned}
|\|\mathbf{y} - h(\mathbf{z})\|_{\mathbb{R}^K} - \|\mathbf{y}' - h(\mathbf{z}')\|_{\mathbb{R}^K}| &\leq \|\mathbf{y} - \mathbf{y}' - (h(\mathbf{z}) - h(\mathbf{z}'))\|_{\mathbb{R}^K} \\
&\leq \|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}^K} + \|h(\mathbf{z}) - h(\mathbf{z}')\|_{\mathbb{R}^K} \\
&\leq \|\mathbf{y} - \mathbf{y}'\|_{\mathbb{R}^K} + L\|\mathbf{z} - \mathbf{z}'\|_{\mathbb{R}^d} \\
&\leq \max\{L, 1\} D(\mathbf{x} = (\mathbf{z}, \mathbf{y}), \mathbf{x}' = (\mathbf{z}', \mathbf{y}')) \,.
\end{aligned}$$

Consequently the loss can be written as the p-th power of a Lipschitz function and the task is p-Wasserstein regular with constant $\max\{L, 1\}$ using Proposition 34 (with W_p computed with the distance D).

This setting encompasses regressors such as multi-layer perceptron (MLP) $h(\mathbf{z}) = f_{\mathrm{MLP}}(\mathbf{z}) = T_J \circ \rho_{J-1} \circ \cdots \circ \rho_1 \circ T_1(\mathbf{z})$ where $T_j(\mathbf{w}) = \mathbf{M}_j\mathbf{w} + \mathbf{b}_j$ is an affine function with bounded weights and ρ_j is a non-linear activation function. Designing Lipschitz-continuous neural networks and computing precisely their Lipschitz constant is an (NP)hard problem and is an active line of research (Virmaux and Scaman, 2018; Fazlyab et al., 2019; Latorre et al., 2020; Kim et al., 2021). However, for fully-connected networks such as MLP with 1-Lipschitz activation functions (*e.g.* ReLU, Leaky ReLU, SoftPlus, Tanh, Sigmoid, ArcTan or Softsign) a simple upper-bound of the Lipschitz constant of f_{MLP} is given by $L = \Pi_{j=1}^J \|\mathbf{M}_j\|_{2\to 2}$ (Virmaux and Scaman, 2018) where $\|\cdot\|_{2\to 2}$ denotes the 2-operator norm for matrices. This bound is not necessarily tight, however we can use it to prove that regression tasks using MLP with bounded parameters and with 1-Lipschitz activation functions is Wasserstein regular as soon as $\forall j \in [\![J]\!], \|\mathbf{M}_j\|_{2\to 2} \leq R$ for some $R > 0$.

3.4.2 Classification Tasks

Binary classifications tasks can also be related to Wasserstein regularity. These problems corresponds to $\mathcal{X} = \mathbb{R}^d \times \{+1, -1\}$ and often rely on *convex surrogates* of the $0-1$ loss such as $\ell(\mathbf{x} = (\mathbf{z}, y), h) = \beta(yh(\mathbf{z}))$ where $y \in \{-1, +1\}$, $h : \mathbb{R}^d \to \mathbb{R}$ and $\beta : \mathbb{R} \to \mathbb{R}_+$ is convex (Bartlett et al., 2006). Well known examples include the logistic loss $\beta(t) = \log(1 + e^{-t})$, the hinge loss $\beta(t) = \max(1 - t, 0)$ or the squared hinge loss $\beta(t) = \max(1 - t, 0)^2$. In all of these cases β can be written as φ^p for some Lipschitz function φ and $p \geq 1$. If the hypothesis space is made of uniformly bounded and Lipschitz classifiers then the previous reasoning also applies. Indeed if $h \in \mathcal{H} \subseteq \mathrm{Lip}_L(\mathbb{R}^d, \mathbb{R})$, with $\|h\|_\infty \leq B$ then, for any $\mathbf{x} = (\mathbf{z}, y), \mathbf{x}' = (\mathbf{z}', y')$, we have

$$\begin{aligned}
|\varphi(yh(\mathbf{z})) - \varphi(y'h(\mathbf{z}'))| &\leq |\varphi(yh(\mathbf{z})) - \varphi(yh(\mathbf{z}'))| + |\varphi(yh(\mathbf{z}')) - \varphi(y'h(\mathbf{z}'))| \\
&\leq L_\varphi(|yh(\mathbf{z}) - yh(\mathbf{z}')| + |yh(\mathbf{z}') - y'h(\mathbf{z}')|) \\
&\leq L_\varphi(|y||h(\mathbf{z}) - h(\mathbf{z}')| + |h(\mathbf{z}')||y - y'|) \\
&\leq L_\varphi \max(L, B)(\|\mathbf{z} - \mathbf{z}'\|_2 + |y - y'|).
\end{aligned}$$

Consequently, by Proposition 34, the task is p-Wasserstein regular with constant $L_\varphi \max(L, B)$ with W_p computed with the distance $D((\mathbf{z}, y), (\mathbf{z}', y')) = \|\mathbf{z} - \mathbf{z}'\|_2 + |y - y'|$. In particular, this

example includes classifiers of the type $h = \rho \circ f_{\mathrm{MLP}}$ where $f_{\mathrm{MLP}} : \mathbb{R}^d \to \mathbb{R}$ as in Section 3.4.1 and $\rho : \mathbb{R} \to [-1, 1]$ is an "output-layer" function that is Lipschitz such as the tanh function (in this case $B = 1$).

4. Application to Compressive Statistical Learning

In the previous sections, we identified conditions allowing to 1) upper bound task-specific metrics by a Wasserstein distance W_p (notion of Wasserstein regularity, Section 3); 2) control W_p by an MMD, modulo an exponent $\delta \in (0, 1]$, under certain conditions on the model set of distributions $\mathfrak{S}$ at stake and the kernel of the MMD (Section 2). We apply in this section these results to the theory of compressive statistical learning. The goal is to establish theoretical guarantees for CSL. This section is organized as follows: we first recall the main concepts and objectives of CSL, then we introduce a generalization of the existing framework (namely the Hölder LRIP) which we finally connect with the results of Section 3 and 2 to establish the guarantees.

4.1 Compressive Statistical Learning

In contrast to the empirical risk minimization approach described in Section 3.1 the principle of compressive statistical learning is to learn a hypothesis $\hat{h}$ by relying on a single *sketch* vector $\mathbf{s} \in \mathbb{R}^m$ instead of the full dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$ (or equivalently the empirical distribution π_n). This sketch aims to summarize the properties of the empirical distribution that are essential for the learning task. The benefits of this approach are numerous. First, as a side effect of its definition, the sketching mechanism is adapted for distributed and streaming scenarios since the sketch of a concatenation of datasets is a simple average of the sketches of those datasets. More importantly, when $m \ll nd$ the data are drastically compressed, which facilitates their storage and transfer. Finally, it has be shown that sketching can preserve privacy (Chatalic, 2020; Balog et al., 2018) since the transformation which turns a dataset into a single vector discard the individual-user informations.

The compressive statistical learning framework requires two steps: 1) to compute a sketch vector $\mathbf{s} \in \mathbb{R}^m$ of size m driven by the complexity of the learning task 2) to address a nonlinear least-squares optimization problem on this sketch to learn the hypothesis $\hat{h}$ that best solves our learning task. As described latter, this step is an inverse problem in the space of measures and can be related to the generalized method of moments (Hall, 2005). We summarize in the following the main concepts related to the CSL theory established in Gribonval et al. (2021a,b) that will be useful to describe our contributions.

4.1.1 The Sketching Operator

Given a collection of data points $\mathbf{X} = (\mathbf{x}_i)_{i \in [\![n]\!]}$ where $\mathbf{x}_i \in \mathcal{X}$, the CSL procedure relies on an operator Φ which maps a sample $\mathbf{x}_i \in \mathcal{X}$ to either $\Phi(\mathbf{x}_i) \in \mathbb{R}^m$ or $\mathbb{C}^m$. Based on this operator, a sketch of a dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$ is defined *via* the *vector*

$$\mathbf{s} := \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{x}_i)\,.$$

The main challenge is to find, depending on the task, an adequate Φ and a reasonable sketch size m to learn the specific task (see Figure 4). As described in the next sections this can be achieved by exploiting links with the formalism of linear inverse problems, compressive sensing, and low complexity recovery. Given Φ, the associated *sketching operator* is

$$\begin{aligned} \mathcal{A} :\ & \mathcal{P}(\mathcal{X}) \to \mathbb{R}^m \text{ or } \mathbb{C}^m \\ & \pi \to \mathcal{A}(\pi) := \int_{\mathcal{X}} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x})\,. \end{aligned} \tag{21}$$

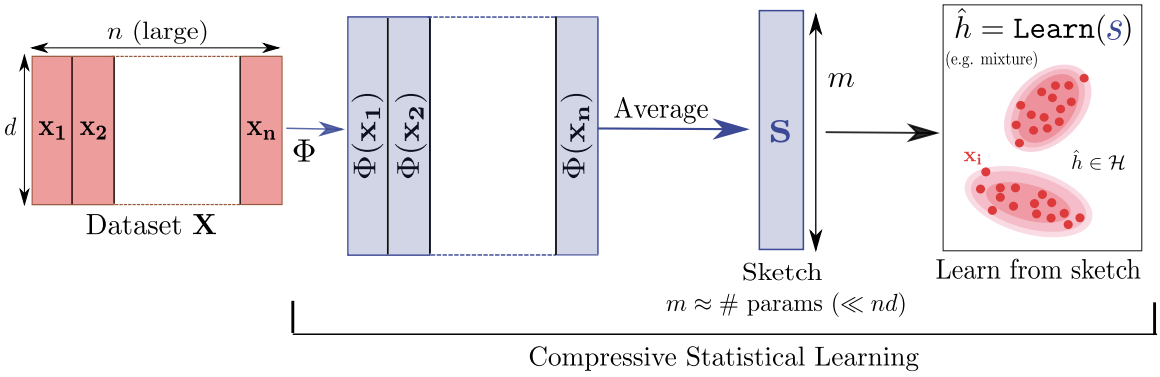

Figure 4: The principle of CSL (when $\mathcal{X} = \mathbb{R}^d$). From a dataset $\mathbf{X}$ with n samples (usually n is large) we push each sample $\mathbf{x}_i \in \mathbb{R}^d$ to either $\mathbb{R}^m$ or $\mathbb{C}^m$ using a well-chosen feature function $\Phi(\mathbf{x}_i)$. The second step is to average all the $\Phi(\mathbf{x}_i)$ to form a *sketch* of of the dataset $\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i)$ (which is convenient for distributed data and data streams). We finally learn a hypothesis $\hat{h} \in \mathcal{H}$ based only the sketch whose size is driven by the learning task and is usually of the order of the number of parameters to learn.

This operator is "linear"[17] in π in that $\mathcal{A}((1-\lambda)\pi + \lambda\pi') = (1-\lambda)\mathcal{A}(\pi) + \lambda\mathcal{A}(\pi')$ for $\lambda \in [0,1]$. When applied to the empirical distribution $\pi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$ we recover the sketch $\mathbf{s}$ as

$$\mathcal{A}(\pi_n) = \mathcal{A}(\frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathbf{s}\,.$$

This sketch can be understood as a the average of generalized empirical moments on the training collection based on the feature function Φ (Hall, 2005).

4.1.2 The Model Set and the Decoder

A central operator in CSL is the *decoder* that is, informally, an operator Δ that goes in the other direction than $\mathcal{A}$: it takes as input a vector and outputs a probability distribution. Ideally we would like to be able to perfectly decode our original distribution from the sketch, *i.e.* to find Δ such that $\Delta \circ \mathcal{A} = \mathrm{id}$. However, as described in Gribonval et al. (2021a), we can not hope to perfectly recover any distribution without assumptions. These assumptions are formalized by the means of a *model set* $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ which describes a subset of probability distributions where the decoding is perfect and robust to noise. A *decoder* is defined very generally as an operator

$$\Delta : \mathbf{s} \to \Delta[\mathbf{s}] \in \mathfrak{S}\,.$$

Suppose for the moment that we know how to sketch and how to decode *i.e.* we know $\mathcal{A}$ and Δ. Given a sketch $\mathbf{s}$ of the dataset and a decoder Δ we can find a hypothesis based on the following risk minimization:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)\,.$$

As such in CSL the risk $\mathcal{R}(\Delta[\mathbf{s}], \cdot)$ acts as a proxy for the empirical risk $\mathcal{R}(\pi_n, \cdot)$, and one hopes to produce a hypothesis which is as good as the one obtained by empirical risk minimization (ERM). At

17. We can extend $\mathcal{A}$ to the space of finite signed measure $\mathcal{M}(\mathcal{X})$ where it is a linear operator in the usual sense.

first sight it seems that solving $\arg\min_{h\in\mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$ is as hard as doing ERM. The crucial point is that, by definition, $\Delta[\mathbf{s}]$ is a probability distribution in the model set $\mathfrak{S}$ and thus usually admits a simple expression. Consequently finding $\hat{h}$ with this procedure is most of the time simpler than doing ERM.

How to obtain statistical guarantees ? Theoretical guarantees of CSL can be derived when the operator $\mathcal{A}$ satisfies the so-called *Lower Restricted Isometric Property* (LRIP) (Gribonval et al., 2021a; Keriven and Gribonval, 2018):

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \,. \tag{22}$$

This property implies that two distributions in the model set $\mathfrak{S}$ (*i.e.* "simple" distributions for which we hope that everything works "fine") have the same sketches then they are equivalent with respect to the task-dependent metric $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$, *i.e.*, they lead to the same risk for every hypothesis. When this condition holds, the following decoder Δ provides many interesting guarantees:

$$\Delta[\mathbf{s}] \in \arg\min_{\pi\in\mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \,. \tag{23}$$

Indeed it can be shown Gribonval et al. (2021a) that this decoder is *ideal* in the sense that it satisfies the *Instance Optimality Property* (IOP) which allows to have a control on the excess risk for *all* probability distributions. We will describe this property more in depth in Section 4.2 and only give now its consequence when we consider any data generating distribution $\pi \in \mathcal{P}(\mathcal{X})$ associated to the optimal hypothesis $h^* \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h)$ and π_n an empirical distribution associated to samples from π. Suppose that we have access only to a sketch $\mathbf{s} = \mathcal{A}(\pi_n)$ of this empirical distribution with $\mathcal{A}$ that satisfies the LRIP. Consider the decoder Δ defined in (23) and $\hat{h}$ such that $\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$. Using the IOP property it can be shown that

$$\|\pi - \Delta[\mathbf{s}]\|_{\mathcal{L}(\mathcal{H})} \lesssim \mathrm{Bias}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2 \,,$$

where $\mathrm{Bias}(\pi, \mathfrak{S})$ is a *bias term* (which will be properly defined latter) which is large when π is far from the model set and vanishes when $\pi \in \mathfrak{S}$. This leads to the following bound on the excess risk:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim \mathrm{Bias}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2 \,.$$

This inequality echoes the well-known risk decomposition in statistical learning: the first term $\mathrm{Bias}(\pi, \mathfrak{S})$ resembles the approximation error coming from the chosen model and $\|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$ resembles the estimation error and typically converges to zero with a $n^{-1/2}$ rate. Consequently, if the model set $\mathfrak{S}$ is such that the bias term is of the order of the true risk $\mathcal{R}(\pi, h^*)$ (this can be ensured for certain learning tasks Gribonval et al., 2021b) then $\mathcal{R}(\pi, \hat{h})$ converges to the order of the true risk as n grows.

