Controlling Wasserstein distances by Kernel norms with application to Compressive Statistical Learning

Titouan Vayer, Rémi Gribonval

Controlling Wasserstein distances by Kernel norms with application to Compressive Statistical Learning

Titouan Vayer TITOUAN.VAYER@ENS-LYON.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

Abstract

Comparing probability distributions is at the crux of many machine learning algorithms. Maximum Mean Discrepancies (MMD) and Optimal Transport distances (OT) are two classes of distances between probability measures 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* (Hölder LRIP) and show that this property comes with interesting guarantees for compressive statistical learning. Based on the relations between the MMD and the Wasserstein distance, we provide guarantees for compressive statistical learning by introducing and studying the concept of *Wasserstein learnability* of the learning task, that is when some task-specific metric between probability distributions can be bounded by a Wasserstein distance.

Keywords: Optimal Transport, Kernel norms, Statistical Learning, Inverse problems

Contents

1. Introduction

Countless methods in machine learning 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 measures π and π'. Divergences and metrics between probability distributions 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). An important issue is the choice of such a metric, 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 following 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, *e.g.* 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 $\mathrm{TaskMetric}$, and 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) (Berlinet and Thomas-Agnan, 2004)) and the 1-Wasserstein distance W_1 (Villani, 2008) (when $\mathcal{G}$ is the class of 1-Lipschitz functions). Both are attracting increasing interest from the machine learning community due to their ability to handle the metric structure of the feature space *e.g.* 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. In particular we study the conditions under which the Wasserstein distance W_p can be upper-bounded by a MMD with a "Hölder" exponent, that is, informally, when:

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

Especially, we are interested in MMD associated to RKHS engendered by *translation invariant (TI) positive semi-definite (p.s.d.) kernels* that are widely used in many machine learning (ML) applications and are at the core of many large-scale learning algorithms (Rahimi and Recht, 2008, 2007). Despite some connections between MMD 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 MMD. We show 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) holds 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 4).

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 MMD 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 learning approaches. One of the main notion in this context is found in the *Lower*

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 $D(x, y) \geq 0$, satsifies 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 $x \neq y$)

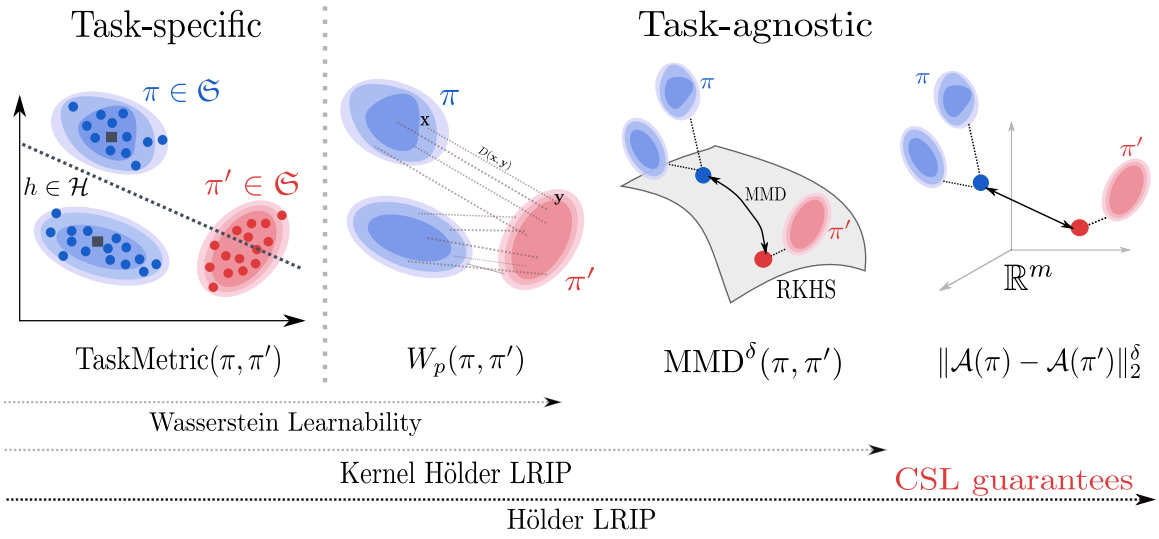

Figure 1: The reasoning used in the paper to obtain compressive statistical learning guarantees. **(left)** Given two distributions π, π' on a model set $\mathfrak{S}$, our goal is to control some task-specific metric $\mathrm{TaskMetric}(\pi, \pi')$ that depends on the learning task (see Section 2). **(middle left)** First, in Section 3, we use an upper-bound $\mathrm{TaskMetric}(\pi, \pi') \lesssim W_p(\pi, \pi')$ by introducing the notion of Wasserstein learnability of the task. **(middle right)** Then, in Section 4 we show how to control the Wasserstein distance by a MMD with a Hölder exponent $\delta \in]0, 1], W_p(\pi, \pi') \lesssim \mathrm{MMD}^\delta(\pi, \pi')$. **(right)** Finally in Section 5 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 (Section 2)

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 translation invariant kernel whose MMD dominates $\mathrm{TaskMetric}$ (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 $\mathrm{TaskMetric}$ by a Wasserstein distance (*Wasserstein learnability*). In particular we prove that many unsupervised learning tasks such as *compression-type tasks* (K-means/medians, PCA (Gribonval et al., 2021a)), generative learning tasks 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 organize our work as follows: we start by presenting in Section 2 the compressive statistical learning framework which motivates our study. We introduce in this section the different notations used in the rest of the paper and define the different concepts at stake in CSL. We study a generalization of the LRIP, namely the Hölder LRIP, and we show that this property has many advantages for CSL. In Section 3 we study the relations between task-specific metrics between probability distributions and the Wasserstein distance. We introduce the concept of *Wasserstein learnability* of the learning task and show that, when it holds, the excess risk can be bounded

by a Wasserstein distance. The goal of Section 4 is to study the relations between Wasserstein and MMD. We provide conditions so that $W_p \lesssim MMD^{\delta}$ holds for some $\delta \in]0,1]$. Although Section 3 and 4 provide many important results for establishing CSL guarantees they are of independent interest and a reader interested primarily in the connections between OT and MMD may skip Section 2 at first reading. We conclude with in Section 5 by giving an instruction for the use of the different results in the case of CSL.

1.1 Notations and definitions

Metric spaces. In this article the space $\mathcal{X}$ will always be considered as a measurable, separable completely metrizable topological 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 functions from a metric space $(\mathcal{X},d_{\mathcal{X}})$ to $(\mathcal{Y},d_{\mathcal{Y}})$ is denoted by $\mathrm{Lip}_L((\mathcal{X},d_{\mathcal{X}}),(\mathcal{Y},d_{\mathcal{Y}}))$ simply by $\mathrm{Lip}_L(\mathcal{X},\mathcal{Y})$ when it is clear from the context. If $f \in \mathrm{Lip}_L((\mathcal{X},d_{\mathcal{X}}),(\mathcal{Y},d_{\mathcal{Y}}))$ we have $\forall \mathbf{x},\mathbf{x}' \in \mathcal{X}, d_{\mathcal{Y}}(f(\mathbf{x}),f(\mathbf{x}')) \leq Ld_{\mathcal{X}}(\mathbf{x},\mathbf{x}')$. The support of a function $f : \mathcal{X} \to \mathbb{R}$ is denoted as $\mathrm{supp}(f)$ and is defined as the closure of $\{\mathbf{x} \in \mathcal{X}, f(\mathbf{x}) \neq 0\}$. In the following, $\|.\|_2$ denotes the ℓ_2 norm, vectors and matrices are written in bold. By abuse of notation a vector in $\mathbf{x} \in \mathbb{R}^d$ is also considered to be a column matrix. On a normed space $(\mathcal{X},\|.\|)$, the ball centered at $\mathbf{x}_0 \in \mathcal{X}$ and with radius $R > 0$ is denoted $B_{\|.\|}(\mathbf{x}_0,R)$ or simply by $B(\mathbf{x}_0,R)$ when it is clear from the context.

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 $\mathcal{X}$ and $\mathcal{M}_+(\mathcal{X})$ is the space of non-negative finite measures on $\mathcal{X}$. For a probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ which admits a density f *w.r.t.* the Lebesgue measure on $\mathbb{R}^d$ we adopt the notation $\pi \ll f\mathrm{d}\mathbf{x}$. 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 measure $T\#\pi \in \mathcal{P}(\mathcal{Y})$ *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 for law $T\#\pi$. The support of a probability distribution is denoted as $\mathrm{supp}(\pi)$ and it is defined as the smallest closet set S such that $\pi(S) = 1$.

Integrability, Fourier transform and Sobolev spaces 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 satisty $\int_{\mathcal{X}} |f(\mathbf{x})|^p \mathrm{d}\mu(\mathbf{x}) < +\infty$. When $\mathcal{X} = \mathbb{R}^d$ we note $L_p(\mathbb{R}^d)$ the space of p-integrable functions with respect to the Lesbegue measure. For a integrable function $f \in L_1(\mathbb{R}^d)$ we note $\hat{f}(\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega}) = \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} f(\mathbf{x})\mathrm{d}\mathbf{x}$ its Fourier transform. 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 $\widehat{\mu}(\boldsymbol{\omega}) := \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\mu(\mathbf{x})$. We recall some concepts related to Sobolev spaces and we refer to (Adams and Fournier, 2003) for a more detailed description. A multi-index is a tuple of non negative integers: $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$ where $\alpha_i \in \mathbb{N}$. We define $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$ the *order* of $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}! = \alpha_1 \cdot \alpha_2 \cdots \alpha_d$. For $\mathbf{x} \in \mathbb{R}^d$ we adopt the notation $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_d^{\alpha_d}$. For a function $\phi \in C^s(\mathbb{R}^d)$ we adopt the notation $\partial^{\boldsymbol{\alpha}}\phi = \frac{\partial^{|\boldsymbol{\alpha}|}\phi}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ for $|\boldsymbol{\alpha}| \leq s$. For a function $u \in L_{1,\mathrm{loc}}(\mathbb{R}^d)$ we use the same notation $\partial^{\boldsymbol{\alpha}} u$ for its weak derivative which satisfies $\int (\partial^{\boldsymbol{\alpha}} u)\phi \mathrm{d}\mathbf{x} = (-1)^{|\boldsymbol{\alpha}|} \int u(\partial^{\boldsymbol{\alpha}}\phi)\mathrm{d}\mathbf{x}$ for all $\phi \in C_c^\infty(\mathbb{R}^d)$ (smooth functions with compact support). Recall that if u is differentiable in the conventional sense then its weak derivative is identical to its conventional derivative. In the following we use the following definition of the Sobolev norm:

$$\|u\|_{W^{s,p}(\mathbb{R}^d)} := \left(\sum_{|\boldsymbol{\alpha}| \leq s} \int_{\mathbb{R}^d} |\partial^{\boldsymbol{\alpha}} u(\mathbf{x})|^p \mathrm{d}\mathbf{x} \right)^{1/p} \tag{3}$$

This norm makes sense, *i.e.* $\partial^{\boldsymbol{\alpha}} u$ exists in the conventional or weak sense, as soon as $u \in L_{1,\mathrm{loc}}(\mathbb{R}^d)$ or when $u \in C^s(\mathbb{R}^d)$. The Sobolev space $W^{s,p}(\mathbb{R}^d)$ of functions is defined as the space of functions $u \in L_p(\mathbb{R}^d)$ such that $\|u\|_{W^{s,p}(\mathbb{R}^d)} < +\infty$.

2. From statistical learning to compressive statistical learning

In this section we present the main objective of the compressive statistical learning theory by first introducing the main concept of statistical learning.

2.1 A gentle introduction to the notations: statistical learning

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 ? For supervised learning problems, *e.g.* in the context of classification, the samples are tuples $\mathbf{x}_i = (\mathbf{z}_i, y_i)$ where $\mathbf{z}_i$ is generally a vector in $\mathbb{R}^d$ and y_i is a label in a space $\mathcal{Y}$ such as $\{+1, -1\}$ in the context of binary classification. The learning algorithm aims here at producing a *classifier i.e.* a function $h \in \mathcal{H}$ which takes as input a sample $\mathbf{z}_i$ and outputs a label, which should be close to y_i, the true label of the sample. For unsupervised learning problems one may desire to faithfully summarize the collection of samples by reducing its size, or to find suitable representatives of this collection. As an illustrative example of unsupervised learning problem we can think of K-means where one wishes to select a set of centroids, $h = (\mathbf{c}_1, \cdots, \mathbf{c}_K)$ with $\mathbf{c}_i \in \mathcal{X} = \mathbb{R}^d$, that best represents our dataset.

In all of these problems the ideal hypothesis minimizes a certain *risk* which provides a performance measure. It is most of the time defined with the help of a loss function $\ell : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$ which takes as input a sample and a hypothesis $(\mathbf{x}, h)$ and returns a scalar value $\ell(\mathbf{x}, h)$. To illustrate, in K-Means this loss is defined by $\ell(\mathbf{x}, h = (\mathbf{c}_1, \cdots, \mathbf{c}_K)) = \min_{i\in[\![k]\!]} \|\mathbf{x} - \mathbf{c}_i\|_2^2$ that is the (squared) distance between $\mathbf{x}$ and its closest centroid. 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 $\mathbf{z} \in \mathbb{R}^d$. 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* which is defined by:

$$\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)] \tag{4}$$

This quantity reflects how effective is h on average on the data-generating distribution. As such, the optimal hypothesis $h^* \in \mathcal{H}$ 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* $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ of the training samples. The *excess risk* , *i.e.* how good $\hat{h}$ behaves compared to h^* is measured by the quantity $\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*)$ where π is the true data-generating distribution. One aims at guaranteeing, with high probability, the following bound on the excess risk:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq \eta_n \tag{5}$$

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[2]:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \leq 2 \sup_{h\in\mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi_n, h)| \tag{6}$$

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

Consequently, being able to control the right term in the previous equation is a central problem in statistical learning and *e.g.* arguments involving Rademacher complexities can lead to the desired bound in (5) (see *e.g.* (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 $\mathrm{TaskMetric}(\pi,\pi')$ in the introduction, defines a central quantity for the rest of the paper and we introduce the following notation for $\pi,\pi'\in\mathcal{P}(\mathcal{X})$:

$$\|\pi-\pi'\|_{\mathcal{L}(\mathcal{H})}:=\sup_{h\in\mathcal{H}}|\mathcal{R}(\pi,h)-\mathcal{R}(\pi',h)| \tag{7}$$

The quantity $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ defines a semi-norm[3] on the space of finite signed measure $\mathcal{M}(\mathcal{X})$ and is an IPM by considering the space of functions $\mathcal{G}=\mathcal{L}(\mathcal{H}):=\{\mathbf{x}\to\ell(\mathbf{x},h);h\in\mathcal{H}\}$. It is important to note that this semi-norm is *task-specific i.e.* that it depends on the learning task via the function 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 described previously, when $\|\pi-\pi_n\|_{\mathcal{L}(\mathcal{H})}\leq\eta_n$ one can control the excess risk (5), thus 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, *i.e.*, when $\ell:\mathcal{X}\times\mathcal{H}\to\mathbb{R}_+$, we introduce for $p\geq1$ the semi-norm:

$$\|\pi-\pi'\|_{\mathcal{L}(\mathcal{H}),p}:=\sup_{h\in\mathcal{H}}|\mathcal{R}^{1/p}(\pi,h)-\mathcal{R}^{1/p}(\pi',h)| \tag{8}$$

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 will often write $\|\pi-\pi_n\|_{\mathcal{L}(\mathcal{H}),p}$ without specifying that the loss function is non-negative and that $p\geq1$ (this will be implicitly assumed).

Remark 1. *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 $\|\pi-\pi'\|_{\Delta\mathcal{L}(\mathcal{H})}:=\sup_{h,h_0\in\mathcal{H}}\left[\{\mathcal{R}(\pi,h)-\mathcal{R}(\pi,h_0)\}-\{\mathcal{R}(\pi',h)-\mathcal{R}(\pi',h_0)\}\right]$ which is related to $\|\pi-\pi'\|_{\mathcal{L}(\mathcal{H})}$ via the inequality $\|\pi-\pi'\|_{\Delta\mathcal{L}(\mathcal{H})}\leq2\|\pi-\pi'\|_{\mathcal{L}(\mathcal{H})}$ (Gribonval et al., 2021a). However in this work we choose to focus only on the quantities defined in (7) and (8) and leave the analysis of $\|\cdot\|_{\Delta\mathcal{L}(\mathcal{H})}$ for further works.*

2.2 Compressive statistical learning

In contrast to the empirical risk minimization approach described in Section 2.1 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 a inverse problem in the space of measures and can 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.

3. A semi-norm $\|.\|$ 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 $\lambda\in\mathbb{R}$, $\|\lambda\mathbf{x}\|=|\lambda|\|\mathbf{x}\|$.

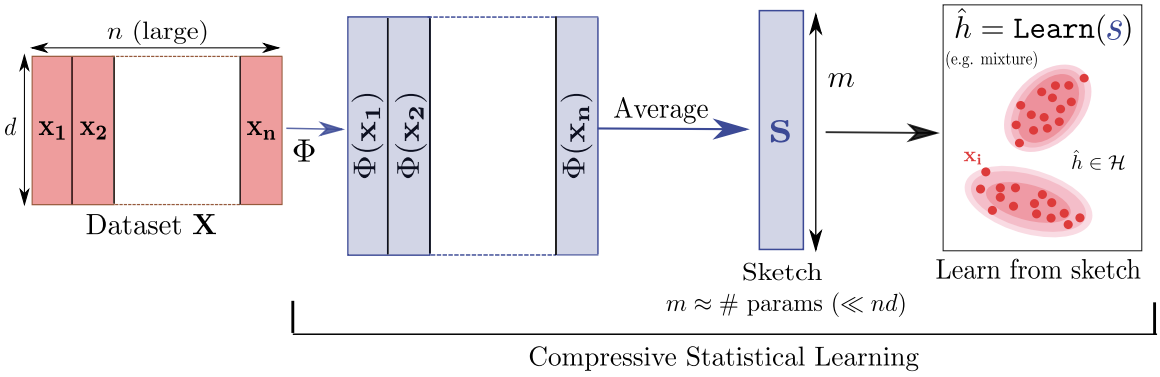

Figure 2: The principle of CSL (when $\mathcal{X} = \mathbb{R}^d$). From a dataset $\mathbf{X}$ with n samples (usually n is large) we push each sample $\mathbf{x}_i \in \mathbb{R}^d$ to either $\mathbb{R}^m$ or $\mathbb{C}^m$ using a well-chosen feature function $\Phi(\mathbf{x}_i)$. The second step 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.

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*:

$$\mathbf{s} := \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) \tag{9}$$

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 2). 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 (Gribonval et al., 2021a,b). Given Φ, the associated *sketching operator* is defined by:

$$\begin{aligned} \mathcal{A}: \ \mathcal{P}(\mathcal{X}) &\to \mathbb{R}^m \text{ or } \mathbb{C}^m \\ \pi &\to \mathcal{A}(\pi) := \int_{\mathcal{X}} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) \end{aligned} \tag{10}$$

This operator is linear in π in that $\mathcal{A}((1-\lambda)\pi + \lambda\pi') = (1-\lambda)\mathcal{A}(\pi) + \lambda\mathcal{A}(\pi')$ for $\lambda \in [0,1]$[4]. When applied to the empirical distribution of the dataset, $\pi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$ we recover the sketch $\mathbf{s}$ as:

$$\mathcal{A}(\pi_n) = \mathcal{A}(\frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathbf{s}. \tag{11}$$

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).

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

4. 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.

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:

$$\Delta : \mathbf{s} \to \Delta[\mathbf{s}] \in \mathfrak{S} \tag{12}$$

Suppose for the moment that we know how to sketch and how to decode *i.e.* we know $\mathcal{A}$ and Δ (we will describe in the next sections how to construct these operators). Given a sketch $\mathbf{s}$ of the dataset and a decoder Δ we can find a hypothesis based on the following risk minimization:

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h) \tag{13}$$

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.

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):

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \tag{14}$$

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 "equal" with respect to the task-dependent metric, *i.e.*, they lead to the same risk for every hypothesis. When this condition holds, the following decoder Δ provides many interesting guarantees:

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \tag{15}$$

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 described this property more in depth in Section 2.3 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 a dataset drawn from π. Suppose that we have access only to a sketch of this empirical distribution *i.e.* $\mathbf{s} = \mathcal{A}(\pi_n)$ with $\mathcal{A}$ that satisfies the LRIP. Consider the decoder Δ defined in (15) and $\hat{h}$ such that $\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$. Then it can be shown using the IOP property (Gribonval et al., 2021a):

$$\|\pi - \Delta[\mathbf{s}]\|_{\mathcal{L}(\mathcal{H})} \lesssim \mathrm{Bias}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2 \tag{16}$$

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:

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^*) \lesssim \mathrm{Bias}(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2 \tag{17}$$

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 in $O(n^{-1/2})$. 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.

A quick overview To summarize, the reasoning in CSL follows four important steps:

(i) Design a well chosen model set $\mathfrak{S}$. This choice is driven by the learning task and can be achieved with prior knowledge on general properties of the true underlying distribution (such as being a mixture of Gaussian distributions (Gribonval et al., 2021b)). The main idea here is to choose $\mathfrak{S}$ so that the bias term $\mathrm{Bias}(\pi, \mathfrak{S})$ will be small or comparable to the true risk $\mathcal{R}(\pi, h^*)$.

(ii) Find a sketching operator $\mathcal{A}$ which satisfies the LRIP (14) on $\mathfrak{S}$. This step is crucial: finding a large class of model sets $\mathfrak{S}$ for which it is possible to find such $\mathcal{A}$ is of particular interest. **Our paper focuses on this step, with a generalized different notion of LRIP for which we establish generalized compressive learning guarantees** (see Section 2.3).

(iii) Solve a Generalized Method of Moments problem associated to the decoder

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

This optimization problem is an inverse problem on the space of measure and is not trivial to solve. When $\mathfrak{S}$ is the space of mixture of k diracs (as in compressive K-means (Keriven et al., 2017; Gribonval et al., 2021b)) this problem echoes to the Beurling LASSO which has an extensive literature (see *e.g.* (Candes and Fernandez-Granda, 2012; Gao and Pavel, 2018; Denoyelle et al., 2018) and references therein) and greedy methods like orthogonal matching pursuit (OMP) were proposed (Keriven et al., 2018; Elvira et al., 2020). We do not cover this aspect of sketching and leave it to further works.

(iv) Find a hypothesis $\hat{h}$ which solves $\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$. At first sight it seems that what is gained by not doing an ERM is lost in solving $\arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$. 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 negligible compared to doing an ERM.

2.3 Extending compressive statistical learning guarantees with Hölder LRIP and Hölder IOP

Our first contribution is to define an extended notion of LRIP, namely the Hölder LRIP, and to show 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 $\delta \in]0, 1]$. We will show that this definition has many advantages, and also some drawbacks that we will discuss. More precisely we consider the following definition:

Definition 1 (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:*

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

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 1 (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:*

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \tag{18}$$

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 (18) *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 A.1. In this paper we always assume that the minimization problem (18) 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 property allows us to have some guarantees on the excess risk as described in the next theorem:

Theorem 1 (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 true data generating distribution and π_n an empirical distribution i.e. $\pi_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ where $\mathbf{x}_i \sim \pi$. Consider a sketch of the dataset $\mathbf{s} = \mathcal{A}(\pi_n)$ and Δ the ideal decoder $\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2$*

Let $h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\pi, h)$ be the optimal hypothesis and $\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\Delta[\mathbf{s}], h)$. Then:

$$\mathcal{R}(\pi, \hat{h})^{1/p} - \mathcal{R}(\pi, h^*)^{1/p} \leq 2\,\mathrm{Bias}(\pi, \mathfrak{S}) + 2C\|\mathcal{A}(\pi) - \mathcal{A}(\pi_n)\|_2^\delta + 2\eta \tag{19}$$

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 1 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 $\|\pi - \Delta[\mathcal{A}(\pi_n)]\|_{\mathcal{L}(\mathcal{H}),p} \leq \mathrm{Bias}(\pi, \mathfrak{S}) + C\,\|\mathcal{A}(\pi_n) - \mathcal{A}(\pi)\|_2^\delta + \eta$. However as described in the introduction we have $\mathcal{R}(\pi, \hat{h})^{1/p} - \mathcal{R}(\pi, h^*)^{1/p} \leq 2\|\pi - \Delta[\mathbf{s}]\|_{\mathcal{L}(\mathcal{H}),p} = 2\|\pi - \Delta[\mathcal{A}(\pi_n)]\|_{\mathcal{L}(\mathcal{H}),p}$ which concludes the proof. ■

This result is essential: it illustrates that if we have carefully designed $\mathfrak{S}$ so that the bias term is of the order of $\mathcal{R}^{1/p}(\pi, h^*)$ and if we know a sketching operator with the Hölder LRIP property then $\mathcal{R}(\pi, \hat{h})$ converges to the true risk as n grows (with some additive term $\eta \geq 0$). 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. This nevertheless comes with a benefit : as we will show the existence of a sketching operator satisfying the Hölder LRIP with $\delta < 1$ is easier to prove than with $\delta = 1$.

2.4 The roadmap to the Hölder LRIP

As described in Theorem 1, guarantees on the excess risk can be achieved with a sketching operator $\mathcal{A}$ that satisfies the Hölder LRIP. It is natural to wonder whether such operators exist at all, and the second main contribution of this paper is to provide conditions ensuring their existence and to exhibit them. 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. We briefly describe here the main notions of this theory and we refer to (Berlinet and Thomas-Agnan, 2004) for more details about kernels.

Kernels and Maximum Mean Discrepancy. The theory of kernel has a long history when it comes to learning problems or more generally to probability and statistics (Aronszajn, 1950; Berlinet and Thomas-Agnan, 2004; Muandet et al., 2017). In the rest of the paper κ will denote a *positive semi-definite* (*p.s.d.*) kernel on a space $\mathcal{X}$[5]. Such kernel 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* (RKHS) and is defined by the reproducing property $f \in \mathcal{H}_\kappa$ if $f(\mathbf{x}) = \langle f, \kappa(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_\kappa}$ for any $\mathbf{x} \in \mathcal{X}$. Such a *p.s.d.* 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 $\pi' \in \mathcal{P}(\mathcal{X})$ via[6]:

$$\|\pi - \pi'\|_\kappa := \left(\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{x}')\sim\pi} [\kappa(\mathbf{x}, \mathbf{x}')] + \mathop{\mathbb{E}}_{(\mathbf{y},\mathbf{y}')\sim\pi} [\kappa(\mathbf{y}, \mathbf{y}')] - 2 \operatorname{Re}(\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\pi} [\kappa(\mathbf{x}, \mathbf{y})]) \right)^{1/2} \tag{20}$$

This quantity defines a semi-norm on the space of probability distribution and can be completed to a proper norm when the kernel is characteristic (Simon-Gabriel et al., 2020; Sriperumbudur et al., 2010), *i.e.* when $\|\pi - \pi'\|_\kappa = 0 \iff \pi = \pi'$. The MMD admits also the characterization $\|\pi - \pi'\|_\kappa = \sup_{\|f\|_{\mathcal{H}_\kappa} \leq 1} |\int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{x}) \mathrm{d}\pi'(\mathbf{x})|$ and can be naturally extended to any finite signed measure $\mu \in \mathcal{M}(\mathcal{X})$ by defining the semi-norm $\|\mu\|_\kappa^2 := \int \int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{y})$.

An important family of kernels, namely *translation invariant (TI) p.s.d. 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 even, continuous *p.s.d.* function $\kappa_0 : \mathbb{R}^d \to \mathbb{C}$ [7]. This family encompasses many popular kernels such as Gaussian, Laplacian or kernels of the Matèrn class (Sriperumbudur et al., 2010). The following characterization of such kernels is due to the celebrated Bochner theorem (see Theorem 6.6 and Theorem 6.11 in (Wendland, 2004)):

Proposition 2 (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 and even, is a p.s.d. kernel if and only if there exists a probability distribution $\Lambda \in \mathcal{P}(\mathbb{R}^d)$ such that:*

$$\forall \mathbf{x} \in \mathbb{R}^d, \ \kappa_0(\mathbf{x}) = \kappa_0(0) \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\Lambda(\boldsymbol{\omega}) \tag{21}$$

If κ_0 is continuous, even and *also in $L_1(\mathbb{R}^d)$ then $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ is a p.s.d. kernel if and only if κ_0 is bounded and $\forall \boldsymbol{\omega} \in \mathbb{R}^d, \widehat{\kappa}_0(\boldsymbol{\omega}) \geq 0$.*

5. A function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a *p.s.d.* kernel if it is *symmetric* 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 $\sum_{i,j=1}^n c_i \overline{c_j} \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.
6. The MMD $\|\pi - \pi'\|_\kappa$ is finite for any probability distributions π, π' for example when the kernel κ is bounded.
7. A function $\kappa_0 : \mathbb{R}^d \to \mathbb{C}$ is *p.s.d.* 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 if it is even *i.e.* $\kappa_0(-\mathbf{x}) = \kappa_0(\mathbf{x})$ then κ_0 is real-valued (Wendland, 2004, Theorem 6.2). Consequently a TI and p.s.d. kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ is real-valued (such κ is symmetric *iff* κ_0 is even).

