Journal of Machine Learning Research 18 (2017) 1-47 Submitted 1/17; Revised 7/17; Published 7/17

Minimax Estimation of Kernel Mean Embeddings

Ilya Tolstikhin ILYA@TUEBINGEN.MPG.DE
[†]*Department of Empirical Inference*
Max Planck Institute for Intelligent Systems
Spemanstraße 38, Tübingen 72076, Germany

Bharath K. Sriperumbudur BKS18@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Krikamol Muandet[†] KRIKAMOL@TUEBINGEN.MPG.DE
Department of Mathematics
Faculty of Science, Mahidol University
272 Rama VI Rd. Rajchathevi, Bangkok 10400, Thailand

Editor: Andreas Christmann

Abstract

In this paper, we study the minimax estimation of the Bochner integral

$$\mu_k(P) := \int_{\mathcal{X}} k(\cdot, x)\, dP(x),$$

also called as the *kernel mean embedding*, based on random samples drawn i.i.d. from P, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel. Various estimators (including the empirical estimator), $\hat{\theta}_n$ of $\mu_k(P)$ are studied in the literature wherein all of them satisfy $\|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} = O_P(n^{-1/2})$ with $\mathcal{H}_k$ being the reproducing kernel Hilbert space induced by k. The main contribution of the paper is in showing that the above mentioned rate of $n^{-1/2}$ is minimax in $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$-norms over the class of discrete measures and the class of measures that has an infinitely differentiable density, with k being a continuous translation-invariant kernel on $\mathbb{R}^d$. The interesting aspect of this result is that the minimax rate is independent of the smoothness of the kernel and the density of P (if it exists).

Keywords: Bochner integral, Bochner's theorem, kernel mean embeddings, minimax lower bounds, reproducing kernel Hilbert space, translation invariant kernel

1. Introduction

Over the last few years, kernel embedding of distributions (Smola et al., 2007; Sriperumbudur et al., 2010) has gained a lot of attention in the machine learning community due to the wide variety of applications it has been employed in. Some of these applications include kernel two-sample testing (Gretton et al., 2007, 2012), kernel independence and conditional independence tests (Gretton et al., 2008; Fukumizu et al., 2008), covariate-shift (Smola et al., 2007), density estimation (Sriperumbudur, 2011), feature selection (Song et al., 2012), causal inference (Lopez-Paz et al., 2015), kernel Bayes' rule (Fukumizu et al., 2013) and distribution regression (Szabó et al., 2015).

Formally, let $\mathcal{H}_k$ be a separable reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950) with a continuous reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on a separable topological space $\mathcal{X}$. Given a Borel probability measure P defined over $\mathcal{X}$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\, dP(x) < \infty$, the kernel mean or the mean element is defined as the Bochner integral

$$\mu_P := \int_{\mathcal{X}} k(\cdot, x)\, dP(x) \in \mathcal{H}_k. \tag{1}$$

We refer the reader to Diestel and Uhl (1977, Chapter 2) and Dinculeanu (2000, Chapter 1) for the definition of a Bochner integral. The mean element in (1) can be viewed as an embedding of P in $\mathcal{H}_k$,

$$\mu_k : M_+^1(\mathcal{X}) \to \mathcal{H}_k, \qquad \mu_k(P) = \mu_P,$$

where $M_+^1(\mathcal{X})$ denotes the set of all Borel probability measures on $\mathcal{X}$. Hence, we also refer to μ_k as the *kernel mean embedding* (KME). The mean embedding can be seen as a generalization of the classical kernel feature map that embeds points of an input space $\mathcal{X}$ as elements in $\mathcal{H}_k$. The mean embedding μ_k can also be seen as a generalization of the classical notions of characteristic function, moment generation function (if it exists), and Weierstrass transform of P (all defined on $\mathbb{R}^d$) to an arbitrary topological space $\mathcal{X}$ as the choice of $k(\cdot, x)$ as $(2\pi)^{-d/2} e^{-\sqrt{-1}\langle \cdot, x\rangle}$, $e^{\langle \cdot, x\rangle}$, and $(4\pi)^{-d/2} e^{-\|\cdot - x\|_2^2}$, $x \in \mathbb{R}^d$ respectively reduces μ_k to these notions. The mean embedding μ_k is closely related to the *maximum mean discrepancy* (MMD) (Gretton et al., 2007), which is the RKHS distance between the mean embeddings of two probability measures. We refer the reader to (Sriperumbudur et al., 2010; Sriperumbudur, 2016) for more details on the properties of μ_k and the corresponding MMD.

In all the above mentioned statistical and machine learning applications, since the underlying distribution P is known only through random samples $X_1, \ldots, X_n$ drawn i.i.d. from it, an estimator of μ_P is employed. The goal of this paper is to study the minimax optimal estimation of μ_P. In the literature, various estimators of μ_P have been proposed. The simplest and most popular is the empirical estimator μ_{P_n}, which is constructed by replacing P by its empirical counterpart, $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x denotes a Dirac measure at $x \in \mathcal{X}$. In fact, all the above mentioned applications deal with the empirical estimator of μ_P because of its simplicity. Using Bernstein's inequality in separable Hilbert spaces (Yurinsky, 1995, Theorem 3.3.4), it follows that for bounded continuous kernels, $\|\mu_{P_n} - \mu_P\|_{\mathcal{H}_k} = O_P(n^{-1/2})$ for any P, i.e., the empirical estimator is a $\sqrt{n}$-consistent estimator of μ_P in $\mathcal{H}_k$-norm. This result is also proved in Smola et al. (2007, Theorem 2), Gretton et al. (2012), and Lopez-Paz et al. (2015) using McDiarmid's inequality, which we improve in Proposition A.1 (also see Remark A.2 in Appendix A) by providing a better constant. Assuming $\mathcal{X} = \mathbb{R}^d$ and P to have a density p, Sriperumbudur (2016, Theorem 4.1) proposed to estimate $\mu_P = \int_{\mathbb{R}^d} k(\cdot, x) p(x)\, dx$ by replacing p with a kernel density estimator, which is then shown to be $\sqrt{n}$-consistent in $\mathcal{H}_k$-norm if k is a bounded continuous *translation invariant kernel*—see Section 2 for its definition—on $\mathbb{R}^d$. Recently, Muandet et al. (2016, Section 2.4, Theorem 7) proposed a non-parametric shrinkage estimator of μ_P and established its $\sqrt{n}$-consistency in $\mathcal{H}_k$-norm for bounded continuous kernels on $\mathcal{X}$. Muandet et al. (2016, Section 3, Theorem 10) also proposed a penalized M-estimator for μ_P

where the penalization parameter is computed in a completely data-driven manner using leave-one-out cross validation and showed that it is also $\sqrt{n}$-consistent in $\mathcal{H}_k$-norm. In fact, the $\sqrt{n}$-consistency of all these estimators is established by showing that they are all within a $\|\cdot\|_{\mathcal{H}_k}$-ball of size $o_P(n^{-1/2})$ around the empirical estimator μ_{P_n}.

In the above discussion, it is important to note that the convergence rate of μ_{P_n} (and also other estimators) to μ_P in $\mathcal{H}_k$-norm does not depend on the smoothness of k or the density, p (if it exists). Under some mild conditions on the kernel (defined on $\mathbb{R}^d$), it can be shown (see Section 4) that $\mathcal{H}_k$ is continuously included in $L^2(\mathbb{R}^d)$ and $\|f\|_{L^2(\mathbb{R}^d)} \leq c_k \|f\|_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$, where c_k is a constant that depends only on the kernel. This means, $\|\cdot\|_{L^2(\mathbb{R}^d)}$ is a weaker norm than $\|\cdot\|_{\mathcal{H}_k}$ and therefore it could be possible that μ_{P_n} converges to μ_P in $L^2(\mathbb{R}^d)$ at a rate faster than $n^{-1/2}$ (depending on the smoothness of k). In Proposition A.1 (also see Remark A.3 in Appendix A) we show that $\|\mu_{P_n} - \mu_P\|_{L^2(\mathbb{R}^d)} = O_P(n^{-1/2})$. Now given these results, it is of interest to understand whether these rates are optimal in a minimax sense, i.e., whether the above mentioned estimators are minimax rate optimal or can they be improved upon? Therefore the goal of this work is to obtain minimax rates for the estimation of μ_P in $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$.

Formally, we would like to find the minimax rate $r_{n,k}(\mathcal{F}, \mathcal{P})$ and a positive constant $c_k(\mathcal{F}, \mathcal{P})$ (independent of n) such that

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ r_{n,k}^{-1}(\mathcal{F}, \mathcal{P}) \|\hat{\theta}_n - \mu_P\|_{\mathcal{F}} \geq c_k(\mathcal{F}, \mathcal{P}) \right\} > 0, \tag{2}$$

where $\mathcal{F}$ is either $\mathcal{H}_k$ or $L^2(\mathbb{R}^d)$, $\mathcal{P}$ is a suitable subset of Borel probability measures on $\mathcal{X}$, and the infimum is taken over all estimators $\hat{\theta}_n$ mapping the i.i.d. sample $X_1, \ldots, X_n$ to $\mathcal{F}$. Suppose $k(x, y) = \langle x, y \rangle$, $x, y \in \mathbb{R}^d$. Norms $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$ match for this choice of k and the corresponding RKHS is finite dimensional, i.e., $\mathcal{H}_k = \mathbb{R}^d$. For a distribution P on $\mathbb{R}^d$ satisfying $\int_{\mathbb{R}^d} \|x\|_2 \, dP(x) < \infty$, this choice of kernel yields $\mu_P = \int x \, dP(x)$ as the mean embedding of P which simply is the mean of P. It is well-known (Lehmann and Casella, 2008, Chapter 5, Example 1.14) that the minimax rate of estimating $\mu_P \in \mathbb{R}^d$ based on $(X_i)_{i=1}^n$ is $r_{n,k}(\mathcal{F}, \mathcal{P}) = n^{-1/2}$ for the class $\mathcal{P}$ of Gaussian distributions on $\mathbb{R}^d$. In fact, this rate is attained by the empirical estimator $\mu_{P_n} = \frac{1}{n} \sum_{i=1}^n X_i$, which is the sample mean. Based on this observation, while one can intuitively argue that the minimax rate of estimating μ_P is $n^{-1/2}$ even if $\mathcal{H}_k$ is an infinite dimensional RKHS, it is difficult to extend the finite dimensional argument in a rigorous manner to the estimation of the infinite dimensional object, μ_P. In this paper, through a key inequality—see (3)— we rigorously show that it is indeed the case.

The main result of the paper is that if k is translation invariant on $\mathcal{X} = \mathbb{R}^d$ (see Theorems 1 and 9 for precise conditions on the kernel) and $\mathcal{P}$ is the set of all Borel discrete probability measures on $\mathbb{R}^d$, then the minimax rate $r_{n,k}(\mathcal{F}, \mathcal{P})$ is $n^{-1/2}$ for both $\mathcal{F} = \mathcal{H}_k$ and $\mathcal{F} = L^2(\mathbb{R}^d)$. Next, we show in Theorems 6 and 12 that the minimax rate for the estimation of μ_P in both $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$ still remains $n^{-1/2}$ even when $\mathcal{P}$ is restricted to the class of Borel probability measures which have densities, p that are continuously infinitely differentiable. The reason for considering such a class of distributions with smooth densities is that μ_P, which is the convolution of k and p, is smoother than k. Therefore one might wonder if it could be possible to estimate μ_P at a rate faster than $n^{-1/2}$ that depends on

the smoothness of k and p. Our result establishes that even for the class of distributions with very smooth densities, the minimax rate is independent of the smoothness of k and the density of P. The key ingredient in the proofs of Theorems 6 and 12 is the non-trivial inequality (see Proposition 3)

$$\|\mu_{G_0} - \mu_{G_1}\|_{\mathcal{F}} \geq c'_{k,\sigma^2}\|\tau_0 - \tau_1\|_2, \tag{3}$$

which relates the $\mathcal{F}$-distance between the mean embeddings of the Gaussian distributions, $G_0 = N(\tau_0, \sigma^2 I)$ and $G_1 = N(\tau_1, \sigma^2 I)$ to the Euclidean distance between the means of these Gaussians, where c'_{k,σ^2} is a constant that depends only on σ^2 and the translation invariant characteristic kernel k. Combining (3) with Le Cam's method (see Appendix B) implies that the estimation of an infinite dimensional object μ_P is as hard as the estimation of finite dimensional mean of a Gaussian distribution, thereby establishing the minimax rate to be $n^{-1/2}$. These results show that the empirical estimator—and other estimators we discussed above—of μ_P is minimax rate optimal.