4.2 Extending Compressive Statistical Learning Guarantees with Hölder LRIP and Hölder IOP

In this section we define an extended notion of LRIP, namely the Hölder LRIP, and show that it can be exploited to control the statistical performance of compressive statistical learning. The Hölder LRIP is basically a relaxation of the LRIP with a Hölder exponant $\delta \in (0, 1]$. To connect with the previous sections, this exponent will also be related to the one found in Section 2 to control W_p by the MMD. We consider the following definition:

Definition 35 (Hölder LRIP and IOP) *Consider a learning task $\mathcal{L}(\mathcal{H})$, an exponent $p \in [1, +\infty)$, and a model set $\mathfrak{S}$. A sketching operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{C}^m$ satisfies the Hölder LRIP for $\delta \in (0, 1]$ with error $\eta \geq 0$ and constant $C > 0$ if*

$$\forall \pi, \pi' \in \mathfrak{S},\ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\ \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^\delta + \eta \,. \tag{Hölder-LRIP}$$

A decoder $\Delta : \mathbb{C}^m \to \mathfrak{S}$ *satisfies the Hölder IOP for* $\delta \in (0, 1]$ *with error* $\eta \geq 0$ *and constant* $C > 0$ *if*

$$\forall \pi \in \mathcal{P}(\mathcal{X}), \forall \mathbf{e} \in \mathbb{C}^m, \|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} \leq \mathrm{Bias}(\pi, \mathfrak{S}) + C \ \|\mathbf{e}\|_2^\delta + \eta\,, \qquad \text{(Hölder-IOP)}$$

where $\mathrm{Bias}(\cdot, \mathfrak{S}) : \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$ *is a function such that* $\forall \pi \in \mathfrak{S},\ \mathrm{Bias}(\pi, \mathfrak{S}) = 0$.

The instance optimality property means that the decoder is able to retrieve (with error η) any probability distribution when the modeling is exact (*i.e.* $\pi \in \mathfrak{S}$ and $\mathbf{e} = 0$). As this condition is rarely met in practice, the IOP property also captures robustness to some noise $\mathbf{e}$ and modeling error. As such, the decoding error $\|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p}$ is bounded by the amplitude of the noise and the bias term. The previous definition generalizes the classical LRIP and IOP property (including their definition with an error term η Gribonval et al., 2021a) since both are met when $\delta = 1$. It turns out that both Hölder LRIP and IOP are equivalent as stated in the next result:

Proposition 36 (Equivalence of Hölder LRIP and IOP) *Consider a learning task* $\mathcal{L}(\mathcal{H})$*, an exponent* $p \in [1, +\infty)$ *, and a model set* $\mathfrak{S}$*.*

(i) If $\mathcal{A}$ *satisfies* (Hölder-LRIP) *with error* $\eta \geq 0$ *and constant* $C > 0$ *then the "ideal" decoder defined by*

$$\Delta[\mathbf{s}] \in \operatorname*{arg\,min}_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2\,, \tag{24}$$

satisfies (Hölder-IOP) *with constant* $2C > 0$*, error* $\eta \geq 0$ *and*

$$\mathrm{Bias}(\pi, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2^\delta\,.$$

(ii) Conversely if the decoder Δ *defined in* (24) *satisfies* (Hölder-IOP) *with error* $\eta \geq 0$*, constant* $C > 0$ *and* $\mathrm{Bias}(\pi, \mathfrak{S})$ *defined above, then* $\mathcal{A}$ *satisfies* (Hölder-LRIP) *with constant* $C > 0$ *and error* 2η*.*

The proof is deferred to Appendix C.1. In this paper we always assume that the minimization problem (24) has at least one solution and, as in Bourrier et al. (2014), the result can be adjusted to handle the case where the arg min defining the ideal decoder is only approximated to a certain accuracy. This proposition states that if the Hölder LRIP is satisfied, then the decoder that returns the element in the model that best matches the measurement $\mathcal{A}(\pi)$ is instance optimal. On the other hand, if some instance optimal decoder exists, then the Hölder LRIP must be satisfied. In other words, when the Hölder LRIP is satisfied, we know that a negligible amount of information is lost when encoding a probability measure in $\mathfrak{S}$. As advertised the Hölder LRIP allows us to have some guarantees on the excess risk as described in the next theorem:

Theorem 37 (Compressed statistical learning guarantees) *Consider a sketching operator* $\mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{C}^m$ *that satisfies the Hölder LRIP with* $\delta \in (0, 1]$*, constant* $C > 0$ *and error* $\eta \geq 0$*. Let* $\pi \in \mathcal{P}(\mathcal{X})$ *be the data generating distribution and* $\mathbf{x}_1, \cdots, \mathbf{x}_n \sim \pi$ *(not necessarily i.i.d.). Consider the empirical distribution* $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ *and a sketch of the dataset* $\mathbf{s} = \mathcal{A}(\pi_n)$*.*

Let $h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi, h)$ *be the optimal hypothesis and* $\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$ *where* $\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2$*. Then*

$$\mathcal{R}(\pi, \hat{h})^{1/p} - \mathcal{R}(\pi, h^*)^{1/p} \leq 2\,\mathrm{Bias}(\pi, \mathfrak{S}) + 2C\|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^\delta + 2\eta\,,$$

where $\mathrm{Bias}(\pi, \mathfrak{S}) = \inf_{\tau \in \mathfrak{S}} \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2^\delta$.

Proof Using Proposition 36 we know that the decoder is instance optimal and satisfies the Hölder IOP (Hölder-IOP). Consider $\mathbf{e} = \mathcal{A}(\pi_n) - \mathcal{A}(\pi)$ we have by definition $\|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} \leq \mathrm{Bias}(\pi, \mathfrak{S}) + C\,\|\mathbf{e}\|_2^\delta + \eta$ which gives $\|\pi - \Delta[\mathcal{A}(\pi_n)]\|_{\mathcal{L}(\mathcal{H}),p} \leq \mathrm{Bias}(\pi, \mathfrak{S}) + C\,\|\mathcal{A}(\pi_n) - \mathcal{A}(\pi)\|_2^\delta + \eta$. We conclude the proof by using $\mathcal{R}(\pi, \hat{h})^{1/p} - \mathcal{R}(\pi, h^*)^{1/p} \leq 2\|\pi - \Delta[\mathbf{s}]\|_{\mathcal{L}(\mathcal{H}),p} = 2\|\pi - \Delta[\mathcal{A}(\pi_n)]\|_{\mathcal{L}(\mathcal{H}),p}$. ■

When the samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are *i.i.d.*[18] This result is essential: it illustrates that if we have carefully designed $\mathfrak{S}$ so that the bias term is *of the order of* $\mathcal{R}(\pi, h^*)^{1/p}$, and if we know a sketching operator with the Hölder LRIP property, then $\mathcal{R}(\pi, \hat{h})^{1/p}$ converges to a constant times the order of the true risk as n grows (when the error term $\eta = 0$). The notable price to pay between this result and the one presented in the context of the LRIP ($\delta = 1$) is that while the usual guaranteed speed of convergence is $O(n^{-1/2})$ here it becomes $O(n^{-\delta/2})$, which is slower. The next section outlines how the various results presented in this work can be applied to establish the Hölder LRIP.

4.3 Connecting the Hölder LRIP with the Results of Section 2 and 3

As described in Theorem 37, guarantees on the excess risk can be achieved with a sketching operator $\mathcal{A}$ that satisfies the Hölder LRIP. In this section, we provide elements to obtain this property. In line with the approach developed in Gribonval et al. (2021a), the core of our reasoning is based on the theory of kernel embedding of probability distributions and random features.

4.3.1 Restricted Wasserstein Regularity is Necessary to the Hölder LRIP

Firstly, a prerequisite for the Hölder LRIP is the Wasserstein regularity condition (Definition 28) of the learning task when restricted to the model set $\mathfrak{S}$. More precisely we have the following result:

Proposition 38 (Restricted Wasserstein regularity is necessary) *Consider $\mathcal{X} = \mathbb{R}^d$ equipped with a norm $\|\cdot\|, p \in [1, +\infty)$, and a model set $\mathfrak{S} \subseteq \mathcal{P}_p(\mathbb{R}^d)$. Consider a sketching operator $\mathcal{A}$ defined by $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ with $\Phi \in \mathrm{Lip}_L\left((\mathbb{R}^d, \|\cdot\|), (\mathbb{R}^m, \|\cdot\|_2)\right)$. If $\mathcal{A}$ satisfies* (Hölder-LRIP) *with error $\eta = 0$, constant $C > 0$ and $\delta = 1$ then*

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq CL\,\mathrm{W}_1(\pi, \pi') \leq CL\,\mathrm{W}_p(\pi, \pi')\,,$$

where the Wasserstein distance is computed with the distance $D(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

The proof is deferred to Appendix C.2 and simply amounts to showing that $\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \leq L\,\mathrm{W}_1(\pi, \pi')$. According to this proposition, if Φ is Lipschitz and $\mathcal{A}$ satisfies the Hölder LRIP with $\delta = 1$ then $\mathcal{L}(\mathcal{H})$ is *necessarily* p-Wasserstein regular when we restrict the Definition 28 to distributions belonging to the model set $\mathfrak{S}$. In particular this proposition applies to the classical LRIP setting of Gribonval et al. (2021a). More importantly the Lipschitz hypothesis encompasses the case where Φ is defined with random Fourier features[19] as usually considered in the compressive statistical learning literature (Gribonval et al., 2021a,b; Belhadji and Gribonval, 2022; Shi et al., 2022a,b). This result thus shows that a *restricted* Wasserstein regularity is necessary for establishing statistical guarantees of CSL through the Hölder LRIP.

Remark 39 *The previous result can be easily generalized to the case where $\delta \in (0, 1)$. Under the same assumptions on Φ, if $\mathcal{A}$ satisfies* (Hölder-LRIP) *with an error of $\eta = 0$, a constant $C > 0$, and*

18. We emphasize that the *i.i.d.* assumption is not required in order to obtain the bound in Theorem 37. It is only used to guarantee that $\|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2 \underset{n \to +\infty}{\to} 0$. the term $\|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2$, which is the empirical estimation error, goes to zero as $n \to +\infty$ with a typical $n^{-1/2}$ rate.

19. In this setting $\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}\left(\sqrt{2}\sin(\mathbf{x}^\top \boldsymbol{\omega}_1), \sqrt{2}\cos(\mathbf{x}^\top \boldsymbol{\omega}_1), \cdots, \sqrt{2}\sin(\mathbf{x}^\top \boldsymbol{\omega}_{m/2}), \sqrt{2}\cos(\mathbf{x}^\top \boldsymbol{\omega}_{m/2})\right)^\top$ for some random draw of $\boldsymbol{\omega}_j$.

$\delta \in (0,1)$, we can show that $\forall \pi, \pi' \in \mathfrak{S}$, $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq CL^\delta \, \mathrm{W}_1(\pi,\pi')^\delta \leq CL^\delta \, \mathrm{W}_p(\pi,\pi')^\delta$. This condition extends the Wasserstein regularity property, and it raises the question of which learning tasks satisfy it.

4.3.2 From Wasserstein Regularity to the Kernel Hölder LRIP and Hölder LRIP

Interestingly, a converse of Proposition 38 is also true. Indeed, as shown in Section 3 many learning tasks are Wasserstein regular, and this, *independently* of the choice of the model set $\mathfrak{S}$. For instance, this is true for compression-type tasks such as K-means/medians, PCA, or supervised learning tasks such as regression and binary classification (see Table 1).

Consequently, if we add the elements of Section 2, namely that $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable (Definition 5), we can obtain, under certain assumptions about $\kappa, \mathfrak{S}$, that the metric associated with the task satisfies the following chain of inequalities:

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \overset{\text{Section 3}}{\lesssim} \mathrm{W}_p(\pi,\pi') \overset{\text{Section 2}}{\lesssim} \|\pi - \pi'\|_\kappa^\delta \,. \tag{25}$$

As shown in Section 2, the last inequality can be obtained with an MMD associated with TI, PSD kernels and under certain assumptions on the moments of the distributions in $\mathfrak{S}$ and their regularity. In other words, by combining the results of Section 2 and 3, our analysis shows that for many learning tasks and with some hypothesis on the kernel $\kappa, \mathfrak{S}$ the task metric is bounded by MMD^δ uniformly on $\mathfrak{S}$. We refer to this property as the *kernel Hölder LRIP*, *i.e.* when there exists $C > 0, \delta \in (0,1], \eta \geq 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\pi - \pi'\|_\kappa^\delta + \eta \,. \tag{26}$$

The echoes the *kernel* LRIP described in Gribonval et al. (2021a) but with a Hölder exponent $\delta \in (0,1]$. Informally, our findings show that a kernel Hölder LRIP is not so difficult to obtain for many learning tasks. Therefore, as long as the MMD can be uniformly controlled on $\mathfrak{S}$ by a distance between finite-dimensional sketches, *i.e.* when

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_\kappa \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \,, \tag{27}$$

we can use all the results from the previous sections to obtain the Hölder LRIP.

The property described in (27) depends only on the operator $\mathcal{A}$, the kernel κ, and the model set $\mathfrak{S}$. To establish it, several strategies have been considered in the literature. For the sake of conciseness, we only provide some intuition here and refer the reader to Gribonval et al. (2021a) for a more detailed discussion. The general idea is to construct, from a kernel κ, a function $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$\forall \mathbf{x}, \mathbf{y}, \ \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathbb{R}^m} \approx \kappa(\mathbf{x}, \mathbf{y}) \,, \tag{28}$$

and to "extend" this approximation to pairs of probability distributions as

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_\kappa^2 \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2 \,, \tag{29}$$

where $\mathcal{A}$ is given by Φ as in (21). Ensuring (28) is a well established area of research and, when κ is TI, PSD, approaches such as random Fourier features (RFF) (Rahimi and Recht, 2007), which rely on Bochner's theorem, can be used (see *e.g.* Liu et al. 2021 for a review). On the other hand, condition (29) is much more challenging to obtain. For TI, PSD kernels RFF can also be used: given a pair π, π', the main strategy is to prove a pointwise control of the form $(1-\rho)\|\pi - \pi'\|_\kappa^2 \leq \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2 \leq (1+\rho)\|\pi - \pi'\|_\kappa^2$ with high probability for $\rho \in (0,1]$, and then being able to control certain *covering numbers* related to $\mathfrak{S}$ to obtain a uniform control (Gribonval et al., 2021a; Belhadji and Gribonval, 2022). Another approach, considered for example in Chatalic et al. (2022), is to construct Φ based on data-dependent Nyström approximation which exploits a small random subset of the dataset (and also requires controlling covering numbers). These approaches ensure that for a sufficiently large but controlled m, the condition (29) is satisfied and therefore also (27).

4.3.3 Discussion

As a consequence, when the task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein regular and the space $(\mathfrak{S}, \mathrm{W}_p)$ is (κ, δ)-embeddable the approach presented in this paper combined with the one of Gribonval et al. (2021a) to obtain (27) show that sketching operators based on random Fourier features are suited for a wide range of tasks and lead to CSL guarantees. With this strategy, the convergence rate of the empirical risk (Theorem 37) is governed by the exponent $\delta \in (0, 1]$ resulting from the comparison between W_p and the MMD. This can be placed in the context of results already obtained in CSL for compressive clustering and compressive mixture modeling.