Bochner's theorem shows that a translation invariant *p.s.d.* 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* (RFF) (the functions $\phi(\cdot, \boldsymbol{\omega}_i)$) to accelerate kernel learning algorithms (Rahimi and Recht, 2008, 2007).

From kernels to sketching operator: the LRIP case So far the construction of a sketching operator that satisfies the LRIP ($\delta = 1$) is mostly based on kernel embeddings of probability distributions. The idea is, given a model set $\mathfrak{S}$ and a task $\mathcal{L}(\mathcal{H})$, to follow these two steps (Gribonval et al., 2021a):

(i) Find a translation invariant p.s.d. kernel κ such that:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi - \pi'\|_\kappa \qquad \text{(Kernel-LRIP)}$$

(ii) Use the random feature expansion of the kernel to define $\mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{C}^m$ using random features. More precisely the sketching operator $\mathcal{A}$ is based on a feature function Φ defined by:

$$\Phi(\mathbf{x}) := \frac{1}{\sqrt{m}}(\phi(\mathbf{x}, \boldsymbol{\omega}_1), \cdots, \phi(\mathbf{x}, \boldsymbol{\omega}_m))^\top$$

where $\boldsymbol{\omega}_i \sim \Lambda$ and $\phi(\cdot, \boldsymbol{\omega}_i) = \exp(-i\langle \cdot, \boldsymbol{\omega}_i\rangle)$. Based on the approximation properties of random features and the "low-dimensionality" of $\mathfrak{S}$ prove that for m sufficiently large:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_\kappa \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \text{ w.h.p} \qquad (22)$$

As a consequence of steps (i) and (ii) we can prove that the sketching operator satisfies the LRIP $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2$ for any $\pi, \pi' \in \mathfrak{S}$. However both (i) and (ii) are quite difficult to establish and highly depend on the model set $\mathfrak{S}$. The first step (i) is called the *Kernel LRIP* and is delicate to prove in general. Most of the existing theoretical work on compressive statistical learning focuses on this property when the model set $\mathfrak{S}$ is the space of well separated K-mixtures of diracs or Gaussian distributions (Gribonval et al., 2021b; Keriven et al., 2018). As for it, the second step (ii) can be proven using arguments from the convergence of empirical estimation of the MMD to the true MMD but also requires a precise control of the covering numbers of the so-called *normalized secant set* of $\mathfrak{S}$ (Gribonval et al., 2021a). This can be also proven based on separability assumptions.

Hölder LRIP case: roadmap We will show that establishing CSL guarantees is easier in the context of the Hölder LRIP. Our strategy is, given a model set $\mathfrak{S}$, to find $\mathcal{A}, \kappa$ and $\delta \in]0, 1]$ such that we can prove the following chain of inequality for $\pi, \pi' \in \mathfrak{S}$:

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \overset{\text{Section 3}}{\lesssim} W_p(\pi, \pi') \overset{\text{Section 4}}{\lesssim} \|\pi - \pi'\|_\kappa^\delta \overset{\text{Section 5}}{\lesssim} \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^\delta \qquad (23)$$

The next Section 3 is devoted to the first step of this roadmap and will be called *Wasserstein learnability*. The motivation is twofold. First, we will show that this bound is quite natural for several learning tasks and even holds universally, *i.e.* for $\mathfrak{S} = \mathcal{P}_p(\mathcal{X})$. Second, in Section 4 we will show that under some assumptions on the kernel and the model set, we can control the Wasserstein distance by an MMD. Together with the Wasserstein learnability results this will prove that the task metric $\|.\|_{\mathcal{L}(\mathcal{H}),p}$ can be controlled uniformly on $\mathfrak{S}$ by a MMD with an Hölder exponent $\delta \in]0, 1]$ *i.e.*

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \lesssim \|\pi - \pi'\|_\kappa^\delta$$

This will establish the so-called *Kernel Hölder LRIP*. Finally, in Section 5, we will conclude that when the latter property holds and under additional assumptions on the model set a sketching operator that satisfies the Hölder LRIP can be found, thus establishing CSL guarantees.

3. Wasserstein learnability

As described, the first problem is to study how to control the task-metric $\|\cdot\|_{\mathcal{L}(\mathcal{H}),p}$ defined in (8) by a Wasserstein distance. We briefly describe the main notions of OT and Wasserstein distance and refer the reader to (Peyré and Cuturi, 2019; Santambrogio, 2015) for a more detailed discussion. Broadly speaking the interest of Optimal Transport (OT) 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 distance. Considering a complete and separable metric space $(\mathcal{X}, D)$, the Wasserstein distance is defined for two probability distributions $\pi, \pi' \in \mathcal{P}(\mathcal{X})$ and $p \in [1, +\infty[$ as:

$$W_p(\pi, \pi') = (\inf_{\gamma \in \Pi(\pi,\pi')} \int_{\mathcal{X}\times\mathcal{X}} D(\mathbf{x},\mathbf{y})^p \mathrm{d}\gamma(\mathbf{x},\mathbf{y}))^{1/p} \tag{24}$$

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:

$$\mathcal{P}_p(\mathcal{X}) = \{\pi \in \mathcal{P}(\mathcal{X}); \int_{\mathcal{X}} D(\mathbf{x}_0, \mathbf{y})^p \mathrm{d}\pi(\mathbf{y}) < +\infty \text{ for some arbitrary } \mathbf{x}_0 \in \mathcal{X}\} \tag{25}$$

with a metric structure (Villani, 2008)[8]. When $(\mathcal{X}, D)$ is a normed space such as $(\mathbb{R}^d, \|.\|_2)$ the space $\mathcal{P}_p(\mathcal{X})$ is the space of probability distributions with p-finite moments $\int_{\mathcal{X}} \|\mathbf{x}\|_2^p \mathrm{d}\pi(\mathbf{x}) < +\infty$. 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 (24) is called *optimal coupling* and it provides a probabilistic matching of the points in the support of the distributions π, π' that can be used to find their correspondences. 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 rely on its dual formulation (Santambrogio, 2015). It allows, among others, to characterize W_1 by consider the following maximization problem:

$$W_1(\pi, \pi') = \sup_{f \in \mathrm{Lip}_1(\mathcal{X},\mathbb{R})} |\int f(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int f(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y})| \tag{26}$$

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

As introduced in the previous sections, our goal is to study compressive statistical learning in a specific context, namely when the task-specific norm $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p}$ (Section 2) can be bounded by the Wasserstein distance between π and π'. We formalize this in the following definition:

Definition 2 (Wasserstein learnability)**.** *Given a model set $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ and $p \in [1, +\infty[$, we say that a task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable w.r.t. $\mathfrak{S}$ if there exists $C > 0$ such that:*

$$\forall \pi, \pi \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} = \sup_{h \in \mathcal{H}} |\mathcal{R}^{1/p}(\pi, h) - \mathcal{R}^{1/p}(\pi', h)| \leq C W_p(\pi, \pi') \tag{27}$$

We simply say that a task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable if the previous holds when $\mathfrak{S} = \mathcal{P}_p(\mathcal{X})$.

At first sight the Wasserstein learnability 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 and that several learning tasks satisfy this property *independently* of the choice of the model set. In other words we will see that many tasks $\mathcal{L}(\mathcal{H})$ are p-Wasserstein learnable. More importantly, the following result shows that this property is a necessary condition for the LRIP when the sketching operator is based on random features:

8. The space $\mathcal{P}_p(\mathcal{X})$ is here to formalize that W_p is finite and thus defines a proper distance.

When do we have $\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \lesssim W_p(\pi, \pi')$ for some $p \geq 1$ and task $\mathcal{L}(\mathcal{H})$?		
Condition on the task	**Condition on p**	**Examples**
Density estimation Hypothesis: $h \in \mathbb{R}^D$, Risk: $\mathcal{R}(\pi, h) = W_1(\pi, \pi_h)$	$p = 1$	GAN, GMM (Section 3.1)
Compression type-tasks Loss: $\ell(\mathbf{x}, h) = D^p(\mathbf{x}, P_h(\mathbf{x}))$, P_h is a projection function	$p \geq 1$	PCA, K-means, K-medians, NMF, Dictionary learning (Section 3.2)
Regression tasks Hypothesis: h Lipschitz function, Loss: square-loss or ℓ_p loss	$p \geq 1$	Linear regression, regression using MLP with bounded params (Section 3.3 and 3.4)
Binary classification Hypothesis: h Lipschitz function, Loss: convex surrogate $\ell(\mathbf{x} = (\mathbf{z}, y), h) = \varphi(yh(\mathbf{z}))$	$p = 1$	MLP classifier (bounded params) + Lipschitz ouput layer (Section 3.5)

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

Proposition 3 (Wasserstein learnability is necessary)**.** *Consider $\mathcal{X} = \mathbb{R}^d$, $p \in [1, +\infty[$, and any model set $\mathfrak{S} \subseteq \mathcal{P}_p(\mathbb{R}^d)$. Consider a sketching operator $\mathcal{A}$ defined using random features $\Phi(\mathbf{x}) = (\phi(\mathbf{x}, \boldsymbol{\omega}_1), \cdots, \phi(\mathbf{x}, \boldsymbol{\omega}_m))^\top$ where $\boldsymbol{\omega}_i \sim \Lambda$. Assume that each $\phi(\cdot, \boldsymbol{\omega}_i)$, $i \in [\![m]\!]$, is L_i-Lipschitz with respect to the metric used to define the Wasserstein distance. If $\mathcal{A}$ satisfies (Hölder-LRIP) with error $\eta = 0$, constant $C > 0$ and $\delta = 1$ then we have:*

$$\forall \pi, \pi \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C' \, W_p(\pi, \pi') \tag{28}$$

where $C' = C\sqrt{\sum_{i=1}^m L_i^2}$. In other words, if $\mathcal{A}$ satisfies the LRIP ($\delta = 1$) then $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable w.r.t. $\mathfrak{S}$.

The proof is deferred to Appendix B.2. In particular this proposition applies when $\phi(\cdot, \boldsymbol{\omega}_i) = \exp(-i\langle \cdot, \boldsymbol{\omega}_i \rangle)$ which is $\|\boldsymbol{\omega}_i\|_2$-Lipschitz with respect to the Euclidean norm. Consequently, the Wasserstein learnability is necessary to the LRIP when Φ is defined using RFF as in (Gribonval et al., 2021a). In other words, if one hopes that an LRIP property ($\delta = 1$) holds, it is of interest to understand what kind of model sets and tasks satisfy the Wasserstein learnability. This is the goal of the next section and we provide a summary of the different results in Table 1.

Remark 2 (Particular case $\mathfrak{S} = \mathcal{P}_p(\mathcal{X})$)**.** *A interesting special case of Definition 2 is when then task is p-Wasserstein learnable. Then 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:*

$$\mathcal{R}^{1/p}(\pi, \hat{h}) - \mathcal{R}^{1/p}(\pi, h^*) \leq 2CW_p(\pi, \pi_n) \tag{29}$$

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[9]. Therefore, the smaller the Wasserstein distance between π_n and π, the better $\hat{h}$ is.

9. This can be proved by noting that $\mathcal{R}^{1/p}(\pi, \hat{h}) - \mathcal{R}^{1/p}(\pi, h^*) = \left\{\mathcal{R}^{1/p}(\pi, \hat{h}) - \mathcal{R}^{1/p}(\pi_n, \hat{h})\right\} + \left\{\mathcal{R}^{1/p}(\pi_n, \hat{h}) - \mathcal{R}^{1/p}(\pi_n, h^*)\right\} + \{\mathcal{R}^{1/p}(\pi_n, h^*) - \mathcal{R}^{1/p}(\pi, h^*)\}$. Since $\mathcal{R}^{1/p}(\pi_n, h^*) - \mathcal{R}^{1/p}(\pi_n, \hat{h}) \leq 0$ by

3.1 Parametric density estimation with a Wasserstein loss

The most straightforward case of Wasserstein Learnability is when the risk *itself* can be rewritten as a Wasserstein distance. In this section we consider the important statistical problem of fitting densities, *i.e.* estimating the parameters of a chosen model that meaningfully fits the observed data. A prime example of such a learning task is *Gaussian Mixture Modeling* (GMM) (Dasgupta, 1999) where one wants to find a linear combination of K Gaussian density functions such that the true distribution π is well described by this combination. We can also cite Generative Adversarial Netwoks (GAN) (Goodfellow et al., 2014) where one wants to fit a distribution which is parametrized by a neural network.

In all of these cases the hypothesis space $\mathcal{H} \subseteq \mathbb{R}^M$ is simply a set of parameters, *e.g.* containing the variances and means of the Gaussian components and the terms of the linear combination for GMM or describing the weights of the neural network for GANs. Formally the goal is to find $h \in \mathbb{R}^M$ (*i.e.* the parameters) such that the parametrized distribution π_h best fits the data generating distribution π. To find the hypothesis $h \in \mathcal{H}$ a principled way is to consider the negative likehood loss function $\ell(\mathbf{x}, h) = -\log(\pi_h(\mathbf{x}))$. In practice this corresponds to minimizing over $h \in \mathcal{H}$ the risk $\mathcal{R}(\pi, h) = \mathrm{KL}(\pi||\pi_h) + H(\pi)$ where KL is the Kullback-Leibler divergence and H the entropy.

While this is the most common strategy, this approach is sometimes flawed so that alternatives to the KL fitting criterion have emerged. As described in many contexts such as generative modeling (Genevay et al., 2018; Frogner et al., 2015; Arjovsky et al., 2017) or deconvolution problems (Nguyen, 2013; Rigollet and Weed, 2018; Caillerie et al., 2011; Scricciolo, 2017; Dedecker and Michel, 2013) the Wasserstein distance, or its entropic regularized counterpart, is quite suited. The problem of density estimation in these cases often boils down to minimize the following risks:

$$\mathcal{R}(\pi, h) = W_1(\pi, \pi_h) \tag{30}$$

which corresponds to finding the parametrized distribution π_h closest to π in Wasserstein distance. Using the metric property of the Wasserstein distance (triangle inequality) it is easy to check that for any $\pi, \pi' \in \mathcal{P}_1(\mathcal{X})$ and $h \in \mathcal{H}$:

$$|\mathcal{R}(\pi, h) - \mathcal{R}(\pi', h)| = |W_1(\pi, \pi_h) - W_1(\pi', \pi_h)| \leq W_1(\pi, \pi') \tag{31}$$

Hence we can conclude that $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq W_1(\pi, \pi')$ for any $\pi, \pi' \in \mathcal{P}_1(\mathcal{X})$. In other words, the problem of density estimation using the Wasserstein distance is Wasserstein learnable with constant 1, and this independently of the model set $\mathfrak{S} = \mathcal{P}_1(\mathbb{R}^d)$.

3.2 Compression-type tasks are Wasserstein learnable

A wide range of unsupervised learning tasks can also be recast as a learning problem which involves the Wasserstein distance. Indeed many unsupervised 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 the Wasserstein distance (Canas and Rosasco, 2012). Such problems will be very connected with *compression-type tasks* as defined below :

Definition 3 (Compression-type tasks (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 i.e. 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 PCA and K-means *compression-type tasks*. These two problems are actually related to a wider class of problems, namely k-dimensional coding schemes which

definition of $\hat{h}$ we have $\mathcal{R}^{1/p}(\pi, \hat{h}) - \mathcal{R}^{1/p}(\pi, h^*) \leq 2\sup_{h \in \mathcal{H}} |\mathcal{R}^{1/p}(\pi, h) - \mathcal{R}^{1/p}(\pi_n, h)| \leq 2CW_p(\pi, \pi_n)$ by hypothesis.

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 $\|.\|$) 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 which defines a implementation of the codebook. It corresponds to a projection function which satisfies $\|\mathbf{x} - P_h(\mathbf{x})\|^2 = \min_{\mathbf{y}\in Y} \|\mathbf{x} - h\mathbf{y}\|^2$. In particular, non-negative matrix factorization (NMF) (Lee and Seung; 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 corresponds 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 1 (Canas and Rosasco (2012)). *Consider $S \subseteq \mathcal{X}$, $p \in [1, +\infty[$ 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 we have:*

$$\mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = W_p^p(\pi, P_S\#\pi) \tag{32}$$

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

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 learnable, *i.e.* that the task-specific norm $\|.\|_{\mathcal{L}(\mathcal{H}),p}$ can be bounded by a Wasserstein distance.

Proposition 4 (Compression-type tasks are Wasserstein learnable). *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 3. Then we have:*

$$\forall h \in \mathcal{H}, \pi \in \mathcal{P}_p(\mathcal{X}), \mathcal{R}(\pi, h) = W_p^p(\pi, P_h\#\pi) \tag{33}$$

$$\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq W_p(\pi, \pi') \tag{34}$$

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 1 we have for $\pi \in \mathcal{P}_p(\mathcal{X})$:

$$\mathcal{R}(\pi, h) = \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x}, h)] = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_h(\mathbf{x}))^p] = W_p^p(\pi, P_h\#\pi) \tag{35}$$

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

$$\begin{aligned}\mathcal{R}(\pi, h)^{1/p} - \mathcal{R}(\pi', h)^{1/p} &= W_p(\pi, P_h\#\pi) - W_p(\pi', P_h\#\pi') \leq W_p(\pi, P_h\#\pi') - W_p(\pi', P_h\#\pi') \\ &\leq W_p(\pi, \pi')\end{aligned} \tag{36}$$

where we used $W_p(\pi, P_h\#\pi) \leq W_p(\pi, \nu)$ if $\mathrm{supp}(\nu) \subseteq S$ (Lemma 1) 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 $|\mathcal{R}(\pi, h)^{1/p} - \mathcal{R}(\pi', h)^{1/p}| \leq W_p(\pi, \pi')$. Taking the supremum over $h \in \mathcal{H}$ concludes. ■

3.3 Linear regression tasks are Wasserstein learnable

Appart from the unsupervised learning tasks discussed above, some supervised tasks are also Wasserstein learnable. To begin with, let us consider the problem of linear regression with a scalar output. This corresponds to $\mathcal{X} = \mathbb{R}^{d+1}$ and a loss function defined by $\ell(\mathbf{x} = (\mathbf{z}, y), \boldsymbol{\theta}) = (y - \boldsymbol{\theta}^\top \mathbf{z})^2$ where $\mathbf{z} \in \mathbb{R}^d, \boldsymbol{\theta} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ is the ouput. To write this loss in a more convinient way consider the hypothesis space $\mathcal{H} = \{\overline{\boldsymbol{\theta}} = (\boldsymbol{\theta}, -1) \in \mathbb{R}^{d+1}; \boldsymbol{\theta} \in \mathbb{R}^d, \|\boldsymbol{\theta}\|_2 \leq R\}$ where $\|.\|_2$ is the ℓ_2 norm in $\mathbb{R}^d$ and $R > 0$. For $\overline{\boldsymbol{\theta}} \in \mathcal{H}$ the loss can be rewritten as $\ell(\mathbf{x}, \overline{\boldsymbol{\theta}}) = (\overline{\boldsymbol{\theta}}^\top \mathbf{x})^2$ and we have $\|\overline{\boldsymbol{\theta}}\|_{2,\mathbb{R}^{d+1}}^2 = \|\boldsymbol{\theta}\|_2^2 + 1$ where $\|.\|_{2,\mathbb{R}^{d+1}}$ is the ℓ_2 norm on $\mathbb{R}^{d+1}$. It is easy to check by Cauchy Swartz

in $\mathbb{R}^{d+1}$ that $\left||\overline{\boldsymbol{\theta}}^\top \mathbf{x}| - |\overline{\boldsymbol{\theta}}^\top \mathbf{x}'|\right| \leq \sqrt{R^2+1}\|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+1}}$ so that the loss $\ell(\cdot,\boldsymbol{\theta})$ is the square of a Lipschitz function. This property will allow us to prove that the task is Wasserstein learnable by using the following lower-bound on the Wasserstein distance:

Proposition 5 (Proposition 7.29 in (Villani, 2008)). *Let $(\mathcal{X}, D)$ be a complete separable metric space, $p \geq 1$ and $\pi, \pi' \in \mathcal{P}_p(\mathcal{X})$. Then for any $\phi \in \mathrm{Lip}_L(\mathcal{X}, \mathbb{R}_+)$:*

$$\left| \left(\int \phi(\mathbf{x})^p \mathrm{d}\pi(\mathbf{x}) \right)^{1/p} - \left(\int \phi(\mathbf{y})^p \mathrm{d}\pi'(\mathbf{y}) \right)^{1/p} \right| \leq L W_p(\pi, \pi') \tag{37}$$

With this theorem in mind, consider for $\overline{\boldsymbol{\theta}} \in \mathcal{H}$ the function $\phi_{\overline{\boldsymbol{\theta}}} : \mathbb{R}^{d+1} \to \mathbb{R}_+$ defined by $\phi_{\overline{\boldsymbol{\theta}}}(\mathbf{x}) = |\overline{\boldsymbol{\theta}}^\top \mathbf{x}|$. By the previous reasoning this function satisfies $|\phi_{\overline{\boldsymbol{\theta}}}(\mathbf{x}) - \phi_{\overline{\boldsymbol{\theta}}}(\mathbf{x}')| \leq \sqrt{R^2+1}\|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+1}}$ and thus is $\sqrt{R^2+1}$-Lipschitz from $(\mathbb{R}^{d+1}, \|.\|_{2,\mathbb{R}^{d+1}})$ to $\mathbb{R}_+$. Using Proposition 5, we obtain:

$$\left| \left(\int \phi_{\overline{\boldsymbol{\theta}}}(\mathbf{x})^2 \mathrm{d}\pi(\mathbf{x}) \right)^{1/2} - \left(\int \phi_{\overline{\boldsymbol{\theta}}}(\mathbf{y})^2 \mathrm{d}\pi'(\mathbf{y}) \right)^{1/2} \right| \leq \sqrt{R^2+1} W_2(\pi, \pi') \tag{38}$$

for any $\pi, \pi' \in \mathcal{P}_2(\mathbb{R}^{d+1})$. It suffices now to notice that $\phi_{\overline{\boldsymbol{\theta}}}(\mathbf{x})^2 = (\overline{\boldsymbol{\theta}}^\top \mathbf{x})^2 = \ell(\mathbf{x}, \overline{\boldsymbol{\theta}})$ so that $|(\int \phi_{\overline{\boldsymbol{\theta}}}^2 \mathrm{d}\pi)^{1/2} - (\int \phi_{\overline{\boldsymbol{\theta}}}^2 \mathrm{d}\pi')^{1/2}| = |\mathcal{R}(\pi, \overline{\boldsymbol{\theta}})^{1/2} - \mathcal{R}(\pi', \overline{\boldsymbol{\theta}})^{1/2}|$. Overall, we have the following result:

Proposition 6 (Linear regression is Wasserstein learnable). *Consider $\mathcal{X} = \mathbb{R}^{d+1}$ and the linear regression loss $\ell(\mathbf{x} = (\mathbf{z}, y), \boldsymbol{\theta}) = |y - \boldsymbol{\theta}^\top \mathbf{z}|^2$ along with the hypothesis space $\mathcal{H} = \{\boldsymbol{\theta} \in \mathbb{R}^d, \|\boldsymbol{\theta}\|_2 \leq R\}$ where $R > 0$. The task $\mathcal{L}(\mathcal{H})$ is 2-Wasserstein learnable with constant $\sqrt{R^2+1}$, i.e. :*

$$\forall \pi, \pi' \in \mathcal{P}_2(\mathbb{R}^{d+1}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),2} \leq \sqrt{R^2+1} W_2(\pi, \pi') \tag{39}$$

where the Wasserstein distance is computed with the distance $D(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+1}}$ where $\|.\|_{2,\mathbb{R}^{d+1}}$ is the ℓ_2 norm on $\mathbb{R}^{d+1}$.

A straightforward extension of this result is when the loss writes instead $\ell(\mathbf{x} = (\mathbf{z}, y), \boldsymbol{\theta}) = |y - \boldsymbol{\theta}^\top \mathbf{z}|^p$ for any $p \geq 1$. By using the same argument it is easy to see that this task is p-Wasserstein learnable *i.e.*

$$\forall \pi, \pi' \in \mathcal{P}_p(\mathbb{R}^{d+1}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \sqrt{R^2+1} W_p(\pi, \pi') \tag{40}$$

Interestingly this result generalizes also for "multi-outputs" regression *i.e.* when we wish to predict a vector $\mathbf{y} = (y_1, \cdots, y_K) \in \mathbb{R}^K$. In this case the hypothesis space is a space of matrices and the loss can be written for $\mathbf{x} = (\mathbf{z}, \mathbf{y}) \in \mathbb{R}^{d+K}$ as $\|\mathbf{y} - \mathbf{M}\mathbf{z}\|_{2,\mathbb{R}^K}^p$ where $\mathbf{M} \in \mathbb{R}^{K \times d}$ and $p \geq 1$. In the same way we can define $\overline{\mathbf{M}} = \begin{pmatrix} \mathbf{M} & -\mathbf{I}_{K,K} \end{pmatrix} \in \mathbb{R}^{K \times (d+K)}$ where $\mathbf{I}_{K,K}$ is the identity matrix of size $\mathbb{R}^{K \times K}$ so that $\ell(\mathbf{x}, \overline{\mathbf{M}}) = \|\overline{\mathbf{M}}\mathbf{x}\|_{2,\mathbb{R}^K}^p$. We have:

Proposition 7 (Multi-outputs Linear regression task is Wasserstein learnable). *Consider $\mathcal{X} = \mathbb{R}^{d+K}$ and the multi-output linear regression loss $\ell(\mathbf{x} = (\mathbf{z}, \mathbf{y}), \mathbf{M}) = \|\mathbf{y} - \mathbf{M}\mathbf{z}\|_{2,\mathbb{R}^K}^p$ for $p \geq 1$ along with the hypothesis space $\mathcal{H} = \{\mathbf{M} \in \mathbb{R}^{K \times d}, \|\mathbf{M}\|_{2\to 2} \leq R\}$ where $R > 0$ and $\|\cdot\|_{2\to 2}$ is the 2-operator norm for matrices. Then the task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable with constant $\sqrt{R^2+1}$, i.e. :*

$$\forall \pi, \pi' \in \mathcal{P}_p(\mathbb{R}^{d+K}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \sqrt{R^2+1} W_p(\pi, \pi') \tag{41}$$

where the Wasserstein distance is computed with the distance $D(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+K}}$ where $\|.\|_{2,\mathbb{R}^{d+K}}$ is the ℓ_2 norm on $\mathbb{R}^{d+K}$.

Proof We have $\left|\|\overline{\mathbf{M}}\mathbf{x}\|_{2,\mathbb{R}^K} - \|\overline{\mathbf{M}}\mathbf{x}'\|_{2,\mathbb{R}^K}\right| \leq \|\overline{\mathbf{M}}(\mathbf{x}-\mathbf{x}')\|_{2,\mathbb{R}^K} \leq \|\overline{\mathbf{M}}\|_{2\to 2}\|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+K}}$. As such if $\|\mathbf{M}\|_{2\to 2} \leq R$ then $\|\overline{\mathbf{M}}\|_{2\to 2}^2 \leq R^2+1$ so that $|\|\overline{\mathbf{M}}\mathbf{x}\|_{2,\mathbb{R}^K} - \|\overline{\mathbf{M}}\mathbf{x}'\|_{2,\mathbb{R}^K}| \leq \sqrt{R^2+1}\|\mathbf{x}-\mathbf{x}'\|_{2,\mathbb{R}^{d+K}}$. We can use the same reasoning as before by defining the function $\phi_{\overline{\mathbf{M}}}(\mathbf{x}) = \|\overline{\mathbf{M}}\mathbf{x}\|_{2,\mathbb{R}^K}$ which is $\sqrt{R^2+1}$-Lipschitz whenever $\|\mathbf{M}\|_{2\to 2} \leq R$, hence by using the bound from Proposition 5 we can conclude. ■

3.4 Regression tasks beyond linear regression: multi-layer perceptron (MLP)

With the previous reasoning in mind, a straightforward calculus shows that if we consider general Lipschitz regressors with uniformly bounded Lipschitz constant then the regression task (with the ℓ_p loss) will be p-Wasserstein learnable:

Lemma 2. *Consider $\mathcal{X} = \mathbb{R}^{d+K}$ and the regression loss $\ell(\mathbf{x}=(\mathbf{z},\mathbf{y}),h) = \|\mathbf{y}-h(\mathbf{z})\|_{2,\mathbb{R}^K}^p$ for $p \geq 1$ along with the hypothesis space $\mathcal{H} \subseteq \mathrm{Lip}_{\mathrm{L}}(\mathbb{R}^d,\mathbb{R}^K)$. Then the task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable with constant $\max\{L,1\}$, i.e. :*

$$\forall \pi,\pi' \in \mathcal{P}_p(\mathbb{R}^{d+K}), \|\pi-\pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \max\{L,1\}W_p(\pi,\pi') \tag{42}$$

where the Wasserstein distance is computed with the distance $D(\mathbf{x}=(\mathbf{z},\mathbf{y}),\mathbf{x}'=(\mathbf{z}',\mathbf{y}')) = \|\mathbf{z}-\mathbf{z}'\|_{2,\mathbb{R}^d} + \|\mathbf{y}-\mathbf{y}'\|_{2,\mathbb{R}^K}$.