Ramdas et al. (2015, Corollary 1) derived a special case of (3) for the Gaussian kernel k by ignoring small terms in the Taylor series expansion of $\|\mu_{G_0} - \mu_{G_1}\|_{\mathcal{H}_k}$ (refer to Remark 4). They used this result to show that the MMD between G_0 and G_1 decreases to zero exponentially/polynomially fast in d even when the Kullback-Leibler divergence between the two is kept constant, which in turn sheds some light on the decaying power of MMD-based hypothesis tests in high dimensions. Proposition 3 is more general, as it holds for *any* translation-invariant kernel k and does not require a truncation of small reminder terms.

The paper is organized as follows. Various notations used throughout the paper and definitions are collected in Section 2. The main results on minimax estimation of μ_P in $\|\cdot\|_{\mathcal{H}_k}$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$ for translation invariant kernels (and also *radial* kernels) on $\mathbb{R}^d$ are presented in Sections 3 and 4 respectively. The proofs of the results are provided in Section 5 while some supplementary results needed in the proofs are collected in appendices.

2. Definitions & Notation

Define $\|a\|_2 := \sqrt{\sum_{i=1}^d a_i^2}$ and $\langle a, b\rangle := \sum_{i=1}^d a_i b_i$, where $a := (a_1, \ldots, a_d) \in \mathbb{R}^d$ and $b := (b_1, \ldots, b_d) \in \mathbb{R}^d$. $C(\mathbb{R}^d)$ (*resp.* $C_b(\mathbb{R}^d)$) denotes the space of all continuous (*resp.* bounded continuous) functions on $\mathbb{R}^d$. $f \in C(\mathbb{R}^d)$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \geq \epsilon\}$ is compact. The class of all continuous f on $\mathbb{R}^d$ which vanish at infinity is denoted as $C_0(\mathbb{R}^d)$. For $f \in C_b(\mathbb{R}^d)$, $\|f\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x)|$ denotes the supremum norm of f. $M_b(\mathbb{R}^d)$ (*resp.* $M_+^b(\mathbb{R}^d)$) denotes the set of all finite (*resp.* finite non-negative) Borel measures on $\mathbb{R}^d$. $\mathrm{supp}(\mu)$ denotes the support of $\mu \in M_b(\mathbb{R}^d)$ which is defined as $\mathrm{supp}(\mu) = \{x \in \mathbb{R}^d \mid \text{for any open set } U \text{ such that } x \in U, |\mu|(U) \neq 0\}$, where $|\mu|$ is the total-variation of μ. $M_+^1(\mathbb{R}^d)$ denotes the set of Borel probability measures on $\mathbb{R}^d$. For $\mu \in M_+^b(\mathbb{R}^d)$, $L^r(\mathbb{R}^d, \mu)$ denotes the Banach space of r-power ($r \geq 1$) μ-integrable functions and we will use $L^r(\mathbb{R}^d)$ for $L^r(\mathbb{R}^d, \mu)$ if μ is a Lebesgue measure on $\mathbb{R}^d$. For $f \in L^r(\mathbb{R}^d, \mu)$, $\|f\|_{L^r(\mathbb{R}^d,\mu)} := \left(\int_{\mathbb{R}^d} |f|^r \, d\mu\right)^{1/r}$ denotes the L^r-norm of f for $1 \leq r < \infty$ and we denote it as $\|\cdot\|_{L^r(\mathbb{R}^d)}$ if μ is the Lebesgue measure. The convolution $f * g$ of two measurable functions

f and g on $\mathbb{R}^d$ is defined as

$$(f * g)(x) := \int_{\mathbb{R}^d} f(y)g(x-y)\,dy,$$

provided the integral exists for all $x \in \mathbb{R}^d$. The Fourier transforms of $f \in L^1(\mathbb{R}^d)$ and $\mu \in M_b(\mathbb{R}^d)$ are defined as

$$f^\wedge(y) := \mathscr{F}[f](y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x)\, e^{-i\langle y,x\rangle}\, dx, \qquad y \in \mathbb{R}^d$$

and

$$\mu^\wedge(y) := \mathscr{F}[\mu](y) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle y,x\rangle}\, d\mu(x), \qquad y \in \mathbb{R}^d$$

respectively, where i denotes the imaginary unit $\sqrt{-1}$.

A kernel $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called *translation invariant* if there exists a symmetric positive definite function, ψ such that $k(x,y) = \psi(x-y)$ for all $x, y \in \mathbb{R}^d$. Bochner's theorem (see Wendland, 2005, Theorem 6.6) provides a complete characterization for a positive definite function ψ: A continuous function $\psi\colon \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of $\Lambda_\psi \in M_+^b(\mathbb{R}^d)$, i.e.,

$$\psi(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle x,w\rangle} d\Lambda_\psi(w), \quad x \in \mathbb{R}^d. \tag{4}$$

A kernel k is called *radial* if there exists $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that $k(x,y) = \phi(\|x-y\|_2^2)$ for all $x, y \in \mathbb{R}^d$. From Schönberg's representation (Schoenberg, 1938; Wendland, 2005, Theorems 7.13 & 7.14) it is known that a kernel k is radial on every $\mathbb{R}^d$ if and only if there exists $\nu \in M_+^b([0,\infty))$ such that the following holds for all $x, y \in \mathbb{R}^d$:

$$k(x,y) = \phi(\|x-y\|^2) = \int_0^\infty e^{-t\|x-y\|^2} d\nu(t). \tag{5}$$

Some examples of reproducing kernels on $\mathbb{R}^d$ (in fact all these are radial) that appear throughout the paper are:

1. *Gaussian:* $k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\eta^2}\right)$, $\eta > 0$;
2. *Mixture of Gaussians:* $k(x,y) = \sum_{i=1}^M \beta_i \exp\left(-\frac{\|x-y\|_2^2}{2\eta_i^2}\right)$, where $M \geq 2$, $\eta_1^2 \geq \eta_2^2 \geq \cdots \geq \eta_M^2 > 0$, and positive constants $\beta_1, \ldots, \beta_M$ such that $\sum_{i=1}^M \beta_i = C_M < \infty$;
3. *Inverse Multiquadrics:* $k(x,y) = (c^2 + \|x-y\|_2^2)^{-\gamma}$, $c, \gamma > 0$;
4. *Matérn:* $k(x,y) = \frac{c^{2\tau-d}}{\Gamma(\tau-\frac{d}{2})2^{\tau-1-d/2}} \left(\frac{\|x-y\|_2}{c}\right)^{\tau-\frac{d}{2}} \mathcal{K}_{\frac{d}{2}-\tau}(c\|x-y\|_2)$, $\tau > d/2$, $c > 0$, where $\mathcal{K}_\alpha$ is the *modified Bessel function of the third kind* of order α and Γ is the Gamma function.

A kernel k is said to be *characteristic* if the mean embedding, $\mu_k : P \to \mu_P$ is injective, where μ_P is defined in (1). Various characterizations for the injectivity of μ_k (or k

being characteristic) are known in literature (for details, see Sriperumbudur et al., 2011 and references therein). If k is a bounded continuous translation invariant positive definite kernel on $\mathbb{R}^d$, a simple characterization can be obtained for it to be characteristic (Sriperumbudur et al., 2010, Theorem 9): k is characteristic if and only if $\mathrm{supp}(\Lambda_\psi) = \mathbb{R}^d$ where Λ_ψ is defined in (4). This characterization implies that the above mentioned examples are characteristic kernels. Examples of non-characteristic kernels of translation invariant type include $k(x,y) = \frac{\sin(x-y)}{x-y}$, $x, y \in \mathbb{R}$ and $k(x,y) = \cos(x-y)$, $x, y \in \mathbb{R}$. More generally, polynomial kernels of any finite order are non-characteristic.

3. Minimax Estimation of μ_P in the RKHS Norm

In this section, we present our main results related to the minimax estimation of kernel mean embeddings (KMEs) in the RKHS norm. As discussed in Section 1, various estimators of $\mu_k(P)$ are known in literature and all these have a convergence rate of $n^{-1/2}$ if the kernel is bounded. The main goal of this section is to show that the rate $n^{-1/2}$ is actually minimax optimal for different choices of $\mathcal{P}$ (see (2)) under some mild conditions on k.

First, choosing $\mathcal{P}$ to be the set of all discrete probability measures on $\mathbb{R}^d$, in Section 3.1 (see Theorem 1 and Corollary 2), we present the minimax lower bounds of order $\Omega(n^{-1/2})$ with constant factors depending only on the properties of the kernel for translation invariant and radial kernels respectively. Next we will show in Section 3.2 that the rate $n^{-1/2}$ remains minimax optimal for translation invariant and radial kernels even if we choose the class $\mathcal{P}$ to contain only probability distributions with infinitely continuously differentiable densities. For translation invariant kernels the result (see Theorem 6) is based on a key inequality, which relates the RKHS distance between embeddings of Gaussian distributions to the Euclidean distance between the mean vectors of these distributions (see Proposition 3). The minimax lower bound for radial kernels (see Theorem 8) is derived using a slightly different argument. Instead of applying the bound of Theorem 6 to the particular case of radial kernels, we will present a direct analysis based on the special properties of radial kernels. This will lead us to the lower bound with almost optimal constant factors, depending only on the shape of Borel measure ν corresponding to the kernel.

Our analysis is based on the following simple idea: if a kernel k is characteristic, there is a one-to-one correspondence between any given set of Borel probability measures $\mathcal{P}$ defined over $\mathbb{R}^d$ and a set $\mu_k(\mathcal{P})$ of their embeddings into the RKHS $\mathcal{H}_k$. This means that distributions in $\mathcal{P}$ are indexed by their embeddings $\Theta := \mu_k(\mathcal{P})$ and so (2) can be equivalently written as

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left\{ r_{n,k}^{-1}(\mathcal{H}_k, \mathcal{P}) \|\hat{\theta}_n - \theta\|_{\mathcal{H}_k} \geq c_k(\mathcal{H}_k, \mathcal{P}) \right\} > 0, \tag{6}$$

where the goal is to find the minimax rate $r_{n,k}(\mathcal{H}_k, \mathcal{P})$ and a positive constant $c_k(\mathcal{H}_k, \mathcal{P})$ (independent of n) such that (6) holds and $\mathbb{P}_\theta = P^n$ when $\theta = \mu_k(P)$. Using this equivalence, we obtain the minimax rates by employing Le Cam's method (Tsybakov, 2008)—see Theorems B.1 and B.2 for a reference.

3.1 Lower Bounds for Discrete Probability Measures

The following result (proved in Section 5.1) presents a minimax rate of $n^{-1/2}$ for estimating $\mu_k(P)$, where k is assumed to be translation invariant on $\mathbb{R}^d$.

Theorem 1 (Translation invariant kernels) *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Assume there exists $z \in \mathbb{R}^d$ and $\beta > 0$, such that $\psi(0) - \psi(z) \geq \beta$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{6}\sqrt{\frac{2\beta}{n}} \right\} \geq \frac{1}{4}.$$

The result is based on Le Cam's method involving two hypotheses (see Theorem B.1), where we choose them to be KMEs of discrete measures, both supported on the same pair of points separated by z in $\mathbb{R}^d$.