Firstly, it is already established that for mixtures of K Diracs (used in compressive K-means) separation assumptions on the centers are necessary to establish the LRIP (Gribonval et al., 2021b, Lemma 3.4.). One might ask if these assumptions can be dispensed at the cost of slower convergence with the Hölder LRIP. In this framework, our results demonstrate that the distance W_p cannot be controlled by the MMD when $\delta > 2/K$ (Corollary 12). This raises the question of whether this rate is indeed achievable without separation assumptions, and if, in such a case, (27) could also be obtained, which would imply the Hölder LRIP with $\delta = 1/(2K)$ *without separation*.

Furthermore, these same separation assumptions are also used for compressive learning of Gaussian mixture (for compressive GMM estimation). Interestingly, in this case, Theorem 15 ensures that we can control W_p by MMD^δ with an exponent δ as close as desired to $\delta = \frac{1}{2p}$ and with a kernel of the Matérn class. Establishing control (27) without separation for these models would enable obtaining learning rates of the order of $n^{-1/(4p)}$ for compressive GMM with relaxed assumptions.

5. Conclusion & Perspectives

The main contributions of this paper are the following. We establish different bounds between metrics between probability distributions. We show that for many learning tasks, the task-related metric can be controlled by a Wasserstein distance. In particular, many supervised and unsupervised tasks fall into this category (PCA, K-Means, GMM learning, linear and nonlinear regression...). We show that the Wasserstein distance can be controlled by kernel norms to the power of a Hölder exponent smaller than 1 and under certain conditions on the regularity of the kernel and of the distributions at stake (by introducing a *model set* of distributions). These different results allow us to establish learning guarantees in the context of *compressive learning* whose goal is to summarized the training data in a single vector, by a so-called *sketching operator*, and to rely solely on this vector to solve the learning task. The different bounds allow us to establish a property called the *Hölder LRIP* that generalizes the LRIP property in compressive learning and provide a control of the excess risk related to the compressive learning procedure. Therefore, one of the contributions of this article is to provide a general framework for obtaining compressive learning guarantees.

This work opens many perspectives. The first one is to use our results for new compressive learning tasks that have been tackled in practice but for which theoretical guarantees are missing. In particular, we envision applications of our framework for learning generative models based on sketching (Schellekens and Jacques, 2020), denoising (Shi et al., 2022a) or for classification tasks (Schellekens and Jacques, 2018). Related to the compressive statistical learning theory, another interesting line of works would be to see if we can construct interesting sketching operators from the different kernels used in this paper for tasks for which there are already compressive learning guarantees. More precisely, for compressive learning tasks such as K-means and GMM one question would be to see if we can obtain compressive learning guarantees without separation assumptions (Gribonval et al., 2021b), possibly at the price of a Hölder exponent $\delta < 1$ hence with reduced rate of convergence with respect to the number of samples. Another interesting perspective concern the bounds between the Wasserstein distance and the MMD. We believe that the different results presented in this paper could be used for specific problems related to the statistical estimation of the Wasserstein distance.

Acknowledgments

This project was supported in part by the AllegroAssai ANR project ANR-19-CHIA-0009. This work was supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

Appendix A. Proofs of Section 2

A.1 Proof of Proposition 2 and Corollary 3

We recall the proposition:

Proposition 2 *Let $(\mathcal{X}, D)$ be a complete separable metric space, $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a PSD kernel, $\mathcal{H}_\kappa$ the associated RKHS and $B_\kappa := \{f \in \mathcal{H}_\kappa : \|f\|_{\mathcal{H}_\kappa} \leq 1\}$ the unit ball in $\mathcal{H}_\kappa$. Consider the Wasserstein distances computed with the metric D. For any $C > 0$ the following statements are equivalent:*

(i)
$$B_\kappa \subseteq \mathrm{Lip}_C((\mathcal{X}, D), \mathbb{R}) \tag{5}$$

(ii)
$$\forall p \in [1, +\infty), \forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_\kappa \leq C\, \mathrm{W}_p(\pi, \pi') \tag{6}$$

(iii)
$$\exists p \in [1, +\infty), \forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_\kappa \leq C\, \mathrm{W}_p(\pi, \pi') \tag{7}$$

(iv)
$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X},\ \kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{y}, \mathbf{y}) - 2\kappa(\mathbf{x}, \mathbf{y}) \leq C^2 D^2(\mathbf{x}, \mathbf{y}) \tag{8}$$

Proof We will prove (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (iv) $\implies$ (i).
(i) $\implies$ (ii). Assuming (i) we prove (ii) for $p = 1$. By monotonicity of the Wasserstein distance with respect to p we have the conclusion for any $p \in [1, +\infty)$. Considering $\pi, \pi' \in \mathcal{P}_1(\mathcal{X})$, we have $\|\pi - \pi'\|_\kappa = \sup_{f \in B_\kappa} |\int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y})|$ (Sriperumbudur et al., 2010). For any $f \in B_\kappa$, by hypothesis (i) we have $\frac{1}{C} f \in \mathrm{Lip}_1((\mathcal{X}, D), (\mathbb{R}, |\cdot|))$ thus by the dual characterization of the 1-Wasserstein distance (2) we obtain $\|\pi - \pi'\|_\kappa \leq C\, \mathrm{W}_1(\pi, \pi')$. The implication **(ii) $\implies$ (iii)** is straightforward.
(iii) $\implies$ (iv). Consider $\pi = \delta_{\mathbf{x}}, \pi' = \delta_{\mathbf{y}}$ for arbitrary $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. We have $\|\pi - \pi'\|_\kappa^2 = \kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{y}, \mathbf{y}) - 2\kappa(\mathbf{x}, \mathbf{y})$ and $\mathrm{W}_p(\pi, \pi') = D(\mathbf{x}, \mathbf{y})$, hence the conclusion.
(iv) $\implies$ (i). Considering $f \in B_\kappa$, we have for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$\begin{aligned} |f(\mathbf{x}) - f(\mathbf{y})|^2 &= |\langle f, \kappa(\mathbf{x}, \cdot)\rangle_{\mathcal{H}_\kappa} - \langle f, \kappa(\mathbf{y}, \cdot)\rangle_{\mathcal{H}_\kappa}|^2 = |\langle f, \kappa(\mathbf{x}, \cdot) - \kappa(\mathbf{y}, \cdot)\rangle_{\mathcal{H}_\kappa}|^2 \\ &\leq \|f\|_{\mathcal{H}_\kappa}^2 \|\kappa(\mathbf{x}, \cdot) - \kappa(\mathbf{y}, \cdot)\|_{\mathcal{H}_\kappa}^2 \leq 1 \cdot (\|\kappa(\mathbf{x}, \cdot)\|_{\mathcal{H}_\kappa}^2 + \|\kappa(\mathbf{y}, \cdot)\|_{\mathcal{H}_\kappa}^2 - 2\kappa(\mathbf{x}, \mathbf{y})) \\ &= \kappa(\mathbf{x}, \mathbf{x}) + \kappa(\mathbf{y}, \mathbf{y}) - 2\kappa(\mathbf{x}, \mathbf{y}) \overset{(iv)}{\leq} C^2 D^2(\mathbf{x}, \mathbf{y})\,. \end{aligned} \tag{30}$$

This gives $|f(\mathbf{x}) - f(\mathbf{y})| \leq C D(\mathbf{x}, \mathbf{y})$ hence f is C-Lipschitz with respect to the metric D. ■

Corollary 3 *Consider $\mathcal{X} = \mathbb{R}^d$ equipped with the Euclidean distance $D(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ and a PSD kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that is normalized, i.e. $\kappa(\mathbf{x}, \mathbf{x}) = 1$ for every $\mathbf{x} \in \mathcal{X}$. Assume that for each $\mathbf{x} \in \mathcal{X}$ the function $\phi_{\mathbf{x}} : \mathbf{y} \mapsto \kappa(\mathbf{x}, \mathbf{y})$ is C^2 in a neighborhood of $\mathbf{x}$, and denote $\mathbf{H}_{\mathbf{x}} = -\nabla^2[\phi_{\mathbf{x}}](\mathbf{x})$ its negative Hessian matrix evaluated at $\mathbf{x}$. Then the following holds:*

(i) Any of the four equivalent properties of Proposition 2 implies

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \lambda_{\max}(\mathbf{H}_{\mathbf{x}}) \leq C^2\,, \tag{9}$$

where $\lambda_{\max}(\mathbf{H}_{\mathbf{x}})$ denotes the largest eigenvalue of $\mathbf{H}_{\mathbf{x}}$.

(ii) If κ is translation invariant, i.e. $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ for every $\mathbf{x},\mathbf{y} \in \mathcal{X}$, then conversely, (6) *holds with $C := \sqrt{\sup_{\mathbf{x}} \lambda_{\max}(\mathbf{H}_{\mathbf{x}})} = \sqrt{\lambda_{\max}(-\nabla^2[\kappa_0](0))}$.*

Proof For the first part (i). Since the kernel is normalized, using formulation (iv) of the four equivalent properties of Proposition 2 and setting $\mathbf{h} = \mathbf{y} - \mathbf{x}$ yields:

$$\forall \mathbf{x}, \mathbf{h} \in \mathbb{R}^d, \kappa(\mathbf{x}, \mathbf{x}+\mathbf{h}) \geq 1 - \frac{C^2}{2}\|\mathbf{h}\|_2^2 \tag{31}$$

Given any $\mathbf{x} \in \mathbb{R}^d$, since $\phi_{\mathbf{x}}$ is C^2 in a neighborhood of $\mathbf{x}$, a Taylor expansion yields:

$$\phi_{\mathbf{x}}(\mathbf{x}+\mathbf{h}) = \phi_{\mathbf{x}}(\mathbf{x}) + \langle \nabla\phi_{\mathbf{x}}(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2}\mathbf{h}^\top \nabla^2\phi_{\mathbf{x}}(\mathbf{x})\mathbf{h} + \|\mathbf{h}\|_2^2 g(\mathbf{x}+\mathbf{h}) \tag{32}$$

where g is a function such that $\lim_{\mathbf{h}\to 0} g(\mathbf{x}+\mathbf{h}) = 0$. Moreover $\phi_{\mathbf{x}}(\mathbf{x}) = \kappa(\mathbf{x},\mathbf{x}) = 1$ and $\nabla\phi_{\mathbf{x}}(\mathbf{x}) = 0$ since the maximum of $\mathbf{y} \to \kappa(\mathbf{x},\mathbf{y})$ is always attained at $\mathbf{y} = \mathbf{x}$ when κ is a PSD kernel. Hence

$$\kappa(\mathbf{x}, \mathbf{x}+\mathbf{h}) = 1 - \frac{1}{2}\mathbf{h}^\top \mathbf{H}_{\mathbf{x}}\mathbf{h} + o_{\|\mathbf{h}\|_2\to 0}(\|\mathbf{h}\|_2^2) \tag{33}$$

Considering an arbitrary unit vector $\mathbf{u}$ and $\mathbf{h} = \epsilon\mathbf{u}$ and using (31) gives: $-\epsilon^2\mathbf{u}^\top\mathbf{H}_{\mathbf{x}}\mathbf{u} \geq -\epsilon^2(C^2 + o_{\epsilon\to 0}(1))$ hence $\mathbf{u}^\top\mathbf{H}_{\mathbf{x}}\mathbf{u} \leq C^2$. Since $\phi_{\mathbf{x}}$ is C^2 in a neighborhood of $\mathbf{x}$, by Schwarz's theorem its Hessian matrix is symmetric hence diagonalizable, and the above property implies that $\lambda_{\max}(\mathbf{H}_{\mathbf{x}}) \leq C^2$. As this holds for every $\mathbf{x}$ we get the desired conclusion.

For (ii), observe first that $\mathbf{H}_{\mathbf{x}} = -\nabla^2\phi_{\mathbf{x}}(\mathbf{x}) = -\nabla^2[\kappa_0](0)$ is independent of $\mathbf{x}$. Since $\phi_{\mathbf{x}}$ is C^2 the matrix $\mathbf{H}_{\mathbf{x}}$ is also symmetric, and since $\phi_{\mathbf{x}}(\mathbf{y})$ is maximum at $\mathbf{y} = \mathbf{x}$, $\mathbf{H}_{\mathbf{x}}$ is also positive semi-definite, hence $\sup_{\mathbf{x}} \lambda_{\max}(\mathbf{H}_{\mathbf{x}}) = \lambda_{\max}(-\nabla^2[\kappa_0](0)) \geq 0$ and $C := \sqrt{\lambda_{\max}(-\nabla^2[\kappa_0](0))}$ is well-defined. Now, by Bochner's theorem, since the kernel is normalized, real-valued, and twice continuously differentiable in the neighborhood of zero, there is a frequency distribution $\Lambda \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\kappa_0(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\cos(\boldsymbol{\omega}^\top\mathbf{x})]$. It follows by standard arguments that the gradient and Hessian can be written as $\nabla\kappa_0(\mathbf{x}) = -\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\sin(\boldsymbol{\omega}^\top\mathbf{x})]$, $\nabla^2\kappa_0(\mathbf{x}) = -\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\boldsymbol{\omega}^\top\cos(\boldsymbol{\omega}^\top\mathbf{x})]$ Consequently, $\mathbf{H}_{\mathbf{x}} = -\nabla^2[\kappa_0](0) = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\boldsymbol{\omega}^\top]$ and $C = \sqrt{\lambda_{\max}(\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\boldsymbol{\omega}^\top])}$. Consider $\mathbf{z} \in \mathbb{R}^d$, we will show that:

$$2(1 - \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\cos(\boldsymbol{\omega}^\top\mathbf{z})]) \leq C^2\|\mathbf{z}\|_2^2 \tag{34}$$

which will prove property (iv) of Proposition 2, and consequently all other equivalent properties. Indeed, using that $1-\cos(t) \leq \frac{t^2}{2}$ for all $t \in \mathbb{R}$ we have $1 - \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\cos(\boldsymbol{\omega}^\top\mathbf{z})] = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[1-\cos(\boldsymbol{\omega}^\top\mathbf{z})] \leq \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\frac{|\boldsymbol{\omega}^\top\mathbf{z}|^2}{2}] = \mathbf{z}^\top\left(\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\boldsymbol{\omega}^\top]\right)\mathbf{z} \leq \lambda_{\max}(\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}[\boldsymbol{\omega}\boldsymbol{\omega}^\top])\|\mathbf{z}\|_2^2 = C^2\|\mathbf{z}\|_2^2$. ■

A.2 Rate of Convergence of the MMD

We have the following result which is a direct consequence of Lemma 2 in Briol et al. (2019):

Lemma 40 *et $\pi \in \mathcal{P}(\mathcal{X})$ and $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ where $\mathbf{x}_i \sim \pi$ i.i.d. Then*

$$\mathbb{E}[\|\pi - \pi_n\|_\kappa^2] = n^{-1}\left(\int \kappa(\mathbf{x},\mathbf{x})\mathrm{d}\pi(\mathbf{x}) - \int\int \kappa(\mathbf{x},\mathbf{y})\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi(\mathbf{y})\right), \tag{35}$$

where the expectation is taken on the draws of the $(\mathbf{x}_i)_{i\in[\![n]\!]}$.