Proof We have for $\mathbf{x}=(\mathbf{z},\mathbf{y}),\mathbf{x}'=(\mathbf{z}',\mathbf{y}')$:

$$\begin{aligned} |\|\mathbf{y}-h(\mathbf{z})\|_{2,\mathbb{R}^K} - \|\mathbf{y}-h(\mathbf{z})\|_{2,\mathbb{R}^K}| &\leq \|\mathbf{y}-\mathbf{y}'-(h(\mathbf{z})-h(\mathbf{z}'))\|_{2,\mathbb{R}^K} \\ &\leq \|\mathbf{y}-\mathbf{y}'\|_{2,\mathbb{R}^K} + L\|h(\mathbf{z})-h(\mathbf{z}')\|_{2,\mathbb{R}^d} \\ &\leq \max\{L,1\}D(\mathbf{x}=(\mathbf{z},\mathbf{y}),\mathbf{x}'=(\mathbf{z}',\mathbf{y}')) \end{aligned} \tag{43}$$

Such that the loss can be written as the p-th power of a Lipschitz function, hence by using the bound from Proposition 5 we can conclude. ■

In particular this situation encompasses regressors such as 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 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 $\Pi_{j=1}^J\|\mathbf{M}_j\|_{2\to 2}$ (Virmaux and Scaman, 2018). 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 learnable as soon as $\|\mathbf{M}_j\|_{2\to 2} \leq R$ for some $R > 0$ which gives a (naive) Lipschitz constant $L = R^J$

3.5 Binary classification tasks with a Lipschitz classifier are Wasserstein learnable

Binary classifications tasks can also be related to Wasserstein learnability. These problems corresponds to $\mathcal{X} = \mathbb{R}^d \times \{+1,-1\}$. In binary classification one often relies on convex surrogates of the $0-1$ loss and considers losses of the type $\ell(\mathbf{x}=(\mathbf{z},y),h) = \varphi(yh(\mathbf{z}))$ where $y \in \{-1,+1\}$, $h : \mathbb{R}^d \to [-1,+1]$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is convex. Well known examples include the logistic loss $\varphi(t) = \log(1+e^{-t})$ the hinge loss $\varphi(t) = \max(1-t,0)$ or the squared hinge loss $\varphi(t) = \max(1-t,0)^2$.

Suppose now that φ is also L_φ-Lipschitz (such as the hinge loss) and that the hypothesis is L-lipschitz. This includes hypothesis of the type $h(\mathbf{z}) = \rho(f_{MLP}(\mathbf{z}))$ where $f_{MLP} : \mathbb{R}^d \to \mathbb{R}$ is an

MLP with bounded parameters, with 1-Lipschitz activation functions as described in Section 3.4 and $\rho : \mathbb{R} \to [-1, 1]$ is an "output-layer" function that is Lipschitz such as the sigmoid function $\rho(t) = \frac{1}{1+e^{-t}}$.

In this setting we can apply the duality argument of the Wasserstein distance and prove the following result:

Proposition 8 (Binary classification tasks with Lipschitz classifier)**.** *Consider a binary classification problem where $\mathcal{X} = \mathbb{R}^d \times \{+1, -1\}$. Consider the hypothesis space $\mathcal{H} \subseteq \mathrm{Lip}_L(\mathbb{R}^d, [-1, 1])$. Consider a convex surrogate loss defined for $\mathbf{x} = (\mathbf{z}, y) \in \mathbb{R}^d \times \{+1, -1\}$ and $h \in \mathcal{H}$ by $\ell(\mathbf{x} = (\mathbf{z}, y), h) = \varphi(yh(\mathbf{z}))$ where $\varphi : \mathbb{R} \to \mathbb{R}_+$ is L_φ-Lipschitz. Then $\mathcal{L}(\mathcal{H})$ is 1-Wasserstein learnable with constant $L_\varphi \max(L, 1)$ i.e. :*

$$\forall \pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d \times \{+1, -1\}), \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq L_\varphi \max(L, 1) W_1(\pi, \pi') \tag{44}$$

where the Wasserstein distance is computed with the distance $D((\mathbf{z}, y), (\mathbf{z}', y')) = \|\mathbf{z} - \mathbf{z}'\|_2 + |y - y'|$ on $\mathbb{R}^d \times \{+1, -1\}$.

Proof For any $\mathbf{x} = (\mathbf{z}, y), \mathbf{x}' = (\mathbf{z}', y')$ we have:

$$\begin{aligned} |\varphi(yh(\mathbf{z})) - \varphi(y'h(\mathbf{z}'))| &\leq |\varphi(yh(\mathbf{z})) - \varphi(yh(\mathbf{z}'))| + |\varphi(yh(\mathbf{z}')) - \varphi(y'h(\mathbf{z}'))| \\ &\leq L_\varphi(|yh(\mathbf{z}) - yh(\mathbf{z}')| + |yh(\mathbf{z}') - y'h(\mathbf{z}')|) \\ &\leq L_\varphi(|y||h(\mathbf{z}) - h(\mathbf{z}')| + |h(\mathbf{z}')||y - y'|) \\ &\leq L_\varphi \max(L, 1)(\|\mathbf{z} - \mathbf{z}'\|_2 + |y - y'|). \end{aligned} \tag{45}$$

Thus, $|\ell(\mathbf{x} = (\mathbf{z}, y), h) - \ell(\mathbf{x}' = (\mathbf{z}', y'), h)| \leq L_\varphi \max(L, 1) D((\mathbf{z}, y), (\mathbf{z}', y'))$ where $D((\mathbf{z}, y), (\mathbf{z}', y')) = \|\mathbf{z} - \mathbf{z}'\|_2 + |y - y'|$ is a distance on $\mathcal{X} = \mathbb{R}^d \times \{-1, +1\}$. In this way for all $h \in \mathcal{H}$ the loss $\ell(\cdot, h)$ is $L_\varphi \max(L, 1)$-Lipchitz and we can use the duality representation of the 1-Wasserstein distance (26) to conclude that the task is 1-Wasserstein learnable with constant $L_\varphi \max(L, 1)$. ■

Remark 4. *It is sufficient that $\varphi = g^p$ for $p \geq 1$ where g is L_g Lipschitz to obtain Wasserstein learnability. Indeed we can use the same reasoning as in the previous section, based on Proposition 5, to obtain that $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable i.e. $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq L_\varphi \max(L, 1) W_p(\pi, \pi')$. In particular this includes the case of the squared hinge loss $\varphi(t) = \max(1 - t, 0)^2$.*

3.6 Conclusion on Wasserstein learnability

We have established in this section various controls of the task-specific metric by a Wasserstein distance. These bounds are based on the Wasserstein learnability property and encompasses tasks where CSL is known to provide guarantees (K-means, GMM) (Gribonval et al., 2021a,b). Interestingly enough, little is known in CSL concerning tasks such as regression and classification tasks that are also considered here. The advantages of the previous results are that they give a bound on $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \lesssim W_p(\pi, \pi')$ *for every* $\pi, \pi' \in \mathcal{P}_p(\mathcal{X})$ for various tasks $\mathcal{L}(\mathcal{H})$. We would like also to emphasize that these bounds *do not require any restriction to a model set* $\mathfrak{S}$.

4. Controlling Wasserstein distances by kernel norms

We focus in this section on the second step of our reasoning, that is comparing Optimal Transport distances (see definition in Section 3) and Maximum Mean Discrepancies (see Section 2). The goal of this section is to find reasonable conditions on a model set of distributions $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ and on a *p.s.d.* kernel κ such that the Wasserstein distance can be controlled with the MMD with kernel κ. To formalize we adopt the following definition:

Definition 3. *Let $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ be a model set, κ a p.s.d. kernel on $\mathcal{X}$ and $\delta \in]0,1]$. We say that the space $(\mathfrak{S}, W_p)$ is (κ,δ)-embeddable with error $\eta \geq 0$ if:*

$$\exists C > 0, \forall \pi, \pi' \in \mathfrak{S}, W_p(\pi,\pi') \leq C\|\pi - \pi'\|_\kappa^\delta + \eta \tag{46}$$

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

Note that the constants C, η, δ in (46) **do not depend** on the probability distributions π, π': we want to bound uniformly on the whole model set $\mathfrak{S}$. Moreover, as advertised in Section 2, we are particularly interested in establishing such an inequality for translation invariant p.s.d. kernels that at the core of the compressive statistical learning theory. Indeed these kernels admit a random feature expansion useful to find a sketching operator based on random features (Gribonval et al., 2021a). Let us emphasize that, unlike in the previous section where the comparisons between task-based metrics $\|.\|_{\mathcal{L}(\mathcal{H}),p}$ and Optimal Transport distances W_p were valid regardless on the choice of the model set (*i.e.*, valid for any $\pi, \pi' \in \mathcal{P}_p(\mathcal{X})$), here and in the next section the comparison is restricted to model sets $\mathfrak{S}$ with certain properties such as regularity. Section 5 will further strengthen the assumptions on these model sets to obtain finite dimensional embeddings.

Remark 5. *An immediate consequence of this definition is that when $(\mathfrak{S}, W_p)$ 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)*

In this section we focus on property (46) with no error *i.e.* $\eta = 0$. First we consider necessary conditions, i.e., we argue that property (46) 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 $(\mathfrak{S}, W_p)$ is (κ,δ)-embeddable. Unless stated otherwise, we consider in the following that $\mathcal{X} = \mathbb{R}^d$ is endowed with the ℓ_2 norm, which is used to define Wasserstein distances.

4.1 Boundedness of the model set is necessary.

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

$$\mathrm{m}(\pi) := \int \mathbf{x} \mathrm{d}\pi(\mathbf{x}) \tag{47}$$

the mean of $\pi \in \mathcal{P}_1(\mathbb{R}^d)$. On the one hand, simple calculus (Lemma 13 in Appendix C.2) shows that for any $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ and $p \in [1, +\infty[$:

$$W_p(\pi,\pi') \geq \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2. \tag{48}$$

On the other hand, if κ is a *bounbed* p.s.d. kernel (*i.e.*, $\sup_{\mathbf{x}} \kappa(\mathbf{x},\mathbf{x}) \leq K < +\infty$) then, by the Cauchy-Schwarz inequality for kernels $|\kappa(\mathbf{x},\mathbf{y})| \leq \sqrt{\kappa(\mathbf{x},\mathbf{x})}\sqrt{\kappa(\mathbf{y},\mathbf{y})} \leq K$ for every $\mathbf{x}, \mathbf{y}$ hence for any $(\pi,\pi') \in \mathfrak{S}$ we have $\|\pi - \pi'\|_\kappa \leq \sqrt{2K}$. As a result, if $\mathfrak{S}$ is unbounded in the sense that $\sup_{\pi,\pi' \in \mathfrak{S}} \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2 = +\infty$ then for each $\delta > 0$ we have

$$\sup_{(\pi,\pi') \in \mathfrak{S}} \frac{W_p(\pi,\pi')}{\|\pi - \pi'\|_\kappa^\delta} = +\infty \tag{49}$$

hence can not hope to have (46) for any $\delta > 0$. In other words, we have shown that:

Lemma 6. *If κ is bounded and $(\mathfrak{S}, W_p)$ is (κ,δ)-embeddable for some $\delta > 0$ then $\mathfrak{S}$ is necessarily bounded in the sense that* $\mathrm{m\text{-}diam}(\mathfrak{S}) := \sup_{\pi,\pi' \in \mathfrak{S}} \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2 < \infty$.

4.2 Bounds on δ due to the convergence rate of empirical measures.

Another obstacle to (46) 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 $\mathbb{E}[W_1(\pi, \pi_n)] \gtrsim n^{-1/d}$ 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, 2017). By monotonicity of W_p in p this is also true for W_p with $p \geq 1$ (since for $p \leq q, W_p(\pi, \pi) \leq W_q(\pi, \pi')$ for any π, π'[10]). On the contrary, it is not difficult to see that if the p.s.d. kernel κ is bounded by K then $\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta] \leq (2K)^{\delta/2} n^{-\delta/2}$ (see Lemma 12 in Appendix C.1). Consequently, even when the model set $\mathfrak{S} \subseteq \mathcal{P}_1(\mathbb{R}^d)$ satisfies $\mathrm{m\text{-}diam}(\mathfrak{S}) < +\infty$ (to avoid the obstacles to (46) already identified in Lemma 6), 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 (46) implies $n^{-1/d} \lesssim n^{-\delta/2}$, hence $\delta \leq 2/d$. An example of such a model set is the set of all probability distributions producing almost surely vectors in a prescribed Euclidean ball, leading to the following result:

Lemma 3. *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 p.s.d. kernel, and $p \in [1, +\infty[$. If $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable then $\delta \leq 2/d$.*

Remark 7. *If $\mathfrak{S} \subseteq \mathfrak{S}'$ where $(\mathfrak{S}', W_p)$ is (κ, δ)-embeddable then $(\mathfrak{S}, W_p)$ 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 context of CSL, as described in Section 2, such $\delta \leq 2/d$ would imply in a very slow convergence rate of the order of $O(n^{-\frac{1}{d}})$ at best. In other words, if the strategy described in Section 2 is followed we would require a exponential amount of samples in order to have reasonable CSL guarantees which is problematic for 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.

4.3 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 can not hope to have (46) with error $\eta = 0$ when $p \cdot \delta > 1$.

Proposition 9. *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 p.s.d. 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 $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable then $\delta \leq 1/p$.*

The result is mostly based on (Niles-Weed and Berthet, 2020). Its proof in Appendix C.3 essentially amounts to showing (49) 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 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 p.s.d. kernel κ, $(\mathfrak{S}, W_p)$ is **not** (κ, δ)-embeddable for $\delta > 1/p$ when $\mathfrak{S}$ contains *e.g.* mixtures of two Diracs or more generally mixtures of two compactly supported distributions.

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

We postulate that a similar result holds when we can find $(\pi_0, \pi_1) \in \mathfrak{S}$ such that only a small faction of the mass of π_0 and π_1 can be put in-between $\mathrm{supp}(\pi_0)$ and $\mathrm{supp}(\pi_1)$. We emphasize that this result does not depend on the dimension of the ambient space and is true for any *p.s.d.* kernel.

4.4 Bound on δ for mixture models and smooth TI kernels

In most of the concrete applications, one often has to compare *discrete* distributions. We will show in this paragraph that the regularity of the kernel plays an important role when trying to control the Wasserstein distance with an MMD as in (46) for model sets made of discrete distributions. In the following we define for $M \in \mathbb{N}^*$ and $\Omega \subseteq \mathcal{X} = \mathbb{R}^d$ a non-empty set the space of mixtures of M diracs located in Ω

$$\mathfrak{S}_K(\Omega) := \{\sum_{i=1}^{M} a_i \delta_{\mathbf{x}_i}; a_i \in \mathbb{R}_+, \sum_{i=1}^{M} a_i = 1, \forall i \in [\![M]\!], \mathbf{x}_i \in \Omega\}. \tag{50}$$

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 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 that there is a trade-off between the Hölder exponent δ and the regularity of the kernel provided that the model set is rich enough to contain discrete distributions with enough diracs.

Theorem 2 *Consider a TI p.s.d. 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 $p \in [1, +\infty[$, $\mathbf{x}_0 \in \mathbb{R}^d$, $R > 0$ and $\Omega = B(\mathbf{x}_0, R)$.*
If $(\mathfrak{S}_{\lfloor \frac{k}{2} \rfloor + 1}(\Omega), W_p)$ 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. The proof in Appendix C.4 amounts to showing (49) as soon as $\delta > 2/k$. Theorem 2 shows 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 1. *Consider a TI p.s.d. 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 $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable, where $p \in [1, +\infty[$, 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, $(\mathfrak{S}, W_p)$ is *not* (κ, δ)-embeddable for any $\delta > 0$. In other words, it shows that finding a absolute constant $C > 0$ such that $W_p(\pi, \pi') \leq C \, \mathrm{MMD}_\kappa^\delta(\pi, \pi')$ 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 $\mathbb{R}^d$. 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 (46). We will show in the next sections how to obtain these types of control with additional hypotheses on the regularity of the distributions in $\mathfrak{S}$. Finally, from a CSL perspective when one considers sketching operators defined with random Fourier features (Section 2), Corollary 1 shows that when the model set is the space of K diracs, the strategy proposed in (23) can only be achieved when $\delta \leq 2/K$ without additional assumptions. Indeed in this case $\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 = \|\pi - \pi'\|_{\kappa_\Phi}$ where κ_Φ is the empirical kernel $\kappa_\Phi(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^m e^{-i \boldsymbol{\omega}_i^\top (\mathbf{x} - \mathbf{y})}$ where $\boldsymbol{\omega}_i \sim \Lambda$ and thus is can be written as $\kappa_\Phi(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ with κ_0 that is smooth. This suggests that the same separability assumptions that the one made in (Gribonval et al., 2021b) may help. The Figure 3 summarizes the necessary conditions established in the previous sections.

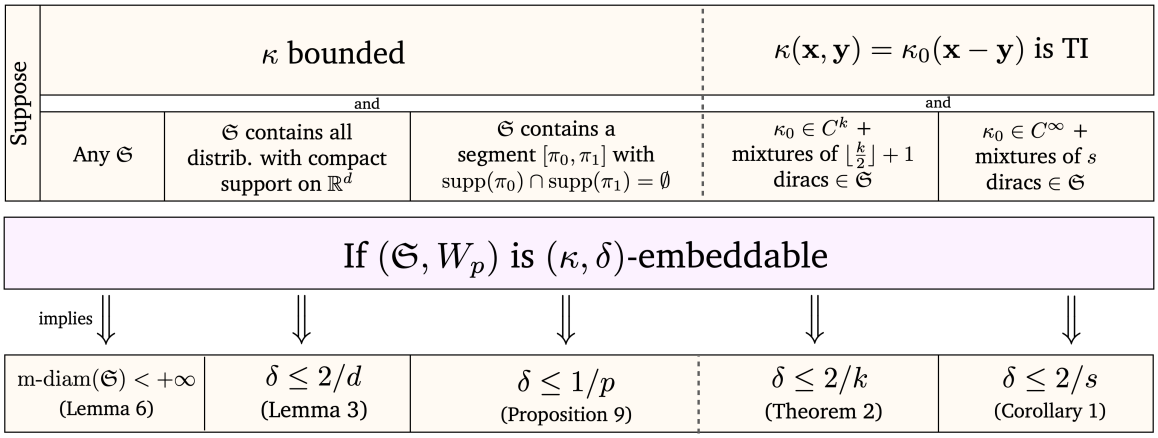

Figure 3: Summary of the established necessary conditions to the (κ,δ)-embedabbility property.

4.5 A study on the real line for TI kernels

In this section we restrict ourselves to the case of probability distributions on the real line $\mathbb{R}$ that admit a density with respect to the Lesbegue measure. We will prove that, under some regularity conditions on the kernel and as long as the probability distributions have the same mean and are regular enough (Sobolev) then we have have the control (46) with error $\eta = 0, \delta = 1/2$ and W_2. This result will be based on the closed-form expression of the Wasserstein distance on the real line which states that if $\pi, \pi' \in \mathcal{P}_p(\mathbb{R})$[12] admit a density with respect to the Lesbesgue measure in $\mathbb{R}$ then (Santambrogio, 2015):

$$W_p(\pi,\pi) = \left(\int_{\mathbb{R}} |F(x) - G(x)|^p \mathrm{d}x \right)^{1/p} \tag{51}$$

where the Wasserstein distance is computed using the distance $D(x,y) = |x-y|$ and F, G stand for the cumulative distribution functions (CDF) of the probability densities of π, π'. We recall that the Fourier transform of a integrable function f is defined by $\hat{f}(\omega) = \int_{\mathbb{R}} e^{-i\omega x} f(x) \mathrm{d}x$. The equation (51) allows us to connect the Wasserstein distance and any MMD associated with a TI p.s.d. kernel on $\mathbb{R}$ as shown in the next Lemma (the proof can be found in Appendix C.5):

Lemma 4. *Consider $\pi, \pi' \in \mathcal{P}_2(\mathbb{R})$ with densities f, g with respect to the Lesbegue measure, i.e. $\pi \ll f\mathrm{d}x, \pi' \ll g\mathrm{d}x$. Let $\kappa(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$. Then we have:*

$$W_2(\pi,\pi') \leq (2\pi)^{-1/4} \left(\int_{\mathbb{R}} \frac{|\hat{f}(\omega) - \hat{g}(\omega)|^2}{|\omega|^4 \hat{\kappa}_0(\omega)} \mathrm{d}\omega \right)^{1/4} \|\pi - \pi'\|_\kappa^{1/2} \tag{52}$$

where the Wasserstein distance is computed using $D(x,y) = |x-y|$.

The integral term in the previous lemma may be infinite and depends on π, π'. Using some additional assumptions on the kernel and on the regularity of the densities we are now able to bound this integral by a constant as described in the next theorem:

Theorem 3 *Let $\kappa(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$ and such that $\widehat{\kappa}_0(\omega) > 0$ for every ω, $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{q_\kappa})$ as $\omega \to 0$ and $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{s_\kappa})$ as $\omega \to +\infty$ for some*

12. We recall that if $\pi \in \mathcal{P}_p(\mathbb{R})$ then $\int_{\mathbb{R}} |x|^p \mathrm{d}\pi(x) < +\infty$ see (25)

$q_\kappa > -1, s_\kappa \in \mathbb{R}_+$. *Consider any* $s \geq \frac{s_\kappa}{2} + 1$, $0 < M < +\infty$ *and the following model set:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}_2(\mathbb{R}) : \exists f \in C^s(\mathbb{R}),\ \pi \ll f\mathrm{d}x,\ \|f\|_{W^{s,1}(\mathbb{R})} \leq M\} \tag{53}$$

1. *There exists* $C = C(M, s, \kappa) > 0$ *such that for every* $1 \leq p \leq 2$*:*

$$\forall \pi, \pi' \in \mathfrak{S},\ \textit{if } \mathrm{m}(\pi) = \mathrm{m}(\pi') \textit{ then } W_p(\pi, \pi') \leq C\|\pi - \pi'\|_\kappa^{1/2} \tag{54}$$

2. *If in addition* κ_0 *is L-Lipschitz continuous, then for every* $1 \leq p \leq 2$

$$\forall \pi, \pi' \in \mathfrak{S},\ W_p(\pi, \pi') \leq C\|\pi - \pi'\|_\kappa^{1/2} + CL^{1/4}|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|. \tag{55}$$

Remark 8. *On the one hand, the assumption* $\widehat{\kappa}_0(\omega) = O(\omega^{s_\kappa})$ *at infinity means that the Fourier transform of* κ_0 *should not decay too fast, i.e., it should not be too regular. On the other hand, the additional Lipschitz assumption essentially corresponds to assuming a sufficiently fast* $O(\omega^{-1})$ *decay of this Fourier transform.*

Following Remark 7, the same conclusion also holds for any model set that is *contained* in the model set $\mathfrak{S}$ defined in (53). The proof can be found in Appendix C.6.

The first claim of this theorem implies that $(\mathfrak{S}', W_p)$ is then $(\kappa, \frac{1}{2})$-embeddable as soon as $1 \leq p \leq 2$, κ is a TI p.s.d. kernel with some regularity, and the distributions in $\mathfrak{S}'$ are sufficiently regular with the same mean. The latter hypothesis is important in the proof in order to have the finiteness of the integral from Lemma 4. Note that thanks to the assumption on the means there is no pair of distributions in $\mathfrak{S}'$ with disjoint supports, hence this result does not contradict Proposition 9 (also $p \cdot \delta = 1$ in this case). Moreover since the distributions in $\mathfrak{S}'$ admit a density then the constraints of Theorem 2 do not apply here and, as such, the kernel can be smooth.

Focusing now our attention on a bounded model set $\mathfrak{S}'$ (in light of Lemma 6), the second claim of the theorem similarly implies that $(\mathfrak{S}', W_p)$ is $(\kappa, \frac{1}{2})$-embeddable *with error*

$$\eta \leq CL^{1/4}\,\text{m-diam}^{1/4}(\mathfrak{S}) + \text{m-diam}(\mathfrak{S}),$$

under slightly stronger regularity assumptions on the TI p.s.d. kernel κ, provided again that the distributions in $\mathfrak{S}'$ are sufficiently regular (but without requiring the same means). As we now show, we can actually obtain a result without error by slightly changing the kernel.

Corollary 2. *Consider* κ_0*,* s*,* M *and* $\mathfrak{S}$ *as in Theorem 3. Assume that* κ_0 *is L-Lipschitz continuous and that* $\mathfrak{S}' \subseteq \mathfrak{S}$ *satisfies* $\text{m-diam}(\mathfrak{S}') < +\infty$*. There is a constant* $C = C(M, s, \kappa, L, \text{m-diam}(\mathfrak{S}'))$ *such that for every* $1 \leq p \leq 2$*:*

$$\forall \pi, \pi' \in \mathfrak{S}',\ W_p(\pi, \pi') \leq C\|\pi - \pi'\|_{\tilde{\kappa}}^{1/2} \tag{56}$$

with the p.s.d. kernel $\tilde{\kappa}(x, y) := \kappa_0(x - y) + xy$*.*

Proof First, we can apply the second claim of Theorem 3 to obtain (59). Second, since $t = t^{1/4}t^{3/4} \leq t^{1/4}T^{3/4}$ for any $0 \leq t \leq T := \text{m-diam}(\mathfrak{S}')$, we have

$$\forall \pi, \pi' \in \mathfrak{S}',\ |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')| \leq |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4} \cdot \text{m-diam}^{3/4}(\mathfrak{S}').$$

hence there is a constant C_1 depending only on $C(M, s, \kappa)$, L, and $\text{m-diam}(\mathfrak{S}')$, such that for every $\pi, \pi' \in \mathfrak{S}'$ the right hand side in (59) is bounded by $C_1(\|\pi - \pi'\|_\kappa^{1/2} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4})$. Since $a + b \leq 2^{3/4}(a^4 + b^4)^{1/4}$ for every $a, b \geq 0$, we have

$$\|\pi - \pi'\|_\kappa^{1/2} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4} \leq 2^{3/4}\left(\|\pi - \pi'\|_\kappa^2 + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^2\right)^{1/4}.$$

To conclude, observe that

$$\begin{aligned}\|\pi-\pi'\|_\kappa^2+|\,\mathrm{m}(\pi)-\mathrm{m}(\pi')|^2 =&\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi}\kappa(x,y)-2\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi'}\kappa(x,y)+\mathbb{E}_{x\sim\pi'}\mathbb{E}_{y\sim\pi'}\kappa(x,y)\\&+\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi}xy-2\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi'}xy+\mathbb{E}_{x\sim\pi'}\mathbb{E}_{y\sim\pi'}xy\\=&\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi}\tilde\kappa(x,y)-2\mathbb{E}_{x\sim\pi}\mathbb{E}_{y\sim\pi'}\tilde\kappa(x,y)+\mathbb{E}_{x\sim\pi'}\mathbb{E}_{y\sim\pi'}\tilde\kappa(x,y)\\=&\|\pi-\pi'\|_{\tilde\kappa}.\end{aligned}$$

■

This shows that if $\mathfrak{S}'$ is made of sufficiently smooth distributions and $\mathrm{m\text{-}diam}(\mathfrak{S}')$ is finite, then for every $1\leq p\leq 2$ $(\mathfrak{S}', W_p)$ is $(\tilde\kappa, \frac{1}{2})$-embeddable. Notice that unlike the initial kernel κ, the modified one $\tilde\kappa$ (which is still p.s.d.) is no longer translation invariant. As described next (the proof can be found in Appendix C.8), the regularity assumptions on $\mathfrak{S}'$ can be met *e.g.* with certain Gaussian mixtures on $\mathbb{R}$.