Remark (Choosing z and β) *As discussed in Sriperumbudur et al. (2010, Section 3.4), if k is translation invariant and characteristic on $\mathbb{R}^d$, then it is also strictly positive definite. This means that $\psi(0) > 0$. Moreover, the following hold: (a) Since ψ is positive definite, we have $|\psi(x)| \leq \psi(0)$ for all $x \in \mathbb{R}^d$ and (b) since ψ is characteristic, it cannot be a constant function. Together these facts show that there always exist $z \in \mathbb{R}^d$ and $\beta > 0$ satisfying the assumptions of Theorem 1. For instance, a Gaussian kernel $k(x,v) = \exp\big(-\|x-v\|_2^2/(2\eta^2)\big)$ satisfies $\psi(0) - \psi(z) \geq \|z\|_2^2/(4\eta^2)$ if $\|z\|_2^2 \leq 2\eta^2$, where we used a simple fact that $1 - e^{-x} \geq x/2$ for $0 \leq x \leq 1$.*

While Theorem 1 dealt with general translation invariant kernels, the following result (proved in Section 5.2) specializes it to radial kernels, i.e., kernels of the form in (5), by providing a simple condition on ν under which Theorem 1 holds.

Corollary 2 (Radial kernels) *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$ and k be radial on $\mathbb{R}^d$, i.e., $k(x,y) = \psi_\nu(x-y) := \int_0^\infty e^{-t\|x-y\|_2^2} d\nu(t)$, where $\nu \in M_+^b([0,\infty))$ such that $\operatorname{supp}(\nu) \neq \{0\}$. Assume there exist $0 < t_1 < \infty$ and $\alpha > 0$ satisfying $\nu([t_1,\infty)) \geq \alpha$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{6}\sqrt{\frac{\alpha}{n}} \right\} \geq \frac{1}{4}.$$

Remark (Choosing t_1 and α) *Since $\operatorname{supp}(\nu) \neq \{0\}$ the assumption of $\nu[t_1,\infty) \geq \alpha$ is always satisfied. For instance, if ν is a probability measure with positive median η then we can set $t_1 = \eta$ and $\alpha = \frac{1}{2}$. Based on this, it is easy to verify (see Appendix D.1) that $\alpha = 1$ for Gaussian, $\alpha = C_M$ for mixture of Gaussian kernels, $\alpha = \frac{c^{-2\gamma}}{2}$ for inverse multiquadrics and $\alpha = \frac{1}{2}$ for Matérn kernels.*

3.2 Lower Bounds for Probability Measures with Smooth Densities

So far, we have shown that the rate $n^{-1/2}$ is minimax optimal for the problem of KME estimation (both for translation invariant and radial kernels). As discussed in Section 1,

since this rate is independent of the smoothness of the estimand (which is determined by the smoothness of the kernel), one might wonder whether the minimax rate can be improved by restricting $\mathcal{P}$ to distributions with smooth densities. We show in this section (see Theorems 6 and 8) that this is not the case by restricting $\mathcal{P}$ to contain only distributions with infinitely continuously differentiable densities and proving the minimax lower bound of order $n^{-1/2}$.

We will start the analysis with translation invariant kernels and present a corresponding lower bound in Theorem 6. The proof of this result is again based on an application of Le Cam's method involving two hypotheses (see Theorem B.1), where this time these hypotheses are chosen to be embeddings of the d-dimensional Gaussian distributions. One of the main steps, when applying Theorem B.1, is to lower bound the distance between these embeddings. This is done in the following result (proved in Section 5.3), which essentially shows that if we take two Gaussian distributions $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ with the mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ which are close enough to each other, then the RKHS distance between the corresponding embeddings can be lower bounded by the Euclidean distance $\|\mu_0 - \mu_1\|_2$.

Proposition 3 *Let $\sigma > 0$. Suppose $k(x, y) = \psi(x - y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Then there exist constants $\epsilon_{\psi,\sigma^2}, c_{\psi,\sigma^2} > 0$ depending only on ψ and σ^2, such that the following condition holds for any $a \in \mathbb{R}^d$ with $\|a\|_2^2 \leq \epsilon_{\psi,\sigma^2}$:*

$$c_{\psi,\sigma^2} \leq \min_{e_z \in S^{d-1}} \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2 \|w\|_2^2} \langle e_z, w \rangle^2 \cos\left(\langle a, w \rangle\right) d\Lambda_\psi(w) < \infty, \tag{7}$$

where S^{d-1} is a unit sphere in $\mathbb{R}^d$ and $\Lambda_\psi \in M_+^b(\mathbb{R}^d)$ is defined in (4). Moreover, for all vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ satisfying $\|\mu_0 - \mu_1\|_2^2 \leq \epsilon_{\psi,\sigma^2}$, the following holds:

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k} \geq \sqrt{\frac{c_{\psi,\sigma^2}}{2}} \|\mu_0 - \mu_1\|_2, \tag{8}$$

where θ_0 and θ_1 are KMEs of Gaussian measures $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ respectively.

Remark 4 (KME expands small distances) *For a Gaussian kernel, it is possible to show (Sriperumbudur et al., 2012, Example 3; Ramdas et al., 2015, Proposition 1) that $\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 = C_1\big(1 - \exp(-C_2 \|\mu_0 - \mu_1\|_2^2)\big)$, where C_1 and C_2 are positive constants that depend only on σ^2 and η^2. This shows that (8) holds for $\|\mu_0 - \mu_1\|_2 \in [0, D]$, where D satisfies $C_1\big(1 - \exp(-C_2 D^2)\big) = \frac{1}{2} D^2 c_{\psi,\sigma^2}$. In other words, Proposition 3 states that the mapping $f_{\sigma^2} \colon \mathbb{R}^d \to \mathcal{H}_k$ defined by $f_{\sigma^2}(x) := \mu_k\big(G(x, \sigma^2 I)\big)$ expands small distances.*

Remark 5 (Computing c_{ψ,σ^2} and ϵ_{ψ,σ^2}) *Generally it may be very hard to compute (or bound) the constants c_{ψ,σ^2} and ϵ_{ψ,σ^2} appearing in the statement of Proposition 3. However, in some cases this may be still possible. In Appendix E we will provide an extensive analysis for the case of radial kernels.*

Based on Proposition 3, the following result shows that the rate of $n^{-1/2}$ remains minimax optimal for the problem of KME estimation with translation invariant kernels, even if we restrict the class of distributions $\mathcal{P}$ to contain only measures with smooth densities.

Theorem 6 (Translation invariant kernels) *Let $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Define $c_\psi := c_{\psi,1}$ and $\epsilon_\psi := \epsilon_{\psi,1}$ where $c_{\psi,1}$ and $\epsilon_{\psi,1}$ are positive constants that satisfy (7) in Proposition 3. Then for any $n \geq \frac{1}{\epsilon_\psi}$, the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{2}\sqrt{\frac{c_\psi}{2n}} \right\} \geq \frac{1}{4}.$$

Proof The proof will be based on Theorem B.1. For this we need to find two probability measures P_0 and P_1 on $\mathbb{R}^d$ and corresponding KMEs θ_0 and θ_1, such that $\|\theta_0 - \theta_1\|_{\mathcal{H}_k}$ is of the order $\Omega(n^{-1/2})$, while $\mathrm{KL}(P_0^n \| P_1^n)$ is upper bounded by a constant independent of n. Here $\mathrm{KL}(P_0\|P_1)$ denotes the Kullback-Leibler divergence between P_0 and P_1, which is defined as $\mathrm{KL}(P_0\|P_1) = \int \log \frac{dP_0}{dP_1}\, dP_0$ where P_0 is absolutely continuous w.r.t. P_1.

Pick two Gaussian distributions $G_0 := G(\mu_0, \sigma^2 I)$ and $G_1 := G(\mu_1, \sigma^2 I)$ for $\mu_0, \mu_1 \in \mathbb{R}^d$, and $\sigma^2 > 0$. It is known that (Tsybakov, 2008, Section 2.4)

$$\mathrm{KL}(G_0^n \| G_1^n) = n \cdot \frac{\|\mu_0 - \mu_1\|_2^2}{2\sigma^2}, \tag{9}$$

where G_0^n and G_1^n are n-fold product distributions. Choose μ_0 and μ_1 such that

$$\|\mu_0 - \mu_1\|_2^2 = \frac{1}{n}.$$

Denote KMEs of G_0 and G_1 using θ_0 and θ_1 respectively. Next we will take $\sigma^2 = 1$ and apply Proposition 3. Since c_ψ and ϵ_ψ satisfy (7) in Proposition 3, it follows from Proposition 3 that for $1/n \leq \epsilon_\psi$,

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 \geq \frac{c_\psi}{2}\|\mu_0 - \mu_1\|_2^2 = \frac{c_\psi}{2n}.$$

This shows that the first condition of Theorem B.1 is satisfied for θ_0 and θ_1 with $s := \frac{1}{2}\sqrt{c_\psi/(2n)}$. Moreover, using (9) we can show that the second condition of Theorem B.1 is satisfied with $\alpha = \frac{1}{2}$. We conclude the proof with an application of Theorem B.1. ■

Remark 7 (Lower bound on the sample size n) *Note that Theorem 6 holds only for large enough sample size n (i.e., $n \geq 1/\epsilon_\psi$). This assumption on n can be dropped if we set $\|\mu_0 - \mu_1\|_2^2 = \epsilon_\psi/n$ in the proof. In this case, the lower bound $\frac{1}{2}\sqrt{c_\psi/(2n)}$ will be replaced with $\frac{1}{2}\sqrt{c_\psi \epsilon_\psi/(2n)}$, while the lower bound on the minimax probability $1/4$ will be replaced with*

$$\max\left(\frac{1}{4}e^{-\frac{\epsilon_\psi}{2}}, \frac{1-\sqrt{\epsilon_\psi/4}}{2}\right).$$

The latter is generally undesirable, especially if ϵ_ψ grows with $d \to \infty$, since we want the minimax probability to be lower bounded by some universal non-zero constant that does not depend on the properties of the problem at hand.

Since radial kernels are particular instances of translation invariant kernels, Theorem 6 can be specialized by explicitly computing the constants c_ψ and ϵ_ψ to derive a minimax lower bound of order $\Omega(n^{-1/2})$. Unfortunately, the resulting lower bound will depend on the dimensionality d in a rather bad way and, as a consequence, is suboptimal in some situations. For instance, if we consider a Gaussian kernel $k(x,y) = \exp\left(-\frac{1}{2\eta^2}\|x-y\|_2^2\right)$, then a straightforward computation of c_ψ shows that the lower bound in Theorem 6 has the form $\sqrt{(1+2/\eta^2)^{-d/2}/n}$ which shrinks to zero as $d \to \infty$, while Proposition A.1 (also see Remark A.2) provides a dimension independent upper bound of the order $O_p(n^{-1/2})$. Therefore, instead of specializing Theorem 6 to radial kernels, we obtain the following result for radial kernels by using a refined analysis which yields a minimax rate of $\Omega(n^{-1/2})$ that matches the upper bound of Proposition A.1 up to constant factors that depend only on the shape of Borel measure ν. In particular, when specialized to the Gaussian kernel, the result matches the upper bound up to a constant factor independent of d.