Lemma 41 *Let $\pi \in \mathcal{P}(\mathcal{X})$ and $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ where $\mathbf{x}_i \sim \pi$ i.i.d. If $\sup_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) \leq K$ then, for any $\delta \in (0,2]$, we have*

$$\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta] \leq (2K)^{\delta/2} n^{-\delta/2}\,. \tag{36}$$

Proof By the previous lemma, since $\sup_{\mathbf{x}} k(\mathbf{x},\mathbf{x}) \leq K$ we have $\mathbb{E}[\|\pi - \pi_n\|_\kappa^2] \leq 2Kn^{-1}$ since for all $\mathbf{x},\mathbf{y} \in \mathcal{X}$ $|k(\mathbf{x},\mathbf{y})| \leq \sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x},\mathbf{x}) \leq K$ because the kernel is positive semi-definite (the maximum value of a PSD kernel is necessarily on the diagonal). The fact that $\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta] \leq (2K)^{\delta/2} n^{-\delta/2}$ is a direct consequence of Jensen's inequality as $(\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta])^{2/\delta} \leq \mathbb{E}[\|\pi - \pi_n\|_\kappa^2])$ when $2/\delta \geq 1$. ■

A.3 Simple Bound Between Wassersein Distance and Distance Between the Means

Lemma 42 *Let $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ and $\|\cdot\|$ a norm on $\mathbb{R}^d$ with the associated dual norm $\|\cdot\|_\star$ defined by $\|\mathbf{z}\|_\star = \sup_{\|\mathbf{x}\|\leq 1}\langle \mathbf{x},\mathbf{z}\rangle$. Then for every $1 \leq p < \infty$ we have*

$$\mathrm{W}_p(\pi,\pi') \geq \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_\star\,, \tag{37}$$

where the Wassertein distance is computed with the distance $D(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$.

Proof Consider $\mathbf{u} \in \mathbb{R}^d$ an arbitrary vector such that $\|\mathbf{u}\| = 1$ and denote $f_{\mathbf{u}}(\mathbf{x}) = \langle \mathbf{u},\mathbf{x}\rangle \in \mathbb{R}$ for any $\mathbf{x} \in \mathbb{R}^d$. Since $\|\mathbf{u}\| = 1$ the function $f_{\mathbf{u}} : \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz with respect to $D(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$, hence by duality of the Wasserstein distance (2)

$$|\langle \mathbf{u}, \mathrm{m}(\pi) - \mathrm{m}(\pi')\rangle| = \left| \int f_{\mathbf{u}}(\mathbf{x})\mathrm{d}\pi(\mathbf{x}) - \int f_{\mathbf{u}}(\mathbf{y})\mathrm{d}\pi(\mathbf{y}) \right| \leq \mathrm{W}_1(\pi,\pi').$$

The supremum with respect to unitary vectors $\mathbf{u}$ yields $\|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_\star \leq \mathrm{W}_1(\pi,\pi')$. The last step uses the fact that $\mathrm{W}_1(\pi,\pi') \leq \mathrm{W}_p(\pi,\pi')$ for any $p \in [1,+\infty)$ which concludes the proof.

■

A.4 Proof of Proposition 10

We will prove the following result:

Proposition 10 *Let $(\mathcal{X}, D)$ be a complete and separable metric space and consider the Wasserstein distances computed with the distance D. Let κ be any PSD kernel. Consider two arbitrary probability distributions $\pi_0, \pi_1 \in \mathcal{P}(\mathcal{X})$ such that $\|\pi_0 - \pi_1\|_\kappa < +\infty$ and $\mathrm{supp}(\pi_0)$ and $\mathrm{supp}(\pi_1)$ are disjoint*[20]*. Consider $\mathfrak{S} := \{(1-t)\pi_0 + t\pi_1 : t \in [0,1]\}$. If $(\mathfrak{S}, \mathrm{W}_p)$ is (κ,δ)-embeddable then $\delta \leq 1/p$.*

In order to prove this proposition we will use the following lemma:

Lemma 43 *(Niles-Weed and Berthet, 2022, Lemma 9) Let $\pi_0, \pi_1 \in \mathcal{P}(\mathbb{R}^d)$ be any probability distributions. Suppose that there exist two compact sets $S, T \subseteq \mathbb{R}^d$ such that $d(S,T) := \inf_{(\mathbf{x},\mathbf{y})\in S\times T} \|\mathbf{x} - \mathbf{y}\|_2 \geq c > 0$ and that the supports of π_0 and π_1 lie in $S \cup T$. Then*

$$\forall p \in [1,+\infty), \mathrm{W}_p(\pi_0,\pi_1) \geq c|\pi_0(S) - \pi_1(S)|^{1/p}\,. \tag{38}$$

Proof [Of Proposition 10] This result is mainly taken from Theorem 9 in Niles-Weed and Berthet (2022) but we rewrite it in our context for completeness. For any $\lambda \in [0,1]$, set

$$\pi_\lambda := \tfrac{1}{2}\left((1+\lambda)\pi_0 + (1-\lambda)\pi_1\right)\,,$$
$$\pi'_\lambda := \tfrac{1}{2}\left((1-\lambda)\pi_0 + (1+\lambda)\pi_1\right)\,.$$

20. We recall that the support $\mathrm{supp}(\pi)$ of a probability distribution $\pi \in \mathcal{P}(\mathcal{X})$ is the smallest closed set S such that $\pi(S) = 1$.

Note that $\pi_\lambda, \pi'_\lambda \in \mathfrak{S}$ by assumption and $\|\pi_\lambda - \pi'_\lambda\|_\kappa = \lambda\|\pi_0 - \pi_1\|_\kappa$. Since the sets $\mathrm{supp}(\pi_0)$ and $\mathrm{supp}(\pi_1)$ are disjoint, there exist two sets S and T and $c > 0$ such that $\mathrm{supp}\,(\pi_0) \subseteq S$ and $\mathrm{supp}\,(\pi_1) \subseteq T$ and $d(\mathbf{x}, \mathbf{y}) \geq c > 0$ for any $\mathbf{x} \in S, \mathbf{y} \in T$. Moreover it is clear by definition that $\mathrm{supp}(\pi_\lambda)$ and $\mathrm{supp}(\pi'_\lambda)$ lie in $S \cup T$. The Lemma 43 gives, for any p,

$$\mathrm{W}_p(\pi_\lambda, \pi'_\lambda) \geq c\,|\pi_\lambda(S) - \pi'_\lambda(S)|^{1/p} = c\lambda^{1/p}\,. \tag{39}$$

We obtain, for $\delta \in (0, 1]$,

$$\sup_{(\pi,\pi')\in\mathfrak{S}} \frac{\mathrm{W}_p(\pi, \pi')}{\|\pi - \pi'\|_\kappa^\delta} \geq \sup_{\lambda\in(0,1)} \frac{\mathrm{W}_p(\pi_\lambda, \pi'_\lambda)}{\|\pi_\lambda - \pi'_\lambda\|_\kappa^\delta} \gtrsim \sup_{\lambda\in[0,1]} \lambda^{1/p-\delta} = +\infty\,.$$

The last equality is true because $p\delta > 1$. ■

A.5 Proof of Theorem 11

We recall that, for $K \in \mathbb{N}^*$ and $\Omega \subseteq \mathbb{R}^d$, the space of mixtures of K diracs located in Ω is defined by

$$\mathfrak{S}_K(\Omega) := \left\{ \sum_{i=1}^K a_i \delta_{\mathbf{x}_i} : a_i \in \mathbb{R}_+, \sum_{i=1}^K a_i = 1, \forall i \in [\![K]\!], \mathbf{x}_i \in \Omega \right\}. \tag{40}$$

The goal of this section is to prove the following theorem:

Theorem 11 *Consider a TI, PSD kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ such that κ_0 is k times differentiable at 0 with $k \in \mathbb{N}^*$. Consider $p \in [1, +\infty)$, a Wasserstein distance W_p based on a norm in $\mathbb{R}^d$, a vector $\mathbf{x}_0 \in \mathbb{R}^d$, $R > 0$ and $\Omega = B(\mathbf{x}_0, R)$. If $(\mathfrak{S}_{\lfloor \frac{k}{2} \rfloor + 1}(\Omega), \mathrm{W}_p)$ is (κ, δ)-embeddable then $\delta \leq 2/k$.*

We will need the following lemma which states that if the kernel is regular at zero and that we can construct some vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ that satisfy certain conditions then we have a constraint on the Hölder exponent δ.

Lemma 44 *Consider a TI, PSD kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ such that κ_0 is k times differentiable at 0 with $k \in \mathbb{N}^*$. Let $M \in \mathbb{N}^*$ and define for $1 \leq s \leq k$ and $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^M$ the function $c_s(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \sum_{i,j=1}^M \beta_i \beta_j (\alpha_i - \alpha_j)^s$. Suppose that there exists $\boldsymbol{\alpha} \in \mathbb{R}^M \setminus \{0\}$ with $\alpha_i \neq \alpha_j$ for $i \neq j$ and $\boldsymbol{\beta} \in \mathbb{R}^M \setminus \{0\}$ with $\sum_{i=1}^M \beta_i = 0$ such that*

$$c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0\,. \tag{41}$$

Define $r(\boldsymbol{\beta}) := \max\{\#T_+(\boldsymbol{\beta}), \#T_-(\boldsymbol{\beta})\}$ where $T_+(\boldsymbol{\beta}) := \{i \in [\![M]\!] : \beta_i \geq 0\}$ and $T_-(\boldsymbol{\beta}) := \{i \in [\![M]\!] : \beta_i < 0\}$.

Consider $\mathfrak{S} = \mathfrak{S}_{r(\boldsymbol{\beta})}(\Omega)$ with $\Omega = B(\mathbf{x}_0, R)$ where $\mathbf{x}_0 \in \mathbb{R}^d, R > 0$ are arbitrary. If $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable, where W_p is based on a norm $\|\cdot\|$ in $\mathbb{R}^d$ with $p \in [1, +\infty)$, then $\delta \leq 2/k$.

Proof Recall that for a finite signed measure $\mu \in \mathcal{M}(\mathbb{R}^d)$ we have $\|\mu\|_\kappa^2 = \int\int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{y})$. Consider $M \in [\![N]\!]^*, \boldsymbol{\beta} \in \mathbb{R}^M$ such that $\sum_{i=1}^M \beta_i = 0$ and $\boldsymbol{\alpha} \in \mathbb{R}^M \setminus \{0\}$ with $\alpha_i \neq \alpha_j$ when $i \neq j$. We define the measure

$$\mu_\varepsilon := \sum_{i=1}^M \beta_i \delta_{\mathbf{x}_0 + \varepsilon\alpha_i \mathbf{u}}\,, \tag{42}$$

where $\mathbf{u} \in \mathbb{R}^d \setminus \{0\}$ and $0 < \varepsilon < \frac{R}{\|\boldsymbol{\alpha}\|_\infty \|\mathbf{u}\|_2}$ is sufficiently small to ensure that $\mathbf{x}_0 + \varepsilon\alpha_i\mathbf{u} \in \Omega = B(\mathbf{x}_0, R)$. We define $T_+ := \{i \in [\![M]\!] : \beta_i \geq 0\}$ and $T_- := \{i \in [\![M]\!] : \beta_i < 0\}$ such that $T_- \cup T_+ = [\![M]\!]$ and $T_- \cap T_+ = \emptyset$. We define also $\rho := \sum_{i\in T_+} \beta_i = -\sum_{i\in T_-} \beta_i > 0$ and

$$\pi_\varepsilon := \sum_{i\in T_+} \frac{\beta_i}{\rho}\delta_{\mathbf{x}_0+\varepsilon\alpha_i\mathbf{u}} \text{ and } \pi'_\varepsilon := \sum_{i\in T_-} -\frac{\beta_i}{\rho}\delta_{\mathbf{x}_0+\varepsilon\alpha_i\mathbf{u}}\,. \tag{43}$$

We have that $\#T_+ \leq r(\boldsymbol{\beta})$ and $\#T_- \leq r(\boldsymbol{\beta})$ by definition of $r(\boldsymbol{\beta})$. Since ε is small enough we have that $\pi_\varepsilon, \pi'_\varepsilon \in \mathfrak{S}_{r(\boldsymbol{\beta})}(\Omega)$. Moreover $\mu_\varepsilon = \frac{1}{\rho}(\pi_\varepsilon - \pi'_\varepsilon)$. Hence

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2\|\mu_\varepsilon\|_\kappa^2 = \rho^2 \sum_{i,j=1}^M \beta_i\beta_j\kappa(\mathbf{x}_0 + \varepsilon\alpha_i\mathbf{u}, \mathbf{x}_0 + \varepsilon\alpha_j\mathbf{u}) = \rho^2 \sum_{i,j=1}^M \beta_i\beta_j\kappa_0(\varepsilon(\alpha_i - \alpha_j)\mathbf{u})\,. \tag{44}$$

Since the kernel is k times differentiable at 0, the function $g : t \mapsto \kappa_0(t\mathbf{u})$ is also k times differentiable at 0. A Taylor expansion yields

$$\kappa_0(\varepsilon\mathbf{u}) := g(\varepsilon) = g(0) + \sum_{n=1}^k \frac{g^{(n)}(0)}{n!}\varepsilon^n + o_{\varepsilon\to 0}(\varepsilon^k)\,, \tag{45}$$

hence

$$\begin{aligned}\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 &= \rho^2 \sum_{i,j=1}^M \beta_i\beta_j \left(g(0) + \sum_{n=1}^k \frac{g^{(n)}(0)}{n!}(\alpha_i - \alpha_j)^n\varepsilon^n + o_{\varepsilon\to 0}(\varepsilon^k)\right) \\ &= \rho^2 \sum_{n=1}^k \left(\sum_{i,j=1}^M \beta_i\beta_j(\alpha_i - \alpha_j)^n\right)\varepsilon^k\frac{g^{(n)}(0)}{n!} + o_{\varepsilon\to 0}(\varepsilon^k)\,,\end{aligned} \tag{46}$$

where we used that $\sum_{i,j=1}^M \beta_i\beta_j g(0) = 0$ since $(\sum_{i=1}^M \beta_i)^2 = 0$. With the notations of the Lemma we have

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2 \sum_{n=1}^k c_n(\boldsymbol{\alpha}, \boldsymbol{\beta})\varepsilon^k\frac{g^{(n)}(0)}{n!} + o_{\varepsilon\to 0}(\varepsilon^k)\,. \tag{47}$$

Now, since by assumption we have

$$c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0\,, \tag{48}$$

we get

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2 c_k(\boldsymbol{\alpha}, \boldsymbol{\beta})\varepsilon^k\frac{g^{(k)}(0)}{k!} + o_{\varepsilon\to 0}(\varepsilon^k) = O_{\varepsilon\to 0}(\varepsilon^k) \tag{49}$$

hence $\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa = O_{\varepsilon\to 0}(\varepsilon^{k/2})$. Moreover, defining for $i \in T_+$ $a_i = \beta_i/\rho$ and for $j \in T_-$ $b_j = -\beta_j/\rho$ we have

$$\mathrm{W}_p^p(\pi_\varepsilon, \pi'_\varepsilon) = \min_{\gamma\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i\in T_+, j\in T_-} \|\varepsilon\alpha_i\mathbf{u} - \varepsilon\alpha_j\mathbf{u}\|^p\gamma_{ij} = \varepsilon^p\|\mathbf{u}\|^p \min_{\gamma\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j|^p\gamma_{ij}\,. \tag{50}$$

Therefore

$$\mathrm{W}_p^p(\pi_\varepsilon, \pi'_\varepsilon) \geq \left(\varepsilon\|\mathbf{u}\| \min_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j|\right)^p\,, \tag{51}$$

hence $\mathrm{W}_p(\pi_\varepsilon, \pi'_\varepsilon) \geq \varepsilon\|\mathbf{u}\| \min_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j|$. When $i \neq j$ we have $\alpha_i \neq \alpha_j$ by assumption. Since $T_+ \cap T_- = \emptyset$ we have $\min_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j| > 0$. This discussion proves that, as soon as the condition (48) holds and $\delta > \frac{2}{k}$, we have

$$\sup_{(\pi,\pi')\in\mathfrak{S}} \frac{\mathrm{W}_p(\pi, \pi')}{\|\pi - \pi'\|_\kappa^\delta} \geq \sup_{\varepsilon>0} \frac{\mathrm{W}_p(\pi_\varepsilon, \pi'_\varepsilon)}{\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^\delta} \gtrsim \sup_{\varepsilon>0} \frac{\varepsilon}{\varepsilon^{\delta k/2}} = \sup_{\varepsilon>0} \varepsilon^{1-\delta k/2} = +\infty\,. \tag{52}$$

Consequently, $(\mathfrak{S}, W_p)$ is not (κ, δ)-embeddable when $\delta > \frac{2}{k}$ which concludes the proof by contraposition. ■

The idea now is to find a couple $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ that satisfy the conditions $\sum_{i=1}^{M} \beta_i = 0$ and $c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$. The following lemma show that it is possible to construct such vectors provided that $M = k + 1$.