Corollary 3 (GMM on $\mathbb{R}$). *Let $\kappa(x,y)=\kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0\in L_1(\mathbb{R})$ and such that $\widehat{\kappa}_0(\omega)>0$ for every ω, $\frac{1}{\widehat{\kappa}_0(\omega)}=O(\omega^{q_\kappa})$ as $\omega\to 0$ and $\frac{1}{\widehat{\kappa}_0(\omega)}=O(\omega^{s_\kappa})$ as $\omega\to+\infty$ for some $q_\kappa>-1, s_\kappa\in\mathbb{R}_+$. For $K\in\mathbb{N}^*$, $\Omega\subset\mathbb{R}$, and $\sigma_{\min}>0$ consider the model set:*

$$\mathfrak{S}_{\texttt{GMM}}(\Omega,K,\sigma_{\min}):=\{\pi=\sum_{k=1}^K\alpha_k\mathcal{N}(c_k,\sigma_k^2),\boldsymbol{\alpha}\in\Delta_K,c_k\in\mathbb{R},\sigma_k\geq\sigma_{\min},\sum_{k=1}^K\alpha_kc_k\in\Omega\}\tag{57}$$

where $\Delta_K=\{\boldsymbol{\alpha}\in\mathbb{R}_+^K,\sum_{k=1}^K\alpha_k=1\}$ is the probability simplex on $\mathbb{R}^K$.

1. *There exists a constant $C=C(\sigma_{\min},K,\kappa)>0$ such that if $\Omega=\{m\}$ is a prescribed mean then:*
$$\forall\pi,\pi'\in\mathfrak{S}_{\texttt{GMM}}(\Omega,K,\sigma_{\min}),W_2(\pi,\pi')\leq C\|\pi-\pi'\|_\kappa^{1/2}\tag{58}$$
2. *If in addition κ_0 is L-Lipschitz and $\mathrm{diam}(\Omega):=\sup_{x,y\in\mathcal{Y}}|x-y|<+\infty$, then for every $1\leq p\leq 2$:*
$$\forall\pi,\pi'\in\mathfrak{S},\ W_p(\pi,\pi')\leq C'\|\pi-\pi'\|_{\tilde\kappa}^{1/2}\tag{59}$$
with $\tilde\kappa(x,y):=\kappa_0(x-y)+xy$. The constant C' depends only on $C(\sigma_{\min},K,\kappa)$, L and $\mathrm{diam}(\Omega)$.

Proof [Sketch of proof] The goal is to show that any probability distribution in $\mathfrak{S}_{\texttt{GMM}}$ has a density which lies in a Sobolev ball with radius M. Indeed we can show that for $s=\lceil k/2+1\rceil$ the density F of any $\pi\in\mathfrak{S}$ satisfies $\|F\|_{W^{s,1}(\mathbb{R})}\leq\max(1,\sigma_{\min}^{1-s})\sum_{n=1}^s\sqrt{n!}$, independently of the choice of π. ■

Interestingly, this corollary shows that the space of GMM with K-mixtures is (κ,δ)-embeddable with $\delta=1/2$ *even without separation assumptions on the mixtures*. Note that this is true only when the means of the GMM mixtures are identical (but the variances can vary and are lower bounded) or for a kernel that is *not TI*. In the compressive learning context, this can be put in perspective with (Gribonval et al., 2021b) where separation assumptions on Dirac/Gaussian mixtures (with identical covariance matrices) are required in order to find a finite dimensional kernel mean embedding of the distributions that satisfy the LRIP.

Admissible TI kernels An important family of TI kernels satisfies the hypothesis of Theorem 3, that is the kernels of the Matérn class (Rasmussen and Williams, 2005, Section 4.2.1). These kernels are 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 admit the following Fourier transform[13] :

$$\widehat{\kappa_0}(\boldsymbol{\omega})=\frac{2^{d+\nu}\pi^{d/2}\Gamma(\nu+d/2)\nu^\nu}{\Gamma(\nu)\sigma^{2\nu}}\left(\frac{2\nu}{\sigma^2}+\|\boldsymbol{\omega}\|_2^2\right)^{-(\nu+d/2)}\tag{60}$$

13. This result be found in (Rasmussen and Williams, 2005, Section 4.2.1) with a slight modification due the conventions of the Fourier transforms.

Interestingly enough, when $\nu = \frac{1}{2}$ it gives 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 when $\nu \to +\infty$ it recovers the RBF kernel see (Rasmussen and Williams, 2005, Section 4.2.1)[14]. The Fourier transform of the RBF kernel however decays too fast to satisfy the assumptions of Theorem 3. In the context of compressive learning, translation invariant kernels are most useful if they can be approximated with random Fourier features (see Section 2) with good concentration properties. An interesting question for future work is thus whether the "slow decay" of the Fourier transform needed to apply Theorem 3 appears as a strong constraint in such a context.

4.6 From the real line to the Euclidean space: the case of compactly supported distributions

From the real line study of the previous section we can derive a control (46) in the $\mathbb{R}^d$ case. As we can no longer exploit the closed-form expression of the Wasserstein distances in terms of cumulative density functions (51), the idea is to use the connections between the Wasserstein distance and the Sliced-Wasserstein distance (SW) (Rabin et al., 2011; Kolouri et al., 2016) that enjoys the same topological properties than the Wasserstein distance for compactly supported measures (Bonnotte, 2013). This distance is defined as follows:

Definition 4 (Sliced-Wasserstein distance)**.** *Let $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$. For $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ (i.e.$\|\boldsymbol{\theta}\|_2 = 1$) we note $P_{\boldsymbol{\theta}}$ the function $P_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$. Let σ be the probability measure associated with the uniform distribution on the sphere $\mathbb{S}^{d-1}$. Then the Sliced-Wasserstein distance between π and π' is defined by:*

$$SW_1(\pi, \pi') := \mathbb{E}_{\boldsymbol{\theta} \sim \sigma}[W_1(P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi')] \tag{61}$$

We recall that $P_{\boldsymbol{\theta}}\#\pi$ is the probability meassure on $\mathbb{R}$ defined by $P_{\boldsymbol{\theta}}\#\pi(A) := \pi(P_{\boldsymbol{\theta}}^{-1}(A))$ for every measurable set $A \subseteq \mathbb{R}$. The key intuition behind SW is to randomly select lines in $\mathbb{R}^d$, to project the measures into these lines and to compute the resulting 1D-Wasserstein distance between $P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi'$ which can be done in closed-form and relies only on simple sorts of the supports (Peyré and Cuturi, 2019). The Sliced-Wasserstein distance admits also a useful alternative definition using the Radon transform (Helgason, 2011):

Definition 5 (Radon transform)**.** *Let $f \in L_1(\mathbb{R}^d)$. The Radon transform of f is defined for $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}^{d-1}$ by $\mathcal{R}[f](t, \boldsymbol{\theta}) = \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle = t} f(\mathbf{x}) \mathrm{d}\mathbf{x} = \int_{\mathbf{y} \in \boldsymbol{\theta}^\perp} f(t\boldsymbol{\theta} + \mathbf{y}) \mathrm{d}\mathbf{y}$.*

Based on this definition, when π and π' admit densities f and g with respect to the Lesbegue measure on $\mathbb{R}^d$ we have[15] (Kolouri et al., 2016; Rabin et al., 2011):

$$SW_1(\pi, \pi') = \mathbb{E}_{\boldsymbol{\theta} \sim \sigma}[W_1(\mathcal{R}[f](\cdot, \boldsymbol{\theta}), \mathcal{R}[g](\cdot, \boldsymbol{\theta}))] \tag{62}$$

As introduced, a fundamental result connects the Wasserstein distance and the Sliced-Wasserstein distance when the measures have compact supports:

Lemma 5 (Lemma 5.1.4 in (Bonnotte, 2013))**.** *Let $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ such that π, π' are supported in $B(0, R)$ for some $R > 0$. There exists a constant $C = C(d, R) > 0$ such that:*

$$W_1(\pi, \pi') \leq C\, SW_1(\pi, \pi')^{1/(d+1)} = C\, (\mathbb{E}_{\boldsymbol{\theta} \sim \sigma}[W_1(P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi')])^{1/(d+1)} \tag{63}$$

14. Likewise we have adapted the conventions of the Fourier transforms.
15. If f is a probability density function on $\mathbb{R}^d$ then $\mathcal{R}[f](\cdot, \boldsymbol{\theta})$ is a probability density function on $\mathbb{R}$. This is a consequence of Fubini's theorem: $\int_{\mathbb{R}} \mathcal{R}[f](t, \boldsymbol{\theta}) \mathrm{d}t = \int_{\mathbb{R}} \int_{\mathbf{y} \in \boldsymbol{\theta}^\perp} f(t\boldsymbol{\theta} + \mathbf{y}) \mathrm{d}\mathbf{y} \mathrm{d}t = \int_{\mathbb{R}^d} f(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$. As such, when $\pi \ll f\mathrm{d}\mathbf{x}$ the density of $P_{\boldsymbol{\theta}}\#\pi$ is exactly $\mathcal{R}[f](\cdot, \boldsymbol{\theta})$. This is clear by considering that, by definition, $P_{\boldsymbol{\theta}}\#\pi(A) = \pi(\{\mathbf{x} : \langle \mathbf{x}, \boldsymbol{\theta} \rangle \in A\}) = \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle \in A} f(\mathbf{x}) \mathrm{d}\mathbf{x}$ for every measurable set $A \subseteq \mathbb{R}$.

The strategy to derive a control of the type (46) is then to consider regular probability distributions with compact support and to use the study on $\mathbb{R}$ (Theorem 3) to upper bound each Wasserstein distance $W_1(P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi')$ with a MMD. This will allow us to define a TI p.s.d. kernel on $\mathbb{R}^d$ whose MMD dominates the Wassertein distance. In order to apply Theorem 3 we need to relate the regularity of the density of π with the one of $P_{\boldsymbol{\theta}}\#\pi$. We can prove that, as soon as the densities on $\mathbb{R}^d$ are regular enough the densities "on each line" are also regular on $\mathbb{R}$. More precisely:

Lemma 6. *Suppose that $d \geq 2$. Let $f \in C^s(\mathbb{R}^d)$ be integrable and compactly supported. For any $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ the Radon transform satisfies $\mathcal{R}[f](\cdot, \boldsymbol{\theta}) \in C^s(\mathbb{R})$ and $\|\mathcal{R}[f](\cdot, \boldsymbol{\theta})\|_{W^{s,1}(\mathbb{R})} \leq d^{s+1}\|f\|_{W^{s,1}(\mathbb{R}^d)}$*

We also need one other technical lemma which exhibits a TI kernel on $\mathbb{R}^d$ and an MMD on $\mathcal{P}(\mathbb{R}^d)$ from a translation kernel on $\mathbb{R}$:

Lemma 7. *Let $\kappa_{\mathbb{R}}(x, y) = \kappa_0(x - y)$ be a TI p.s.d. kernel on $\mathbb{R}$ where κ_0 is continuous. Consider the kernel κ on $\mathbb{R}^d$ defined by $\kappa(\mathbf{x}, \mathbf{y}) := \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top \mathbf{x}, \boldsymbol{\theta}^\top \mathbf{y})] = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_0(\boldsymbol{\theta}^\top(\mathbf{x} - \mathbf{y}))]$. Then κ is TI, continuous, bounded and positive semi-definite. Moreover we have for any $(\pi, \pi') \in \mathcal{P}(\mathbb{R}^d)$:*

$$\|\pi - \pi'\|_\kappa^2 = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^2] \tag{64}$$

The proof of these two lemmas can be found in Appendix C.9. Lemma 7 exhibits a way of constructing an MMD on $\mathcal{P}(\mathbb{R}^d)$ from slices of the distributions and an MMD on $\mathcal{P}(\mathbb{R})$, in the exact same manner of the Sliced-Wasserstein distance (as in (Nadjahi et al., 2020)). Based on these results we can prove the main theorem of this section:

Theorem 4 *Let $\kappa_{\mathbb{R}}(x, y) = \kappa_0(x - y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$ and such that $\widehat{\kappa}_0(\omega) > 0$ for every ω, $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{q_\kappa})$ as $\omega \to 0$ and $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{s_\kappa})$ as $\omega \to +\infty$ for some $q_\kappa > -1, s_\kappa \in \mathbb{R}_+$. Consider for any $s \geq \frac{s_\kappa}{2} + 1$, $0 < M, R < +\infty$, and the model set:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}_2(\mathbb{R}^d) : \exists f \in C^s(\mathbb{R}^d), \pi \ll f\mathrm{d}\mathbf{x}, \|f\|_{W^{s,1}(\mathbb{R}^d)} \leq M, \mathrm{supp}(f) \subseteq B(0, R)\}$$

1. *There exists a constant $C = C(R, M, s, d) > 0$ such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, \textit{if } \mathrm{m}(\pi) = \mathrm{m}(\pi') \textit{ then } W_1(\pi, \pi') \leq C\|\pi - \pi'\|_\kappa^{\frac{1}{2(d+1)}} \tag{65}$$

   where κ is a translation invariant p.s.d. kernel defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) := \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top \mathbf{x}, \boldsymbol{\theta}^\top \mathbf{y})] = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_0(\boldsymbol{\theta}^\top(\mathbf{x} - \mathbf{y}))] \tag{66}$$

2. *If in addition κ_0 is L-Lipschitz then there is $C = C(R, M, s, d, L)$ such that*

$$\forall \pi, \pi' \in \mathfrak{S}, W_1(\pi, \pi') \leq C\|\pi - \pi'\|_{\tilde{\kappa}}^{\frac{1}{2(d+1)}} \tag{67}$$

   where $\tilde{\kappa}$ (which is no longer translation invariant) is a p.s.d. kernel defined by:

$$\tilde{\kappa}(\mathbf{x}, \mathbf{y}) := \kappa(\mathbf{x}, \mathbf{y}) + \frac{1}{d}\langle \mathbf{x}, \mathbf{y}\rangle \tag{68}$$

Proof Let $\pi, \pi' \in \mathfrak{S}$ with densities f, g. By hypothesis f, g are supported on $B(0, R)$ so $\mathrm{m}(\pi), \mathrm{m}(\pi') \in B(0, R)$ and $\text{m-diam}(\mathfrak{S}) \leq 2R$ and by Lemma 5 we have $W_1(\pi, \pi') \leq C\ SW_1(\pi, \pi')^{1/(d+1)}$ for a constant $C > 0$ that only depends on the dimension and on R. Moreover since the distributions admit a density, by (62) we can write $SW_1(\pi, \pi') = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[W_1(\mathcal{R}[f](\cdot, \boldsymbol{\theta}), \mathcal{R}[g](\cdot, \boldsymbol{\theta}))]$. Then for each $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ we can use Lemma 6 to prove that $\mathcal{R}[f](\cdot, \boldsymbol{\theta}), \mathcal{R}[g](\cdot, \boldsymbol{\theta}) \in C^s(\mathbb{R})$ and $\|\mathcal{R}[f](\cdot, \boldsymbol{\theta})\|_{W^{s,1}(\mathbb{R})} \leq d^{s+1}M$

since $\|f\|_{W^{s,1}}(\mathbb{R}^d) \leq M$ by hypothesis (same for g). Moreover, since $\pi, \pi' \in \mathcal{P}_2(\mathbb{R}^d)$ we have also that $P_{\boldsymbol{\theta}}\#\pi$ and $P_{\boldsymbol{\theta}}\#\pi'$ are in $\mathcal{P}_2(\mathbb{R})$. Indeed, $\int |x|^2 \mathrm{d}P_{\boldsymbol{\theta}}\#\pi(x) = \int |\langle \mathbf{x}, \boldsymbol{\theta}\rangle|^2 \mathrm{d}\pi(\mathbf{x}) \leq \int \|\mathbf{x}\|_2^2 \mathrm{d}\pi(\mathbf{x}) < +\infty$. Overall this proves that $P_{\boldsymbol{\theta}}\#\pi$ and $P_{\boldsymbol{\theta}}\#\pi'$ with densities $\mathcal{R}[f](\cdot, \boldsymbol{\theta}), \mathcal{R}[g](\cdot, \boldsymbol{\theta})$ belong to the following set:

$$\mathfrak{S}_{\boldsymbol{\theta}} := \{\pi \in \mathcal{P}_2(\mathbb{R}) : \exists h \in C^s(\mathbb{R}), \pi \ll h\mathrm{d}x, \|h\|_{W^{s,1}}(\mathbb{R}) \leq d^{s+1}M\} \tag{69}$$

We also have $\mathrm{m}(\pi) = \mathbb{E}_{x\sim P_{\boldsymbol{\theta}}\#\pi}[x] = \mathbb{E}_{\mathbf{x}\sim\pi}[\langle \boldsymbol{\theta}, \mathbf{x}\rangle] = \langle \boldsymbol{\theta}, \mathbb{E}_{\mathbf{x}\sim\pi}[\mathbf{x}]\rangle = \langle \boldsymbol{\theta}, \mathrm{m}(\pi)\rangle$ by linearity of the expectation, and similarly $\mathrm{m}(\pi') = \langle \boldsymbol{\theta}, \mathrm{m}(\pi')\rangle$.

For the first claim, since $\mathrm{m}(\pi) = \mathrm{m}(\pi')$ we have $\mathrm{m}(P_{\boldsymbol{\theta}}\#\pi) = \mathrm{m}(P_{\boldsymbol{\theta}}\#\pi')$. We can thus apply the first claim of Theorem 3 to $P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi'$ with the kernel $\kappa_{\mathbb{R}}$ so that we have:

$$W_1(P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi') \leq C_2 \|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^{1/2} \tag{70}$$

where C_2 depends only on M, s and the kernel κ (and not on $\boldsymbol{\theta}$ since the constant does not depend on the mean). For the second claim, since $\|\boldsymbol{\theta}\|_2 = 1$ we have $|\,\mathrm{m}(P_{\boldsymbol{\theta}}\#\pi) - \mathrm{m}(P_{\boldsymbol{\theta}}\#\pi)| = |\langle \boldsymbol{\theta}, \mathrm{m}(\pi) - \mathrm{m}(\pi')\rangle| \leq \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2 \leq \text{m-diam}(\mathfrak{S}) \leq 2R$, hence $\text{m-diam}(\mathfrak{S}_{\boldsymbol{\theta}}) \leq 2R$ and we can apply Corollary 2 to $P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi'$ to obtain:

$$W_1(P_{\boldsymbol{\theta}}\#\pi, P_{\boldsymbol{\theta}}\#\pi') \leq C_2 \|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\tilde{\kappa}_{\mathbb{R}}}^{1/2} \tag{71}$$

where $\tilde{\kappa}_{\mathbb{R}}(x, y) := \kappa_0(x - y) + xy$ and C_2 is a constant that only depends on M, s, L, R and κ. Consequently, since C_2 does not depend on $\boldsymbol{\theta}$, we have:

$$W_1(\pi, \pi') \leq C \cdot C_2^{1/(d+1)} \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^{1/2}]^{1/(d+1)} \tag{72}$$

Using Hölder inequality we get:

$$\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^{1/2}] \leq \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^{2}]^{1/4} \tag{73}$$

Finally we can apply Lemma 7 to conclude for the first claim. For the second claim, the last two inequalities can be repeated with $\kappa_{\mathbb{R}}$ replaced by $\tilde{\kappa}_{\mathbb{R}}$. Since $\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\langle \boldsymbol{\theta}, \mathbf{u}\rangle^2] = \mathbf{u}^\top \left(\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]\right) \mathbf{u} = \|\mathbf{u}\|_2^2/d$ for every vector $\mathbf{u}$, we have

$$\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[|\,\mathrm{m}(P_{\boldsymbol{\theta}}\#\pi) - \mathrm{m}(P_{\boldsymbol{\theta}}\#\pi')|^2] = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\langle \boldsymbol{\theta}, \mathrm{m}(\pi) - \mathrm{m}(\pi')\rangle^2] = \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2^2/d$$

hence

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\tilde{\kappa}_{\mathbb{R}}}^2] &= \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^2] + \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[|\,\mathrm{m}(P_{\boldsymbol{\theta}}\#\pi) - \mathrm{m}(P_{\boldsymbol{\theta}}\#\pi')|^2] \\ &= \|\pi - \pi'\|_\kappa^2 + \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2^2/d = \|\pi - \pi'\|_{\tilde{\kappa}}^2\end{aligned}$$

where we applied again Lemma 7 and used the same arguments as in the proof of Corollary 2 for the last inequality. ■

As a corollary of this result we can deduce that under the same hypothesis the W_p distance satisfies the same control for any $p \in [1, +\infty[$:

Corollary 9. *Consider a kernel $\kappa_{\mathbb{R}}(x, y) = \kappa_0(x - y)$ on $\mathbb{R}$, $\mathfrak{S} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ and $\kappa, \tilde{\kappa}$ as in Theorem 4. If κ_0 is L-Lipschitz then for each $p \geq 1$ there exists a constant $C = C(R, M, s, d, L, p) > 0$ such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, W_p(\pi, \pi') \leq C\|\pi - \pi'\|_{\tilde{\kappa}}^{\frac{1}{2p(d+1)}}. \tag{74}$$

Proof When the probability distributions are supported on $B(0, R)$ we have (see (Santambrogio, 2015, Section 5.1)) $W_p(\pi, \pi') \leq (2R)^{\frac{p-1}{p}} W_1(\pi, \pi')^{1/p}$ which concludes. ■

This theorem proves that for such $\mathfrak{S}, \tilde{\kappa}$ we have that $(\mathfrak{S}, W_p)$ is $(\tilde{\kappa}, \frac{1}{2p(d+1)})$-embeddable. We would like to emphasize that, as discussed in Section 2, we can only expect to have slow rates for CSL by using this bound with a curse of dimensionality phenomenon.

4.7 The case of non-compactly supported distributions

The case of non-compactly supported measures on $\mathbb{R}^d$ is more delicate to study, as one can not rely anymore on the closed-form of the Wasserstein distance on $\mathbb{R}$. We will however prove that, at the price of an arbitrary small additive term $\eta > 0$, we have the control (46) 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 6 (Regularizer). *We say that a function $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ is a* regularizer *if it is a* non-negative, continuous, even *and bounded function such that $\int_{\mathbb{R}^d} \Phi = 1$ and $\Phi \in L_2(\mathbb{R}^d)$. We say that the regularizer has p-finite moments if $\int \|\mathbf{z}\|^p \Phi(\mathbf{z}) \mathrm{d}\mathbf{z} < +\infty$ for some $p \geq 1$.*

When considering a regularizer Φ and a probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ (not necessarily regular) the convolution $\Phi * \pi$ defines a probability density function[16] on $\mathbb{R}^d$ *via* $\Phi * \pi(\mathbf{x}) = \int_{\mathbb{R}^d} \Phi(\mathbf{x} - \mathbf{y}) \mathrm{d}\pi(\mathbf{y})$. In the following we will note π_Φ the probability distribution associated to the density $\Phi * \pi$. Note that π_Φ is usually regular by imposing that Φ is (such as when Φ is the Gaussian density). The interpretation behind π_Φ is the following: if $X \sim \pi$ and Y_Φ is a random variable independant of X and whose distribution has density Φ then the random variable $X + Y_\Phi$ has distribution π_Φ. 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, 2020; Goldfeld and Greenewald, 2020; Nguyen, 2013). We have the following lemma which relates W_p to its regularized counterpart:

Lemma 8. *Consider a regularizer Φ with p-finite moments where $p \geq 1$. We have:*

$$\forall \pi, \pi' \in \mathcal{P}(\mathbb{R}^d),\ W_p(\pi, \pi') \leq W_p(\pi_\Phi, \pi'_\Phi) + 2 \left(\int \|\mathbf{z}\|_2^p \Phi(\mathbf{z}) \mathrm{d}\mathbf{z} \right)^{1/p} \tag{75}$$

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

When Φ is the density of the Gaussian $\mathcal{N}(0, \sigma^2 \mathbf{I})$ the distance $W_p(\pi_\Phi, \pi'_\Phi)$ 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, $W_p(\pi_\Phi, \pi'_\Phi)$ can be controlled by an MMD associated to a translation invariant kernel. We will use the following lemma:

Lemma 9. *Let $s > 1$. Assume that $\pi, \pi' \in \mathcal{P}_s(\mathbb{R}^d)$ have densities f, g with respect to the Lebesgue measure. Then for any $1 \leq p < s$ we have:*

$$W_p(\pi, \pi') \leq 2^{\frac{1}{p} + 1 - \frac{1}{s}} V_d^{\frac{s-p}{(d+2s)p}} \left(\mathbb{E}_{\mathbf{x} \sim \pi}[\|\mathbf{x}\|_2^s] + \mathbb{E}_{\mathbf{y} \sim \pi'}[\|\mathbf{y}\|_2^s] \right)^{\frac{2p+d}{(d+2s)p}} \|f - g\|_{L_2(\mathbb{R}^d)}^{\frac{2(s-p)}{(d+2s)p}} \tag{76}$$

with $V_d = \pi^{d/2} \Gamma(d/2 + 1)$ the volume of the d-dimensional unit sphere.

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

The proof of this result can be found in Appendix C.10. To connect with the MMD we will use the following result whose proof is in Appendix C.11:

Lemma 10. *Let Φ be a regularizer and $\kappa_0 := \Phi * \Phi$. Then $\kappa_0 \in L_1(\mathbb{R}^d)$ is even, bounded, continuous and has non-negative Fourier transform. Consider the kernel $\kappa(\mathbf{x}, \mathbf{y}) := \kappa_0(\mathbf{x} - \mathbf{y})$. We have that κ defines a TI p.s.d. kernel. Moreover, for $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$:*

$$\|\pi - \pi'\|_\kappa = \|\Phi * \pi - \Phi * \pi'\|_{L_2(\mathbb{R}^d)} \tag{77}$$

Based on these results we have the following upper-bound on $W_p(\pi_\sigma, \pi'_\sigma)$ using the MMD associated to a TI p.s.d. kernel (the proof can be found in Appendix C.11):

Proposition 10. *Let $s > 1$. Consider a regularizer Φ with s-finite moments. Consider the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 := \Phi * \Phi$. It defines a TI p.s.d kernel by Lemma 10. Moreover, we have for any $\pi, \pi' \in \mathcal{P}_s(\mathbb{R}^d)$ and $1 \leq p < s$:*

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

where $C_{d,s,p} = 2^{\frac{1}{p}+1-\frac{1}{s}} V_d^{\frac{s-p}{(d+2s)p}}$ is a constant.

As a corollary of Proposition 10 and Lemma 8 we are now able to prove the main theorem of this section (the proof is in Appendix C.11):

Theorem 5 *Let $s > 1$. Consider a regularizer Φ with s-bounded moments. Consider the kernel $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ where $\kappa_0 := \Phi * \Phi$. It defines a TI p.s.d kernel by Lemma 10. We consider the following model set:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}_{\mathbf{x} \sim \pi}[\|\mathbf{x}\|_2^s] \leq M\} \tag{78}$$

Then for any $1 \leq p < s$ there exists a constant $C = C_{d,s,p} > 0$ such that:

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

This theorem has many implications. First it shows that, for a wide range of TI p.s.d. kernels, and under mild assumptions, $(\mathfrak{S}, W_p)$ is $(\kappa, \delta = \frac{2(s-p)}{(d+2s)p})$-embeddable with error $\eta > 0$. We will see with the example in Section 4.8 how this error term can be controlled. We emphasize that few assumptions on $\mathfrak{S}$ are required: the distributions in the model set must have uniformly bounded s-moment, *i.e.* $\sup_{\pi \in \mathfrak{S}} \mathbb{E}_{\mathbf{x} \sim \pi}[\|\mathbf{x}\|_2^s] < +\infty$. This assumption is verified when, for example if $\mathfrak{S}$ is the space of GMMs whose parameters are in a compact subspace as considered in CSL (Keriven et al., 2018). Interestingly, when s is big compared to d, p, *i.e.* the model set contains sufficiently well-behaved distributions, then $\delta \approx \frac{1}{p}$. In the context of CSL (Section 2), this would give a rate of convergence of the empirical sketch to the true sketch of the order of $O(n^{-\frac{1}{2p}})$, which is reasonable compared to the $O(n^{-\frac{1}{2}})$ of the LRIP when p is small (as in Section 3).

Condition on the TI kernel κ The condition $\kappa_0 = \Phi * \Phi$ can be met in two ways. First, fixing a regularizer Φ with s-bounded moments gives a TI p.s.d. kernel such that Theorem 5 holds. This can be achieved for example by considering a p.s.d. function $\Phi \in L_1(\mathbb{R}^d)$ with a sufficient number of bounded moments that is even, continuous and positive (continuous, integrable and p.s.d. functions are bounded (Wendland, 2004)). A simple normalization $\Phi \leftarrow \Phi / \int \Phi$ will then produce a suitable Φ. We give an example in Section 4.8 of such function Φ by considering the Gaussian density that produces the Gaussian kernel. 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 $\kappa_0 = \Phi * \Phi$ with Φ a regularizer with s-bounded moments and $\widehat{\Phi} \geq 0$ so

that Theorem 4 will hold. 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).