Theorem 8 (Radial kernels) *Let k be radial on $\mathbb{R}^d$, i.e., $k(x,y) = \int_0^\infty e^{-t\|x-y\|_2^2}\, d\nu(t)$, where $\nu \in M_+^b([0,\infty))$ and $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable. Assume that $\mathrm{supp}(\nu) \neq \{0\}$ and there exist $0 < t_0 \leq t_1 < \infty$, $0 < \beta < \infty$ such that $\nu([t_0,t_1]) \geq \beta$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{50}\sqrt{\frac{1}{n} \cdot \frac{\beta t_0}{t_1 e}\left(1 - \frac{2}{2+d}\right)} \right\} \geq \frac{1}{5}.$$

Proof The proof, which is presented in Section 5.4, is based on an application of Le Cam's method involving multiple hypotheses (see Theorem B.2), where we use exponential (in d) number of Gaussian distributions with variances decaying as $\frac{1}{d}$. ■

Remark (Non-trivial lower bound as $d \to \infty$) *The proof of Theorem 8 is based on Gaussian distributions with variances decaying as $1/d$. As $d \to \infty$, it is obvious the densities of these distributions do not have uniformly bounded Lipschitz constants, i.e., they are arbitrarily "peaky". Hence, if we choose $\mathcal{P}$ to be class of distributions with infinitely differentiable densities that have uniformly bounded Lipschitz constants, then as $d \to \infty$, the densities considered in the proof of Theorem 8 do not belong to $\mathcal{P}$. On the other hand, the densities considered in the proof of Theorem 6 still belong to $\mathcal{P}$ but yielding an uninteresting result since $c_\psi \to 0$ when $d \to \infty$. Therefore, it is an open question whether a non-trivial lower bound can be obtained for radial kernels (or any other translation invariant kernels) if we choose $\mathcal{P}$ to contain only distributions with densities having uniformly bounded Lipschitz constants.*

Remark (Alternative Proof) *For completeness, we also present an alternative proof of Theorem 8 in Appendix E. It is based on Proposition 3, which holds for any translation invariant kernel. As a result, this proof leads to slightly worse constants compared to Theorem 8 (where we used an analysis specific to radial kernels), as well as a superfluous condition on the minimal sample size n.*

In Appendix D.2, we compute the positive constant $B_k := \frac{\beta t_0}{t_1}$ that appears in the lower bound in Theorem 8 in a closed form for Gaussian, mixture of Gaussian, inverse multiquadric and Matérn kernels.

4. Minimax Estimation of μ_P in the $L^2(\mathbb{R}^d)$ Norm

So far, we have discussed the minimax estimation of the kernel mean embedding (KME) in the RKHS norm. In this section, we investigate the minimax estimation of KME in $L^2(\mathbb{R}^d)$ norm. The reason for this investigation is as follows. Let $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$, where $\psi \in L^1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ is strictly positive definite. The corresponding RKHS is given by (see Wendland, 2005, Theorem 10.12)

$$\mathcal{H}_k = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \,:\, \int_{\mathbb{R}^d} \frac{|f^\wedge(\omega)|^2}{\psi^\wedge(\omega)}\, d\omega < \infty \right\}, \tag{10}$$

which is endowed with the inner product $\langle f, g\rangle_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{f^\wedge(\omega)\overline{g^\wedge(\omega)}}{\psi^\wedge(\omega)}\, d\omega$ with $f^\wedge$ being the Fourier transform of f in the L^2-sense. It follows from (10) that for any $f \in \mathcal{H}_k$,

$$\|f\|^2_{L^2(\mathbb{R}^d)} \overset{(\star)}{=} \|f^\wedge\|^2_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \left|f^\wedge(\omega)\right|^2 \, d\omega = \int_{\mathbb{R}^d} \frac{|f^\wedge(\omega)|^2}{\psi^\wedge(\omega)} \psi^\wedge(\omega)\, d\omega \overset{(\dagger)}{\leq} \|\psi^\wedge\|_\infty \|f\|^2_{\mathcal{H}_k} \overset{(\ddagger)}{<} \infty, \tag{11}$$

where $(\star)$ follows from Plancherel theorem (Wendland, 2005, Corollary 5.25), $\|f\|_{\mathcal{H}_k}$ is defined in (10), $(\dagger)$ follows from Hölder's inequality, and $(\ddagger)$ holds since $\psi^\wedge \in C_0(\mathbb{R}^d)$ (by Riemann-Lebesgue lemma, Folland, 1999, Theorem 8.22). Note that $\psi^\wedge$ is non-negative (Wendland, 2005, Theorem 6.11) and so the inequality in $(\dagger)$ is valid. It therefore follows from (11) that $\mathcal{H}_k$ is continuously included in $L^2(\mathbb{R}^d)$ and $\|\cdot\|_{L^2(\mathbb{R}^d)}$ is a weaker norm than $\|\cdot\|_{\mathcal{H}_k}$.[1] This means it is possible that the minimax rate of estimating μ_P in $\|\cdot\|_{L^2(\mathbb{R}^d)}$ could be faster than its RKHS counterpart with the rate possibly depending on the smoothness of k. Hence, it is of interest to analyze the minimax rates of estimating μ_P in $\|\cdot\|_{L^2(\mathbb{R}^d)}$. Interestingly, we show in this section that the minimax rate in the L^2 setting is still $n^{-1/2}$.

The analysis in the L^2 setting follows ideas similar to those of the RKHS setting wherein, first, in Section 4.1, we consider the minimax rate of estimating μ_P for translation invariant and radial kernels when $\mathcal{P}$ is the set of all Borel discrete probability measures on $\mathbb{R}^d$ (see Theorem 9 and Corollary 10). Next, in Section 4.2, we choose $\mathcal{P}$ to be the set of all probability distributions that have infinitely continuously differentiable densities and study the question of minimax rates for translation invariant (see Theorem 12) and radial kernels (see Theorem 13). For both these choices of $\mathcal{P}$, we show that the rate is $n^{-1/2}$ irrespective of the smoothness of k. Exploiting the injectivity of mean embedding for characteristic kernels (see the paragraph below and the paragraph around (6)), these results are derived using Le Cam's method (see Theorems B.1 and B.2). Combined with Proposition A.1 (also see Remark A.3), these results show that the empirical estimator, μ_{P_n} is minimax optimal. Finally, in Section 4.3 we discuss the relation between our results and some classical results of nonparametric density estimation, particularly, those of the kernel density estimator.

1. The continuous inclusion of $\mathcal{H}_k$ in $L^2(\mathbb{R}^d)$ is known for Gaussian kernels on $\mathbb{R}^d$ (e.g., see Vert and Vert, 2006, Lemma 11). Similar result is classical for Sobolev spaces in general (e.g., see Folland, 1999, Section 9.3, p. 302) and particularly for those induced by Matérn kernels. Steinwart and Christmann (2008, Theorem 4.26) provides a general result for continuous inclusion of $\mathcal{H}_k$ in $L^2(\mu)$ assuming $\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mu(x) < \infty$ where μ is a σ-finite measure. However, the result does not hold for translation invariant kernels on $\mathbb{R}^d$ as the integrability condition is violated.

Before we proceed to the main results of this section, we briefly discuss the difference between estimation in RKHS and $L^2(\mathbb{R}^d)$ norms. Suppose $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ where $\psi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ is positive definite and characteristic. It is easy to verify that $\mu_P \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Since $\mu_P = \psi * P$, (10) implies

$$\|\mu_P\|^2_{\mathcal{H}_k} = \int_{\mathbb{R}^d} \frac{|(\psi * P)^\wedge|^2}{\psi^\wedge(\omega)}\, d\omega = \int_{\mathbb{R}^d} |\phi_P(\omega)|^2 \psi^\wedge(\omega)\, d\omega = \|\phi_P\|^2_{L^2(\mathbb{R}^d, \psi^\wedge)} \tag{12}$$

whereas

$$\|\mu_P\|^2_{L^2(\mathbb{R}^d)} \overset{(\star)}{=} \int_{\mathbb{R}^d} |\mu_P^\wedge(\omega)|^2\, d\omega = \int_{\mathbb{R}^d} |\phi_P(\omega)|^2 (\psi^\wedge)^2(\omega)\, d\omega = \|\phi_P\|^2_{L^2(\mathbb{R}^d, (\psi^\wedge)^2)}, \tag{13}$$

where $\phi_P(\omega) := \int e^{-i\omega^T x}\, dP(x)$ is the characteristic function of P and $(\star)$ follows from Plancherel's theorem. It follows from (12) and (13) that the RKHS norm emphasizes the high frequencies of ϕ_P compared to that of the L^2-norm. Since ψ is characteristic, i.e., $P \mapsto \mu_k(P) \in \mathcal{H}_k$ is injective, which is guaranteed if and only if $\mathrm{supp}(\psi^\wedge) = \mathbb{R}^d$ (Sriperumbudur et al., 2010, Theorem 9), it follows from (13) that $P \mapsto \mu_k(P) \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ is injective. Therefore (2) can be equivalently written as (6) by replacing $\|\cdot\|_{\mathcal{H}_k}$ with $\|\cdot\|_{L^2(\mathbb{R}^d)}$ (see the discussion around (6)) and we obtain minimax rates by employing Le Cam's method as we did in the previous section.

4.1 Lower Bounds for Discrete Probability Measures

The following result (proved in Section 5.5) for translation invariant kernels is based on an application of Le Cam's method involving two hypotheses (see Theorem B.1), where we choose them to be KMEs of discrete measures, both supported on the same pair of points separated by a vector z in $\mathbb{R}^d$.

Theorem 9 (Translation invariant kernels) *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$. Suppose $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ where $\psi \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ is positive definite and k is characteristic. Define*

$$C_z^\psi := 2\left(\|\psi\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi(y)\psi(y+z)dy\right) \tag{14}$$

for some $z \in \mathbb{R}^d \setminus \{0\}$. Then $C_z^\psi > 0$ and

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{L^2(\mathbb{R}^d)} \geq \frac{1}{6}\sqrt{\frac{C_z^\psi}{n}} \right\} \geq \frac{1}{4}.$$

Using Cauchy-Schwartz inequality, the constant C_z^ψ in Theorem 9 can be shown (see the proof of Lemma 15 in Section 5.5) to be positive for every $z \in \mathbb{R}^d \setminus \{0\}$ if k is characteristic, i.e., $\mathrm{supp}(\Lambda_\psi) = \mathbb{R}^d$ (see (4) for Λ_ψ). The following result (proved in Section 5.6) specializes Theorem 9 to radial kernels.

Corollary 10 (Radial kernels) *Let $\mathcal{P}$ be the set of all Borel discrete probability measures on $\mathbb{R}^d$ and k be radial on $\mathbb{R}^d$, i.e., $k(x,y) = \psi_\nu(x-y) := \int_0^\infty e^{-t\|x-y\|_2^2} d\nu(t)$, where $\nu \in M_+^b([0,\infty))$ such that $\operatorname{supp}(\nu) \neq \{0\}$ and*

$$\int_0^\infty t^{-d/2} d\nu(t) < \infty. \tag{15}$$

Assume that there exist $0 < \delta_0 \leq \delta_1 < \infty$ and $\beta > 0$ such that $\nu([\delta_0, \delta_1]) \geq \beta$. Then the following holds:

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{L^2(\mathbb{R}^d)} \geq \frac{\beta}{6} \sqrt{\frac{1}{n} \left(\frac{\pi}{2\delta_1} \right)^{d/2}} \right\} \geq \frac{1}{4}.$$

In Corollary 10, since $\operatorname{supp}(\nu) \neq \{0\}$, the assumption of $\nu([\delta_0, \delta_1]) \geq \beta$ is always satisfied. In addition, the condition (15) on ν is satisfied by Gaussian, mixture of Gaussians, inverse multiquadric (while (15) is satisfied for $\gamma > d/2$, the result in Corollary 10 holds for $\gamma > d/4$) and Matérn kernels—refer to Remark A.3 for more details. Also, for these examples of kernels, the positive constant $A_k := \beta^2 \delta_1^{-d/2}$ in the lower bound in Corollary 10 can be computed in a closed form (see Appendix D.3 for details).