Lemma 45 *Consider a TI, PSD kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ such that κ_0 is k times differentiable at 0 with $k \in \mathbb{N}^*$. With the same notations $c_s(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $r(\boldsymbol{\beta})$ as in Lemma 44, there exists $\boldsymbol{\alpha} \in \mathbb{R}^{k+1} \setminus \{0\}$ with $\alpha_i \neq \alpha_j$ for $i \neq j$ and $\boldsymbol{\beta} \in \mathbb{R}^{k+1} \setminus \{0\}$ with $\sum_{i=1}^{k+1} \beta_i = 0$ such that*

$$c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0\,. \tag{53}$$

Also if k is odd then $\#T_+(\boldsymbol{\beta}) = \#T_-(\boldsymbol{\beta}) = \frac{k+1}{2}$ and if k is even $\#T_+(\boldsymbol{\beta}) = \frac{k}{2} + 1$ and $\#T_-(\boldsymbol{\beta}) = \frac{k}{2}$. Overall for any $k \in \mathbb{N}^*$ we have $r(\boldsymbol{\beta}) \leq \lfloor \frac{k}{2} \rfloor + 1$.

Proof The condition $c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ writes $\sum_{i,j=1}^{k+1} \beta_i \beta_j (\alpha_i - \alpha_j) = 0$ which is true for any $\boldsymbol{\alpha} \in \mathbb{R}^{k+1}$ when $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ satisfies $\sum_{i=1}^{k+1} \beta_i = 0$. Indeed $\sum_{i,j=1}^{k+1} \beta_i \beta_j (\alpha_i - \alpha_j) = (\sum_{j=1}^{k+1} \beta_j) \sum_{i}^{k+1} \beta_i \alpha_i - (\sum_{i=1}^{k+1} \beta_i) \sum_{j}^{k+1} \beta_j \alpha_j = 0$. The condition $c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ writes $\sum_{i,j=1}^{k+1} \beta_i \beta_j (\alpha_i - \alpha_j)^2 = 0$. However $\sum_{i,j=1}^{k+1} \beta_i \beta_j (\alpha_i - \alpha_j)^2 = \sum_{i,j=1}^{k+1} \beta_i \beta_j (\alpha_i^2 + \alpha_j^2 - 2\alpha_i \alpha_j)$. The term $\sum_{i,j=1}^{k+1} \beta_i \beta_j \alpha_i \alpha_j$ vanishes as soon as $\sum_{i=1}^{k+1} \beta_i \alpha_i = 0$. The other terms $\sum_{i,j=1}^{k+1} \beta_i \beta_j \alpha_i^2$ and $\sum_{i,j=1}^{k+1} \beta_j \beta_i \alpha_j^2$ as soon as $\sum_{i=1}^{k+1} \beta_i = 0$. With an immediate recurrence by using the Binomial formula we see that $c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ as soon as

$$\sum_{i=1}^{k+1} \beta_i = \sum_{i=1}^{k+1} \beta_i \alpha_i = \sum_{i=1}^{k+1} \beta_i \alpha_i^2 = \cdots = \sum_{i=1}^{k+1} \beta_i \alpha_i^{k-1} = 0\,. \tag{54}$$

Define $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ by for all $1 \leq i \leq k+1, \beta_i = (-1)^{i-1} \binom{k}{i-1}$ and $\boldsymbol{\alpha} \in \mathbb{R}^{k+1}$ by $\alpha_i = i$. Then the α_i's are pairwise distinct and

$$0 = \sum_{i=0}^{k} (-1)^i \binom{k}{i} = \sum_{i=1}^{k+1} (-1)^{i-1} \binom{k}{i-1} = \sum_{i=1}^{k+1} \beta_i\,. \tag{55}$$

Then for any $1 \leq s \leq k-1$ we have

$$\sum_{i=1}^{k+1} \beta_i \alpha_i^s = \sum_{i=1}^{k+1} (-1)^{i-1} \binom{k}{i-1} i^s = \sum_{i=0}^{k} (-1)^i \binom{k}{i} (i+1)^s = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left(\sum_{l=0}^{s} \binom{s}{l} i^l \right)\,. \tag{56}$$

Consquently

$$\sum_{i=1}^{k+1} \beta_i \alpha_i^s = \sum_{l=0}^{s} \binom{s}{l} \left(\sum_{i=0}^{k} (-1)^i \binom{k}{i} i^l \right)\,. \tag{57}$$

But for $0 \leq l \leq s$ we have

$$\sum_{i=0}^{k} (-1)^i \binom{k}{i} i^l = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{k-i} (k-i)^l = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} (k-i)^l = (-1)^k \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^l\,, \tag{58}$$

so $\sum_{i=0}^{k} (-1)^i \binom{k}{i} i^l = (-1)^k k! S_2(l, k)$ where $S_2(l, k)$ is the Stirling number of the second kind which is zero as soon as $l < k$. Since $l \leq s \leq k-1 < k$ by hypothesis we have that $\sum_{i=0}^{k} (-1)^i \binom{k}{i} i^l = 0$

and thus $\sum_{i=1}^{k+1} \beta_i \alpha_i^s = 0$ for all $1 \leq s \leq k-1$ and $\sum_{i=1}^{k+1} \beta_i = 0$. So this implies that $c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$. For such $\boldsymbol{\beta}$ we have that $\#T_+(\boldsymbol{\beta}) = \#T_-(\boldsymbol{\beta}) = \frac{k+1}{2}$ for k odd. If k is even then $\#T_+(\boldsymbol{\beta}) = \frac{k}{2} + 1$ and $\#T_-(\boldsymbol{\beta}) = \frac{k}{2}$. ■

With this results we can now prove Theorem 11.

Proof [Proof of Theorem 11] Define $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as in Lemma 45. Then we have $c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$ and $r(\boldsymbol{\beta}) \leq \lfloor \frac{k}{2} \rfloor + 1$ which proves the theorem by using Lemma 44 with $M = k + 1$.

■

A.6 Proof of Proposition 13

Proposition 13 is an immediate corollary of the following variation of its statement:

Proposition 46 *Consider any $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ having densities f, g with respect to the Lebesgue measure, i.e. $\pi = f\mathrm{d}\mathbf{x}, \pi' = g\mathrm{d}\mathbf{x}$. Denote $V_d = \pi^{d/2}/\Gamma(d/2+1)$ the volume of the unit d -dimensional unit sphere.*

(i) Consider $1 \leq p < r$. If $\mathrm{M}_r[\pi], \mathrm{M}_r[\pi']$ are finite then

$$\mathrm{W}_p(\pi, \pi') \leq c_{d,p,r} (\mathrm{M}_r^r[\pi] + \mathrm{M}_r^r[\pi'])^{\frac{d+2p}{p(d+2r)}} \left(\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \right)^{\frac{r-p}{(d+2r)p}},$$

where $0 < c_{d,p,r} \leq 2(\max\{V_d, 1\})^{\frac{1}{2p}}$.

(ii) Consider $1 \leq p < r$. If $\max\{\mathrm{M}_r[\pi], \mathrm{M}_r[\pi']\} \leq M$ where $M > 0$ then

$$\mathrm{W}_p(\pi, \pi') \leq 2c_{d,p,r} M^{\frac{r(d+2p)}{p(d+2r)}} \left(\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \right)^{\frac{r-p}{(d+2r)p}},$$

(iii) If π, π' are supported in some Euclidean ball centered at 0 of radius $M > 0$ then, for any $p \in [1, +\infty)$,

$$\mathrm{W}_p(\pi, \pi') \leq 2^{\frac{p-1}{p}} V_d^{\frac{1}{2p}} M^{\frac{2p+d}{2p}} \left(\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \right)^{\frac{1}{2p}} \tag{59}$$

Proof As a preliminary observe that by Villani (2008, Theorem 6.15) the Wasserstein distance is bounded by a weighted Total Variation distance:

$$\mathrm{W}_p^p(\pi, \pi') \leq 2^{p-1} \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^p \, \mathrm{d}|\pi - \pi'|(\mathbf{x}) = 2^{p-1} \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^p \, |f(\mathbf{x}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{x}.$$

Given any $R > 0$, write $\int_{\mathbb{R}^d} \|\mathbf{x}\|_2^p \, |f(\mathbf{x}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{x} = \int_{\|\mathbf{x}\|_2 \leq R} \|\mathbf{x}\|_2^p \, |f(\mathbf{x}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{x} + \int_{\|\mathbf{x}\|_2 > R} \|\mathbf{x}\|_2^p \, |f(\mathbf{x}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{x}$. By Cauchy-Schwarz inequality the first term of this decomposition is bounded as

$$\int_{\|\mathbf{x}\|_2 \leq R} \|\mathbf{x}\|_2^p \, |f(\mathbf{x}) - g(\mathbf{x})| \, \mathrm{d}\mathbf{x} \leq \sqrt{\int_{\|\mathbf{x}\|_2 \leq R} \|\mathbf{x}\|_2^{2p} \, \mathrm{d}\mathbf{x}} \sqrt{\int_{\|\mathbf{x}\|_2 \leq R} |f(\mathbf{x}) - g(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}} \leq C \|f - g\|_{L_2(\mathbb{R}^d)}$$

where

$$C \coloneqq \sqrt{\int_{\|\mathbf{x}\|_2 \leq R} \|\mathbf{x}\|_2^{2p} \, \mathrm{d}\mathbf{x}} = \sqrt{\int_{\|\mathbf{u}\|_2 \leq 1} \|R\mathbf{u}\|_2^{2p} \, R^d \mathrm{d}\mathbf{u}} = \sqrt{R^{2p+d} \int_{\|\mathbf{u}\|_2 \leq 1} \|\mathbf{u}\|_2^{2p} \, \mathrm{d}\mathbf{u}} \leq R^{\frac{2p+d}{2}} \sqrt{V_d}.$$

The second term is bounded as

$$\int_{\|\mathbf{x}\|_2>R} \|\mathbf{x}\|_2^p \, |f(\mathbf{x})-g(\mathbf{x})| \, \mathrm{d}\mathbf{x} = \int_{\|\mathbf{x}\|_2>R} \|\mathbf{x}\|_2^{p-r}\|\mathbf{x}\|_2^r \, |f(\mathbf{x})-g(\mathbf{x})| \, \mathrm{d}\mathbf{x}$$
$$\overset{r>p}{\leq} R^{p-r}\int_{\|\mathbf{x}\|_2>R} \|\mathbf{x}\|_2^r \, |f(\mathbf{x})-g(\mathbf{x})| \, \mathrm{d}\mathbf{x} \leq R^{p-r}(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])$$

hence

$$\forall p\in[1,r),\ \mathrm{W}_p^p(\pi,\pi') \leq 2^{p-1}\Big(V_d^{1/2}\|f-g\|_{L_2(\mathbb{R}^d)}R^{\frac{2p+d}{2}} + (\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])R^{p-r}\Big). \tag{60}$$

We now have the ingredients to prove the three points.

For the first point, with $R := \left(\frac{\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi']}{V_d^{1/2}}\right)^{\frac{2}{d+2r}}\|f-g\|_{L_2(\mathbb{R}^d)}^{-\frac{2}{d+2r}}$ we have $V_d^{1/2}\|f-g\|_{L_2(\mathbb{R}^d)}R^{\frac{2p+d}{2}} = (\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])R^{p-r}$ hence by (60) we have for each $p\in[1,r)$

$$\mathrm{W}_p^p(\pi,\pi') \leq 2^p(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])R^{p-r} = 2^p(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])\left(\frac{\mathrm{M}_r[\pi]+\mathrm{M}_r[\pi']}{V_d^{1/2}}\right)^{\frac{2(p-r)}{d+2r}}\|f-g\|_{L_2(\mathbb{R}^d)}^{\frac{2(r-p)}{d+2r}}.$$

Taking the p-th root yields the first claim once we check that $2^p(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])\left(\frac{\mathrm{M}_r[\pi]+\mathrm{M}_r[\pi']}{V_d^{1/2}}\right)^{\frac{2(p-r)}{d+2r}} \leq c_{d,p,r}^p(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])^{(d+2p)/(d+2r)}$ where $0<c_{d,p,r}\leq 2(\max\{V_d,1\})^{\frac{1}{2p}}$. Since

$$2^p(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])\left(\frac{\mathrm{M}_r[\pi]+\mathrm{M}_r[\pi']}{V_d^{1/2}}\right)^{\frac{2(p-r)}{d+2r}} = 2^p V_d^{\frac{r-p}{d+2r}}(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])^{\frac{d+2p}{d+2r}}$$

it is enough to bound $c_{d,p,r} \coloneqq 2V_d^{\frac{r-p}{p(d+2r)}}$.