4.8 An application with the RBF kernel

As an example of use of Theorem 5 consider the Gaussian function $\varphi(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\|\mathbf{x}\|_2^2/2)$. Define for $\sigma > 0$ the regularizer $\Phi(\mathbf{x}) = \sigma^{-d}\varphi(\frac{\mathbf{x}}{\sigma})$. We have that Φ is continuous, even, bounded, all s-moments are finite, $\int_{\mathbb{R}^d}\Phi = 1$. 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 $s > 1$. The error term $2\int\|\mathbf{z}\|_2\Phi(\mathbf{z})\mathrm{d}\mathbf{z} = 2\sigma\int\|\mathbf{z}\|_2\varphi(\mathbf{z})\mathrm{d}\mathbf{z}$ can be controlled as:

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

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 $2\sigma(\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I})}[\|\mathbf{x}\|_2^2])^{1/2} = 2\sigma\sqrt{d}$. We have also that $\int\|\mathbf{z}\|_2^s\Phi(\mathbf{z})\mathrm{d}\mathbf{z} = \sigma^s\int\|\mathbf{z}\|_2^s\varphi(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I})}[\|\mathbf{x}\|_2^s] = 2^{s/2}\frac{\Gamma(\frac{s+d}{2})}{\Gamma(\frac{s}{2})}$ (it is the s-th moment of a χ_2 distribution). Then using Theorem 5 we have:

Corollary 4. *Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) = \pi^{d/2}\sigma^{-d}\exp(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{4\sigma^2})$ for $\sigma > 0$ and two numbers $s > 1, M > 0$. Consider the model set $\mathfrak{S}$ defined as:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}(\mathbb{R}^d),\ \mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^s] \leq M\}. \tag{79}$$

We have

$$\forall\pi,\pi' \in \mathfrak{S},\ W_1(\pi,\pi') \leq C\left(M + 2^{s/2}\sigma^s\frac{\Gamma(\frac{s+d}{2})}{\Gamma(\frac{s}{2})}\right)^{\frac{d+2}{d+2s}}\|\pi-\pi'\|_\kappa^{\frac{2(s-1)}{d+2s}} + 2\sigma\sqrt{d} \tag{80}$$

with $C = C_{d,s,1}$ defined in Proposition 10.

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 $\sigma > 0$). Consequently, in the context of CSL (Section 2) this error term can always be chosen smaller compared to the bias term $\mathrm{Bias}(\pi,\mathfrak{S})$. Moreover, when s is big compared to d then $\delta = \frac{2(s-1)}{d+2s} \approx 1$, such that, if the model set contains distributions with enough bounded moments, δ is close to the $\delta = 1$ case of the LRIP.

4.9 Conclusion and related works

We have shown in this section various controls of the form $W_p \lesssim \mathrm{MMD}_\kappa^\delta$ that depend on $\delta \in]0,1]$, the properties of the model set and the kernel κ. All these results are summarized in Figure 4. All were obtained for translation invariant p.s.d. kernels on $\mathbb{R}^d$, that are at the core of the CSL framework when using RFF (Section 2), and for Wasserstein distances based on the metric $D(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$ that corresponds to the results of Section 3. Some other connections between MMD and the Wasserstein distance have been explored in the literature when the latter constraints are relaxed. 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.* $D(\mathbf{x},\mathbf{y}) = \|\kappa(\cdot,\mathbf{x}) - \kappa(\cdot,\mathbf{y})\|_\mathcal{H}$. 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).

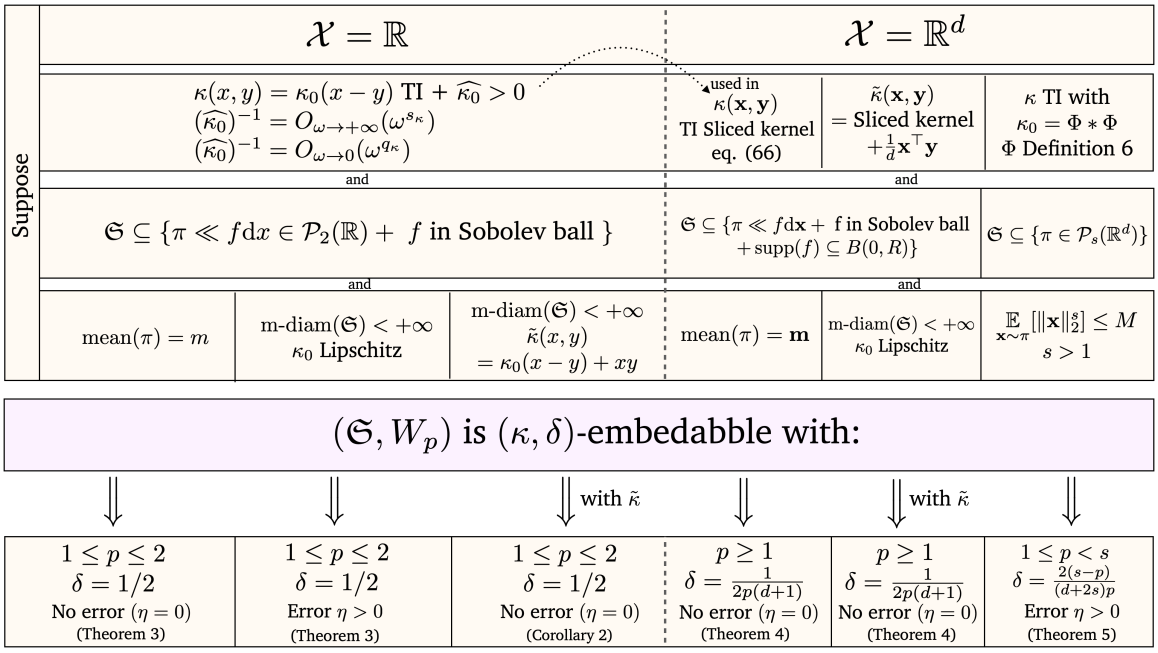

Figure 4: Summary of the different results of Section 4. TI= Translation invariant kernel as $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$. The mention "with error" means that the relation holds when adding an error $\eta > 0$ that does not depends on the measures. $\pi \ll f\mathrm{d}\mathbf{x}$ means that the measure has density f with respect to the Lesbegue measure.

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

$$\forall \pi, \pi' \in \mathfrak{S}_\kappa, W_p(\pi_\Phi, \pi'_\Phi) \le 2\sigma\|\pi - \pi'\|_\kappa^{1/p} \tag{81}$$

where $\mathfrak{S}_\kappa := \{\pi \in \mathcal{P}(\mathbb{R}^d); \int \sqrt{\kappa(\mathbf{x},\mathbf{x}')}\mathrm{d}\pi(\mathbf{x}) < +\infty\}$. With the same type of arguments than those presented in Lemma 8 we can prove that for any $\pi, \pi' \in \mathfrak{S}_\kappa$ we have $W_p(\pi,\pi') \le 2\sigma\|\pi - \pi'\|_\kappa^{1/p} + \eta$ where $\eta = 2\left(\int \|\mathbf{z}\|_2^p \Phi(\mathbf{z})\mathrm{d}\mathbf{z}\right)^{1/p}$. As a corollary, for this kernel that is not TI we can use the result of (Zhang et al., 2021) to prove that $(\mathfrak{S}_\kappa, W_p)$ is $(\kappa, \frac{1}{p})$-embeddable with error $\eta = 2\left(\int \|\mathbf{z}\|_2^p \Phi(\mathbf{z})\mathrm{d}\mathbf{z}\right)^{1/p}$ that will behave as $O(\sigma)$ as shown in Section 4.8. 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 MMD. In (Nietert et al., 2021b) authors control the Wasserstein distance with an MMD whose kernel, which is not TI, is defined by $\kappa(\mathbf{x},\mathbf{y}) = -\sigma^2\,\mathrm{Ein}(-\langle\mathbf{x},\mathbf{y}\rangle/\sigma^2)$ where $\mathrm{Ein} = \int_0^z \frac{(1-e^{-t})}{t}\mathrm{d}t$. 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 $W_p(\pi_\Phi, \pi'_\Phi)$ by controlling it with a MMD based on *specific* kernel but instead to control $W_p(\pi,\pi')$ by kernel norms for *many* types of TI kernels.

Relaxing the *p.s.d.* assumption on the kernel Beyond *p.s.d.* 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 Section 2 but share similar topological properties. For example, by considering the *conditionally p.s.d.* [17] kernel $\kappa(\mathbf{x},\mathbf{y}) = -\|\mathbf{x}-\mathbf{y}\|_2^\beta$ for $\beta \in]0,2]$ the term $\|\pi - \pi'\|_\kappa$ is non-negative for $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ and defines a valid MMD which is called the energy (or *Cramer*) distance (Székely and Rizzo, 2017; Szekely and Rizzo, 2004; Sejdinovic et al., 2013). 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:

$$\kappa_0(\mathbf{x}) := \begin{cases} \log \frac{1}{\|\mathbf{x}\|_2} & \text{if } d = 2 \\ \frac{1}{\|\mathbf{x}\|_2^{d-2}} & \text{if } d \geq 3 \end{cases} \tag{82}$$

In this case, for compactly supported $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$ with $\|\pi\|_\kappa, \|\pi'\|_\kappa < +\infty$, the quantity $\|\pi - \pi'\|_\kappa$ is well defined, finite, non-negative, 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 as described in (Chafaï et al., 2016). More precisely consider, for $K \subseteq \mathbb{R}^d$ *compact*, a model set $\mathfrak{S}$ such that:

$$\mathfrak{S} = \{\pi \in \mathcal{P}(\mathbb{R}^d), \mathrm{supp}(\pi) \subseteq K, \|\pi\|_\kappa < +\infty\}$$

Then (Chafaï et al., 2016, Theorem 1) shows that there exists $C = C(K) > 0$ such that:

$$\forall \pi, \pi' \in \mathfrak{S}, W_1(\pi,\pi') \leq C\|\pi - \pi'\|_\kappa \tag{83}$$

This result shows that, with the above $\mathfrak{S}$, $(\mathfrak{S}, W_1)$ is $(\kappa, \delta = 1)$-embeddable with no error. It is remarkable in the sense few assumptions on the model set are required and in particular the distributions can be discrete. An important remark is that the kernel is TI but not p.s.d. and, consequently, this result is not in contradiction with Theorem 2. This also suggests that finite dimensional approximations of TI conditionally p.s.d. kernels (as done in (Sun, 1993; Narcowich et al., 2007)) could lead to interesting feature maps for CSL.

5. From kernel embeddings of distributions to sketching operators

At this point, we propose to refocus on CSL and to make a brief summary of the different notions used in this paper to establish CSL learning guarantees. As described in Section 2, when one finds a sketching operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^m$ that satisfies the Hölder LRIP with constant $C > 0$ and $\eta \geq 0$ on a model set $\mathfrak{S}$ (Definition 1) then we can control the excess risk for any probability distribution $\pi \in \mathcal{P}(\mathcal{X})$ (Theorem 1). We conclude this paper by outlining a possible strategy to find such a sketching operator using the tools developed in the previous sections. We have seen in Section 3 that several learning tasks $\mathcal{L}(\mathcal{H})$ satisfy the Wasserstein learnability property:

$$\forall \pi, \pi' \in \mathcal{P}_p(\mathcal{X}), \quad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C_1 W_p(\pi,\pi') \tag{84}$$

In Section 4 we also exhibited conditions on a model set $\mathfrak{S}$ and a TI p.s.d. kernel κ under which $(\mathfrak{S}, W_p)$ is (κ,δ)-embedabble with (potentially) an error $\mathrm{err} \geq 0$ *i.e.*

$$\forall \pi, \pi' \in \mathfrak{S}, W_p(\pi,\pi') \leq C_2\|\pi - \pi'\|_\kappa^\delta + \mathrm{err} \tag{85}$$

17. A conditionally *p.s.d.* kernel on $\mathcal{X}$ is such that $\sum_{i,j=1}^n c_i c_j \kappa(\mathbf{x}_i,\mathbf{x}_j) \geq 0$ 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)

By combining both properties we obtain the *Kernel Hölder-LRIP* with error $\eta = C_1 \times \text{err} \geq 0$, constant $C = C_1 \times C_2 > 0$ and $\delta \in]0, 1]$:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\pi - \pi'\|_\kappa^\delta + \eta \qquad \text{(Kernel Hölder LRIP)}$$

The goal of this section is to further constrain the model set $\mathfrak{S}$ in order to find a finite dimension m and a sketching operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}^m$ that satisfies the Hölder LRIP.

5.1 From the kernel Hölder LRIP to the Hölder LRIP: existence of a sketching operator

Interestingly enough we will see that as soon as we have the Kernel Hölder LRIP with a *bounded kernel* (not necessarily translation invariant) with a compact model set with respect to the total variation norm then there exists a sketching operator that satisfies the Hölder LRIP. To establish this we will rely on general Banach spaces embedding results from (Robinson, 2010). We first introduce some definitions:

Definition 7. *Let $(\mathcal{X}, D)$ be a (pseudo)metric space with (pseudo)metric D. Consider a set $S \subseteq \mathcal{X}$ and $\varepsilon > 0$. We say that $(x_1, \cdots, x_N) \in S^N$ is an ε-net of S of size N if $S \subseteq \cup_{i \in [\![N]\!]} B_D(x_i, \varepsilon)$ where $B_D(x_i, \varepsilon)$ is the closed ball centered at x_i of radius ε. Equivalently there is an ε-net of S of size $N > 0$ if:*

$$\exists (x_1, \cdots, x_N) \in S^N, \forall x \in S, \exists i \in [\![N]\!], D(x, x_i) \leq \varepsilon \tag{86}$$

The ε-covering number of S is defined by:

$$\mathcal{N}(S, D, \varepsilon) = \min\{N : \exists \varepsilon - \textit{net of } S \textit{ of size } N\} \tag{87}$$

The covering number of a set S is the minimal number of closed balls that we need to have to cover the whole set S. It allows to define a notion of "dimension" of a set S which is given by the upper box-counting dimension:

$$d_B(S) := \limsup_{\varepsilon \to 0} \frac{\log(\mathcal{N}(S, D, \varepsilon))}{-\log(\varepsilon)} \tag{88}$$

In the following we consider the Banach space of finite signed measure $\mathcal{M}(\mathcal{X})$ on $\mathcal{X}$ equipped with the total variation norm $\|.\|_{\text{TV}}$ *i.e.* $\|\mu\|_{\text{TV}} = |\mu|(\mathcal{X})$ (see (Halmos, 1976)). We are now ready to prove that, under some mild assumptions on $\mathfrak{S}$ and when we have the Kernel Hölder LRIP there exists an operator $\mathcal{A}$ that satisfies the Hölder LRIP. The following result is a consequence of Theorem 8.1 in (Robinson, 2010) and its proof is deferred to Appendix D.1:[18]

Theorem 6 (Existence of a sketching operator when the Kernel Hölder LRIP holds) *Consider a model set $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ such that $\mathfrak{S}$ is* compact *in $(\mathcal{M}(\mathcal{X}), \|.\|_{\text{TV}})$ and has* finite upper box-counting dimension*, i.e. $d_B(\mathfrak{S}) < +\infty$. Suppose that there exists $\beta > 0, C > 0$ and $\eta \geq 0$ such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\pi - \pi'\|_\kappa^\beta + \eta \tag{89}$$

for some bounded kernel *κ (i.e. such that $\sup_{\mathbf{x} \in \mathcal{X}} \kappa(\mathbf{x}, \mathbf{x}) \leq K$). Then for any finite dimension $m > 2d_B(\mathfrak{S})$ there exists $0 < \delta < \beta, C' > 0$, and a prevalent set of bounded linear maps[a] $\mathcal{A} : \mathcal{M}(\mathcal{X}) \to \mathbb{R}^m$ such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C'\|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^\delta + \eta \tag{90}$$

18. When carefully looking at the proof of Theorem 6 it seems at first sight that the assumption that $\mathfrak{S}$ is compact in $(\mathcal{M}(\mathcal{X}), \|.\|_{\text{TV}})$ may be replaced by a compactness assumption in the metric space $(\mathcal{M}(\mathcal{X}), \|.\|_\kappa)$ (with κ bounded). In order to do the same reasoning, the space $(\mathcal{M}(\mathcal{X}), \|.\|_\kappa)$ should however be a complete normed space, which is only possible when κ is characteristic to $\mathcal{M}(\mathcal{X})$ (thus $\|.\|_\kappa$ defines a pre-Hilbert norm on $\mathcal{M}(\mathcal{X})$). However, it was shown in (Steinwart and Ziegel, 2017, Theorem 3.1) that $(\mathcal{M}(\mathcal{X}), \|.\|_\kappa)$ is complete (hence a Hilbert space) only when $\|.\|_\kappa$ and $\|.\|_{\text{TV}}$ are equivalent. Consequently we choose to keep the compactness assumption with respect to $\|.\|_{\text{TV}}$

In other words if $\beta \leq 1$ there exists a sketching operator that satisfies the Hölder LRIP with some $0 < \delta < 1$ and error $\eta \geq 0$.

a. A prevalent set is a set whose complement is *shy* that is, informally, negligeable: in the case of Euclidean space it is a space whose complement has Lesbegue measure zero.

In particular, this result states that when the Kernel Hölder LRIP holds with no error ($\eta = 0$) there exists a finite dimensional embedding of our distributions in $\mathfrak{S}$ and a sketching operator $\mathcal{A}$ that satisfies the Hölder LRIP with error $\eta = 0$. This result can be put in contrast to the LRIP case ($\delta = 1$) where such existence theorem do not exists. The previous discussion leads to the following informal proposition that summarizes the different contributions:

Proposition 11. *Consider a task $\mathcal{L}(\mathcal{H})$ and a model set $\mathfrak{S} \subset \mathcal{P}(\mathcal{X})$ and $p \in [1, +\infty[$. Suppose that:*

- *The task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable with constant $C_1 > 0$ (Definition 2, Section 3)*
- *The space $(\mathfrak{S}, W_p)$ is (κ, β)-embeddable with constant $C_2 > 0$ and error $\mathrm{err} \geq 0$ for some bounded kernel κ on $\mathcal{X}$ and $\beta \in]0, 1]$ (Definition 3, Section 4)*
- *$\mathfrak{S}$ is* compact *in $(\mathcal{M}(\mathcal{X}), \|.\|_{TV})$ and has* finite upper box-counting dimension*, i.e. $d_B(\mathfrak{S}) < +\infty$.*

Then for $m > 2d_B(\mathfrak{S})$ there exists a sketching operator that satisfies the Hölder LRIP with constant $C = C_1 \times C_2 > 0$, error $\eta = C_1 \times \mathrm{err} \geq 0$ and $0 < \delta < \beta \leq 1$.

This result gives necessary conditions under which we theoretically have CSL guarantees via the Hölder LRIP and based on the different results presented in this paper. However it comes with limitations. Indeed, this proposition proves the existence of a sketching operator that comes with CSL guarantees but does not give its concrete, calculable expression. Moreover the finite dimension $m > 2d_B(\mathfrak{S})$ could be potentially very large and the Hölder exponent δ may be very close to zero implying slow learning rates. Fortunately the different results established in (Gribonval et al., 2021a) are directly applicable to our setting since most of the kernels considered in Section 4 are TI. For the sake of conciseness we only give here some intuitions and we refer the reader to (Gribonval et al., 2021a) for a more detailed discussion. Informally these results allow to prove that, for a controlled dimension m, a TI p.s.d. kernel κ and $\mathcal{A}$ defined with RFF (Section 2) we have:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_\kappa \approx \|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \tag{91}$$

This property is valid when certain covering numbers of the *normalized secant-set* of $\mathfrak{S}$ are controlled (Gribonval et al., 2021a). As such, when the task $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable and the space $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable the approach presented in this paper combined with the one of (Gribonval et al., 2021a) show that sketching operators based on RFF are suited for a wide range of tasks and lead to CSL guarantees.

6. Conclusion & perspectives

The main contributions of this paper are the following. We establish different bounds between metrics between probability distributions. First, 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 then show that this 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 statistical learning* (CSL) 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. We show that the different bounds allow us to establish a property called the *Hölder LRIP* that generalizes the LRIP property in CSL and allows, for the given sketching operator, to control 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 CSL 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) or for classification tasks (Schellekens and Jacques, 2018). Related to the CSL 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 CSL guarantees. More precisely, for compressive learning tasks such as K-means and GMM one question would be to see if we can obtain CSL 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. An interesting question would be to see if these bounds can be used to mitigate the curse of dimensionality of the Wasserstein distance when the distributions are constrained to a certain model set (since the MMD do not suffer from the curse of dimensionality). Finally, these bounds are valid in one direction (Wasserstein controlled by the MMD) and further works could be devoted to finding bounds in the other direction and to prove that the MMD and the Wasserstein distance induce the same topology on some subspaces of probability distributions.

Acknowledgements 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 1

We recall the result here:

Proposition 1 (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:*

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \tag{18}$$

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 (18) *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:

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq \|\pi - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} + \|\pi' - \Delta[\mathcal{A}(\pi)]\|_{\mathcal{L}(\mathcal{H}),p} \tag{92}$$

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:

$$\Delta[\mathbf{s}] \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{s}\|_2 \tag{93}$$

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 $\tau \in \mathfrak{S}$. We have:

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

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

$$\begin{aligned}
\|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} &\leq \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\tau) - (\mathcal{A}(\pi) + \mathbf{e})\|_2^\delta + \eta \\
&\leq \|\pi - \tau\|_{\mathcal{L}(\mathcal{H}),p} + 2C\|\mathcal{A}(\tau) - \mathcal{A}(\pi)\|_2^\delta + 2C\|\mathbf{e}\|_2^\delta + \eta. \\
\|\pi - \Delta[\mathcal{A}(\pi) + \mathbf{e}]\|_{\mathcal{L}(\mathcal{H}),p} &\overset{*}{\leq} \mathrm{Bias}(\pi, \mathfrak{S}) + 2C\|\mathbf{e}\|_2^\delta + \eta
\end{aligned} \tag{95}$$

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

Appendix B. Proofs of Section 3

B.1 Proof of Lemma 1

Lemma 1 (Canas and Rosasco (2012)). *Consider $S \subseteq \mathcal{X}$, $p \in [1, +\infty[$ 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 we have:*

$$\mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = W_p^p(\pi, P_S\#\pi) \tag{32}$$

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

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

$$W_p^p(\pi, P_S\#\pi) \leq \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma(\mathbf{x}, \mathbf{y}) = \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] \tag{96}$$

Conversely, if γ^* is an optimal coupling for $W_q(\pi, P_S\#\pi)$ 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:

$$W_p^p(\pi, P_S\#\pi) = \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \geq \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) = \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) \tag{97}$$

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

$$\begin{aligned} W_p^p(\pi, \nu) &= \int D^p(\mathbf{x}, \mathbf{y})\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \overset{\mathrm{supp}(\nu)\subseteq S}{\geq} \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\gamma^*(\mathbf{x}, \mathbf{y}) \\ &= \int D^p(\mathbf{x}, P_S(\mathbf{x}))\mathrm{d}\pi(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim\pi}[D(\mathbf{x}, P_S(\mathbf{x}))^p] = W_p^p(\pi, P_S\#\pi) \end{aligned} \tag{98}$$

■

B.2 Proof of Proposition 3

Proposition 3 (Wasserstein learnability is necessary). *Consider $\mathcal{X} = \mathbb{R}^d$, $p \in [1, +\infty[$, and any model set $\mathfrak{S} \subseteq \mathcal{P}_p(\mathbb{R}^d)$. Consider a sketching operator $\mathcal{A}$ defined using random features $\Phi(\mathbf{x}) = (\phi(\mathbf{x}, \boldsymbol{\omega}_1), \cdots, \phi(\mathbf{x}, \boldsymbol{\omega}_m))^\top$ where $\boldsymbol{\omega}_i \sim \Lambda$. Assume that each $\phi(\cdot, \boldsymbol{\omega}_i)$, $i \in [\![m]\!]$, is L_i-Lipschitz with respect to the metric used to define the Wasserstein distance. If $\mathcal{A}$ satisfies (Hölder-LRIP) with error $\eta = 0$, constant $C > 0$ and $\delta = 1$ then we have:*

$$\forall \pi, \pi \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C' \, W_p(\pi, \pi') \tag{28}$$

where $C' = C\sqrt{\sum_{i=1}^m L_i^2}$. In other words, if $\mathcal{A}$ satisfies the LRIP ($\delta = 1$) then $\mathcal{L}(\mathcal{H})$ is p-Wasserstein learnable w.r.t. $\mathfrak{S}$.

Proof Under the hypothesis of the proposition we have:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \tag{99}$$

for some $C > 0$. We will show that the duality property of the Wasserstein distance implies $\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \leq W_1(\pi, \pi')$. Indeed we have for $\pi, \pi' \in \mathfrak{S}$:

$$\begin{aligned}\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2 &= \| \int_{\mathbb{R}^d} \Phi(\mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int_{\mathbb{R}^d} \Phi(\mathbf{y}) \mathrm{d}\pi'(\mathbf{y})\|_2^2 \\ &= \sum_{i=1}^m \left| \int \phi(\mathbf{x}, \boldsymbol{\omega}_i) \mathrm{d}\pi(\mathbf{x}) - \int \phi(\mathbf{y}, \boldsymbol{\omega}_i) \mathrm{d}\pi'(\mathbf{y}) \right|^2 \\ &\overset{\star}{\leq} \sum_{i=1}^m [L_i W_1(\pi, \pi')]^2 = \sum_{i=1}^m L_i^2 [W_1(\pi, \pi')]^2\end{aligned} \tag{100}$$

Where in $(\star)$ we used that ϕ is L_i-Lipschitz so that $|\int \phi(\mathbf{x}, \boldsymbol{\omega}_i) \mathrm{d}\pi(\mathbf{x}) - \int \phi(\mathbf{y}, \boldsymbol{\omega}_i) \mathrm{d}\pi'(\mathbf{y})| \leq L_i W_1(\pi, \pi')$ using the duality of the Wasserstein distance. Overall we have:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C \sqrt{\sum_{i=1}^m L_i^2} \cdot W_1(\pi, \pi') \tag{101}$$

Finally we use that $W_1(\pi, \pi') \leq W_p(\pi, \pi')$ since $p \in [1, +\infty[$ (Santambrogio, 2015, Section 5.1). ■

Appendix C. Proofs of Section 4

C.1 Convergence of finite samples

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

Lemma 11. *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:*

$$\mathbb{E}[\|\pi - \pi_n\|_\kappa^2] = n^{-1} \left(\int \kappa(\mathbf{x}, \mathbf{x}) \mathrm{d}\pi(\mathbf{x}) - \int \int \kappa(\mathbf{x}, \mathbf{y}) \mathrm{d}\pi(\mathbf{x}) \mathrm{d}\pi(\mathbf{y}) \right) \tag{102}$$

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

Lemma 12. *Let $\pi \in \mathcal{P}(\mathcal{X})$ and $\pi_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ where $\mathbf{x}_i \overset{i.i.d}{\sim} \pi$. If $\sup_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}) \leq K$ then for each $\delta \in (0, 2]$ we have:*

$$\mathbb{E}[\|\pi - \pi_n\|_\kappa^\delta] \leq (2K)^{\delta/2} n^{-\delta/2} \tag{103}$$

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 p.s.d. 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 $2/\delta \geq 1$. ■

C.2 Simple bound between Wassersein and distance between the means

Lemma 13. *Let $\pi, \pi' \in \mathcal{P}(\mathbb{R}^d)$. Then for every $1 \leq p < \infty$ we have:*

$$W_p(\pi, \pi') \geq \| \mathrm{m}(\pi) - \mathrm{m}(\pi') \|_2. \tag{104}$$

Proof Consider $\mathbf{u} \in \mathbb{R}^d$ an arbitrary unitary vector 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 $\|\mathbf{u}\|_2 = 1$ the function $f_{\mathbf{u}} : \mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz with respect to the Euclidean norm, hence by duality of the Wasserstein distance (cf (26))

$$|\langle \mathbf{u}, \mathrm{m}(\pi) - \mathrm{m}(\pi') \rangle| = \left| \int f_{\mathbf{u}}(\mathbf{x}) d\pi(\mathbf{x}) - \int f_{\mathbf{u}}(\mathbf{y}) d\pi(\mathbf{y}) \right| \leq W_1(\pi, \pi').$$

The supremum with respect to unitary vectors $\mathbf{u}$ yields $\|\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2 \leq W_1(\pi, \pi')$. The last step uses the fact that $W_1(\pi, \pi') \leq W_p(\pi, \pi')$ for any $p \in [1, +\infty[$ which concludes the proof. ■

C.3 Proof of Proposition 9

We will prove the following result:

Proposition 9. *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 p.s.d. 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*[19]*. Consider $\mathfrak{S} := \{(1-t)\pi_0 + t\pi_1, t \in [0,1]\}$. If $(\mathfrak{S}, W_p)$ is (κ, δ)-embeddable then $\delta \leq 1/p$.*

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

Lemma 14. *(Niles-Weed and Berthet, 2020, 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:*

$$\forall p \in [1, +\infty[, W_p(\pi_0, \pi_1) \geq c|\pi_0(S) - \pi_0(S)|^{1/p} \tag{105}$$

Proof [Of Proposition 9] This result is mainly taken from Theorem 9 in (Niles-Weed and Berthet, 2020) but we rewrite it in our context for completeness. For any $\lambda \in [0,1]$, set:

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

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 14 gives for any p:

$$W_p(\pi_\lambda, \pi'_\lambda) \geq c\,|\pi_\lambda(S) - \pi'_\lambda(S)|^{1/p} = c\lambda^{1/p} \tag{106}$$

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

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

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

C.4 Proof of Theorem 2

We recall that that for $M \in \mathbb{N}^*$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ the space of mixtures of M diracs located in $\mathcal{Y}$ is defined by:

$$\mathfrak{S}_M(\mathcal{Y}) := \{\sum_{i=1}^M a_i \delta_{\mathbf{x}_i}; a_i \in \mathbb{R}_+, \sum_{i=1}^M a_i = 1, \forall i \in [\![M]\!], \mathbf{x}_i \in \mathcal{Y}\} \tag{107}$$

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

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

Theorem 2 *Consider a TI p.s.d. 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 $p \in [1,+\infty[$, $\mathbf{x}_0 \in \mathbb{R}^d$, $R > 0$ and $\Omega = B(\mathbf{x}_0, R)$.*
If $(\mathfrak{S}_{\lfloor \frac{k}{2} \rfloor + 1}(\Omega), W_p)$ 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.