4.2 Lower Bounds for Probability Measures with Smooth Densities

Next, as we did in Section 3.2, we choose $\mathcal{P}$ to be the set of all probability measures that have infinitely continuously differentiable densities and show that the minimax rate of estimating μ_P in L^2-norm for translation invariant (see Theorem 12) and radial kernels (see Theorem 13) is $n^{-1/2}$. The proof of these results are again based on an application of Le Cam's method involving two (see Theorem B.1) and multiple hypotheses (see Theorem B.2), where these hypotheses are chosen to be embeddings of the d-dimensional Gaussian distributions. As in Section 3.2, the results of this section are based on the following result (proved in Section 5.7), which is conceptually similar to that of Proposition 3.

Proposition 11 *Let $\sigma > 0$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Then there exist constants $\epsilon_{\psi,\sigma^2}, c_{\psi,\sigma^2} > 0$ depending only on ψ and σ^2, such that the following condition holds for any $a \in \mathbb{R}^d$ with $\|a\|_2^2 \leq \epsilon_{\psi,\sigma^2}$:*

$$c_{\psi,\sigma^2} \leq \min_{e_z \in S^{d-1}} 2 \int_{\mathbb{R}^d} e^{-\sigma^2 \|w\|_2^2} \langle e_z, w \rangle^2 \cos\left(\langle a, w \rangle\right) \left(\psi^\wedge(w)\right)^2 dw < \infty, \tag{16}$$

where S^{d-1} is a unit sphere in $\mathbb{R}^d$. Moreover, for all vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ satisfying $\|\mu_0 - \mu_1\|_2^2 \leq \epsilon_{\psi,\sigma^2}$, the following holds:

$$\|\theta_0 - \theta_1\|_{L^2(\mathbb{R}^d)} \geq \sqrt{\frac{c_{\psi,\sigma^2}}{2}} \|\mu_0 - \mu_1\|_2,$$

where θ_0 and θ_1 are KMEs of the Gaussian measures $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ respectively.

The following result for translation invariant kernels is established using the above result wherein the proof is exactly the same as that of Theorem 6 except for an application of Proposition 11 in place of Proposition 3.

Theorem 12 (Translation invariant kernels) *Let $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Define $c_\psi := c_{\psi,1}$ and $\epsilon_\psi := \epsilon_{\psi,1}$ where $c_{\psi,1}$ and $\epsilon_{\psi,1}$ are positive constants that satisfy (16) in Proposition 11. Then for any $n \geq \frac{1}{\epsilon_\psi}$, the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{L^2(\mathbb{R}^d)} \geq \frac{1}{2}\sqrt{\frac{c_\psi}{2n}} \right\} \geq \frac{1}{4}.$$

As discussed in Remark 7, it is possible to remove the requirement of minimal sample size in Theorem 12. Also, as discussed in Remark 5 and in the paragraph following Remark 7, the constants c_ψ and ϵ_ψ appearing in the bound in Theorem 12 are not only difficult to compute but also may depend on the dimensionality d in a sup-optimal manner, particularly as $d \to \infty$. Therefore, similar to what was done in Section 3.2, we will not specialize Theorem 12 to radial kernels but instead present the following result (proved in Section 5.8 and the proof closely follows that of Theorem 8), which is based on a direct analysis involving the properties of radial kernels. For the particular case of a Gaussian kernel, this lower bound matches the upper bound of Proposition A.1 (also see Remark A.3) up to a constant factor independent of d.

Theorem 13 (Radial kernels) *Let k be radial on $\mathbb{R}^d$, i.e., $k(x,y) = \int_0^\infty e^{-t\|x-y\|_2^2}\, d\nu(t)$, where $\nu \in M_+^b([0,\infty))$ and $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable. Assume that (15) holds,* $\operatorname{supp}(\nu) \neq \{0\}$ *and there exist $0 < \delta_0 \leq \delta_1 < \infty$, $0 < \beta < \infty$ such that $\nu([\delta_0, \delta_1]) \geq \beta$. Then the following holds:*

$$\inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{L^2(\mathbb{R}^d)} \geq \frac{1}{50}\sqrt{\frac{1}{n}\left(\frac{\pi}{2\delta_1}\right)^{d/2} \frac{\beta^2\delta_0}{\delta_1 e}\left(1 - \frac{2}{2+d}\right)} \right\} \geq \frac{1}{5}.$$

The constant $B_k := \beta^2\delta_0\delta_1^{-\frac{d+2}{2}}$ in the lower bound in the above result can be computed in a closed form for Gaussian, mixture of Gaussian, inverse multiquadric, and Matérn kernels (see Appendix D.4 for details). The factor $(\pi/2)^{d/4}$ can be eliminated from the lower bound by considering a rescaled kernel $(\pi/2)^{-d/4}\psi(x-y)$. Nevertheless, the bound will still depend on d exponentially as captured by the constant B_k. This can be further overcome by using the normalized kernel $k(x,y)/\|\psi\|_{L^2(\mathbb{R}^d)}$. In the particular case of normalized Gaussian kernels $(\pi\eta^2)^{-d/2}\exp\bigl(-\frac{1}{2\eta^2}\|x-y\|_2^2\bigr)$ this will lead to dimension-free lower bounds.

4.3 Relation to Kernel Density Estimation

In this section, we discuss the relation between the estimation of μ_P and density estimation. The problem of density estimation deals with estimating an unknown density, p based on

random samples $(X_i)_{i=1}^n$ drawn i.i.d. from it. One of the popular non-parametric methods for density estimation is kernel density estimation (KDE), where the estimator is of the form (Tsybakov, 2008, Section 1.2)

$$\hat{p}_n(x_1, \ldots, x_d) = \frac{1}{n \prod_{i=1}^d h_i} \sum_{i=1}^n K\left(\frac{X_{i,1} - x_1}{h_1}, \cdots, \frac{X_{i,d} - x_d}{h_d}\right).$$

Here $K \colon \mathbb{R}^d \to \mathbb{R}$ is the *smoothing kernel* (this kernel should not be confused with the reproducing kernel k which we used throughout the paper), $h_1, \ldots, h_d > 0$ are bandwidths, and $X_{i,j}$ is the j-th coordinate of the i-th sample point. Assuming $p \in L^2(\mathbb{R}^d)$, the consistency of $\hat{p}_n$ is usually studied in the sense of *mean integrated squared error* (MISE) $\mathbb{E}\|\hat{p}_n - p\|^2_{L^2(\mathbb{R}^d)}$, which can be decomposed into variance and bias terms as:

$$\mathbb{E}\|\hat{p}_n - p\|^2_{L^2(\mathbb{R}^d)} = \mathbb{E}\big\|\hat{p}_n - \mathbb{E}[\hat{p}_n]\big\|^2_{L^2(\mathbb{R}^d)} + \big\|p - \mathbb{E}[\hat{p}_n]\big\|^2_{L^2(\mathbb{R}^d)}. \tag{17}$$

Assume K to be bounded and $h_1 = \cdots = h_d = h$. Define $K_h := h^{-d} K(\cdot/h)$. Then for any fixed $x \in \mathbb{R}^d$,

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) = \int_{\mathbb{R}^d} K_h(z - x)\, dP_n(z)$$

and

$$\mathbb{E}[\hat{p}_n(x)] = \frac{1}{h^d} \int_{\mathbb{R}^d} K\left(\frac{z - x}{h}\right) p(z) dz = (K_h * p)(x).$$

This shows that $\hat{p}_n = \mu_{K_h}(P_n)$ and $\mathbb{E}[\hat{p}_n] = \mu_{K_h}(P)$ where P is the distribution with p as its density w.r.t. the Lebesgue measure and P_n is the empirical measure constructed based on samples $(X_i)_{i=1}^n$ drawn from p. Therefore the results of Section 4 (and more generally of this paper) are about the minimax rates for $\mathbb{E}[\hat{p}_n]$. However, note that K_h need not be positive definite (and therefore need not be the reproducing kernel of some RKHS). On the other hand, K has to be positive, i.e., $K(x) \geq 0$, $\forall\, x \in \mathbb{R}^d$ and normalized, i.e., $\int_{\mathbb{R}^d} K(x)\, dx = 1$ to yield an estimator that is a valid density, unlike in kernel mean estimation where k need not be positive nor normalized. The minimax rate of $n^{-1/2}$ for estimating $\mathbb{E}[\hat{p}_n]$ is achieved by the kernel density estimator $\hat{p}_n$ (which is nothing but the empirical estimator of $\mu_{K_h}(P)$) as it is known (based on a straightforward generalization of Tsybakov, 2008, Proposition 1.4 for multiple dimensions) that

$$\mathbb{E}\big\|\hat{p}_n - \mathbb{E}[\hat{p}_n]\big\|^2_{L^2(\mathbb{R}^d)} \leq \frac{\|K\|^2_{L^2(\mathbb{R}^d)}}{nh^d},$$

where we assume $K \in L^2(\mathbb{R}^d)$. The bandwidth parameter h is immaterial in the estimation of $\mu_{K_h}(P)$ and can be treated as a constant (independent of n) unlike in the problem of estimating p where h should decay to zero at an appropriate rate for the bias $\big\|p - \mathbb{E}[\hat{p}_n]\big\|_{L^2(\mathbb{R}^d)}$ to converge to zero as $n \to \infty$. In particular, if p lies in a Sobolev space of smoothness index s, then the bias-squared term in (17) behaves as h^{2s}, which combined with the above bound on the variance yields a rate of $n^{-\frac{2s}{2s+1}}$ for $h = n^{-\frac{1}{2s+1}}$. This rate is known to be minimax optimal for the problem of estimating p while our rates are minimax optimal for the problem of smoothed density estimation where the smoothing is carried out by the kernel.

5. Proofs

In this section we present all the missing proofs of results of Sections 3 and 4.

5.1 Proof of Theorem 1

Pick two discrete distributions $P_0 = p_0\delta_x + (1-p_0)\delta_v$ and $P_1 = p_1\delta_x + (1-p_1)\delta_v$, where $x, v \in \mathbb{R}^d$, $0 < p_0 < 1$, $0 < p_1 < 1$ and δ_x denotes a Dirac measure supported at x. Define $\theta_0 = \mu_k(P_0)$ and $\theta_1 = \mu_k(P_1)$. Since $\|\theta_0\|^2_{\mathcal{H}_k} = \int\int k(x,y)\,dP_0(x)\,dP_0(y)$, which follows from the reproducing property of k, it is easy to verify that

$$\|\theta_0 - \theta_1\|^2_{\mathcal{H}_k} = \mathbb{E}[k(\xi,\xi')] + \mathbb{E}[k(\eta,\eta')] - 2\,\mathbb{E}[k(\xi,\eta)],$$

where ξ and η are random variables distributed according to P_0 and P_1 respectively, and ξ' and η' are independent copies of ξ and η. Since k is translation invariant, we have $k(v,v) = k(x,x) = \psi(0)$ and $k(x,v) = k(v,x) = \psi(x-v)$, which imply

$$\|\theta_0 - \theta_1\|^2_{\mathcal{H}_k} = 2(p_0-p_1)^2\big(\psi(0) - \psi(x-v)\big). \tag{18}$$

Also note that

$$\begin{aligned}
\mathrm{KL}(P_0\|P_1) &= p_0 \log\frac{p_0}{p_1} + (1-p_0)\log\frac{1-p_0}{1-p_1} \\
&= p_0 \log\left(1 + \frac{p_0-p_1}{p_1}\right) + (1-p_0)\log\left(1 + \frac{p_1-p_0}{1-p_1}\right) \\
&\overset{(*)}{\leq} \log\left\{p_0\left(1 + \frac{p_0-p_1}{p_1}\right) + (1-p_0)\left(1 + \frac{p_1-p_0}{1-p_1}\right)\right\} \\
&= \log\left(1 + (p_0-p_1)\left(\frac{p_0}{p_1} - \frac{1-p_0}{1-p_1}\right)\right),
\end{aligned}$$

where we used Jensen's inequality in $(*)$ for the logarithmic function, which is concave. Next, using a simple inequality $\log(1+x) \leq x$, which holds for all $x > -1$, we get

$$\mathrm{KL}(P_0\|P_1) \leq (p_0-p_1)\left(\frac{p_0}{p_1} - \frac{1-p_0}{1-p_1}\right) = \frac{(p_0-p_1)^2}{p_1(1-p_1)}.$$

Note that a maximal value of denominator is achieved when $p_1 = \frac{1}{2}$. Setting $p_1 = \frac{1}{2}$ we get the following upper bound: $\mathrm{KL}(P_0\|P_1) \leq 4\left(p_0 - \frac{1}{2}\right)^2$, which when used in the chain rule of KL-divergence yields

$$\mathrm{KL}(P_0^n\|P_1^n) \leq 4n\left(p_0 - \frac{1}{2}\right)^2.$$

Choosing p_0 such that $(p_0 - \frac{1}{2})^2 = \frac{1}{9n}$ yields $\mathrm{KL}(P_0^n\|P_1^n) \leq \frac{4}{9}$ and $\|\theta_0 - \theta_1\|^2_{\mathcal{H}_k} = \frac{2}{9n}\big(\psi(0) - \psi(x-v)\big)$. Choose x and v in such a way that $x - v = z$, where $z \in \mathbb{R}^d$ is a point for which $\psi(0) - \psi(z) \geq \beta$ and $\beta > 0$. This yields

$$\|\theta_0 - \theta_1\|^2_{\mathcal{H}_k} \geq \frac{2\beta}{9n},$$

which shows that the assumptions of Theorem B.1 are satisfied with $s := \frac{1}{6}\sqrt{\frac{2\beta}{n}}$ and $\alpha := \frac{4}{9}$. The result follows from an application of Theorem B.1 by noticing that $\frac{1-\sqrt{\alpha/2}}{2} > 1/4$.