Indeed, since the function $r\mapsto \frac{r-p}{d+2r} = \frac{1}{2}-\frac{d+2p}{2(d+2r)}$ is monotonically increasing and $p<r<\infty$, we have $0<\frac{r-p}{d+2r}<\lim_{r'\to+\infty}\frac{r'-p}{d+2r'}=\frac{1}{2}$, hence we have as claimed

$$c_{d,p,r} = 2V_d^{\frac{r-p}{p(d+2r)}} \leq 2(\max\{V_d,1\})^{\frac{r-p}{p(d+2r)}} \leq 2(\max\{V_d,1\})^{\frac{1}{2p}}.$$

The second point is an immediate consequence of the first one. Since $1\leq p<r$ we have $\frac{d+2p}{p(d+2r)}\leq\frac{1}{p}\leq 1$, hence using that $\max\{\mathrm{M}_r[\pi],\mathrm{M}_r[\pi']\}\leq M$ we get

$$(\mathrm{M}_r^r[\pi]+\mathrm{M}_r^r[\pi'])^{\frac{d+2p}{p(d+2r)}} \leq 2^{\frac{d+2p}{p(d+2r)}}M^{\frac{r(d+2p)}{p(d+2r)}} \leq 2M^{\frac{r(d+2p)}{p(d+2r)}}.$$

For the last point we have $\forall r>1, \max\{\mathrm{M}_r[\pi],\mathrm{M}_r[\pi']\}\leq M$ and thus (60) gives for any choice of $R>0$:

$$\forall r>1,\forall p\in[1,r),\ \mathrm{W}_p^p(\pi,\pi') \leq 2^{p-1}\Big(V_d^{1/2}\|f-g\|_{L_2(\mathbb{R}^d)}R^{\frac{2p+d}{2}} + 2\Big(\frac{M}{R}\Big)^r R^p\Big). \tag{61}$$

Consider any $R>M$. We can take the limit as $r\to+\infty$ in (61) which gives

$$\forall R>M,\forall p\in[1,+\infty),\ \mathrm{W}_p^p(\pi,\pi') \leq 2^{p-1}\Big(V_d^{1/2}\|f-g\|_{L_2(\mathbb{R}^d)}R^{\frac{2p+d}{2}}\Big).$$

since $\lim_{r\to+\infty}(\frac{M}{R})^r=0$. Since this is true for any $R>M$ we can conclude that

$$\forall p\in[1,+\infty),\ \mathrm{W}_p^p(\pi,\pi') \leq 2^{p-1}V_d^{1/2}M^{\frac{2p+d}{2}}\left(\int_{\mathbb{R}^d}|f(\mathbf{x})-g(\mathbf{x})|^2\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}}.$$

Taking the p-th root yields the conclusion.

■

A.7 Proof of Theorem 14 and 15

We first prove the following result:

Theorem 14 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ such that $\kappa_0 \in L_1(\mathbb{R}^d)$, $\widehat{\kappa_0}(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$. For $B, M, r \geq 0$, denote*

$$\mathfrak{S}_{B,M,r,\kappa} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \ \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{\mathcal{H}_\kappa} \leq B \text{ and } \ \mathrm{M}_r[\pi] \leq M\right\} \subset \mathcal{P}_r(\mathbb{R}^d)\,. \tag{14}$$

If $r > 1$ then for each $1 \leq p < r$ we have

$$\forall \pi, \pi' \in \mathfrak{S}_{B,M,r,\kappa},\ \mathrm{W}_p(\pi,\pi') \leq C' \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,,$$

where $C' = 8(\max\{V_d, 1\})^{\frac{1}{2p}} B^{\frac{r-p}{(d+2r)p}} M^{\frac{(d+2p)r}{(d+2r)p}}$.

Proof Take any $\pi, \pi' \in \mathfrak{S}_{B,M,r,\kappa}$ and recall that this implies notably that $\mathrm{M}_r[\pi] \leq M$ (and similarly for π'). By Proposition 46 we have, with $C_1 := 2c_{d,p,r} M^{\frac{r(d+2p)}{p(d+2r)}}$:

$$\mathrm{W}_p(\pi,\pi') \leq C_1 \left(\int |f(\mathbf{x}) - g(\mathbf{x})|^2 \mathrm{d}\mathbf{x}\right)^{\frac{r-p}{p(d+2r)}}\,. \tag{62}$$

Since $\kappa_0 \in L_1(\mathbb{R}^d)$ it has a Fourier transform $\widehat{\kappa_0}$, which is non-negative by Bochner's theorem. Consequently:

$$\begin{aligned}
\mathrm{W}_p(\pi,\pi') &\overset{\star}{\leq} C_1 \left((2\pi)^{-d} \int |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{p(d+2r)}} = (2\pi)^{\frac{-d(r-p)}{p(d+2r)}} C_1 \left(\int |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{p(d+2r)}} \\
&= (2\pi)^{\frac{-d(r-p)}{p(d+2r)}} C_1 \left(\int \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|}{\sqrt{\widehat{\kappa_0}(\boldsymbol{\omega})}} \sqrt{\widehat{\kappa_0}(\boldsymbol{\omega})} |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})| \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{p(d+2r)}} \\
&\overset{\star\star}{\leq} (2\pi)^{\frac{-d(r-p)}{p(d+2r)}} C_1 \left(\int \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{2p(d+2r)}} \left(\int \widehat{\kappa_0}(\boldsymbol{\omega}) |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{2p(d+2r)}} \\
&\overset{\star\star\star}{\leq} (2\pi)^{\frac{-d(r-p)}{p(d+2r)}} C_1 \left(\int \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{2p(d+2r)}} (2\pi)^{\frac{d(r-p)}{2p(d+2r)}} \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}} \\
&= (2\pi)^{\frac{-d(r-p)}{2p(d+2r)}} C_1 \left(\int \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega}\right)^{\frac{r-p}{2p(d+2r)}} \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}} \\
&= C_1 \|f - g\|_{\mathcal{H}_\kappa}^{\frac{r-p}{p(d+2r)}} \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,,
\end{aligned} \tag{63}$$

where in $(\star)$ we used the Plancherel formula, in $(\star\star)$ we used the Cauchy–Schwarz inequality and in $(\star\star\star)$ we relied on Lemma 48 whose proof is postponed below. In the last step we used $(\int \frac{|\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega})^{1/2} = (2\pi)^{d/2} \|f - g\|_{\mathcal{H}_\kappa}$. We used Theorem 10.12 in Wendland 2004 where we adapted the conventions on the Fourier transform. We can apply this theorem since κ_0 is continuous (by hypothesis), and its Fourier transform $\widehat{\kappa_0} > 0$ thus κ_0 is positive definite (Wendland, 2004, Corollary 6.9). Finally $\max\{\|f\|_{\mathcal{H}_\kappa}, \|g\|_{\mathcal{H}_\kappa}\} \leq B$ by hypothesis. Thus $\|f - g\|_{\mathcal{H}_\kappa}^{\frac{r-p}{(d+2r)p}} \leq (B+B)^{\frac{r-p}{(d+2r)p}} \leq 2B^{\frac{r-p}{(d+2r)p}}$ since $\frac{r-p}{(d+2r)p} \leq 1$. This concludes the proof with $C := 2B^{\frac{r-p}{(d+2r)p}} C_1 = 4c_{d,p,r} B^{\frac{r-p}{(d+2r)p}} M^{\frac{r(d+2p)}{p(d+2r)}}$.

■

As a consequence we have the theorem:

Theorem 15 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ such that $\kappa_0 \in L_1(\mathbb{R}^d)$, $\widehat{\kappa_0}(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$, and assume there is $s_\kappa > 0$ such that*

$$\frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})} = O(\|\boldsymbol{\omega}\|_2^{s_\kappa}) \text{ as } \|\boldsymbol{\omega}\|_2 \to +\infty\,. \tag{15}$$

For $r, B, M, s \geq 0$, denote

$$\mathfrak{S}_{B,M,r,s} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) :\ \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{H^s(\mathbb{R}^d)} \leq B \text{ and } \ \mathrm{M}_r[\pi] \leq M\right\} \subset \mathcal{P}_r(\mathbb{R}^d)\,. \tag{16}$$

If $s \geq s_\kappa/2$ and $r > 1$ then for each $1 \leq p < r$ there exists $C = C(B, M, r, s, d, \kappa, p) > 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}_{B,M,r,s},\ \mathrm{W}_p(\pi,\pi') \leq C\|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,.$$

Proof This is a direct consequence of Theorem 14 once we establish that, under the assumptions on κ, we have $\mathfrak{S}_{B,M,r,s} \subseteq \mathfrak{S}_{CB,M,r,\kappa}$ where $C = C(d,s,\kappa)$ is the constant from Lemma 47 below. Indeed, consider $\pi = f\mathrm{d}\mathbf{x} \in \mathfrak{S}_{B,M,r,s}$. By hypothesis we have $\mathrm{M}_r[\pi] \leq M$ and $\|f\|_{H^s(\mathbb{R}^d)} \leq B$. With the hypothesis on the kernel κ we can use Lemma 47 below to prove that there is a constant $C = C(d,s,\kappa)$ such that $\|f\|_{\mathcal{H}_\kappa} \leq C\|f\|_{H^s(\mathbb{R}^d)} \leq CB$, which shows that $\pi \in \mathfrak{S}_{CB,M,r,\kappa}$ as claimed. Thus, by Theorem 14, with $c_{d,p,r}$ the constant defined in Proposition 13 we have:

$$\forall \pi, \pi' \in \mathfrak{S}_{B,M,r,s}, \mathrm{W}_p(\pi,\pi') \leq 4c_{d,p,r}2(C(d,s,\kappa)B)^{\frac{r-p}{(d+2r)p}} M^{\frac{(d+2p)r}{(d+2r)p}} \|\pi - \pi'\|_\kappa^{\frac{r-p}{p(d+2r)}}\,, \tag{64}$$

which concludes the proof. ■

Lemma 47 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ with $\kappa_0 \in L_1(\mathbb{R}^d)$ such that $\widehat{\kappa_0}(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$ and $\frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})} = O(\|\boldsymbol{\omega}\|_2^{s_\kappa})$ as $\|\boldsymbol{\omega}\|_2 \to +\infty$ for some $s_\kappa \in \mathbb{R}_+$. For any $s \geq s_\kappa/2$, there exists a constant $C = C(d,s,\kappa) > 0$ such that for every $f \in H^s(\mathbb{R}^d)$ we have*

$$\|f\|_{\mathcal{H}_\kappa} \leq C\|f\|_{H^s(\mathbb{R}^d)}\,. \tag{65}$$

Proof Given any $R > 0$ we write $\int_{\mathbb{R}^d} \frac{|\hat{f}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega} = \int_{\|\boldsymbol{\omega}\|_2 \leq R} \frac{|\hat{f}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega} + \int_{\|\boldsymbol{\omega}\|_2 > R} \frac{|\hat{f}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega}$ and use the shorthand $I_{\|\boldsymbol{\omega}\|_2 \leq R}$ and $I_{\|\boldsymbol{\omega}\|_2 > R}$ for the two terms. Since $\kappa_0 \in L_1(\mathbb{R}^d)$ the Fourier transform $\widehat{\kappa_0}$ is continuous. It is also positive and thus and the term $I_{\|\boldsymbol{\omega}\|_2 < R}$ can be bounded as

$$I_{\|\boldsymbol{\omega}\|_2 \leq R} \leq \left(\sup_{\|\boldsymbol{\omega}\|_2 \leq R} \widehat{\kappa_0}(\boldsymbol{\omega})^{-1}\right) \int_{\|\boldsymbol{\omega}\|_2 \leq R} |\hat{f}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \leq \left(\sup_{\|\boldsymbol{\omega}\|_2 \leq R} \widehat{\kappa}_0(\boldsymbol{\omega})^{-1}\right) \|f\|^2_{H^s(\mathbb{R}^d)}\,. \tag{66}$$

Now consider $I_{\|\boldsymbol{\omega}\|_2 > R}$ and take $s \geq \frac{s_\kappa}{2}$. We have:

$$\begin{aligned}
\int_{\|\boldsymbol{\omega}\|_2 > R} |\hat{f}(\boldsymbol{\omega})|^2 \frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega} &= \int_{\|\boldsymbol{\omega}\|_2 > R} (1 + \|\boldsymbol{\omega}\|_2^2)^s |\hat{f}(\boldsymbol{\omega})|^2 (1 + \|\boldsymbol{\omega}\|_2^2)^{-s} \frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega} \\
&\leq \sup_{\|\boldsymbol{\omega}\|_2 > R} \left(\frac{(1 + \|\boldsymbol{\omega}\|_2^2)^{-s}}{\widehat{\kappa_0}(\boldsymbol{\omega})}\right) \int_{\|\boldsymbol{\omega}\|_2 > R} (1 + \|\boldsymbol{\omega}\|_2^2)^s |\hat{f}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\
&\leq \sup_{\|\boldsymbol{\omega}\|_2 > R} \left(\frac{(1 + \|\boldsymbol{\omega}\|_2^2)^{-s}}{\widehat{\kappa_0}(\boldsymbol{\omega})}\right) \|f\|^2_{H^s(\mathbb{R}^d)}
\end{aligned} \tag{67}$$

By hypothesis $\frac{\|\boldsymbol{\omega}\|_2^{-2s}}{\widehat{\kappa_0}(\boldsymbol{\omega})} = O_{\|\boldsymbol{\omega}\|_2 \to +\infty}(\frac{1}{\|\boldsymbol{\omega}\|_2^{2s-s_\kappa}})$. Since $s \geq \frac{s_\kappa}{2}$ we have $2s - s_\kappa \geq 0$ thus the quantity $\sup_{\|\boldsymbol{\omega}\|_2 > R} \left(\frac{(1+\|\boldsymbol{\omega}\|_2^2)^{-s}}{\widehat{\kappa_0}(\boldsymbol{\omega})} \right)$ is finite. The previous reasoning gives, for any $R > 0$,

$$\|f\|_{\mathcal{H}_\kappa}^2 = (2\pi)^{-d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega} \leq (2\pi)^{-d} \left(\sup_{\|\boldsymbol{\omega}\|_2 \leq R} \frac{1}{\widehat{\kappa_0}(\boldsymbol{\omega})} + \sup_{\|\boldsymbol{\omega}\|_2 > R} \frac{(1+\|\boldsymbol{\omega}\|_2^2)^{-s}}{\widehat{\kappa_0}(\boldsymbol{\omega})} \right) \|f\|_{H^s(\mathbb{R}^d)}^2 . \tag{68}$$

The infimum over $R > 0$ yields a constant $C(d, s, \kappa)$ such that $\|f\|_{\mathcal{H}_\kappa} \leq C(d, s, \kappa) \|f\|_{H^s(\mathbb{R}^d)}$. ■

Lemma 48 *Let $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d \times \mathbb{R}^d$ where $\kappa_0 \in L_1(\mathbb{R}^d)$. Then for $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ we have the formula*

$$\|\pi - \pi'\|_\kappa^2 = (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} . \tag{69}$$

In particular when π, π' have densities f, g with respect to the Lebesgue measure we have

$$\|\pi - \pi'\|_\kappa^2 = (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} . \tag{70}$$

Proof This result can be found in Sriperumbudur et al. (2010) but we rewrite the proof for completeness. Since κ_0 is a continuous PSD function and $\kappa_0 \in L_1(\mathbb{R}^d)$ then by Bochner's theorem $\widehat{\kappa_0} \geq 0$. So κ_0 is even (κ is symmetric), integrable, continuous (in particular at 0) and has nonnegative Fourier transform so $\widehat{\kappa_0} \in L_1(\mathbb{R}^d)$ (Stein and Weiss, 2016). Then by Fourier inversion theorem

$$\forall \mathbf{x} \in \mathbb{R}^d, \ \kappa_0(\mathbf{x}) = (2\pi)^{-d} \int e^{i\boldsymbol{\omega}^\top \mathbf{x}} \widehat{\kappa_0}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} . \tag{71}$$