Lemma 15. *Consider a TI p.s.d. 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:*

$$c_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = c_2(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \tag{108}$$

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 $p \in [1,+\infty[$, 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:

$$\mu_\varepsilon := \sum_{i=1}^M \beta_i \delta_{\mathbf{x}_0 + \varepsilon \alpha_i \mathbf{u}} \tag{109}$$

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:

$$\pi_\varepsilon := \sum_{i \in T_+} \frac{\beta_i}{\rho} \delta_{\mathbf{x}_0 + \varepsilon \alpha_i \mathbf{u}} \text{ and } \pi'_\varepsilon := \sum_{i \in T_-} -\frac{\beta_i}{\rho} \delta_{\mathbf{x}_0 + \varepsilon \alpha_i \mathbf{u}} \tag{110}$$

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 we have $\mu_\varepsilon = \frac{1}{\rho}(\pi_\varepsilon - \pi'_\varepsilon)$. Hence:

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2 \|\mu_\varepsilon\|_\kappa^2 = \rho^2 \sum_{i,j=1}^M \beta_i \beta_j \kappa(\mathbf{x}_0 + \varepsilon \alpha_i \mathbf{u}, \mathbf{x}_0 + \varepsilon \alpha_j \mathbf{u}) = \rho^2 \sum_{i,j=1}^M \beta_i \beta_j \kappa_0(\varepsilon(\alpha_i - \alpha_j)\mathbf{u}) \tag{111}$$

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

$$\kappa_0(\varepsilon \mathbf{u}) = g(\varepsilon) = g(0) + \sum_{n=1}^k \frac{g^{(n)}(0)}{n!} \varepsilon^n + o_{\varepsilon \to 0}(\varepsilon^k) \tag{112}$$

hence

$$\begin{aligned} \|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 &= \rho^2 \sum_{i,j=1}^M \beta_i \beta_j \left(g(0) + \sum_{n=1}^k \frac{g^{(n)}(0)}{n!} (\alpha_i - \alpha_j)^n \varepsilon^n + o_{\varepsilon \to 0}(\varepsilon^k) \right) \\ &= \rho^2 \sum_{n=1}^k \left(\sum_{i,j=1}^M \beta_i \beta_j (\alpha_i - \alpha_j)^n \right) \varepsilon^k \frac{g^{(n)}(0)}{n!} + o_{\varepsilon \to 0}(\varepsilon^k) \end{aligned} \tag{113}$$

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:

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2 \sum_{n=1}^{k} c_n(\boldsymbol{\alpha},\boldsymbol{\beta})\varepsilon^k \frac{g^{(n)}(0)}{n!} + o_{\varepsilon\to 0}(\varepsilon^k) \tag{114}$$

Now, since by assumption we have

$$c_1(\boldsymbol{\alpha},\boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 0, \tag{115}$$

we get

$$\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^2 = \rho^2 c_k(\boldsymbol{\alpha},\boldsymbol{\beta})\varepsilon^k \frac{g^{(k)}(0)}{k!} + o_{\varepsilon\to 0}(\varepsilon^k) = O_{\varepsilon\to 0}(\varepsilon^k) \tag{116}$$

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:

$$W_p^p(\pi_\varepsilon,\pi'_\varepsilon) = \min_{\gamma\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i\in T_+, j\in T_-} \|\varepsilon\alpha_i\mathbf{u} - \varepsilon\alpha_j\mathbf{u}\|^p\gamma_{ij} = \varepsilon^p\|\mathbf{u}\|^p \min_{\gamma\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j|^p\gamma_{ij} \tag{117}$$

therefore

$$W_p^p(\pi_\varepsilon,\pi'_\varepsilon) \geq \left(\varepsilon\|\mathbf{u}\| \min_{i\in T_+, j\in T_-} |\alpha_i - \alpha_j|\right)^p \tag{118}$$

hence $W_p(\pi_\varepsilon,\pi'_\varepsilon) \geq \varepsilon\|\mathbf{u}\| \min_{i\in T_+,j\in T_-} |\alpha_i - \alpha_j|$. When $i \neq j$ we have that $\alpha_i \neq \alpha_j$ by assumption. Hence since $T_+ \cap T_- = \emptyset$ we have that $\min_{i\in T_+,j\in T_-} |\alpha_i - \alpha_j| > 0$. This discussion proves that, as soon as the condition (115) holds and $\delta > \frac{2}{k}$, we have:

$$\sup_{(\pi,\pi')\in\mathfrak{S}} \frac{W_p(\pi,\pi')}{\|\pi - \pi'\|_\kappa^\delta} \geq \sup_{\varepsilon>0} \frac{W_p(\pi_\varepsilon,\pi'_\varepsilon)}{\|\pi_\varepsilon - \pi'_\varepsilon\|_\kappa^\delta} \gtrsim \sup_{\varepsilon>0} \frac{\varepsilon}{\varepsilon^{\delta k/2}} = \sup_{\varepsilon>0} \varepsilon^{1-\delta k/2} = +\infty \tag{119}$$

So $(\mathfrak{S}, W_p)$ 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 16. *Consider a TI p.s.d. 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 15, 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:*

$$c_1(\boldsymbol{\alpha},\boldsymbol{\beta}) = c_2(\boldsymbol{\alpha},\boldsymbol{\beta}) = \cdots = c_{k-1}(\boldsymbol{\alpha},\boldsymbol{\beta}) = 0 \tag{120}$$

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:

$$\sum_{i=1}^{k+1}\beta_i = \sum_{i=1}^{k+1}\beta_i\alpha_i = \sum_{i=1}^{k+1}\beta_i\alpha_i^2 = \cdots = \sum_{i=1}^{k+1}\beta_i\alpha_i^{k-1} = 0 \tag{121}$$

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:

$$0 = \sum_{i=0}^{k}(-1)^i\binom{k}{i} = \sum_{i=1}^{k+1}(-1)^{i-1}\binom{k}{i-1} = \sum_{i=1}^{k+1}\beta_i \tag{122}$$

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

$$\sum_{i=1}^{k+1}\beta_i\alpha_i^s = \sum_{i=1}^{k+1}(-1)^{i-1}\binom{k}{i-1}i^s = \sum_{i=0}^{k}(-1)^i\binom{k}{i}(i+1)^s = \sum_{i=0}^{k}(-1)^i\binom{k}{i}\left(\sum_{l=0}^{s}\binom{s}{l}i^l\right) \tag{123}$$

So:

$$\sum_{i=1}^{k+1}\beta_i\alpha_i^s = \sum_{l=0}^{s}\binom{s}{l}\left(\sum_{i=0}^{k}(-1)^i\binom{k}{i}i^l\right) \tag{124}$$

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

$$\sum_{i=0}^{k}(-1)^i\binom{k}{i}i^l = \sum_{i=0}^{k}(-1)^{k-i}\binom{k}{k-i}(k-i)^l = \sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}(k-i)^l = (-1)^k\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^l \tag{125}$$

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 2.

Proof [Proof of Theorem 2] Define $(\boldsymbol{\alpha},\boldsymbol{\beta})$ as in Lemma 16. 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 15 with $M = k+1$.

■

C.5 Proof of Lemma 4

We recall the Lemma:

Lemma 4. *Consider $\pi, \pi' \in \mathcal{P}_2(\mathbb{R})$ with densities f, g with respect to the Lesbegue measure, i.e. $\pi \ll f\mathrm{d}x, \pi' \ll g\mathrm{d}x$. Let $\kappa(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$. Then we have:*

$$W_2(\pi,\pi') \leq (2\pi)^{-1/4}\left(\int_{\mathbb{R}}\frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4\hat{\kappa}_0(\omega)}\mathrm{d}\omega\right)^{1/4}\|\pi-\pi'\|_\kappa^{1/2} \tag{52}$$

where the Wasserstein distance is computed using $D(x,y) = |x-y|$.

Proof This proof is inspired from (Carrillo and Toscani, 2008, Theorem 2.21). With the previous notations we have $F' = f$ where F is the CDF. We have moreover $\widehat{F'}(\omega) = i\omega\hat{F}(\omega)$ (same with g). So $i\omega\hat{F}(\omega) = \hat{f}(\omega)$. This gives:

$$\widehat{F-G}(\omega) = \frac{\hat{f}(\omega)-\hat{g}(\omega)}{i\omega} \tag{126}$$

Moreover the Wasserstein distance in this case is given by $W_p(\pi,\pi') = (\int_{\mathbb{R}} |F(x)-G(x)|^p dx)^{1/p}$. For $p=2$ this gives:

$$W_2^2(\pi,\pi') = \int_{\mathbb{R}} |F(x)-G(x)|^2 \mathrm{d}x \tag{127}$$

Based on the Plancherel formula with the convention of the Fourier transform $\hat{f}(\omega) = \int_{\mathbb{R}} f(x) e^{-i\omega x} \mathrm{d}x$ we have:

$$\int_{\mathbb{R}} |F(x)-G(x)|^2 \mathrm{d}x = \frac{1}{2\pi} \int_{\mathbb{R}} |\hat{F}(\omega)-\hat{G}(\omega)|^2 \mathrm{d}\omega \tag{128}$$

which gives, using Cauchy-Schwarz inequality,

$$\begin{aligned} W_2^2(\pi,\pi') &= \frac{1}{2\pi} \int_{\mathbb{R}} |\omega|^{-2} |\hat{f}(\omega)-\hat{g}(\omega)|^2 \mathrm{d}\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{|\omega|^{-2}}{\sqrt{\hat{\kappa}_0(\omega)}} |\hat{f}(\omega)-\hat{g}(\omega)| \sqrt{\hat{\kappa}_0(\omega)} |\hat{f}(\omega)-\hat{g}(\omega)| \mathrm{d}\omega \\ &\leq \frac{1}{2\pi} \left(\int_{\mathbb{R}} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4 \hat{\kappa}_0(\omega)}) \mathrm{d}\omega \right)^{1/2} \left(\int_{\mathbb{R}} \hat{\kappa}_0(\omega) |\hat{f}(\omega)-\hat{g}(\omega)|^2 \mathrm{d}\omega \right)^{1/2} \\ &\overset{*}{=} (2\pi)^{-1/2} \left(\int_{\mathbb{R}} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4 \hat{\kappa}_0(\omega)}) \mathrm{d}\omega \right)^{1/2} \|\pi-\pi'\|_\kappa \end{aligned} \tag{129}$$

where in (*) we used the Lemma 17 below. ■

We recall that 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 $\widehat{\mu}(\boldsymbol{\omega}) := \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\mu(\mathbf{x})$. We have the following result:

Lemma 17. *Let $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ be a TI p.s.d. kernel on $\mathbb{R}^d \times \mathbb{R}^d$ where $\kappa_0 \in L_1(\mathbb{R}^d)$. Then for $\pi,\pi' \in \mathcal{P}(\mathbb{R}^d)$. Then:*

$$\|\pi-\pi'\|_\kappa^2 = (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \tag{130}$$

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

$$\|\pi-\pi'\|_\kappa^2 = (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\hat{f}(\boldsymbol{\omega}) - \hat{g}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \tag{131}$$

Proof This result can be found in (Sriperumbudur et al., 2010) but we rewrite the proof for completeness. Since κ_0 is a continuous p.s.d. function and $\kappa_0 \in L_1(\mathbb{R}^d)$ then by Bochner 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:

$$\forall \mathbf{x} \in \mathbb{R}^d,\ \kappa_0(\mathbf{x}) = (2\pi)^{-d} \int e^{i\boldsymbol{\omega}^\top \mathbf{x}} \widehat{\kappa_0}(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \tag{132}$$

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 theorem). We have:

$$\begin{aligned}
\|\pi - \pi'\|_\kappa^2 &= \int\int \kappa_0(\mathbf{x}-\mathbf{y})\mathrm{d}(\pi-\pi')(\mathbf{x})\mathrm{d}(\pi-\pi')(\mathbf{y}) \\
&\overset{\star}{=} \int\int\int e^{i\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{y})}\mathrm{d}\Lambda(\boldsymbol{\omega})\mathrm{d}(\pi-\pi')(\mathbf{x})\mathrm{d}(\pi-\pi')(\mathbf{y}) \\
&= \int \left(\int e^{i\boldsymbol{\omega}^\top\mathbf{x}}\mathrm{d}(\pi-\pi')(\mathbf{x})\right)\left(\int e^{-i\boldsymbol{\omega}^\top\mathbf{y}}\mathrm{d}(\pi-\pi')(\mathbf{y})\right)\mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= \int (\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega}))\overline{(\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega}))}\mathrm{d}\Lambda(\boldsymbol{\omega}) = \int |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&= (2\pi)^{-d}\int \widehat{\kappa_0}(\boldsymbol{\omega})|\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}
\end{aligned} \tag{133}$$

where in $(\star)$ we used (132) and Fubini theorem. ■

C.6 Proof of Theorem 3

We recall the theorem

Theorem 3 *Let $\kappa(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$ and such that $\widehat{\kappa}_0(\omega) > 0$ for every ω, $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{q_\kappa})$ as $\omega \to 0$ and $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{s_\kappa})$ as $\omega \to +\infty$ for some $q_\kappa > -1, s_\kappa \in \mathbb{R}_+$. Consider any $s \geq \frac{s_\kappa}{2} + 1$, $0 < M < +\infty$ and the following model set:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}_2(\mathbb{R}) : \exists f \in C^s(\mathbb{R}),\ \pi \ll f\mathrm{d}x,\ \|f\|_{W^{s,1}(\mathbb{R})} \leq M\} \tag{53}$$

1. *There exists $C = C(M,s,\kappa) > 0$ such that for every $1 \leq p \leq 2$:*

$$\forall \pi, \pi' \in \mathfrak{S},\ \text{if } \mathrm{m}(\pi) = \mathrm{m}(\pi') \text{ then } W_p(\pi,\pi') \leq C\|\pi - \pi'\|_\kappa^{1/2} \tag{54}$$

2. *If in addition κ_0 is L-Lipschitz continuous, then for every $1 \leq p \leq 2$*

$$\forall \pi, \pi' \in \mathfrak{S},\ W_p(\pi,\pi') \leq C\|\pi - \pi'\|_\kappa^{1/2} + CL^{1/4}|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|. \tag{55}$$

Proof We prove the results for $p = 2$. The results then also immediately hold for W_p, $1 \leq p \leq 2$ under the same hypothesis on $\mathfrak{S}$ since $W_p(\pi,\pi') \leq W_2(\pi,\pi')$ for any π, π'.

Case of distributions with the same mean (first claim). To prove the first result for $p = 2$ we prove the finiteness of the integral $\int_{\mathbb{R}} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4\widehat{\kappa}_0(\omega)}\mathrm{d}\omega$ from Lemma 4. Given any $R > 0$ we decompose it as $\int_{\mathbb{R}} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4\widehat{\kappa}_0(\omega)}\mathrm{d}\omega = \int_{|\omega|<R} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4\widehat{\kappa}_0(\omega)}\mathrm{d}\omega + \int_{|\omega|\geq R} \frac{|\hat{f}(\omega)-\hat{g}(\omega)|^2}{|\omega|^4\widehat{\kappa}_0(\omega)}\mathrm{d}\omega$ and use the shorthand $I_{|\omega|<R}$ and $I_{|\omega|\geq R}$ for the two terms. Since π, π' are in $\mathcal{P}_2(\mathbb{R})$ the Fourier transform of f, g are at least twice differentiable. Then, by a Taylor expansion at zero we have:

$$\hat{f}(\omega) = \sum_{n=0}^{K} \frac{\hat{f}^{(n)}(0)}{n!}\omega^n + H_{K,\hat{f}}(\omega) \tag{134}$$

where $H_{K,\hat{f}}(\omega)$ is the remainder (same for $\hat{g}$). We have moreover that $\hat{f}^{(0)}(0) = \hat{g}^{(0)}(0) = 1$ since f, g are probability density functions. Also using the properties of the Fourier transform we have $\hat{f}^{(1)}(0) = i^{-1}\mathbb{E}_{X\sim\alpha}[X]$ hence $\hat{f}^{(1)}(0) = \hat{g}^{(1)}(0)$ since $\mathrm{m}(\pi) = \mathrm{m}(\pi')$. So we have using Taylor with $K = 1$:

$$|\hat{f}(\omega) - \hat{g}(\omega)| = |H_{1,\hat{f}}(\omega) - H_{1,\hat{g}}(\omega)| \tag{135}$$

Moreover the remainder can be written as $H_{1,\hat{f}}(\omega) = \int_0^\omega \hat{f}^{(2)}(\omega)(\omega - t)\mathrm{d}t$ so that $|H_{1,\hat{f}}(\omega) - H_{1,\hat{g}}(\omega)| \leq \int_0^\omega |\hat{f}^{(2)}(t) - \hat{g}^{(2)}(t)||\omega - t|\mathrm{d}t$. Now note that for all $t \in \mathbb{R}$ $|\hat{f}^{(2)}(t)| = |\int f^{(2)}(x)e^{-itx}\mathrm{d}x| \leq \|f^{(2)}\|_{L_1(\mathbb{R})} \leq \sum_{j=0}^s \|f^{(j)}\|_{L_1(\mathbb{R})}$. Hence $|\hat{f}^{(2)}(t)| \leq \|f\|_{W^{s,1}(\mathbb{R})}$ (and same for g) and we obtain

$$|H_{1,\hat{f}}(\omega) - H_{1,\hat{g}}(\omega)| \leq \left(\|f\|_{W^{s,1}(\mathbb{R})} + \|g\|_{W^{s,1}(\mathbb{R})}\right) \int_0^\omega |\omega - t|\mathrm{d}t \leq 2M\frac{1}{2}|\omega|^2 = M|\omega|^2. \tag{136}$$

This gives $I_{|\omega|<R} \leq M^2 \int_{|\omega|<R} \frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega$. Since $\kappa_0 \in L^1(\mathbb{R})$ is a p.s.d kernel, its Fourier transform is non-negative, continuous, and decays to zero at infinity. Since $1/\widehat{\kappa_0}(\omega) = O(\omega^{q_\kappa})$ as $\omega \to 0$ with $q_\kappa > -1$ the integral $\int_{|\omega|<R} \frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega$ is finite (and obviously does not depends on f, g). Now consider $I_{|\omega|\geq R}$. We have:

$$\begin{aligned}
\int_{|\omega|\geq R} |\omega|^{-4}|\hat{f}(\omega) - \hat{g}(\omega)|^2 \frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega &= \int_{|\omega|\geq R} |\omega|^{2s-4}|\hat{f}(\omega) - \hat{g}(\omega)|^2|\omega|^{-2s}\frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega \\
&= \int_{|\omega|\geq R} (|\omega|^2)^{s-2}|\hat{f}(\omega) - \hat{g}(\omega)|^2|\omega|^{-2s}\frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega \\
&\leq \sup_{|\omega|\geq R}\left(|\omega|^{s-2}|\hat{f}(\omega) - \hat{g}(\omega)|\right)^2 \int_{|\omega|\geq R} \frac{|\omega|^{-2s}}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega
\end{aligned} \tag{137}$$

By hypothesis $\frac{|\omega|^{-2s}}{\widehat{\kappa_0}(\omega)} = O_{\omega\to+\infty}(|\omega|^{s_\kappa - 2s})$, since $s \geq \frac{s_\kappa}{2} + 1$ then $s_\kappa - 2s \leq -2$ so that $\frac{|\omega|^{-2s}}{\widehat{\kappa_0}(\omega)} = O_{\omega\to+\infty}(|\omega|^{-2})$ and $\int_{|\omega|\geq R} \frac{|\omega|^{-2s}}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega < +\infty$. Moreover:

$$\begin{aligned}
|\omega^{s-2}\hat{f}(\omega)| = |\widehat{f^{(s-2)}}(\omega)| = |\int f^{(s-2)}(x)e^{-i\omega x}\mathrm{d}x| &\leq \|f^{(s-2)}\|_{L_1(\mathbb{R})} \leq \sum_{j=0}^{s-2}\|f^{(j)}\|_{L_1(\mathbb{R})} \\
&\leq \sum_{j=0}^{s}\|f^{(j)}\|_{L_1(\mathbb{R})}
\end{aligned} \tag{138}$$

Hence $|\omega^{s-2}\hat{f}(\omega)| \leq \|f\|_{W^{s,1}(\mathbb{R})}$ and same with g. So:

$$|\omega|^{s-2}|\hat{f}(\omega) - \hat{g}(\omega)| \leq \|f\|_{W^{s,1}(\mathbb{R})} + \|g\|_{W^{s,1}(\mathbb{R})} \leq 2M \tag{139}$$

This gives $(|\omega|^{s-2}|\hat{f}(\omega) - \hat{g}(\omega)|)^2 \leq 4M^2$. Hence $\sup_{|\omega|\geq R}(|\omega|^{s-2}|\hat{f}(\omega) - \hat{g}(\omega)|)^2 \leq 4M^2$. This gives $C = C(M, s, \kappa, R) = M^2 \int_{|\omega|<R} \frac{1}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega + 4M^2 \int_{|\omega|\geq R} \frac{|\omega|^{-2s}}{\widehat{\kappa_0}(\omega)}\mathrm{d}\omega$. Since this is true for any $R > 0$ we can take the infimum over R to have a constant $C(M, s, \kappa)$.

Case of distributions with different means, with a Lipschitz kernel. Denote $T_{\mathbf{x}}$ the translation function $T_{\mathbf{x}}(\mathbf{y}) = \mathbf{x} + \mathbf{y}$ and $\mathrm{m}(\pi)$ the mean of a distribution π so that the distribution $T_{-\mathrm{m}(\pi)}\#\pi$ is centered. We will use the following lemma (see Appendix C.7):

Lemma 18 (Translation properties of W_2 and the MMD)**.** *For any $\pi, \pi' \in \mathcal{P}_2(\mathbb{R}^d)$ we have:*

$$W_2^2(\pi, \pi') = W_2^2(T_{-\mathrm{m}(\pi)}\#\pi, T_{-\mathrm{m}(\pi')}\#\pi') + \|\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2^2 \tag{140}$$

Moreover, if $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ is a p.s.d TI kernel on $\mathbb{R}^d$ such that κ_0 is L-Lipschitz continuous with respect to the Euclidean norm, then

$$\|\pi - \pi'\|_\kappa^2 = \|T_{-\mathrm{m}(\pi)}\#\pi - T_{-\mathrm{m}(\pi')}\#\pi'\|_\kappa^2 + R_{\pi,\pi',\kappa} \tag{141}$$

where $|R_{\pi,\pi',\kappa}| \leq L\|\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2$.

Let $\pi, \pi' \in \mathfrak{S}$. Using the translation properties of W_2 described in Lemma 18 we have

$$W_2^2(\pi, \pi') = W_2^2(T_{-\,\mathrm{m}(\pi)}\#\pi, T_{-\,\mathrm{m}(\pi')}\#\pi') + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^2$$

Since $T_{-\,\mathrm{m}(\pi)}\#\pi$ and $T_{-\,\mathrm{m}(\pi')}\#\pi'$ are both centered and as smooth as π, π', by the first claim of Theorem 3 (that we just proved) we have

$$W_2(T_{-\,\mathrm{m}(\pi)}\#\pi, T_{-\,\mathrm{m}(\pi')}\#\pi') \leq C\|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^{1/2}$$

where $C = C(M, s, \kappa) > 0$. Further, by the second part of Lemma 18 (with $d = 1$)

$$\|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^2 = \|\pi - \pi'\|_\kappa^2 - R_{\pi,\pi',\kappa} \leq \|\pi - \pi'\|_\kappa^2 + L|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')| \tag{142}$$

Since $\sqrt{a^2 + b^2} \leq a + b$ and $(a + b)^{1/4} \leq a^{1/4} + b^{1/4}$ for every $a, b \in \mathbb{R}_+$, we get

$$\begin{aligned} W_2(\pi, \pi') &\leq W_2(T_{-\,\mathrm{m}(\pi)}\#\pi, T_{-\,\mathrm{m}(\pi')}\#\pi') + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')| \\ &\leq C\|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^{1/2} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')| \\ &\leq C\|\pi - \pi'\|_\kappa^{1/2} + C \cdot L^{1/4}|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|^{1/4} + |\,\mathrm{m}(\pi) - \mathrm{m}(\pi')|. \end{aligned}$$

■

C.7 Proof of Lemma 18

Lemma 18 (Translation properties of W_2 and the MMD). *For any $\pi, \pi' \in \mathcal{P}_2(\mathbb{R}^d)$ we have:*

$$W_2^2(\pi, \pi') = W_2^2(T_{-\,\mathrm{m}(\pi)}\#\pi, T_{-\,\mathrm{m}(\pi')}\#\pi') + \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2^2 \tag{140}$$

Moreover, if $\kappa(\mathbf{x}, \mathbf{y}) = \kappa_0(\mathbf{x} - \mathbf{y})$ is a p.s.d TI kernel on $\mathbb{R}^d$ such that κ_0 is L-Lipschitz continuous with respect to the Euclidean norm, then

$$\|\pi - \pi'\|_\kappa^2 = \|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^2 + R_{\pi,\pi',\kappa} \tag{141}$$

where $|R_{\pi,\pi',\kappa}| \leq L\|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2$.

Proof For the first point see (Auricchio et al., 2020, Lemma 2). For the second point, using the translation invariance property of the kernel, it is easy to check that:

$$\|\pi - \pi'\|_\kappa^2 - \|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^2 = 2\mathbb{E}_{\mathbf{x}\sim\pi,\mathbf{y}\sim\pi'}[\kappa_0((\mathbf{x}-\mathbf{y})-(\mathrm{m}(\pi)-\mathrm{m}(\pi')))-\kappa_0(\mathbf{x}-\mathbf{y})] \tag{143}$$

Since κ_0 is L-Lipschitz with respect to the Euclidean norm we have for any $\mathbf{x}, \mathbf{y}$

$$|\kappa_0((\mathbf{x} - \mathbf{y}) - (\mathrm{m}(\pi) - \mathrm{m}(\pi'))) - \kappa_0(\mathbf{x} - \mathbf{y})| \leq L\|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2,$$

hence

$$\left|\|\pi - \pi'\|_\kappa^2 - \|T_{-\,\mathrm{m}(\pi)}\#\pi - T_{-\,\mathrm{m}(\pi')}\#\pi'\|_\kappa^2\right| \leq L \cdot \|\,\mathrm{m}(\pi) - \mathrm{m}(\pi')\|_2 \tag{144}$$

which proves (141) by defining $R_{\pi,\pi',\kappa} = 2\mathbb{E}_{\mathbf{x}\sim\pi,\mathbf{y}\sim\pi'}[\kappa_0((\mathbf{x} - \mathbf{y}) - (\mathrm{m}(\pi) - \mathrm{m}(\pi'))) - \kappa_0(\mathbf{x} - \mathbf{y})]$. ■

C.8 Proof of Corollary 3

Corollary 3 (GMM on $\mathbb{R}$). *Let $\kappa(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ with $\kappa_0 \in L_1(\mathbb{R})$ and such that $\widehat{\kappa}_0(\omega) > 0$ for every ω, $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{q_\kappa})$ as $\omega \to 0$ and $\frac{1}{\widehat{\kappa}_0(\omega)} = O(\omega^{s_\kappa})$ as $\omega \to +\infty$ for some $q_\kappa > -1, s_\kappa \in \mathbb{R}_+$. For $K \in \mathbb{N}^*$, $\Omega \subset \mathbb{R}$, and $\sigma_{\min} > 0$ consider the model set:*

$$\mathfrak{S}_{\texttt{GMM}}(\Omega, K, \sigma_{\min}) := \{\pi = \sum_{k=1}^K \alpha_k \mathcal{N}(c_k, \sigma_k^2), \boldsymbol{\alpha} \in \Delta_K, c_k \in \mathbb{R}, \sigma_k \geq \sigma_{\min}, \sum_{k=1}^K \alpha_k c_k \in \Omega\} \tag{57}$$

where $\Delta_K = \{\boldsymbol{\alpha} \in \mathbb{R}_+^K, \sum_{k=1}^K \alpha_k = 1\}$ is the probability simplex on $\mathbb{R}^K$.