Remark (Measures with bounded support) *It is evident from the above proof that exactly the same lower bound holds if we restrict $\mathcal{P}$ to contain only probability measures with bounded support. We can proceed further and assume that for each $P \in \mathcal{P}$ the radius of* $\mathrm{supp}(P)$ *is upper bounded by some positive constant R. In this case the same reasoning will work as long as ψ is not "flat" on the ball of radius R centered around origin.*

5.2 Proof of Corollary 2

The proof is based on application of Theorem 1. Since $\mathrm{supp}(\nu) \neq \{0\}$, it follows from (Sriperumbudur et al., 2011, Proposition 5) that k is characteristic. We now show that there exist $z \in \mathbb{R}^d$ and $\beta > 0$, such that $\psi_\nu(0) - \psi_\nu(z) \geq \beta$. Note that for any $x \in \mathbb{R}^d$

$$\begin{aligned}\psi_\nu(0) - \psi_\nu(x) &= \int_0^\infty \left(1 - e^{-t\|x\|_2^2}\right) d\nu(t) \geq \int_{t_1}^\infty \left(1 - e^{-t\|x\|_2^2}\right) d\nu(t) \\ &\geq \int_{t_1}^\infty \left(1 - e^{-t_1\|x\|_2^2}\right) d\nu(t) = \nu([t_1,\infty)) \left(1 - e^{-t_1\|x\|_2^2}\right) \\ &\geq \alpha \left(1 - e^{-t_1\|x\|_2^2}\right) \geq \frac{\alpha t_1}{2}\|x\|_2^2,\end{aligned}$$

where the last inequality holds whenever $\|x\|_2^2 \leq \frac{1}{t_1}$. Choosing z such that $\|z\|_2^2 = \frac{1}{t_1}$ yields $\psi_\nu(0) - \psi_\nu(z) \geq \frac{\alpha}{2}$. The result therefore follows from Theorem 1 by choosing $\beta = \frac{\alpha}{2}$.

5.3 Proof of Proposition 3

Before we prove Proposition 3, first we will derive a closed form expression for the RKHS distance between KMEs of two d-dimensional Gaussian distributions with the kernel being translation invariant, i.e., $k(x,y) = \psi(x-y)$. Throughout this section Λ_ψ will denote a finite non-negative Borel measure corresponding to the positive-definite function ψ from (4).

Lemma 14 *Let θ_0 and θ_1 be KME of Gaussian measures $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ for $\mu_0, \mu_1 \in \mathbb{R}^d$ and $\sigma^2 > 0$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in C_b(\mathbb{R}^d)$ is positive definite. Then*

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \left(1 - \cos\left(\langle \mu_0 - \mu_1, w\rangle\right)\right) d\Lambda_\psi(w). \tag{19}$$

Proof Note that

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 = \|\theta_0\|_{\mathcal{H}_k}^2 + \|\theta_1\|_{\mathcal{H}_k}^2 - 2\langle \theta_0, \theta_1\rangle_{\mathcal{H}_k}, \tag{20}$$

where $\langle \theta_0, \theta_1\rangle_{\mathcal{H}_k} = \mathbb{E}_X \mathbb{E}_Y[k(X,Y)]$ with $X \sim G(\mu_0, \sigma^2 I)$ and $Y \sim G(\mu_1, \sigma^2 I)$. We will now derive the closed form for the inner product:

$$\langle \theta_0, \theta_1\rangle_{\mathcal{H}_k} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \psi(x-y) \frac{1}{(2\pi\sigma^2)^d} e^{-\frac{1}{2\sigma^2}\|x-\mu_0\|_2^2 - \frac{1}{2\sigma^2}\|y-\mu_1\|_2^2} dx dy$$

$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle x-y,w\rangle} d\Lambda_\psi(w) \frac{1}{(2\pi\sigma^2)^d} e^{-\frac{1}{2\sigma^2}\|x-\mu_0\|_2^2 - \frac{1}{2\sigma^2}\|y-\mu_1\|_2^2} dx\, dy,$$

where we used (4). The function appearing under the integral is absolutely integrable and so by Tonelli-Fubini theorem (Dudley, 2002, Theorem 4.4.5) we obtain

$$\begin{aligned}
\langle \theta_0, \theta_1 \rangle_{\mathcal{H}_k} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{e^{i\langle y,w\rangle} e^{-\frac{1}{2\sigma^2}\|y-\mu_1\|_2^2}}{(2\pi\sigma^2)^{d/2}(2\pi)^{d/2}} \left\{ \int_{\mathbb{R}^d} \frac{e^{-i\langle x,w\rangle}}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|x-\mu_0\|_2^2} dx \right\} dy\, d\Lambda_\psi(w) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \frac{1}{(2\pi)^{d/2}} e^{i\langle y,w\rangle} e^{-\frac{1}{2\sigma^2}\|y-\mu_1\|_2^2} e^{-i\langle \mu_0,w\rangle - \frac{\sigma^2\|w\|_2^2}{2}} dy\, d\Lambda_\psi(w) \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left\{ \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} e^{i\langle y,w\rangle} e^{-\frac{1}{2\sigma^2}\|y-\mu_1\|_2^2} dy \right\} e^{-i\langle \mu_0,w\rangle - \frac{\sigma^2\|w\|_2^2}{2}} d\Lambda_\psi(w) \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle \mu_1,w\rangle - \frac{\sigma^2\|w\|_2^2}{2}} e^{-i\langle \mu_0,w\rangle - \frac{\sigma^2\|w\|_2^2}{2}} d\Lambda_\psi(w),
\end{aligned}$$

where we used Lemma C.1 to compute the Fourier transform for a Gaussian density. Using Euler's formula and the fact that Λ_ψ is symmetric according to Lemma C.2, while $\sin(x)$ is an odd function, we get

$$\langle \theta_0, \theta_1 \rangle_{\mathcal{H}_k} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \cos\left(\langle \mu_0 - \mu_1, w\rangle\right) e^{-\sigma^2\|w\|_2^2} d\Lambda_\psi(w). \tag{21}$$

The result in (19) follows by using (21) in (20). ■

Proof of Proposition 3: Define $a := \mu_0 - \mu_1$ and

$$G(a) := \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \left(1 - \cos\langle a, w\rangle\right) d\Lambda_\psi(w).$$

Note that $G(0) = 0$. Next, since for any $i = 1, \ldots, d$

$$\left| \frac{\partial}{\partial a_i} e^{-\sigma^2\|w\|_2^2} (1 - \cos\langle a, w\rangle) \right| = \left| e^{-\sigma^2\|w\|_2^2} w_i \sin\langle a, w\rangle \right| \leq \left| e^{-\sigma^2\|w\|_2^2} w_i \right| \in L_1(\Lambda_\psi),$$

we can differentiate G under the integral sign (Folland, 1999, Theorem 2.27) and get $\nabla G(0) = 0$.

If a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is *strongly convex* with parameter $m > 0$ on some set $A \subseteq \mathbb{R}^d$, then for all $x, y \in A$:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle + \frac{m}{2}\|x - y\|_2^2.$$

If we can show that G is strongly convex on $B_\epsilon := \{b \in \mathbb{R}^d\colon \|b\|_2^2 \leq \epsilon\}$ for some $\epsilon > 0$, then we can apply previous inequality with $y = 0$ and $x = a$ to obtain

$$G(a) \geq \frac{m}{2}\|a\|_2^2, \ \forall a \in B_\epsilon.$$

It is known that a twice continuously differentiable function f is strongly convex on $A \subseteq \mathbb{R}^d$ with parameter $m > 0$ if the matrix $\nabla^2 f(x) - m \cdot I$ is positive definite for all $x \in A$, where

$\nabla^2 f$ is the Hessian and $I \in \mathbb{R}^{d\times d}$ is an identity matrix. Next we compute the Hessian of G by once again employing differentiation under the integral sign (justified in the similar way as above) to obtain

$$\frac{\partial^2 G(a)}{\partial a_i \partial a_j} = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} w_i w_j \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w), \quad 0 \le i, j \le d.$$

Thus

$$\nabla^2 G(a) = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} w w^T \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w).$$

In order to prove that G is strongly convex on $B_\epsilon \subseteq \mathbb{R}^d$ we need to show that $\nabla^2 G(a) - m \cdot I$ is positive definite for each $a \in B_\epsilon$ and some $m > 0$. In other words, we need to show that there is $m > 0$ such that for each $z \in \mathbb{R}^d \setminus \{0\}$ and $a \in B_\epsilon$ the following holds:

$$\langle z, \nabla^2 G(a) z\rangle \ge m\|z\|_2^2,$$

or, equivalently,

$$\frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 e^{-i\langle a, w\rangle} d\Lambda_\psi(w) \ge m, \tag{22}$$

where $e_z := z/\|z\|_2 \in \mathbb{R}^d$ is a vector of unit length pointed in the direction of z. Note that l.h.s. of (22) is the Fourier transform of a measure $\mathcal{T}_z$ on $\mathbb{R}^d$, which is absolutely continuous with respect to Λ_ψ with Radon-Nikodym derivative $2e^{-\sigma^2\|w\|_2^2}\langle e_z, w\rangle^2$.

Fix any $z \in \mathbb{R}^d$. We will first show that we can apply Bochner's Theorem (see (4)) for the measure $\mathcal{T}_z$. For this we need to check that it is (a) non-negative and (b) finite. Part (a) is apparent from the facts that Λ_ψ is non-negative and $\mathcal{T}_z$ has a non-negative density with respect to Λ_ψ. To check (b) we write

$$\int_{\mathbb{R}^d} d\mathcal{T}_z(x) = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 d\Lambda_\psi(w) < \infty,$$

as $e^{-\sigma^2\|w\|_2^2}\langle e_z, w\rangle^2$ is positive and bounded for any $z \in \mathbb{R}^d$, while Λ_ψ is finite. We conclude from Bochner's Theorem, that function $\tilde{\psi}_z(x)$ defined in the following way:

$$\tilde{\psi}_z(x) = \int_{\mathbb{R}^d} e^{-i\langle x, w\rangle} d\mathcal{T}_z(w),$$

is positive-definite. Moreover it is well known (Dudley, 2002, Theorem 9.4.4) that $\tilde{\psi}_z \in C_b(\mathbb{R}^d)$ as $\tilde{\psi}_z$ is the characteristic function of $\mathcal{T}_z$. Finally, it follows from the discussion in (Sriperumbudur et al., 2011, Section 3.3), that if $\mathrm{supp}(\mathcal{T}_z) = \mathbb{R}^d$, then a bounded and continuous function $\tilde{\psi}_z(x)$ is *strictly* positive definite. To check the condition $\mathrm{supp}(\mathcal{T}_z) = \mathbb{R}^d$ we note that $\mathrm{supp}(\Lambda_\psi) = \mathbb{R}^d$ since ψ is characteristic (Sriperumbudur et al., 2010, Theorem 9), and no open sets of $\mathbb{R}^d$ are contained in the region where $e^{-\sigma^2\|w\|_2^2}\langle e_z, w\rangle^2 = 0$.