In the following we define the measure Λ by $\mathrm{d}\Lambda(\boldsymbol{\omega}) := (2\pi)^{-d} \widehat{\kappa_0}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega}$ (which is a non-negative finite measure thanks to Bochner's theorem). We have:

$$\begin{aligned}
\|\pi - \pi'\|_\kappa^2 &= \int \int \kappa_0(\mathbf{x} - \mathbf{y}) \mathrm{d}(\pi - \pi')(\mathbf{x}) \mathrm{d}(\pi - \pi')(\mathbf{y}) \\
&\overset{\star}{=} \int \int \int e^{i\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{y})} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(\pi - \pi')(\mathbf{x}) \mathrm{d}(\pi - \pi')(\mathbf{y}) \\
&= \int \left(\int e^{i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}(\pi - \pi')(\mathbf{x}) \right) \left(\int e^{-i\boldsymbol{\omega}^\top \mathbf{y}} \mathrm{d}(\pi - \pi')(\mathbf{y}) \right) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= \int (\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})) \overline{(\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega}))} \mathrm{d}\Lambda(\boldsymbol{\omega}) = \int |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} ,
\end{aligned} \tag{72}$$

where in $(\star)$ we used (71) and Fubini theorem. ■

A.8 The Compactly Supported Case

We will prove the following result:

Lemma 49 *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI, PSD kernel on $\mathbb{R}^d$ with $\kappa_0 \in L_1(\mathbb{R}^d)$ such that $\widehat{\kappa}_0(\boldsymbol{\omega}) > 0$ for every $\boldsymbol{\omega}$ and $\frac{1}{\widehat{\kappa}_0(\boldsymbol{\omega})} = O(\|\boldsymbol{\omega}\|_2^{s_\kappa})$ as $\|\boldsymbol{\omega}\|_2 \to +\infty$ for some $s_\kappa \in \mathbb{R}_+$. Consider $0 < M, B < +\infty$, $s \geq s_\kappa/2$ and the following model set*

$$\mathfrak{S}_{B,M,s} := \left\{\pi \in \mathcal{P}(\mathbb{R}^d) : \ \pi = f\mathrm{d}\mathbf{x},\ \|f\|_{H^s(\mathbb{R}^d)} \leq B \text{ and } \ \mathrm{supp}(\pi) \subseteq B(0,M)\right\}, \tag{73}$$

where $B(0,M)$ is the Euclidean ball centered at zero with radius M. For any $p \in [1,+\infty)$, there exists a constant $C = C(d,p,M,B,\kappa,s) > 0$ such that

$$\forall \pi,\pi' \in \mathfrak{S}_{B,M,s}, \mathrm{W}_p(\pi,\pi') \leq C\|\pi-\pi'\|_\kappa^{\frac{1}{2p}}. \tag{74}$$

Proof By the third point of Proposition 46 there is a constant $C = C(d,p,M) > 0$ such that

$$\mathrm{W}_p(\pi,\pi') \leq C\left(\int_{\mathbb{R}^d} |f(\mathbf{x})-g(\mathbf{x})|^2\mathrm{d}\mathbf{x}\right)^{\frac{1}{2p}} \tag{75}$$

for every $\pi,\pi' \in \mathfrak{S}_{B,M,s}$. Then, with the same strategy as in the proof of Theorem 15 we have

$$\mathrm{W}_p(\pi,\pi') \leq C_1\left(\int \frac{|\widehat{f}(\boldsymbol{\omega})-\widehat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega}\right)^{\frac{1}{4p}} \|\pi-\pi'\|_\kappa^{\frac{1}{2p}}. \tag{76}$$

for some constant $C_1 > 0$ which depends on d,p,M. By Lemma 47 there exists a constant $C_2 = C_2(\kappa,s,B,d)$ such that $\int \frac{|\widehat{f}(\boldsymbol{\omega})-\widehat{g}(\boldsymbol{\omega})|^2}{\widehat{\kappa_0}(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega} \leq C_2$. This concludes the proof. ■

A.9 Proof of Lemma 22, Proposition 23 and Theorem 24

Lemma 22 *Let α be a regularizer and $\kappa_0 := \alpha * \alpha$. Then $\kappa_0 \in L_1(\mathbb{R}^d)$ is even, bounded, continuous and has non-negative Fourier transform. Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) := \kappa_0(\mathbf{x}-\mathbf{y})$. Then κ defines a TI, PSD kernel. Moreover, for $\pi,\pi' \in \mathcal{P}(\mathbb{R}^d)$,*

$$\|\pi-\pi'\|_\kappa = \|\alpha*\pi - \alpha*\pi'\|_{L_2(\mathbb{R}^d)}.$$

Proof We first prove that the kernel in this proposition defines a TI, PSD kernel. It is clearly translation invariant by definition and symmetric since the convolution of even functions is even thus κ_0 is even. Also κ_0 is continuous and bounded since α is continuous and bounded. Since α is even its Fourier transform is real-valued hence $\widehat{\kappa}_0 = \hat{\alpha}^2 = |\hat{\alpha}|^2 \geq 0$ so the Fourier transform of κ_0 is non negative. Finally $\kappa_0 \in L_1(\mathbb{R}^d)$ as the convolution of two integrable functions. Using Bochner's theorem (see Theorem 1) shows that the kernel κ is a TI, PSD kernel. Moreover:

$$\|\alpha*\pi - \alpha*\pi\|_{L_2(\mathbb{R}^d)}^2 = \int |\alpha*\pi(\mathbf{x}) - \alpha*\pi'(\mathbf{x})|^2\mathrm{d}\mathbf{x} \overset{\star}{=} (2\pi)^{-d}\int |\widehat{\alpha*\pi}(\boldsymbol{\omega}) - \widehat{\alpha*\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}, \tag{77}$$

where in $(\star)$ we used Plancherel formula which is possible since $\alpha*\pi \in L_2(\mathbb{R}^d)$ because $\alpha \in L_2(\mathbb{R}^d)$ (same for $\alpha*\pi'$). So using that $\widehat{\alpha*\pi} = \widehat{\alpha}\times\widehat{\pi}$ (α is a probability density function and π a probability distribution):

$$\|\alpha*\pi-\alpha*\pi\|_{L_2(\mathbb{R}^d)}^2 = (2\pi)^{-d}\int |\widehat{\alpha}(\boldsymbol{\omega})\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\alpha}(\boldsymbol{\omega})\widehat{\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega} = (2\pi)^{-d}\int |\widehat{\alpha}(\boldsymbol{\omega})|^2|\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}. \tag{78}$$

Finally, since $\widehat{\kappa_0} = |\widehat{\alpha}|^2$ we get

$$\|\alpha*\pi-\alpha*\pi\|_{L_2(\mathbb{R}^d)}^2 = (2\pi)^{-d}\int \widehat{\kappa_0}(\boldsymbol{\omega})|\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega} \overset{\star\star}{=} \|\pi-\pi'\|_\kappa^2, \tag{79}$$

where in $(\star\star)$ we used Lemma 48. This concludes the proof. ■

Proposition 23 *Let $r > 1$. Consider a regularizer α with r-finite moments and the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 := \alpha * \alpha$. It defines a TI, PSD kernel by Lemma 22. Moreover, for any $\pi, \pi' \in \mathcal{P}_r(\mathbb{R}^d)$ and $1 \leq p < r$, W_p defined with the Euclidean norm on $\mathbb{R}^d$ satisfies*

$$\mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq C_{d,r,p} \left(\mathbb{E}_{\mathbf{x} \sim \pi_\alpha}[\|\mathbf{x}\|_2^r] + \mathbb{E}_{\mathbf{y} \sim \pi'_\alpha}[\|\mathbf{y}\|_2^r] \right)^{\frac{2p+d}{(d+2r)p}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}},$$

for some constant $C_{d,r,p} > 0$.

Proof In order to prove the proposition we will apply the first point of Proposition 46 with π_α and π'_α that admit the densities $f = \alpha * \pi$ and $g = \alpha * \pi'$ and thus the term $\|f - g\|_{L_2(\mathbb{R}^d)}$ in Proposition 46 becomes $\|f - g\|_{L_2} = \|\alpha * \pi - \alpha * \pi'\|_{L_2(\mathbb{R}^d)}$. To apply Proposition 46 we need to show that π_α, π'_α have r-finite moments which will be true by using that π, π' and α have r-finite moments. Indeed

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim \pi_\alpha} \|\mathbf{x}\|_2^r &= \int \|\mathbf{x}\|_2^r (\alpha * \pi)(\mathbf{x}) \mathrm{d}\mathbf{x} = \int \|\mathbf{x}\|_2^r \left(\int \alpha(\mathbf{x} - \mathbf{y}) \mathrm{d}\pi(\mathbf{y}) \right) \mathrm{d}\mathbf{x} \\ &\overset{\star}{=} \int \int \|\mathbf{x}\|_2^r \alpha(\mathbf{x} - \mathbf{y}) \mathrm{d}\mathbf{x} \mathrm{d}\pi(\mathbf{y}) = \int \left(\int \|\mathbf{x}\|_2^r \alpha(\mathbf{x} - \mathbf{y}) \mathrm{d}\mathbf{x} \right) \mathrm{d}\pi(\mathbf{y}), \end{aligned} \tag{80}$$

where in $(\star)$ we used the Fubini theorem (α is non-negative). Moreover, for any $\mathbf{y} \in \mathbb{R}^d$,

$$\int \|\mathbf{x}\|_2^r \alpha(\mathbf{x} - \mathbf{y}) \mathrm{d}\mathbf{x} = \int \|\mathbf{y} + \mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \leq 2^{r-1} \left(\|\mathbf{y}\|_2^r \int \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right), \tag{81}$$

where in the last inequality we used $\|\mathbf{z} + \mathbf{y}\|_2^r \leq 2^{r-1}(\|\mathbf{z}\|_2^r + \|\mathbf{y}\|_2^r)$. Moreover since $\int \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} = 1$ we have:

$$\mathbb{E}_{\mathbf{x} \sim \pi_\alpha} \|\mathbf{x}\|_2^r \leq 2^{r-1} \left(\int \|\mathbf{y}\|_2^r \mathrm{d}\pi(\mathbf{y}) + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right) < +\infty \tag{82}$$

So by using the first point of Proposition 46 we have

$$\mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq C_{d,p,r} \left(\mathbb{E}_{\mathbf{x} \sim \pi_\alpha} \|\mathbf{x}\|_2^r + \mathbb{E}_{\mathbf{y} \sim \pi'_\alpha} \|\mathbf{y}\|_2^r \right)^{\frac{2p+d}{(d+2r)p}} \|\alpha * \pi - \alpha * \pi'\|_{L_2(\mathbb{R}^d)}^{\frac{2(r-p)}{(d+2r)p}}, \tag{83}$$

for some constant $C_{d,p,r} > 0$. Finally, to relate the term $\|\alpha * \pi - \alpha * \pi'\|_{L_2(\mathbb{R}^d)}$ with the MMD we use the Lemma 22.

■

Finally we can prove the following theorem:

Theorem 24 *Let $r > 1$. Consider a regularizer α with r-bounded moments. Consider the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 := \alpha * \alpha$. It defines a TI, PSD kernel by Lemma 22. We consider the model set*

$$\mathfrak{S}_M := \{\pi \in \mathcal{P}(\mathbb{R}^d) : \mathrm{M}_r[\pi] \leq M\} \subset \mathcal{P}_r(\mathbb{R}^d).$$

Then for any $1 \leq p < r$ there exists a constant $C' = C'_{d,r,p} > 0$ such that

$$\forall \pi, \pi' \in \mathfrak{S}, \mathrm{W}_p(\pi, \pi') \leq C' \left(M^r + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{\frac{2p+d}{p(d+2r)}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}} + 2 \left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{1/p}.$$

Proof With the notations of the theorem we have, by Proposition 23,

$$\mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq C_{d,r,p} \left(\mathbb{E}_{\mathbf{x}\sim\pi_\alpha}[\|\mathbf{x}\|_2^r] + \mathbb{E}_{\mathbf{y}\sim\pi'_\alpha}[\|\mathbf{y}\|_2^r] \right)^{\frac{2p+d}{(d+2r)p}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}}, \tag{84}$$

where $C_{d,r,p}$ is defined in Proposition 23. We can control both terms $\mathbb{E}_{\mathbf{x}\sim\pi_\alpha}[\|\mathbf{x}\|_2^r], \mathbb{E}_{\mathbf{y}\sim\pi'_\alpha}[\|\mathbf{y}\|_2^r]$ as in the proof of Proposition 23 so that

$$\mathbb{E}_{\mathbf{x}\sim\pi_\alpha}[\|\mathbf{x}\|_2^r] \leq 2^r \left(\int \|\mathbf{y}\|_2^r \mathrm{d}\pi(\mathbf{y}) + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z})\mathrm{d}\mathbf{z} \right) \leq 2^r (M^r + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z})\mathrm{d}\mathbf{z}), \tag{85}$$

since $\pi \in \mathfrak{S}$ (and in the same way for $\mathbb{E}_{\mathbf{y}\sim\pi'_\alpha}[\|\mathbf{y}\|_2^r]$). Consequently:

$$\mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq C_{d,r,p} 2^{(r+1)(\frac{2p+d}{(d+2r)p})} (M^r + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z})\mathrm{d}\mathbf{z})^{\frac{2p+d}{(d+2s)p}} \|\pi - \pi'\|_\kappa^{\frac{2(r-p)}{(d+2r)p}}. \tag{86}$$

By defining $C'_{d,r,p} = 2^{(r+1)\frac{2p+d}{(d+2r)p}} C_{d,r,p}$ and using Lemma 21 we have

$$\mathrm{W}_p(\pi_\alpha, \pi'_\alpha) \leq C'_{d,r,p} (M^r + \int \|\mathbf{z}\|_2^r \alpha(\mathbf{z})\mathrm{d}\mathbf{z})^{\frac{2p+d}{(d+2r)p}} + 2 \left(\int \|\mathbf{z}\|_2^p \alpha(\mathbf{z})\mathrm{d}\mathbf{z} \right)^{1/p}, \tag{87}$$

which concludes the proof. ■

Appendix B. Proofs of Section 3

B.1 Proof of Lemma 31

Lemma 31 (Canas and Rosasco, 2012) *Let $S \subseteq \mathcal{X}$, $p \in [1, +\infty)$ and $\pi \in \mathcal{P}_p(\mathcal{X})$. Consider $P_S : \mathcal{X} \to S$, measurable, such that $D(\mathbf{x}, P_S(\mathbf{x})) \leq D(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in S$. Then*

$$\mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = \mathrm{W}_p^p(\pi, P_S\#\pi).$$

Moreover for any $\nu \in \mathcal{P}_p(\mathcal{X})$ such that $\mathrm{supp}(\nu) \subseteq S$ we have $\mathrm{W}_p(\pi, P_S\#\pi) \leq \mathrm{W}_p(\pi, \nu)$.