1. *There exists a constant $C = C(\sigma_{\min}, K, \kappa) > 0$ such that if $\Omega = \{m\}$ is a prescribed mean then:*

$$\forall \pi, \pi' \in \mathfrak{S}_{\texttt{GMM}}(\Omega, K, \sigma_{\min}), W_2(\pi, \pi') \leq C \|\pi - \pi'\|_\kappa^{1/2} \tag{58}$$

2. *If in addition κ_0 is L-Lipschitz and $\mathrm{diam}(\Omega) := \sup_{x,y\in\mathcal{Y}} |x-y| < +\infty$, then for every $1 \leq p \leq 2$:*

$$\forall \pi, \pi' \in \mathfrak{S},\ W_p(\pi, \pi') \leq C' \|\pi - \pi'\|_{\tilde{\kappa}}^{1/2} \tag{59}$$

with $\tilde{\kappa}(x,y) := \kappa_0(x-y) + xy$. The constant C' depends only on $C(\sigma_{\min}, K, \kappa)$, L and $\mathrm{diam}(\Omega)$.

Proof Given the assumption on the kernel, it is sufficient to show that there exists $M = M(\sigma_{\min}) > 0$ such that $\mathfrak{S}_{\texttt{GMM}}(\Omega, K, \sigma_{\min}) \subseteq \mathfrak{S}$ where $\mathfrak{S}$ is defined in Theorem 3 with for $s := \lceil k/2+1 \rceil$. By Theorem 3 the conclusion will then hold using $C = C(M, s, \kappa)$.

The goal is thus to exhibit a radius M such that any probability distribution in $\mathfrak{S}_{\texttt{GMM}}(\Omega, K, \sigma_{\min})$ admits a density which lies in the Sobolev ball with radius M. Let $f(x|c,\sigma)$ be the density function for the Gaussian $\mathcal{N}(c, \sigma^2)$. In the following we note $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ and $H_n(x) = (-1)^n e^{x^2/2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}e^{-x^2/2}$ the Hermite polynomial. For any $n \in \mathbb{N}$, we have the relation $\phi^{(n)}(x) = (-1)^n H_n(x)\phi(x)$. Moreover $f(x|c,\sigma) = \phi(\frac{x-c}{\sigma})$ so:

$$\begin{aligned} \|f^{(n)}(\cdot|c,\sigma)\|_{L_1(\mathbb{R})} &= \int_{\mathbb{R}} |f^{(n)}(x|c,\sigma)|\mathrm{d}x = \sigma^{-n}\int_{\mathbb{R}} |\phi^{(n)}(\frac{x-c}{\sigma})|\mathrm{d}x = \sigma^{1-n}\int_{\mathbb{R}} |\phi^{(n)}(x)|\mathrm{d}x \\ &= \sigma^{1-n}\int_{\mathbb{R}} |H_n(x)|\phi(x)\mathrm{d}x = \sigma^{1-n}\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|H_n(X)|]. \end{aligned} \tag{145}$$

Now consider $\pi \in \mathfrak{S}_{\texttt{GMM}}(\Omega, K, \sigma_{\min})$ with density $F = \sum_{k=1}^K \alpha_k f(\cdot|c_k,\sigma_k)$. We have:

$$\begin{aligned} \|F^{(n)}\|_{L_1(\mathbb{R})} &= \int_{\mathbb{R}} |\sum_{k=1}^K \alpha_k f^{(n)}(x|c_k,\sigma_k)|\mathrm{d}x \leq \sum_{k=1}^K \alpha_k \int_{\mathbb{R}} |f^{(n)}(x|c_k,\sigma_k)|\mathrm{d}x \\ &= \sum_{k=1}^K \alpha_k \sigma_k^{1-n}\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|H_n(X)|] \end{aligned} \tag{146}$$

By Cauchy-Swartz inequality we have $\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|H_n(X)|] \leq (\mathbb{E}_{X\sim\mathcal{N}(0,1)}[|H_n(X)|^2])^{1/2} = \sqrt{n!}$ (Goldfeld et al., 2020, Lemma 1) so that we have for any integer $s \geq 1$:

$$\|F\|_{W^{s,1}(\mathbb{R})} = \sum_{n=0}^s \|F^{(n)}\|_{L_1(\mathbb{R})} \leq \sum_{n=0}^s \sum_{k=1}^K \alpha_k \sigma_k^{1-n}\sqrt{n!} \leq \max(1, \sigma_{\min}^{1-s}) \sum_{n=1}^s \sqrt{n!} \tag{147}$$

where in the last inequality we used that $\sigma_k \geq \sigma_{\min}$ and $a^{1-n} \leq \max(1, a^{1-s})$ for every $a \in \mathbb{R}_+^*$ and $1 \leq n \leq s$. This bounds is independent of F and true for any $s \geq 1$ hence we can set $M := \max(1, \sigma_{\min}^{1-s}) \sum_{n=1}^{\lceil k/2+1 \rceil} \sqrt{n!}$. ■

C.9 Proof of Lemma 6 and 7

In order to prove Lemma 6 we will need the following result:

Lemma 10. *Let $f \in C^s(\mathbb{R}^d)$ be integrable and compactly supported. Then for any $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ we have $\mathcal{R}[f](\cdot, \boldsymbol{\theta}) \in C^s(\mathbb{R}^d)$ and for any $|\boldsymbol{\alpha}| \leq s, \mathcal{R}[\partial^{\boldsymbol{\alpha}} f](\cdot, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\boldsymbol{\alpha}} \partial^{|\boldsymbol{\alpha}|} \mathcal{R}[f](\cdot, \boldsymbol{\theta})$*

Proof The proof of this result can be found in (Evans, 2010, Theorem 3) for $f \in C^\infty$ with compact support. For completeness we write the proof for our context. By definition, for each $i \in [\![d]\!]$,

$$\partial_i f = \frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \tag{148}$$

First we prove that for any $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}^{d-1}$ the partial derivative $\partial_t \mathcal{R}[f](t, \boldsymbol{\theta})$ exists and we have $\mathcal{R}[\partial_i f](t, \boldsymbol{\theta}) = \theta_i \partial_t \mathcal{R}[f](t, \boldsymbol{\theta})$ for each $i \in [\![d]\!]$. By induction it is easy to check that this implies for $|\boldsymbol{\alpha}| \leq s$:

$$\mathcal{R}[\partial^{\boldsymbol{\alpha}} f](\cdot, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\boldsymbol{\alpha}} \partial^{|\boldsymbol{\alpha}|} \mathcal{R}[f](\cdot, \boldsymbol{\theta}) \tag{149}$$

hence the claimed result.

First (we slightly postpone the justification of step $(\star)$ below), for $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{S}^{d-1}$ we have

$$\begin{aligned} \mathcal{R}[\partial_i f](t, \boldsymbol{\theta}) &= \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle = t} \partial_i f(\mathbf{x}) \mathrm{d}\mathbf{x} \overset{\star}{=} \lim_{h \to 0} \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle = t} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \mathrm{d}\mathbf{x} \\ &= \lim_{h \to 0} \mathcal{R}[\mathbf{x} \mapsto \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}](t, \boldsymbol{\theta}). \end{aligned} \tag{150}$$

Second, for any $h \neq 0$, since

$$\begin{aligned} \mathcal{R}[\mathbf{x} \mapsto f(\mathbf{x} + h\mathbf{e}_i)](t, \boldsymbol{\theta}) &= \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle = t} f(\mathbf{x} + h\mathbf{e}_i) \mathrm{d}\mathbf{x} = \int_{\mathbf{y}:\langle \mathbf{y} - h\mathbf{e}_i, \boldsymbol{\theta} \rangle = t} f(\mathbf{y}) \mathrm{d}\mathbf{y} \\ &= \int_{\mathbf{y}:\langle \mathbf{y}, \boldsymbol{\theta} \rangle = t + h\langle \mathbf{e}_i, \boldsymbol{\theta} \rangle} f(\mathbf{y}) \mathrm{d}\mathbf{y} = \int_{\mathbf{y}:\langle \mathbf{y}, \boldsymbol{\theta} \rangle = t + h\theta_i} f(\mathbf{y}) \mathrm{d}\mathbf{y} \\ &= \mathcal{R}[f](t + h\theta_i, \boldsymbol{\theta}) \end{aligned} \tag{151}$$

we obtain by linearity of the Radon transform (with respect to f)

$$\mathcal{R}[\mathbf{x} \mapsto \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}](t, \boldsymbol{\theta}) = \frac{\mathcal{R}[\mathbf{x} \mapsto f(\mathbf{x} + h\mathbf{e}_i)](t, \boldsymbol{\theta}) - \mathcal{R}[f](t, \boldsymbol{\theta})}{h} = \frac{\mathcal{R}[f](t + h\theta_i, \boldsymbol{\theta}) - \mathcal{R}[f](t, \boldsymbol{\theta})}{h} \tag{152}$$

Overall we have proven that for all $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ and $i \in [\![d]\!]$, $\lim_{h \to 0} \frac{\mathcal{R}[f](t + h\theta_i, \boldsymbol{\theta}) - \mathcal{R}[f](t, \boldsymbol{\theta})}{h}$ exists and is equal to $\mathcal{R}[\partial_i f](t, \boldsymbol{\theta})$. Since $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, there is at least one index $i^* \in [\![d]\!]$ such that $\theta_{i^*} \neq 0$. The fact that this limit exists for this index implies that $\partial_t \mathcal{R}[f](t, \boldsymbol{\theta})$ is well defined and satisfies $\mathcal{R}[\partial_{i^*} f](t, \boldsymbol{\theta}) = \theta_{i^*} \partial_t \mathcal{R}[f](t, \boldsymbol{\theta})$. It similarly follows that for every $i \in [\![d]\!]$ we have $\mathcal{R}[\partial_i f](t, \boldsymbol{\theta}) = \theta_i \partial_t \mathcal{R}[f](t, \boldsymbol{\theta})$ as claimed.

To finish the proof we have to justify that we can do $(\star)$. We will use the dominated convergence theorem to justify it. First see that f is compactly supported so there exists $R > 0$ such that $\mathrm{supp}(f) \subseteq B(0, R)$. Then as soon as $\|\mathbf{x}\|_2 > R + h$ we have that $\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} = 0$. Indeed in this case $\|\mathbf{x}\|_2 > R$ and $f(\mathbf{x}) = 0$. Moreover, $\|\mathbf{x} + h\mathbf{e}_i\|_2 \geq |\|\mathbf{x}\|_2 - \| - h\mathbf{e}_i\|_2| \geq \|\mathbf{x}\|_2 - h > R + h - h = R$ and thus $f(\mathbf{x} + h\mathbf{e}_i) = 0$. By noting $H_{\boldsymbol{\theta}, t} = \{\mathbf{x} : \langle \mathbf{x}, \boldsymbol{\theta} \rangle = t\}$ we have:

$$\begin{aligned} \int_{\mathbf{x}:\langle \mathbf{x}, \boldsymbol{\theta} \rangle = t} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} \mathrm{d}\mathbf{x} &= \int_{H_{\boldsymbol{\theta}, t}} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \partial_i f(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_{H_{\boldsymbol{\theta}, t}} \partial_i f(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &= \int_{H_{\boldsymbol{\theta}, t}} \mathbf{1}_{\|\mathbf{x}\|_2 \leq R + h}(\mathbf{x}) \left(\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \partial_i f(\mathbf{x}) \right) \mathrm{d}\mathbf{x} \\ &+ \int_{H_{\boldsymbol{\theta}, t}} \partial_i f(\mathbf{x}) \mathrm{d}\mathbf{x} \end{aligned} \tag{153}$$

We will show that the first term vanishes as $h \to 0$. Consider $\mathbf{x} = (x_1, \cdots, x_d) \in H_{\boldsymbol{\theta},t}$ and the function $g_{\mathbf{x}} : h \to \mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h}(\mathbf{x}) \left(\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} - \partial_i f(\mathbf{x}) \right)$. Then $g_{\mathbf{x}}$ converges pointwise to zero as $h \to 0$ since $\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h}$ tends to $\partial_i f(\mathbf{x})$ and $\mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h}(\mathbf{x})$ tends to $\mathbf{1}_{\|\mathbf{x}\|_2 \leq R}(\mathbf{x}) \leq 1$. Now take $0 < h < h_0$ sufficiently small and that does not depend on $\mathbf{x}$. We can first bound $|g_{\mathbf{x}}(h)| \leq \mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h_0}(\mathbf{x}) | \frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} - \partial_i f(\mathbf{x})|$.

If we note $F = f(x_1, \cdots, x_{i-1}, \cdot, x_{i+1}, \cdots, x_d)$ then by the mean value theorem there exists $t \in]x_i, x_i + h[$ such that $\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} = \frac{F(x_i+h)-F(x_i)}{h} = F'(t)$. By defining:

$$\mathbf{z}_{\mathbf{x},h} = (x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_d)$$

(which depends on $\mathbf{x}, h$) we have $\frac{f(\mathbf{x}+h\mathbf{e}_i)-f(\mathbf{x})}{h} = \partial_i f(\mathbf{z}_{\mathbf{x},h})$. Then

$$|g_{\mathbf{x}}(h)| \leq \mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h_0}(\mathbf{x}) |\partial_i f(\mathbf{z}_{\mathbf{x},h}) - \partial_i f(\mathbf{x})|.$$

Note that $\|\mathbf{z}_{\mathbf{x},h} - \mathbf{x}\|_2 = |t - x_i| \leq h < h_0$ so $\|\mathbf{z}_{\mathbf{x},h}\|_2 \leq \|x\|_2 + h_0 \leq R + 2h_0$ and $|\partial_i f(\mathbf{z}_{\mathbf{x},h})| \leq \sup_{\|\mathbf{y}\|_2 \leq R+2h_0} |\partial_i f(\mathbf{y})|$. As a consequence:

$$\mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h_0}(\mathbf{x}) |\partial_i f(\mathbf{z}_{\mathbf{x},h}) - \partial_i f(\mathbf{x})| \leq 2\mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h_0}(\mathbf{x}) \sup_{\|\mathbf{x}\|_2 \leq R+2h_0} |\partial_i f(\mathbf{x})| \tag{154}$$

Using the fact that $\partial_i f$ is continuous and $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq R + 2h_0\}$ is compact the supremum is finite and smaller that some $M = M(f, R, h_0) > 0$. Overall for all $\mathbf{x} \in H_{\boldsymbol{\theta},t}$ and $0 < h < h_0$ we have $|g_{\mathbf{x}}(h)| \leq 2M\mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h_0}(\mathbf{x})$ which is an integrable function on $H_{\boldsymbol{\theta},t}$ and does not depend on h. Using the dominated convergence theorem we have then:

$$\lim_{h \to 0} \int_{H_{\boldsymbol{\theta},t}} \mathbf{1}_{\|\mathbf{x}\|_2 \leq R+h}(\mathbf{x}) \left(\frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h} - \partial_i f(\mathbf{x}) \right) \mathrm{d}\mathbf{x} = 0 \tag{155}$$

■

Lemma 6. *Suppose that $d \geq 2$. Let $f \in C^s(\mathbb{R}^d)$ be integrable and compactly supported. For any $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ the Radon transform satisfies $\mathcal{R}[f](\cdot, \boldsymbol{\theta}) \in C^s(\mathbb{R})$ and $\|\mathcal{R}[f](\cdot, \boldsymbol{\theta})\|_{W^{s,1}(\mathbb{R})} \leq d^{s+1} \|f\|_{W^{s,1}(\mathbb{R}^d)}$*

Proof By Lemma 10, since f has bounded support and continuous partial derivatives of order s, each of its Radon slices $t \mapsto \mathcal{R}[f](t, \boldsymbol{\theta})$ is s-times differentiable in t and we have for each $|\boldsymbol{\alpha}| \leq s, \mathcal{R}[\partial^{\boldsymbol{\alpha}} f](\cdot, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\boldsymbol{\alpha}} \partial^{|\boldsymbol{\alpha}|} \mathcal{R}[f](\cdot, \boldsymbol{\theta})$ hence

$$\|\boldsymbol{\theta}^{\boldsymbol{\alpha}} \partial^{|\boldsymbol{\alpha}|} \mathcal{R}[f](\cdot, \boldsymbol{\theta})\|_{L_1(\mathbb{R})} = \|\mathcal{R}[\partial^{\boldsymbol{\alpha}} f](\cdot, \boldsymbol{\theta})\|_{L_1(\mathbb{R})} \leq \|\partial^{\boldsymbol{\alpha}} f\|_{L_1(\mathbb{R}^d)} \tag{156}$$

where in the last inequality we used that the $L_1(\mathbb{R})$ norm of the radon transform is smaller than the $L_1(\mathbb{R}^d)$ norm of the function. Indeed, for all $\boldsymbol{\theta}$ and all functions $g \in L_1(\mathbb{R}^d)$ we have:

$$\begin{aligned} \|\mathcal{R}[g](\cdot, \boldsymbol{\theta})\|_{L_1(\mathbb{R})} &= \int_{\mathbb{R}} |\mathcal{R}[g](t, \boldsymbol{\theta})| \mathrm{d}t = \int_{\mathbb{R}} | \int_{\mathbf{y} \in \boldsymbol{\theta}^\perp} g(t\boldsymbol{\theta} + \mathbf{y}) \mathrm{d}\mathbf{y}| \mathrm{d}t \\ &\leq \int_{\mathbb{R}} \int_{\mathbf{y} \in \boldsymbol{\theta}^\perp} |g(t\boldsymbol{\theta} + \mathbf{y})| \mathrm{d}\mathbf{y} \mathrm{d}t = \int_{\mathbb{R}^d} |g(\mathbf{x})| \mathrm{d}\mathbf{x} = \|g\|_{L_1(\mathbb{R}^d)} \end{aligned} \tag{157}$$

where for the last equality we used a orthogonal change of variable and used that $\mathbf{x} \in \mathbb{R}^d$ can be written as the projection on the line supported by $\boldsymbol{\theta}$ plus an vector that is orthogonal to this line $\mathbf{x} = t\boldsymbol{\theta} + \mathbf{y}$ where $\mathbf{y} \in \boldsymbol{\theta}^\perp, t \in \mathbb{R}$. As a result, for any integer $0 \leq k \leq s$ we have:

$$\sum_{\boldsymbol{\alpha} : |\boldsymbol{\alpha}| = k} \|\boldsymbol{\theta}^{\boldsymbol{\alpha}} \partial^k \mathcal{R}[f](\cdot, \boldsymbol{\theta})\|_{L_1(\mathbb{R})} = \sum_{\boldsymbol{\alpha} : |\boldsymbol{\alpha}| = k} \|\mathcal{R}[\partial^{\boldsymbol{\alpha}} f](\cdot, \boldsymbol{\theta})\|_{L_1(\mathbb{R})} \leq \sum_{\boldsymbol{\alpha} : |\boldsymbol{\alpha}| = k} \|\partial^{\boldsymbol{\alpha}} f\|_{L_1(\mathbb{R}^d)} \tag{158}$$

hence

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=k} \|\boldsymbol{\theta}^{\boldsymbol{\alpha}}\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} \leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq s} \|\partial^{\boldsymbol{\alpha}} f\|_{L_1(\mathbb{R}^d)} = \|f\|_{W^{s,1}(\mathbb{R}^d)}. \tag{159}$$

In addition, for each $0 \leq k \leq s$ we also have

$$\begin{aligned}\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=k} \|\boldsymbol{\theta}^{\boldsymbol{\alpha}}\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} &= \left(\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=k} |\boldsymbol{\theta}^{\boldsymbol{\alpha}}|\right) \|\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} \\ &\overset{\star}{\geq} \frac{\|\boldsymbol{\theta}\|_2^k}{d^k} \|\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} \overset{\|\boldsymbol{\theta}\|_2=1}{=} d^{-k} \|\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})}\end{aligned} \tag{160}$$

Where in $(\star)$ we used Lemma 20 in Section C.12. Combining (159) and (160) we obtain that for each $0 \leq k \leq s$

$$\|\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} \leq d^k \|f\|_{W^{s,1}(\mathbb{R}^d)}, \tag{161}$$

hence

$$\|\mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{W^{s,1}(\mathbb{R})} = \sum_{k=0}^{s} \|\partial^k \mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{L_1(\mathbb{R})} \leq \|f\|_{W^{s,1}(\mathbb{R}^d)} (\sum_{k=0}^{s} d^k) = \frac{d^{s+1}-1}{d-1} \|f\|_{W^{s,1}(\mathbb{R}^d)} \tag{162}$$

Then if $d \geq 2$ we have $\|\mathcal{R}[f](\cdot,\boldsymbol{\theta})\|_{W^{s,1}(\mathbb{R})} \leq d^{s+1} \|f\|_{W^{s,1}(\mathbb{R}^d)}$ which concludes the proof. ■

Lemma 7. *Let $\kappa_{\mathbb{R}}(x,y) = \kappa_0(x-y)$ be a TI p.s.d. kernel on $\mathbb{R}$ where κ_0 is continuous. Consider the kernel κ on $\mathbb{R}^d$ defined by $\kappa(\mathbf{x},\mathbf{y}) := \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top\mathbf{x},\boldsymbol{\theta}^\top\mathbf{y})] = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_0(\boldsymbol{\theta}^\top(\mathbf{x}-\mathbf{y}))]$. Then κ is TI, continuous, bounded and positive semi-definite. Moreover we have for any $(\pi,\pi') \in \mathcal{P}(\mathbb{R}^d)$:*

$$\|\pi - \pi'\|_\kappa^2 = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^2] \tag{64}$$

Proof The fact that κ is translation invariant is obvious as it can be written $\kappa(\mathbf{x},\mathbf{y}) = f(\mathbf{x}-\mathbf{y})$ with $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\cdot) := \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_0(\langle\boldsymbol{\theta},\cdot\rangle)]$. To show that κ is continuous we show that f is continuous: take a sequence $(\mathbf{z}_n)_{n\in\mathbb{N}}$ such that $\mathbf{z}_n \underset{n\to+\infty}{\to} \mathbf{z}$. Then:

$$|f(\mathbf{z}_n) - f(\mathbf{z})| \leq \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[|\kappa_0(\langle\boldsymbol{\theta},\mathbf{z}_n\rangle) - \kappa_0(\langle\boldsymbol{\theta},\mathbf{z}\rangle)|] \tag{163}$$

For $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, we have $\lim_{n\to+\infty} |\kappa_0(\langle\boldsymbol{\theta},\mathbf{z}_n\rangle) - \kappa_0(\langle\boldsymbol{\theta},\mathbf{z}\rangle)| = 0$ since κ_0 is continuous. Also, for $n \in \mathbb{N}, \boldsymbol{\theta} \in \mathbb{S}^{d-1}, |\kappa_0(\langle\boldsymbol{\theta},\mathbf{z}_n\rangle) - \kappa_0(\langle\boldsymbol{\theta},\mathbf{z}\rangle)| \leq 2\kappa_0(0)$ since κ_0 is a p.s.d. function (Wendland, 2004, Theorem 6.2). Since $2\kappa_0(0)$ is σ-integrable we can apply the dominated convergence theorem that gives $\lim_{n\to+\infty} \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[|\kappa_0(\langle\boldsymbol{\theta},\mathbf{z}_n\rangle) - \kappa_0(\langle\boldsymbol{\theta},\mathbf{z}\rangle)|] = 0$. So f is continuous. Moreover it is bounded since $|\kappa_0(\mathbf{x})| \leq \kappa_0(0)$ thus $|\kappa(\mathbf{x},\mathbf{y})| \leq \kappa_0(0)$. Now take $\mathbf{x}_1,\cdots,\mathbf{x}_n \in \mathbb{R}^d$ and $c_1,\cdots,c_n \in \mathbb{R}$. Then:

$$\sum_{i,j=1}^{n} c_i c_j \kappa(\mathbf{x}_i,\mathbf{x}_j) = \sum_{i,j=1}^{n} c_i c_j \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top\mathbf{x}_i,\boldsymbol{\theta}^\top\mathbf{x}_j)] = \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\sum_{i,j=1}^{n} c_i c_j \kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top\mathbf{x}_i,\boldsymbol{\theta}^\top\mathbf{x}_j)]$$

which is ≥ 0 since $\kappa_{\mathbb{R}}$ is a *p.s.d.* kernel. So the kernel κ defines a valid MMD. Moreover we have by definition:

$$\begin{aligned}\int \kappa(\mathbf{x},\mathbf{y})\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi'(\mathbf{y}) &= \int \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top\mathbf{x},\boldsymbol{\theta}^\top\mathbf{y})]\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi'(\mathbf{y}) \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}\left[\int \kappa_{\mathbb{R}}(\boldsymbol{\theta}^\top\mathbf{x},\boldsymbol{\theta}^\top\mathbf{y})\mathrm{d}\pi(\mathbf{x})\mathrm{d}\pi'(\mathbf{y})\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}\left[\int \kappa_{\mathbb{R}}(x,y)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi(x)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi'(y)\right]\end{aligned} \tag{164}$$

Hence using the definition of the MMD:

$$\begin{aligned}\|\pi-\pi'\|_\kappa^2 &= \int\int \kappa \mathrm{d}\pi\mathrm{d}\pi + \int\int \kappa \mathrm{d}\pi'\mathrm{d}\pi' - 2\int\int \kappa \mathrm{d}\pi\mathrm{d}\pi' \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim\sigma}\left[\int \kappa_{\mathbb{R}}(x,y)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi(x)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi(y) + \int \kappa_{\mathbb{R}}(x,y)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi'(x)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi'(x)\right. \\ &\left. -2\int \kappa_{\mathbb{R}}(x,y)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi(x)\mathrm{d}P_{\boldsymbol{\theta}}\#\pi'(y)\right]\end{aligned} \tag{165}$$

which is by definition $\mathbb{E}_{\boldsymbol{\theta}\sim\sigma}[\|P_{\boldsymbol{\theta}}\#\pi - P_{\boldsymbol{\theta}}\#\pi'\|_{\kappa_{\mathbb{R}}}^2]$. ■

C.10 Proof of Lemma 9

With some abuse of notations when f is a probability density function we will note $\mathbf{x}\sim f$ which means $\mathbf{x}\sim\pi$ where π is the probability distribution associated to f. We use the following result:

Lemma 19 (Lemma 6 in (Nguyen, 2013)). *Assume that f and g are two probability density functions on $(\mathbb{R}^d, \|.\|_2)$ with bounded s -moments. Then for $t\in\mathbb{R}$ such that $0<t<s$,*

$$\int \|\mathbf{x}\|_2^t\,|f(\mathbf{x})-g(\mathbf{x})|\,\mathrm{d}\mathbf{x} \leq 2\,\|f-g\|_{L_1(\mathbb{R}^d)}^{(s-t)/s}\left(\mathbb{E}_{\mathbf{x}\sim f}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim g}\|\mathbf{y}\|_2^s\right)^{t/s}$$

Let $V_d=\pi^{d/2}\Gamma(d/2+1)$ denote the volume of the d -dimensional unit sphere. Then,

$$\|f-g\|_{L_1(\mathbb{R}^d)} \leq 2V_d^{s/(d+2s)}\left(\mathbb{E}_{\mathbf{x}\sim f}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim g}\|\mathbf{y}\|_2^s\right)^{d/(d+2s)}\|f-g\|_{L_2(\mathbb{R}^d)}^{2s/(d+2s)}$$

We recall the statement of Lemma 9 which is to be proved:

Lemma 9. *Let $s>1$. Assume that $\pi,\pi'\in\mathcal{P}_s(\mathbb{R}^d)$ have densities f,g with respect to the Lebesgue measure. Then for any $1\leq p<s$ we have:*

$$W_p(\pi,\pi') \leq 2^{\frac{1}{p}+1-\frac{1}{s}}V_d^{\frac{s-p}{(d+2s)p}}\left(\mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^s] + \mathbb{E}_{\mathbf{y}\sim\pi'}[\|\mathbf{y}\|_2^s]\right)^{\frac{2p+d}{(d+2s)p}}\|f-g\|_{L_2(\mathbb{R}^d)}^{\frac{2(s-p)}{(d+2s)p}} \tag{76}$$

with $V_d=\pi^{d/2}\Gamma(d/2+1)$ the volume of the d-dimensional unit sphere.