Summarizing, we have established that the l.h.s. of (22) is equal to $\tilde{\psi}_z(a)$, where $\tilde{\psi}_z \colon \mathbb{R}^d \to \mathbb{R}$ is a bounded, continuous and strictly positive definite function for each

$z \in \mathbb{R}^d \setminus \{0\}$. In particular, we have $\tilde{\psi}_z(0) > 0$ for all $z \in \mathbb{R}^d$. Note that $\tilde{\psi}_z(0)$ depends on z only through its direction. Next we want to show that

$$\inf_{z \in S_d} \tilde{\psi}_z(0) > 0, \tag{23}$$

where the infimum is over the unit sphere $S_d := \{b \in \mathbb{R}^d \colon \|b\|_2^2 = 1\}$. Note that the function $F \colon z \to \tilde{\psi}_z(0)$ defined on S_d is continuous. Since S_d is closed and bounded, we know that F attains its minimum on it. In other words, there is $z^* \in S_d$, such that

$$\inf_{z \in S_d} \tilde{\psi}_z(0) = \tilde{\psi}_{z^*}(0).$$

Thus, if $\inf_{z \in S_d} \tilde{\psi}_z(0) = 0$, we will also get $\tilde{\psi}_{z^*}(0) = 0$, which will contradict the fact that $\tilde{\psi}_z(0) > 0$ for each $z \in \mathbb{R}^d \setminus \{0\}$. This proves (23). Using Lemma C.3 we also conclude that $\inf_{z \in S_d} \tilde{\psi}_z \colon \mathbb{R}^d \to \mathbb{R}$ is a continuous function. Now we may finally conclude that there are constants $c_{\psi,\sigma^2}, \epsilon_{\psi,\sigma^2} > 0$ such that

$$\inf_{z \in S_d} \tilde{\psi}_z(a) \geq c_{\psi,\sigma^2}$$

for all $a \in B_{\epsilon_{\psi,\sigma^2}}$. Finally, we take $m = c_{\psi,\sigma^2}$ and this concludes the proof.

5.4 Proof of Theorem 8

The proof is based on application of Theorem B.2 where we choose $\theta_0, \ldots, \theta_M$ to be KMEs of d-dimensional Gaussian measures with variances decaying to zero as $d \to \infty$.

Let $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ be two d-dimensional Gaussian distributions with mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ and variance $\sigma^2 > 0$. Define θ_0 and θ_1 to be the embeddings of $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ respectively.

(A) Deriving a closed form expression for $\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2$.

Using Lemma 14, presented in Section 5.3, we have

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2 \|w\|_2^2} \left(1 - \cos\left(\langle \mu_0 - \mu_1, w \rangle\right)\right) d\Lambda_\psi(w), \tag{24}$$

where Λ_ψ is a finite non-negative Borel measure from the Bochner's Theorem corresponding to the kernel k. We now show that Λ_ψ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and has the following density:

$$\lambda_\psi(w) = \int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t), \quad w \in \mathbb{R}^d.$$

Indeed, by noticing that

$$\int_0^\infty \left(\int_{\mathbb{R}^d} \left| e^{-i\langle w, x \rangle} \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} \right| dw \right) d\nu(t) < \infty$$

we may apply Tonelli-Fubini theorem (Dudley, 2002, Theorem 4.4.5) to interchange the order of integration and get

$$\int_{\mathbb{R}^d} \frac{e^{-i\langle w,x\rangle}}{(2\pi)^{d/2}} \left(\int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t) \right) dw = \int_0^\infty \left(\int_{\mathbb{R}^d} \frac{e^{-i\langle w,x\rangle}}{(2\pi)^{d/2}} \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} dw \right) d\nu(t)$$
$$= \int_0^\infty e^{-t\|x\|_2^2} d\nu(t).$$

Substituting the form of λ_ψ into (24) we can write

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 = \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_0^\infty e^{-\sigma^2\|w\|_2^2} \left(1 - \cos\left(\langle \mu_0 - \mu_1, w\rangle\right)\right) \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} \, d\nu(t) \, dw.$$

Applying Tonelli-Fubini theorem once again and using Lemma C.1 together with Euler's formula we obtain

$$\begin{aligned}
\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 &= \frac{2}{(2\pi)^{d/2}} \int_0^\infty \frac{1}{(2t)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{\|w\|_2^2}{2}(2\sigma^2 + \frac{1}{2t})} \left(1 - \cos\left(\langle \mu_0 - \mu_1, w\rangle\right)\right) dw \, d\nu(t) \\
&= \int_0^\infty \frac{2}{(4\sigma^2 t + 1)^{d/2}} d\nu(t) - \int_0^\infty \frac{2}{(4\sigma^2 t + 1)^{d/2}} e^{-\frac{t\|\mu_0 - \mu_1\|_2^2}{4\sigma^2 t + 1}} d\nu(t) \\
&= \int_0^\infty 2 \left(\frac{1}{1 + 4t\sigma^2} \right)^{d/2} \left(1 - \exp\left(-\frac{t\|\mu_0 - \mu_1\|_2^2}{1 + 4t\sigma^2} \right) \right) d\nu(t). \qquad (25)
\end{aligned}$$

(B) Lower bounding $\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2$ in terms of $\|\mu_0 - \mu_1\|_2^2$.

It follows from (25) that

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 \geq \int_{t_0}^{t_1} 2 \left(\frac{1}{1 + 4t\sigma^2} \right)^{d/2} \left(1 - \exp\left(-\frac{t\|\mu_0 - \mu_1\|_2^2}{1 + 4t\sigma^2} \right) \right) d\nu(t),$$

where $0 < t_0 \leq t_1 < \infty$. Note that $1 - e^{-x} \geq \frac{x}{2}$ for $0 \leq x \leq 1$. Using this we get

$$1 - \exp\left(-\frac{t\|\mu_0 - \mu_1\|_2^2}{1 + 4t\sigma^2} \right) \geq \frac{t\|\mu_0 - \mu_1\|_2^2}{2(1 + 4t\sigma^2)}, \quad \forall t \in [t_0, t_1]$$

as long as

$$t_1 \|\mu_0 - \mu_1\|_2^2 \leq 1 + 4t_1\sigma^2. \qquad (26)$$

Thus, as long as (26) holds, we can lower bound the RKHS distance as:

$$\begin{aligned}
\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 &\geq \int_{t_0}^{t_1} \left(\frac{1}{1 + 4t\sigma^2} \right)^{d/2} \frac{t\|\mu_0 - \mu_1\|_2^2}{1 + 4t\sigma^2} \, d\nu(t) \\
&= \|\mu_0 - \mu_1\|_2^2 \int_{t_0}^{t_1} \frac{t}{(1 + 4t\sigma^2)^{(d+2)/2}} \, d\nu(t). \qquad (27)
\end{aligned}$$

Note that the function $t \mapsto \frac{t}{(1+4t\sigma^2)^{(d+2)/2}}$ monotonically increases on $[0, \frac{1}{2d\sigma^2}]$, reaches its global maximum at $t = \frac{1}{2d\sigma^2}$ and then decreases on $[\frac{1}{2d\sigma^2}, \infty)$. Thus we have

$$\int_{t_0}^{t_1} \frac{t}{(1+4t\sigma^2)^{(d+2)/2}}\, d\nu(t) \geq \beta \min\left\{ \frac{t_0}{(1+4t_0\sigma^2)^{(d+2)/2}}, \frac{t_1}{(1+4t_1\sigma^2)^{(d+2)/2}} \right\}.$$

Setting

$$\sigma^2 = \frac{1}{2t_1 d} \tag{28}$$

yields that $t = t_1$ is the global maximum of the function $t \mapsto \frac{t}{\left(1+\frac{2t}{t_1 d}\right)^{(d+2)/2}}$, in which case

$$\frac{t_0}{(1+4t_0\sigma^2)^{(d+2)/2}} \leq \frac{t_1}{(1+4t_1\sigma^2)^{(d+2)/2}}.$$

With this choice of σ^2 we get

$$\begin{aligned}\int_{t_0}^{t_1} \frac{t}{(1+4t\sigma^2)^{(d+2)/2}}\, d\nu(t) &\geq \frac{\beta t_0}{\left(1+\frac{2t_0}{t_1 d}\right)^{(d+2)/2}} \geq \frac{\beta t_0}{\left(1+\frac{2}{d}\right)^{(d+2)/2}} \\ &= \beta t_0 \left(1 - \frac{2}{2+d}\right)^{(d+2)/2} \geq \frac{\beta t_0}{e}\left(1 - \frac{2}{2+d}\right), \end{aligned} \tag{29}$$

where we used the fact that $(1-\frac{1}{x})^{x-1}$ monotonically decreases to $\frac{1}{e}$. Using (29) in (27), we obtain

$$\|\theta_0 - \theta_1\|_{\mathcal{H}_k}^2 \geq \frac{\beta t_0}{e}\left(1 - \frac{2}{2+d}\right)\|\mu_0 - \mu_1\|_2^2. \tag{30}$$

(C.1) Application of Theorem B.2: Choosing $\theta_0, \ldots, \theta_M$.

Now we are going to apply Theorem B.2. First of all, we need to choose $M+1$ embeddings. Recall that Theorem B.2 requires these embeddings to be sufficiently distant from each other, while the corresponding distributions should be close. We will choose the embeddings $\{\theta_0, \ldots, \theta_M\}$ to be KMEs of Gaussian distributions $G(\mu_i, \sigma^2 I)$ for specific choice of $\sigma^2 > 0$ and $\mu_i \in \mathbb{R}^d$, $i = 0, \ldots, M$. Mean vectors $\{\mu_i\}_{i=0}^M$ will be constrained to live in the ball $B(c_\nu, n) := \left\{x \in \mathbb{R}^d \colon \|x\|_2^2 \leq c_\nu/n\right\}$, where c_ν is a positive constant to be specified later. This guarantees that KL-divergences between the Gaussian distributions will remain small. At the same time, it was shown in (30) that the RKHS distance between embeddings θ_i and θ_j is lower bounded by the Euclidean distance between μ_i and μ_j. In other words, in order for the embeddings $\theta_0, \ldots, \theta_M$ to be sufficiently separated we need to make sure that the mean vectors $\mu_0, \ldots, \mu_M$ are not too close to each other. Summarizing, we face the problem of choosing a finite collection of pairwise distant points in the Euclidean ball. This question is closely related to the concepts of *packing* and *covering numbers*.