Proof The proof in mainly taken from Canas and Rosasco (2012) but we rewrite it in our context. Considering the admissible coupling $\gamma = (id \times P_S)\#\pi \in \Pi(\pi, P_S\#\pi)$, then

$$\mathrm{W}_p^p(\pi, P_S\#\pi) \leq \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma(\mathbf{x}, \mathbf{y}) = \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p]. \tag{88}$$

Conversely, if γ^* is an optimal coupling for $\mathrm{W}_p(\pi, P_S\#\pi)$ then for all $(\mathbf{x}, \mathbf{y}) \in \mathrm{supp}(\gamma^*)$ we have that $\mathbf{y} \in \mathrm{supp}(P_S\#\pi)$ by definition of a coupling which means that $\mathbf{y} \in S$ and so by hypothesis $D^p(\mathbf{x}, \mathbf{y}) \geq D^p(\mathbf{x}, P_S(\mathbf{x}))$. Therefore,

$$\mathrm{W}_p^p(\pi, P_S\#\pi) = \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \geq \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) = \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}). \tag{89}$$

Hence $\mathrm{W}_p^p(\pi, P_S\#\pi) \geq \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p]$. The last inequality can be proved in the same way by considering an optimal coupling γ^* between π and ν:

$$\begin{aligned} \mathrm{W}_p^p(\pi, \nu) &= \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \overset{\mathrm{supp}(\nu)\subseteq S}{\geq} \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \\ &= \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = \mathrm{W}_p^p(\pi, P_S\#\pi). \end{aligned} \tag{90}$$

■

Appendix C. Proofs of Section 4

C.1 Proof of Proposition 36

We recall the result here:

Proposition 36 (Equivalence of Hölder LRIP and IOP) *Consider a learning task $\mathcal{L}(\mathcal{H})$, an exponent $p \in [1, +\infty)$, and a model set $\mathfrak{S}$.*

(i) If $\mathcal{A}$ satisfies (Hölder-LRIP) *with error $\eta \geq 0$ and constant $C > 0$ then the "ideal" decoder defined by*

$$\Delta[\mathbf{s}] \in \operatorname*{arg\,min}_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \,, \tag{24}$$

satisfies (Hölder-IOP) *with constant $2C > 0$, error $\eta \geq 0$ and*

$$\mathrm{Bias}(\pi, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2^\delta \,.$$

(ii) Conversely if the decoder Δ defined in (24) *satisfies* (Hölder-IOP) *with error $\eta \geq 0$, constant $C > 0$ and* $\mathrm{Bias}(\pi, \mathfrak{S})$ *defined above, then $\mathcal{A}$ satisfies* (Hölder-LRIP) *with constant $C > 0$ and error 2η.*

Proof For the proof we will need that if $(a, b) \in \mathbb{R}_+$ and $\delta \in [0, 1]$ then $(a + b)^\delta \leq a^\delta + b^\delta$.

IOP $\implies$ LRIP Suppose that Δ satisfies (Hölder-IOP). Let $\pi, \pi' \in \mathfrak{S}$. Then by the triangle inequality:

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \|\pi - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} + \|\pi' - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} \,. \tag{91}$$

For the first term $\|\pi - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p}$ we can apply the Hölder IOP with $\mathbf{e} = 0$ which gives $\|\pi - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} \leq \eta$ since $\pi \in \mathfrak{S}$ so $\mathrm{Bias}(\pi, \mathfrak{S}) = 0$. For the second term see that $\mathcal{A}(\pi) = \mathcal{A}(\pi') + (\mathcal{A}(\pi) - \mathcal{A}(\pi'))$ so we can apply the IOP with $\mathbf{e} = \mathcal{A}(\pi) - \mathcal{A}(\pi')$ which gives $\|\pi' - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} = \|\pi' - \Delta[\mathcal{A}(\pi') + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} \leq 0 + C\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^\delta + \eta$ and finally we have (Hölder-LRIP) with constant C and error 2η.

LRIP $\implies$ IOP Suppose that $\mathcal{A}$ satisfies (Hölder-LRIP). Consider the decoder

$$\Delta[\mathbf{s}] \in \operatorname*{arg\,min}_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \,, \tag{92}$$

which means that $\|\mathcal{A}(\Delta[\mathbf{s}]) - \mathbf{s}\|_2 \leq \|\mathcal{A}(\tau) - \mathbf{s}\|_2$ for any $\tau \in \mathfrak{S}$. We define

$$\mathrm{Bias}(\pi, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \left(\|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\tau) - \mathcal{A}(\pi)\|_2^\delta \right) .$$

We show that this decoder satisfies (Hölder-IOP) with this Bias term. Let $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathbf{e} \in \mathbb{C}^m$. Consider any $\tau \in \mathfrak{S}$. Then

$$\begin{aligned}
\|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} &\leq \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + \|\tau - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} \\
&\overset{*}{\leq} \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + C\|\mathcal{A}(\tau) - \mathcal{A}(\Delta[\mathcal{A}(\pi) + \mathbf{e}])\|_2^\delta + \eta \\
&\overset{**}{\leq} \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + C\|\mathcal{A}(\tau) - (\mathcal{A}(\pi) + \mathbf{e})\|_2^\delta \\
&\quad + C\|(\mathcal{A}(\pi) + \mathbf{e}) - \mathcal{A}(\Delta[\mathcal{A}(\pi) + \mathbf{e}])\|_2^\delta + \eta \,,
\end{aligned} \tag{93}$$

where in (*) we use the LRIP since τ and $\Delta[\mathcal{A}(\pi) + \mathbf{e}]$ are in $\mathfrak{S}$. In (**) we use the triangle inequality and the property $(a + b)^\delta \leq a^\delta + b^\delta$. By the properties of the decoder we have $\|(\mathcal{A}(\pi) +$

$\mathbf{e}) - \mathcal{A}(\Delta[\mathcal{A}(\pi)+\mathbf{e}])\|_2 \leq \|(\mathcal{A}(\pi)+\mathbf{e}) - \mathcal{A}(\tau)\|_2$. Consequently:

$$\begin{aligned}\|\pi - \Delta[\mathcal{A}(\pi)+\mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} &\leq \|\pi-\tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\tau) - (\mathcal{A}(\pi)+\mathbf{e})\|_2^\delta + \eta \\ &\leq \|\pi-\tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\tau) - \mathcal{A}(\pi)\|_2^\delta + 2C\|\mathbf{e}\|_2^\delta + \eta\,. \\ \|\pi - \Delta[\mathcal{A}(\pi)+\mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} &\overset{*}{\leq} \mathrm{Bias}(\pi,\mathfrak{S}) + 2C\|\mathbf{e}\|_2^\delta + \eta\,,\end{aligned} \tag{94}$$

where in (*) we used the definition of $\mathrm{Bias}(\pi,\mathfrak{S})$ since the previous was true for any $\tau \in \mathfrak{S}$. ■

C.2 Proof of Proposition 38

Proposition 38 (Restricted Wasserstein regularity is necessary) *Consider $\mathcal{X} = \mathbb{R}^d$ equipped with a norm $\|\cdot\|, p \in [1,+\infty)$, and a model set $\mathfrak{S} \subseteq \mathcal{P}_p(\mathbb{R}^d)$. Consider a sketching operator $\mathcal{A}$ defined by $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ with $\Phi \in \mathrm{Lip}_L\left((\mathbb{R}^d, \|\cdot\|), (\mathbb{R}^m, \|\cdot\|_2)\right)$. If $\mathcal{A}$ satisfies* (Hölder-LRIP) *with error $\eta = 0$, constant $C > 0$ and $\delta = 1$ then*

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq CL\,\mathrm{W}_1(\pi,\pi') \leq CL\,\mathrm{W}_p(\pi,\pi')\,,$$

where the Wasserstein distance is computed with the distance $D(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$.

Proof Under the hypothesis of the proposition we have $\Phi \in \mathrm{Lip}(\mathbb{R}^d, \mathbb{R}^m)$ and

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2\,, \tag{95}$$

for some $C > 0$. As shown in Gribonval et al. (2021a, Appendix D, Proof of Lemma 3.2 and Lemma 3.4) the duality property of the Wasserstein distance implies $\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \leq L\,\mathrm{W}_1(\pi,\pi')$. The argument is the following: for $\pi, \pi' \in \mathfrak{S}$,

$$\begin{aligned}\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 &= \sup_{\mathbf{u}\in\mathbb{R}^m : \|\mathbf{u}\|_2 \leq 1} |\langle \mathbf{u}, \mathcal{A}(\pi) - \mathcal{A}(\pi')\rangle| \\ &= \sup_{\mathbf{u}\in\mathbb{R}^m : \|\mathbf{u}\|_2 \leq 1} \left|\int \langle \mathbf{u}, \Phi(\mathbf{x})\rangle \mathrm{d}\pi(\mathbf{x}) - \int \langle \mathbf{u}, \Phi(\mathbf{y})\rangle \mathrm{d}\pi'(\mathbf{y})\right| \\ &= \sup_{\mathbf{u}\in\mathbb{R}^m : \|\mathbf{u}\|_2 \leq 1} \left|\int \Phi_{\mathbf{u}}(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int \Phi_{\mathbf{u}}(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y})\right|,\end{aligned} \tag{96}$$

where we define $\Phi_{\mathbf{u}}(\cdot) = \langle \mathbf{u}, \Phi(\cdot)\rangle$. Moreover, for any $\mathbf{u} \in \mathbb{R}^m$ with $\|\mathbf{u}\|_2 \leq 1$ we have $\Phi_{\mathbf{u}} \in \mathrm{Lip}_L(\mathbb{R}^d, \mathbb{R})$ since $\Phi \in \mathrm{Lip}(\mathbb{R}^d, \mathbb{R}^m)$. Consequently, using the duality property of the Wasserstein distance:

$$\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \leq \sup_{f \in \mathrm{Lip}_L(\mathbb{R}^d,\mathbb{R})} \left|\int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y})\right| = L\,\mathrm{W}_1(\pi,\pi')\,. \tag{97}$$

Combining with (95) we have

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq CL\,\mathrm{W}_1(\pi,\pi')\,. \tag{98}$$

Finally to conclude we use $\mathrm{W}_1(\pi,\pi') \leq \mathrm{W}_p(\pi,\pi')$ since $p \in [1,+\infty)$ (Santambrogio, 2015, Section 5.1).

■

References

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Elsevier Science, 2003.

Roman Razmikovich Akopyan and Andrey Efimov. Boas–Kac roots of positive definite functions of several variables. *Analysis Mathematica*, 43, 2017.

Michael Arbel, Danica J Sutherland, Mikoł aj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 2017.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks, 2017.

N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68, 1950.

Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially Private Database Release via Kernel Mean Embeddings. In *International Conference on Machine Learning (ICML)*, 2018.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 2006.

Ayoub Belhadji and Rémi Gribonval. Revisiting RIP guarantees for sketching operators on mixture models, 2022.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Anthony Bourrier, Mike E. Davies, Tomer Peleg, Patrick Pérez, and Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 2014.

Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical Inference for Generative Models with Maximum Mean Discrepancy, 2019.

Guillermo Canas and Lorenzo Rosasco. Learning Probability Measures with respect to Optimal Transport Metrics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Djalil Chafaï, A. Hardy, and Mylène Maïda. Concentration for Coulomb gases and Coulomb transport inequalities. *Journal of Functional Analysis*, 275, 2016.

Antoine Chatalic. *Efficient and privacy-preserving compressive learning*. Thesis, Université Rennes 1, 2020.

Antoine Chatalic, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Mean nyström embeddings for adaptive compressive learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Sanjoy Dasgupta. Learning Mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99, 1999.

Jérôme Dedecker and Bertrand Michel. Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 122, 2013.

R. M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. *Annals of Mathematical Statistics*, 40, 1969.

Werner Ehm, Tilmann Gneiting, and Donald Richards. Convolution Roots of Radial Positive Definite Functions with Compact Support. *Transactions of the American Mathematical Society*, 356, 2004.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Ziv Goldfeld and Kristjan H. Greenewald. Gaussian-Smoothed Optimal Transport: Metric Structure and Statistical Efficiency. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63, 2020.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining Fairness using Optimal Transport Theory. In *International Conference on Machine Learning (ICML)*, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research (JMLR)*, 13, 2012.

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive Statistical Learning with Random Feature Moments. *Mathematical Statistics and Learning*, 3, 2021a.

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling. *Mathematical Statistics and Learning*, 3, 2021b.

A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.

Nicolas Keriven and Rémi Gribonval. Instance Optimal Decoding and the Restricted Isometry Property. *Journal of Physics: Conference Series*, 1131, 02 2018.

Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval. Compressive K-means. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for Large-Scale Learning of Mixture Models. *Information and Inference*, 7, 2018.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021.

S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34, 2017.

Fabian Latorre, Paul Rolland, and Volkan Cevher. Lipschitz constant estimation of Neural Networks via sparse polynomial optimization. In *International Conference on Learning Representations (ICLR)*, 2020.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401, 1999.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a Unified Analysis of Random Fourier Features. *Journal of Machine Learning Research (JMLR)*, 22, 2021.

F. Liese and I. Vajda. On Divergences and Informations in Statistics and Information Theory. 52, 2006.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 2021.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online Dictionary Learning for Sparse Coding. In *International Conference on Machine Learning (ICML)*, 2009a.

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised Dictionary Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009b.

Andreas Maurer and Massimiliano Pontil. K-Dimensional Coding Schemes in Hilbert Spaces. *IEEE Transactions on Information Theory*, 56, 2010.

Thibault Modeste and Clément Dombry. Characterization of translation invariant mmd on $\mathbb{R}^d$ and connections with wasserstein distances, 2022.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, 10, 2017.

A. Mueller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29, 1997.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41, 2013.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-Wasserstein Distance: Structure, Empirical Approximation, and Statistical Applications. In *International Conference on Machine Learning (ICML)*, 2021a.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. From Smooth Wasserstein Distance to Dual Sobolev Norm: Empirical Approximation and Statistical Applications, 2021b.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50, 2022.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11, 2019.

Jr. R. P. Boas and M. Kac. Inequalities for Fourier transforms of positive functions. *Duke Mathematical Journal*, 12, 1945.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing Minimization with Randomization in Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Mark D. Reid and Robert C. Williamson. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research (JMLR)*, 12, 2011.

Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356, 2018.

E.B. Saff and V. Totik. *Logarithmic Potentials with External Fields*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015.

Vincent Schellekens and Laurent Jacques. Compressive Classification (Machine Learning without learning), 2018.

Vincent Schellekens and Laurent Jacques. Compressive Learning of Generative Networks, 2020.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Hui Shi, Yann Traonmilin, and Jean François Aujol. Compressive learning of deep regularization for denoising. 2022a.

Hui Shi, Yann Traonmilin, and Jean-Francois Aujol. Compressive learning for patch-based image denoising. *SIAM Journal on Imaging Sciences*, 15(3), 2022b.

Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing Weak Convergence with Maximum Mean Discrepancies, 2020.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein Propagation for Semi-Supervised Learning. In *International Conference on Machine Learning (ICML)*, 2014.

Bharath Sriperumbudur and Zoltan Szabo. Optimal Rates for Random Fourier Features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, phi-divergences and binary classification, 2009.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research (JMLR)*, 11, 2010.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6, 2012.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, 2016.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Danica J. Sutherland and Jeff Schneider. On the Error of Random Fourier Features, 2015.

Gabor Szekely and Maria Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

Gábor J. Székely and Maria L. Rizzo. The Energy of Data. *Annual Review of Statistics and Its Application*, 4, 2017.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized Low Rank Models. *Foundations and Trends in Machine Learning*, 9, 2016.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25, 2019.

Holger Wendland. Scattered Data Approximation. *Cambridge University Press*, 2004.

Yixing Zhang, Xiuyuan Cheng, and Galen Reeves. Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.