Proof First, we use the fact (Villani, 2008, Theorem 6.15) that the Wasserstein distance is bounded by a weighted Total Variation distance

$$W_p^p(\pi,\pi') \leq 2^{p-1}\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^p\mathrm{d}|\pi-\pi'|(\mathbf{x}) = 2^{p-1}\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^p|f(\mathbf{x})-g(\mathbf{x})|\mathrm{d}\mathbf{x}. \tag{166}$$

Second, using Lemma 19 we obtain

$$\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^p|f(\mathbf{x})-g(\mathbf{x})|\mathrm{d}\mathbf{x} \leq 2\|f-g\|_{L_1(\mathbb{R}^d)}^{\frac{s-p}{s}}\left(\mathbb{E}_{\mathbf{x}\sim f}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim g}\|\mathbf{y}\|_2^s\right)^{p/s} \tag{167}$$

as well as

$$\|f-g\|_{L_1(\mathbb{R}^d)} \leq 2V_d^{s/(d+2s)}\left(\mathbb{E}_{\mathbf{x}\sim\pi}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim\pi'}\|\mathbf{y}\|_2^s\right)^{d/(d+2s)}\|f-g\|_{L_2(\mathbb{R}^d)}^{2s/(d+2s)}. \tag{168}$$

Combining both inequalities yields the following bound, which allows to conclude:

$$\begin{aligned}\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^p|f(\mathbf{x})-g(\mathbf{x})|\mathrm{d}\mathbf{x} \leq 2^{2-\frac{p}{s}}V_d^{\frac{s-p}{(d+2s)}}&\left(\mathbb{E}_{\mathbf{x}\sim\pi}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim\pi'}\|\mathbf{y}\|_2^s\right)^{\frac{p}{s}+\left(1-\frac{p}{s}\right)\frac{d}{d+2s}} \\ &\times\|f-g\|_{L_2(\mathbb{R}^d)}^{\frac{2(s-p)}{d+2s}}\end{aligned} \tag{169}$$

Using (166) (*i.e.* multiplying by 2^{p-1} and taking the power $1/p$):

$$W_p(\pi,\pi') \leq 2^{\frac{s+p(s-1)}{ps}} V_d^{\frac{s-p}{(d+2s)p}} \left(\mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^s] + \mathbb{E}_{\mathbf{y}\sim\pi'}[\|\mathbf{y}\|_2^s]\right)^{\frac{2p+d}{(d+2s)p}} \|f-g\|_{L_2(\mathbb{R}^d)}^{\frac{2(s-p)}{p(d+2s)}} \tag{170}$$

■

C.11 Proof of Lemma 10, Proposition 10 and Theorem 5

We recall that 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 $\widehat{\mu}(\boldsymbol{\omega}) := \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{x}} \mathrm{d}\mu(\mathbf{x})$.

Lemma 10. *Let Φ be a regularizer and $\kappa_0 := \Phi * \Phi$. Then $\kappa_0 \in L_1(\mathbb{R}^d)$ is even, bounded, continuous and has non-negative Fourier transform. Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) := \kappa_0(\mathbf{x}-\mathbf{y})$. We have that κ defines a TI p.s.d. kernel. Moreover, for $\pi,\pi' \in \mathcal{P}(\mathbb{R}^d)$:*

$$\|\pi - \pi'\|_\kappa = \|\Phi * \pi - \Phi * \pi'\|_{L_2(\mathbb{R}^d)} \tag{77}$$

Proof We first prove that the kernel in this proposition defines a TI *p.s.d.* 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 Φ is continuous and bounded. Moreover since Φ is even its Fourier transform is real-valued hence $\widehat{\kappa}_0 = \hat{\Phi}^2 = |\hat{\Phi}|^2 \geq 0$ so the Fourier transform of κ_0 is non negative. Finally we have $\kappa_0 \in L_1(\mathbb{R}^d)$ as the convolution of two integrable functions. Using Bochner's theorem (see Theorem 2) shows that the kernel κ is a TI p.s.d. kernel. Moreover we have:

$$\|\Phi * \pi - \Phi * \pi\|_{L_2(\mathbb{R}^d)}^2 = \int |\Phi * \pi(\mathbf{x}) - \Phi * \pi'(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \overset{\star}{=} (2\pi)^{-d} \int |\widehat{\Phi * \pi}(\boldsymbol{\omega}) - \widehat{\Phi * \pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \tag{171}$$

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

$$\|\Phi*\pi - \Phi*\pi\|_{L_2(\mathbb{R}^d)}^2 = (2\pi)^{-d} \int |\widehat{\Phi}(\boldsymbol{\omega})\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\Phi}(\boldsymbol{\omega})\widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} = (2\pi)^{-d} \int |\widehat{\Phi}(\boldsymbol{\omega})|^2 |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}. \tag{172}$$

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

$$\|\Phi * \pi - \Phi * \pi\|_{L_2(\mathbb{R}^d)}^2 = (2\pi)^{-d} \int \widehat{\kappa_0}(\boldsymbol{\omega}) |\widehat{\pi}(\boldsymbol{\omega}) - \widehat{\pi'}(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \overset{\star\star}{=} \|\pi - \pi'\|_\kappa^2 \tag{173}$$

where in $(\star\star)$ we used Lemma 17. This concludes the proof. ■

Proposition 10. *Let $s > 1$. Consider a regularizer Φ with s-finite moments. Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ where $\kappa_0 := \Phi * \Phi$. It defines a TI p.s.d kernel by Lemma 10. Moreover, we have for any $\pi,\pi' \in \mathcal{P}_s(\mathbb{R}^d)$ and $1 \leq p < s$:*

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

where $C_{d,s,p} = 2^{\frac{1}{p}+1-\frac{1}{s}} V_d^{\frac{s-p}{(d+2s)p}}$ is a constant.

Proof In order to prove the proposition we will apply the Lemma 9 with π_Φ and π'_Φ that admit the densities $f = \Phi * \pi$ and $g = \Phi * \pi'$ and thus the term $\|f-g\|_{L_2(\mathbb{R}^d)}$ in Lemma 9 becomes $\|f-g\|_{L_2} = \|\Phi * \pi - \Phi * \pi'\|_{L_2(\mathbb{R}^d)}$. To apply Lemma 9 we need to show that π_Φ, π'_Φ have s-finite moments which will be true by using that π, π' and Φ have s-finite moments. Indeed:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}\|\mathbf{x}\|_2^s &= \int \|\mathbf{x}\|_2^s (\Phi * \pi)(\mathbf{x})\mathrm{d}\mathbf{x} = \int \|\mathbf{x}\|_2^s \left(\int \Phi(\mathbf{x}-\mathbf{y})\mathrm{d}\pi(\mathbf{y})\right)\mathrm{d}\mathbf{x} \\ &\overset{\star}{=} \int\int \|\mathbf{x}\|_2^s \Phi(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\pi(\mathbf{y}) = \int \left(\int \|\mathbf{x}\|_2^s \Phi(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{x}\right)\mathrm{d}\pi(\mathbf{y})\end{aligned} \tag{174}$$

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

$$\int \|\mathbf{x}\|_2^s \Phi(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{x} = \int \|\mathbf{y}+\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z} \leq 2^s \left(\|\mathbf{y}\|_2^s \int \Phi(\mathbf{z})\mathrm{d}\mathbf{z} + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z}\right) \tag{175}$$

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

$$\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}\|\mathbf{x}\|_2^s \leq 2^s \left(\int \|\mathbf{y}\|_2^s \mathrm{d}\pi(\mathbf{y}) + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z}\right) < +\infty \tag{176}$$

So by applying Lemma 9 we have:

$$W_p(\pi,\pi') \leq 2^{\frac{1}{p}+1-\frac{1}{s}} V_d^{\frac{s-p}{(d+2s)p}} \left(\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}\|\mathbf{x}\|_2^s + \mathbb{E}_{\mathbf{y}\sim\pi'_\Phi}\|\mathbf{y}\|_2^s\right)^{\frac{2p+d}{(d+2s)p}} \|\Phi*\pi - \Phi*\pi'\|_{L_2(\mathbb{R}^d)}^{\frac{2(s-p)}{(d+2s)p}} \tag{177}$$

Finally, to relate the term $\|\Phi * \pi - \Phi * \pi'\|_{L_2(\mathbb{R}^d)}$ with the MMD we use the Lemma 10

■

Finally we can prove the following theorem:

Theorem 5 *Let $s > 1$. Consider a regularizer Φ with s-bounded moments. Consider the kernel $\kappa(\mathbf{x},\mathbf{y}) = \kappa_0(\mathbf{x}-\mathbf{y})$ where $\kappa_0 := \Phi * \Phi$. It defines a TI p.s.d kernel by Lemma 10. We consider the following model set:*

$$\mathfrak{S} := \{\pi \in \mathcal{P}(\mathbb{R}^d), \mathbb{E}_{\mathbf{x}\sim\pi}[\|\mathbf{x}\|_2^s] \leq M\} \tag{78}$$

Then for any $1 \leq p < s$ there exists a constant $C = C_{d,s,p} > 0$ such that:

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

Proof With the notations of the theorem we have by Proposition 10:

$$W_p(\pi_\Phi,\pi'_\Phi) \leq C_{d,s,p}\left(\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}[\|\mathbf{x}\|_2^s] + \mathbb{E}_{\mathbf{y}\sim\pi'_\Phi}[\|\mathbf{y}\|_2^s]\right)^{\frac{2p+d}{(d+2s)p}} \|\pi-\pi'\|_\kappa^{\frac{2(s-p)}{(d+2s)p}} \tag{178}$$

where $C_{d,s,p}$ is defined in Proposition 10. We can control both terms $\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}[\|\mathbf{x}\|_2^s], \mathbb{E}_{\mathbf{y}\sim\pi'_\Phi}[\|\mathbf{y}\|_2^s]$ as in the proof of Proposition 10 so that:

$$\mathbb{E}_{\mathbf{x}\sim\pi_\Phi}[\|\mathbf{x}\|_2^s] \leq 2^s\left(\int \|\mathbf{y}\|_2^s \mathrm{d}\pi(\mathbf{y}) + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z}\right) \leq 2^s(M + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z}) \tag{179}$$

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

$$W_p(\pi_\Phi,\pi'_\Phi) \leq C_{d,s,p} 2^{(s+1)(\frac{2p+d}{(d+2s)p})}(M + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z})\mathrm{d}\mathbf{z})^{\frac{2p+d}{(d+2s)p}} \|\pi-\pi'\|_\kappa^{\frac{2(s-p)}{(d+2s)p}} \tag{180}$$

Consequently, by defining $C'_{d,s,p} = 2^{(s+1)\frac{2p+d}{(d+2s)p}} C_{d,s,p}$ and using Lemma 8 we have:

$$W_p(\pi, \pi') \leq C'_{d,s,p}(M + \int \|\mathbf{z}\|_2^s \Phi(\mathbf{z}) \mathrm{d}\mathbf{z})^{\frac{2p+d}{(d+2s)p}} + 2\left(\int \|\mathbf{z}\|_2^p \Phi(\mathbf{z}) \mathrm{d}\mathbf{z}\right)^{1/p} \tag{181}$$

which concludes the proof. ■

C.12 Postponed results

Lemma 20. *Consider a vector* $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ *and* $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}^d$ *a multi index. For any* $k \in \mathbb{N}$ *we have:*

$$\|\mathbf{x}\|_2^k \leq d^k \sum_{|\boldsymbol{\alpha}|=k} |\mathbf{x}^{\boldsymbol{\alpha}}| \tag{182}$$

Proof We have $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$ hence for $k \in \mathbb{N}$:

$$\|\mathbf{x}\|_2^k \leq (\sum_{i=1}^d |x_i|)^k = \sum_{|\boldsymbol{\alpha}|=k} \frac{k!}{\boldsymbol{\alpha}!} |\mathbf{x}^{\boldsymbol{\alpha}}| \tag{183}$$

This implies that:

$$\|\mathbf{x}\|_2^k \leq \max_{|\boldsymbol{\alpha}|=k} (\frac{k!}{\boldsymbol{\alpha}!}) \sum_{|\boldsymbol{\alpha}|=k} |\mathbf{x}^{\boldsymbol{\alpha}}| \leq (\sum_{|\boldsymbol{\beta}|=k} \frac{k!}{\boldsymbol{\beta}!})(\sum_{|\boldsymbol{\alpha}|=k} |\mathbf{x}^{\boldsymbol{\alpha}}|) = d^k \sum_{|\boldsymbol{\alpha}|=k} |\mathbf{x}^{\boldsymbol{\alpha}}| \tag{184}$$

where we used that $\max_{|\boldsymbol{\alpha}|=k}(\frac{k!}{\boldsymbol{\alpha}!}) \leq \sum_{|\boldsymbol{\beta}|=k} \frac{k!}{\boldsymbol{\beta}!} = (\underbrace{1 + \cdots + 1}_{d})^k = d^k$. ■

Appendix D. Proofs of Section 5

D.1 Proof of Theorem 6

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

Theorem 6 (Existence of a sketching operator when the Kernel Hölder LRIP holds) *Consider a model set* $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ *such that* $\mathfrak{S}$ *is* compact *in* $(\mathcal{M}(\mathcal{X}), \|.\|_{\mathrm{TV}})$ *and has* finite upper box-counting dimension*, i.e.* $d_B(\mathfrak{S}) < +\infty$. *Suppose that there exists* $\beta > 0, C > 0$ *and* $\eta \geq 0$ *such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\pi - \pi'\|_\kappa^\beta + \eta \tag{89}$$

for some bounded kernel κ *(i.e. such that* $\sup_{\mathbf{x}\in\mathcal{X}} \kappa(\mathbf{x}, \mathbf{x}) \leq K$*). Then for any finite dimension* $m > 2d_B(\mathfrak{S})$ *there exists* $0 < \delta < \beta, C' > 0$*, and a prevalent set of bounded linear maps*[a] $\mathcal{A} : \mathcal{M}(\mathcal{X}) \to \mathbb{R}^m$ *such that:*

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C'\|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^\delta + \eta \tag{90}$$

In other words if $\beta \leq 1$ *there exists a sketching operator that satisfies the Hölder LRIP with some* $0 < \delta < 1$ *and error* $\eta \geq 0$.

a. A prevalent set is a set whose complement is *shy* that is, informally, negligeable: in the case of Euclidean space it is a space whose complement has Lesbegue measure zero.

We will need some definitions and results. In particular we introduce the notion of "dual thickness" (Robinson, 2010):

Definition 8. *Given $\theta > 0$, let $n_\theta(S, \varepsilon)$ denote the lowest dimension of any linear subspace V of $\mathcal{B}^*$ such that for any $x, y \in S$ with $\|x - y\| \geq \varepsilon$ there exists an element $\psi \in V$ such that $\|\psi\|_* = 1$ and*

$$|\psi(x - y)| \geq \varepsilon^{1+\theta}$$

Set

$$\tau_\theta^*(S) = \limsup_{\varepsilon \to 0} \frac{\log n_\theta(S, \varepsilon)}{-\log \varepsilon}$$

and define the dual thickness $\tau^*(S)$ by

$$\tau^*(S) = \lim_{\theta \to 0} \tau_\theta^*(S)$$

We have that $\tau^*(S) \geq 0$ and if S is a compact subspace of $\mathcal{B}$ we have $\tau^*(S) \leq d_B(S)$ (Robinson, 2010, Corollary 7.8). We recall the following important result:

Proposition 12 (Theorem 8.1 in (Robinson, 2010))**.** *Let S be a compact subset of a real Banach space $(\mathcal{B}, \|\cdot\|)$ with $d_B(S) < +\infty$. Let any integer $m > 2d_B(S)$ and any δ with:*

$$0 < \delta < \frac{m - 2d_B(S)}{m(1 + \alpha\tau^*(S))} \tag{185}$$

where $\alpha = 1/2$ if $\mathcal{B}$ is a Hilbert space and $\alpha = 1$ if it is a general Banach space. Then there exists a prevalent set of bounded linear maps $L : \mathcal{B} \to \mathbb{R}^m$ such that:

$$\forall (x, y) \in S, \ \|x - y\| \leq C_L \|Lx - Ly\|_2^\delta \tag{186}$$

Based on Proposition 12 we prove:

Proposition 13. *Consider a model set $\mathfrak{S} \subseteq \mathcal{P}(\mathcal{X})$ such that $\mathfrak{S}$ is compact in $(\mathcal{M}(\mathcal{X}), \|.\|_{\mathrm{TV}})$ and has finite upper box-counting dimension, i.e. $d_B(\mathfrak{S}) < +\infty$.*

Suppose that the following property holds for some $C' > 0, \beta > 0$ and $\eta \geq 0$:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C' \|\pi - \pi'\|_{\mathrm{TV}}^\beta + \eta \tag{187}$$

Then for any integer $m > 2d_B(S)$ there exists $0 < \delta < \beta$, a prevalent set of bounded linear maps $\mathcal{A} : \mathcal{M}(\mathcal{X}) \to \mathbb{R}^m$ and $C = C(\mathcal{A}, \beta) > 0$ such that:

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C \|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^\delta + \eta \tag{188}$$

Proof By hypothesis $\mathfrak{S}$ is compact in $\mathcal{B} := (\mathcal{M}(\mathcal{X}), \|.\|_{\mathrm{TV}})$ which is a Banach space and we have $d_B(\mathfrak{S}) < +\infty$, so we can apply Proposition 12 with the total variation norm. Take any $m = 2d_B(\mathfrak{S}) + k$ for $k \geq 1$ and any θ with:

$$0 < \theta < \frac{m - 2d_B(\mathfrak{S})}{m(1 + \tau^*(\mathfrak{S}))} = \frac{k}{(2d_B(\mathfrak{S}) + k)(1 + \tau^*(\mathfrak{S}))} \leq (1 + \tau^*(\mathfrak{S}))^{-1} \leq 1 \tag{189}$$

By Proposition 12 there exists a prevalent set of bounded linear maps $\mathcal{A} : \mathcal{M}(\mathcal{X}) \to \mathbb{R}^m$ and $C = C(\mathcal{A})$ such that:

$$\forall \pi, \pi' \in \mathfrak{S}, \ \|\pi - \pi'\|_{\mathrm{TV}} \leq C \|\mathcal{A}\pi - \mathcal{A}\pi'\|_2^\theta \tag{190}$$

The fact that $\|\cdot\|_{\mathcal{L}(\mathcal{H}),p}$ is dominated by $\|\cdot\|_{\mathrm{TV}}^\beta$ on $\mathfrak{S}$ for some $\beta \in]0, 1]$ concludes with the Hölder exponent being $\delta = \beta\theta$. Since $\theta < 1$ we have $0 < \delta < \beta$. ■

Proof [Proof of Theorem 6] By hypothesis we have for some $\beta > 0$ and $C > 0$:

$$\forall \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}),p} \leq C\|\pi - \pi'\|_\kappa^\beta + \eta \tag{191}$$

However since the kernel is bounded by K we have $\|\pi - \pi'\|_\kappa \leq \sqrt{K}\|\pi - \pi'\|_{\mathrm{TV}}$ for any $\pi, \pi' \in \mathcal{P}(\mathcal{X})$ (see Theorem 21 in (Sriperumbudur et al., 2010)). In particular we can apply Proposition 13 since $\|\cdot\|_{\mathcal{L}(\mathcal{H}),p}$ is in this case dominated on the secant set $\mathfrak{S} - \mathfrak{S}$ by $CK^{\beta/2}\|\cdot\|_{\mathrm{TV}}^\beta + \eta$ which gives the desired result. ■

References

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. ISSN. Elsevier Science, 2003. ISBN 9780080541297. URL https://books.google.fr/books?id=R5A65Koh-EoC.

Roman Razmikovich Akopyan and Andrey Efimov. Boas–kac roots of positive definite functions of several variables. *Analysis Mathematica*, 43:359–369, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68 (3):337–404, 1950. URL http://dx.doi.org/10.2307/1990404.

Gennaro Auricchio, Andrea Codegoni, Stefano Gualandi, Giuseppe Toscani, and Marco Veneroni. The equivalence of fourier-based and wasserstein metrics on imaging problems, 2020.

Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially private database release via kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 414–422. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balog18a.html.

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 Space in Probability and Statistics*. 01 2004. ISBN 978-1-4613-4792-7. doi: 10.1007/978-1-4419-9096-9.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, 2013.

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*, pages 7928–7946, December 2014. doi: 10.1109/TIT.2014.2364403. URL https://hal.archives-ouvertes.fr/hal-00908358.

Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy, 2019.

Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the Wasserstein Metric and Geometric Inference. *Electronic journal of statistics* , 5:1394–1423, November 2011. doi: 10.1214/11-EJS646. URL https://hal.inria.fr/inria-00607806.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c54e7837e0cd0ced286cb5995327d1ab-Paper.pdf.

Emmanuel Candes and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution, 2012.

J. Carrillo and Giuseppe Toscani. Contractive probability metrics and asymptotic behavior of dissipative kinetic equations. 01 2008.

Djalil Chafaï, A. Hardy, and Mylène Maïda. Concentration for coulomb gases and coulomb transport inequalities. *Journal of Functional Analysis*, 275:1447–1483, 2016.

Antoine Chatalic. *Efficient and privacy-preserving compressive learning*. Theses, Université Rennes 1, November 2020. URL https://tel.archives-ouvertes.fr/tel-03023287.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99, page 634, USA, 1999. IEEE Computer Society. ISBN 0769504094.

Jérôme Dedecker and Bertrand Michel. Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension, 2013.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy, 2018.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 02 1969. doi: 10.1214/aoms/1177697802. URL https://doi.org/10.1214/aoms/1177697802.

Werner Ehm, Tilmann Gneiting, and Donald Richards. Convolution roots of radial positive definite functions with compact support. *Transactions of the American Mathematical Society*, 356(11): 4655–4685, 2004. ISSN 00029947. URL http://www.jstor.org/stable/3844939.

Clément Elvira, Rémi Gribonval, Charles Soussen, and Cédric Herzet. When does omp achieve exact recovery with continuous dictionaries?, 2020.

Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010. ISBN 9780821849743 0821849743.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/95e1533eb1b20a97777749fb94fdb944-Paper.pdf.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/feydy19a.html.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2053–2061. Curran Associates, Inc., 2015.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2018.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/genevay18a.html.

Ziv Goldfeld and Kristjan H. Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *AISTATS*, 2020.

Ziv Goldfeld, Kristjan Greenewald, and Kengo Kato. Asymptotic guarantees for generative modeling based on the smooth wasserstein distance. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2527–2539. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1ac978c8020be6d7212aa71d4f040fc3-Paper.pdf.

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

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/gordaliza19a.html.

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

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive Statistical Learning with Random Feature Moments, August 2021a. URL https://hal.inria.fr/hal-01544609. Main novelties between version 1 and version 2: improved concentration bounds, improved sketch sizes for compressive k-means and compressive GMM that now scale linearly with the ambient dimensionMain novelties of version 3: all content on compressive clustering and compressive GMM is now developed in the companion paper hal-02536818; improved statistical guarantees in a generic framework with illustration of the improvements on compressive PCA.

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling, August 2021b. URL https://hal.inria.fr/hal-02536818. This preprint results from a split and profound restructuring and improvements of of https://hal.inria.fr/hal-01544609v2It is a companion paper to https://hal.inria.fr/hal-01544609v3.

A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005. ISBN 9780198775218. URL https://books.google.fr/books?id=8YkSDAAAQBAJ.

P.R. Halmos. *Measure Theory*. Graduate Texts in Mathematics. Springer New York, 1976. ISBN 9780387900889. URL https://books.google.fr/books?id=-Rz7q4jikxUC.

Sigurdur Helgason. *Integral Geometry and Radon Transforms*. 01 2011. ISBN 978-1-4419-6054-2. doi: 10.1007/978-1-4419-6055-9.

Nicolas Keriven and Rémi Gribonval. Instance optimal decoding and the restricted isometry property. *Journal of Physics: Conference Series*, 1131, 02 2018. doi: 10.1088/1742-6596/1131/1/012002.

Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval. Compressive K-means. In *ICASSP 2017 - IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, United States, March 2017. URL https://hal.inria.fr/hal-01386077.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for Large-Scale Learning of Mixture Models. *Information and Inference*, 7(3):447–508, September 2018. doi: 10.1093/imaiai/iax015. URL https://hal.inria.fr/hal-01329195. to appear in Information and Inference, a journal of the IMA (available online since December 2017).

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 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(4):43–59, 2017.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced wasserstein kernels for probability distributions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Fabian Latorre, Paul Rolland, and Volkan Cevher. Lipschitz constant estimation of neural networks via sparse polynomial optimization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJe4_xSFDB.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 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*, volume 19. MIT Press, 2007. URL https://proceedings.neurips.cc/paper/2006/file/2d71b2ae158c7c5912cc0bbde2bb9d95-Paper.pdf.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021. URL http://jmlr.org/papers/v22/20-1369.html.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theor.*, 52(10):4394–4412, October 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.881731. URL https://doi.org/10.1109/TIT.2006.881731.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 689–696, New York, NY, USA, 2009a. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553463. URL https://doi.org/10.1145/1553374.1553463.

Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009b. URL https://proceedings.neurips.cc/paper/2008/file/c0f168ce8900fa56e57789e2a2f2c9d0-Paper.pdf.

Andreas Maurer and Massimiliano Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Trans. Inf. Theor.*, 56(11):5839–5846, November 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2069250. URL https://doi.org/10.1109/TIT.2010.2069250.

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(1-2):1–141, 2017. ISSN 1935-8245. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

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

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences, 03 2020.

Francis Narcowich, Xinping Sun, and Joseph Ward. Approximation power of rbfs and their associated sbfs: A connection. *Advances in Computational Mathematics*, 27:107–124, 07 2007. doi: 10.1007/s10444-005-7506-1.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370 – 400, 2013. doi: 10.1214/12-AOS1065. URL https://doi.org/10.1214/12-AOS1065.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8172–8183. PMLR, 18–24 Jul 2021a. URL http://proceedings.mlr.press/v139/nietert21a.html.

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, 2020.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11:355–607, 2019.

Jr. R. P. Boas and M. Kac. Inequalities for Fourier transforms of positive functions. *Duke Mathematical Journal*, 12(1):189 – 206, 1945. doi: 10.1215/S0012-7094-45-01215-4. URL https://doi.org/10.1215/S0012-7094-45-01215-4.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, page 1313–1320, Red Hook, NY, USA, 2008. Curran Associates Inc. ISBN 9781605609492.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12(22):731–817, 2011. URL http://jmlr.org/papers/v12/reid11a.html.

Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11):1228–1235, 2018. ISSN 1631-073X. doi: https://doi.org/10.1016/j.crma.2018.10.010. URL https://www.sciencedirect.com/science/article/pii/S1631073X18302802.

J. Robinson. Dimensions, embeddings, and attractors. 2010.

E.B. Saff and V. Totik. *Logarithmic Potentials with External Fields*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013. ISBN 9783662033296. URL https://books.google.fr/books?id=ltPqCAAAQBAJ.

Filippo Santambrogio. Optimal transport for applied mathematicians, 2015.

Vincent Schellekens and Laurent Jacques. Compressive classification (machine learning without learning), 2018.

Vincent Schellekens and Laurent Jacques. Compressive learning of generative networks, 2020.

Catia Scricciolo. Bayes and maximum likelihood for l^1-wasserstein deconvolution of laplace mixtures, 2017.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263 – 2291, 2013. doi: 10.1214/13-AOS1140. URL https://doi.org/10.1214/13-AOS1140.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

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 Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 306–314, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/solomon14.html.

Bharath Sriperumbudur and Zoltan Szabo. Optimal rates for random fourier features. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/d14220ee66aeec73c49038385428ec4c-Paper.pdf.

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. *J. Mach. Learn. Res.*, 11:1517–1561, August 2010. ISSN 1532-4435.

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(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722. URL https://doi.org/10.1214/12-EJS722.

Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces (PMS-32)*. Princeton University Press, 2016. ISBN 9781400883899. doi: doi:10.1515/9781400883899. URL https://doi.org/10.1515/9781400883899.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces, 2017.

X.P. Sun. Conditionally positive definite functions and their application to multivariate interpolations. *Journal of Approximation Theory*, 74(2):159–180, 1993. ISSN 0021-9045. doi: https://doi.org/10.1006/jath.1993.1059. URL https://www.sciencedirect.com/science/article/pii/S0021904583710592.

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, 11 2004.

Gábor J. Székely and Maria L. Rizzo. The energy of data. *Annual Review of Statistics and Its Application*, 4(1):447–479, 2017. doi: 10.1146/annurev-statistics-060116-054026. URL https://doi.org/10.1146/annurev-statistics-060116-054026.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Found. Trends Mach. Learn.*, 9(1):1–118, June 2016. ISSN 1935-8237. doi: 10.1561/2200000055. URL https://doi.org/10.1561/2200000055.

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 S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

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, 2021.