For any set $A \in \mathbb{R}^d$ its *ϵ-packing number* $M(A, \epsilon)$ is the largest number of points in A separated from each other by at least a distance of ϵ. An *ϵ-covering number* $N(A, \epsilon)$ of A

is the minimal number of balls of radius ϵ needed to cover A. Packing numbers are lower bounded by the covering numbers (Dudley, 1999, Theorem 1.2.1):

$$N(A, \epsilon) \leq M(A, \epsilon).$$

Also, it is well known that the ϵ-covering number of a unit d-dimensional Euclidean ball is lower bounded by $\lfloor \epsilon^{-d} \rfloor$. Together, these facts state that we can find at least $\lfloor \epsilon^{-d} \rfloor$ points in the d-dimensional unit ball, which are at least ϵ away from each other. Similarly (just by a simple scaling) we can find at least $\lfloor \epsilon^{-d} \rfloor$ points in the d-dimensional ball of radius $R > 0$, which are at least $R \cdot \epsilon$ away from each other. Applying this fact to $B(c_\nu, n)$ we can finally argue that there are at least N^d points in $B(c_\nu, n)$ which are at least $N^{-1}\sqrt{c_\nu/n}$ away from each other, where $N \geq 3$ is an integer to be specified later. Now, take $M = N^d - 1 \geq 2$ (which explains the lower bound $N \geq 3$) and fix $\mu_0, \ldots, \mu_M$ to be these $M + 1$ points.

(C.2) Application of Theorem B.2: Lower bounding $\|\theta_i - \theta_j\|_{\mathcal{H}_k}$.

With this choice of parameters $\mu_0, \ldots, \mu_M$, for any $0 \leq i < j \leq M$, we have

$$\|\theta_i - \theta_j\|_{\mathcal{H}_k}^2 \geq \frac{c_\nu}{N^2 n} \frac{\beta t_0}{e} \left(1 - \frac{2}{2+d}\right),$$

where we used (30) and the lower bound on $\|\mu_i - \mu_j\|_2$. Setting

$$c_\nu = \frac{C}{\beta t_0} \tag{31}$$

for some $C > 0$ we obtain

$$\|\theta_i - \theta_j\|_{\mathcal{H}_k}^2 \geq \frac{C}{N^2 e n} \left(1 - \frac{2}{2+d}\right).$$

This satisfies the first assumption of Theorem B.2 with $s := \frac{1}{2N}\sqrt{\frac{C}{en}\left(1 - \frac{2}{2+d}\right)}$.

(C.3) Application of Theorem B.2: Upper bounding $\mathrm{KL}(P_{\theta_i} \| P_{\theta_j})$.

Note that for any $0 \leq i < j \leq M$ we have

$$\mathrm{KL}\big(G^n(\mu_i, \sigma^2 I) \| G^n(\mu_j, \sigma^2 I)\big) = n \cdot \frac{\|\mu_i - \mu_j\|_2^2}{2\sigma^2} \leq \frac{2c_\nu}{\sigma^2} = 4C\frac{t_1 d}{\beta t_0},$$

where the inequality holds since $\mu_i \in B(c_\nu, n)$ and the equality follows from (28) and (31). Here we used the fact that for any points x and y contained in a ball of radius R we obviously have $\|x - y\| \leq 2R$. Also note that we chose $M = N^d - 1 \geq 2$ and thus

$$\log(M) = d\log(N) + \log(1 - N^{-d}) \geq d\log(N) + 1 - \frac{N^d}{N^d - 1} \geq d\log(N) - \frac{1}{N-1}$$

for $d \geq 1$, where we used the inequality $\log(x) \geq 1 - \frac{1}{x}$ which holds for $x \geq 0$. Taking

$$C = \frac{\beta t_0}{32 t_1} \left(\log N - \frac{1}{N-1} \right) \tag{32}$$

we get

$$4C \frac{t_1 d}{\beta t_0} = \frac{1}{8} \left(d \log N - \frac{d}{N-1} \right) \leq \frac{1}{8} \left(d \log N - \frac{1}{N-1} \right) \leq \frac{1}{8} \log(M)$$

for any $d \geq 1$. Concluding, we get

$$\mathrm{KL}\big(G^n(\mu_i, \sigma^2 I) \| G^n(\mu_j \| \sigma^2 I)\big) \leq \frac{1}{8} \log(M)$$

and thus the second assumption of Theorem B.2 is satisfied with $\alpha = \frac{1}{8}$. Finally, it is easy to check that if we take $N = 5$ then condition (26) will be satisfied. Indeed,

$$t_1 \|\mu_0 - \mu_1\|_2^2 \leq \frac{4 t_1 c_\nu}{n} = \frac{4 t_1}{\beta t_0 n} \frac{\beta t_0}{32 t_1} \left(\log N - \frac{1}{N-1} \right) = \frac{1}{8n} \left(\log N - \frac{1}{N-1} \right),$$

while

$$1 + 4 t_1 \sigma^2 \geq 1.$$

Thus (26) holds whenever

$$\log N - \frac{1}{N-1} \leq 8n,$$

which obviously holds for $N = 5$ and any $n \geq 1$.

To conclude the proof we insert (32) into $s := \frac{1}{2N} \sqrt{\frac{C}{en} \left(1 - \frac{2}{2+d}\right)}$, lower bound this value using $\frac{1}{8} \left(\log N - \frac{1}{N-1} \right) \geq \frac{4}{25}$, and notice that

$$\frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(M)}} \right) \geq \frac{1}{5}.$$

5.5 Proof of Theorem 9

The proof is based on the following result, which gives a closed form expression for the $L^2(\mathbb{R}^d)$ distance between embeddings of two discrete distributions supported on the same pair of points in $\mathbb{R}^d$.

Lemma 15 *Suppose $k(x, y) = \psi(x - y)$, where $\psi \in L^2(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$ is positive definite and k is characteristic. Define $P_0 = p_0 \delta_x + (1 - p_0)\delta_v$ and $P_1 = p_1 \delta_x + (1 - p_1)\delta_v$, where $0 < p_0 < 1$, $0 < p_1 < 1$, $x, v \in \mathbb{R}^d$, and $x \neq v$. Then*

$$\|\mu_k(P_0) - \mu_k(P_1)\|^2_{L^2(\mathbb{R}^d)} = C^\psi_{x,v} (p_0 - p_1)^2,$$

where

$$C^\psi_{x,v} := 2 \left(\|\psi\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi(y) \psi(y + x - v) dy \right) > 0.$$

Proof We have

$$\begin{aligned}
\|\mu_k(P_0) - \mu_k(P_1)\|^2_{L^2(\mathbb{R}^d)} &= (p_0 - p_1)^2 \, \|k(x,\cdot) - k(v,\cdot)\|^2_{L^2(\mathbb{R}^d)} \\
&= (p_0 - p_1)^2 \int_{\mathbb{R}^d} \big(\psi(x-y) - \psi(v-y)\big)^2 dy \\
&\overset{(\star)}{=} (p_0 - p_1)^2 \int_{\mathbb{R}^d} \big(\psi(y) - \psi(y+x-v)\big)^2 dy \\
&= 2(p_0 - p_1)^2 \int_{\mathbb{R}^d} \psi(y)\big(\psi(y) - \psi(y+x-v)\big) dy \\
&= 2(p_0 - p_1)^2 \left(\|\psi\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi(y)\psi(y+x-v) dy \right),
\end{aligned}$$

where we used the symmetry of ψ in $(\star)$. Cauchy-Schwartz inequality states that

$$\int_{\mathbb{R}^d} \psi(y)\psi(y+x-v)dy \overset{(\star)}{\leq} \sqrt{\int_{\mathbb{R}^d} \psi^2(y)dy} \sqrt{\int_{\mathbb{R}^d} \psi^2(y+x-v)dy} = \|\psi\|^2_{L^2(\mathbb{R}^d)},$$

where the equality in $(\star)$ holds if and only if $\psi(y) = \lambda\, \psi(y+x-v)$ for some constant λ and all $y \in \mathbb{R}^d$. If we take $y = v - x$ the above condition implies $\psi(v-x) = \lambda\, \psi(0)$ and with $y = 0$ we get $\psi(0) = \lambda\, \psi(x-v)$. Together these identities show that $\psi(0) = \lambda^2\, \psi(0)$. Since the kernel is characteristic and translation invariant, ψ is strictly positive definite, which means $\psi(0) > 0$. We conclude that $\lambda = \pm 1$. Assume that $\lambda = -1$. In this case $\psi(x-v) = -\psi(0) < 0$. Repeating the argument we can show that $\psi\big(2(x-v)\big) = \psi(0)$ and generally $\psi\big(m(x-v)\big) = (-1)^m \psi(0)$ for all $m \in \mathbb{N}$. Since $\psi \in L^2(\mathbb{R}^d)$ we need ψ^2 to be integrable on $\mathbb{R}^d$. Summarizing, we showed that a non-negative, integrable, and continuous function takes the same strictly positive value $\psi^2(0) > 0$ infinitely many times, leading to a contradiction. Arguing similarly for $\lambda = 1$ will result in a contradiction. This means the equality in $(\star)$ is never attained which concludes the proof. ■

The proof of Theorem 9 is carried out by simply repeating the proof of Theorem 1 but replacing (18) with the result in Lemma 15 and using $x - v := z$.

5.6 Proof of Corollary 10

The proof will be based on Theorem 9. The moment condition (15) on ν is sufficient for $\psi_\nu \in L^2(\mathbb{R}^d)$ to hold (see Remark A.3). Thus we only need to compute the expression $\|\psi_\nu\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi_\nu(y)\psi_\nu(y+z)dy$ appearing in (14). Note that

$$\int_{\mathbb{R}^d} \psi_\nu(y)\psi_\nu(y+z)dy = \int_{\mathbb{R}^d} \int_0^\infty \int_0^\infty e^{-t_1\|y\|_2^2 - t_2\|y+z\|_2^2} d\nu(t_1)d\nu(t_2)dy. \tag{33}$$

Since

$$\begin{aligned}
\int_0^\infty \int_0^\infty \int_{\mathbb{R}^d} e^{-t_1\|y\|_2^2 - t_2\|y+z\|_2^2} dy\, d\nu(t_1)d\nu(t_2) &= \int_0^\infty \int_0^\infty \left(\frac{\pi}{t_1+t_2}\right)^{d/2} e^{-\frac{t_1 t_2 \|z\|_2^2}{t_1+t_2}} d\nu(t_1)d\nu(t_2) \\
&\leq \nu([0,\infty)) \int_0^\infty \left(\frac{\pi}{t_1}\right)^{d/2} d\nu(t_1) < \infty,
\end{aligned}$$

we may apply Tonelli-Fubini theorem to switch the order of integration in (33) and get

$$\int_{\mathbb{R}^d} \psi_\nu(y)\psi_\nu(y+z)dy = \int_0^\infty\int_0^\infty \left(\frac{\pi}{t_1+t_2}\right)^{d/2} e^{-\frac{t_1t_2\|z\|_2^2}{t_1+t_2}} d\nu(t_1)d\nu(t_2).$$

Using this we get

$$\begin{aligned}\|\psi_\nu\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi_\nu(y)\psi_\nu(y+z)dy &= \int_0^\infty\int_0^\infty \left(\frac{\pi}{t_1+t_2}\right)^{d/2} \left(1 - e^{-\frac{t_1t_2\|z\|_2^2}{t_1+t_2}}\right) d\nu(t_1)d\nu(t_2) \\ &\geq \left(\frac{\pi}{2\delta_1}\right)^{d/2} \int_{\delta_0}^{\delta_1}\int_{\delta_0}^{\delta_1} \left(1 - e^{-\frac{t_1t_2\|z\|_2^2}{t_1+t_2}}\right) d\nu(t_1)d\nu(t_2).\end{aligned}$$

Since t_1 and t_2 are bounded below by $\delta_0 > 0$ we may take $\|z\|_2$ large enough so that the following will hold:

$$\|\psi_\nu\|^2_{L^2(\mathbb{R}^d)} - \int_{\mathbb{R}^d} \psi_\nu(y)\psi_\nu(y+z)dy \geq \frac{\beta^2}{2}\left(\frac{\pi}{2\delta_1}\right)^{d/2}.$$

5.7 Proof of Proposition 11

The proof is based on the following result, which provides a closed form expression for the $L^2(\mathbb{R}^d)$ distance between the embeddings of Gaussian measures.

Lemma 16 *Let θ_0 and θ_1 be KME of Gaussian measures $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ for $\mu_0, \mu_1 \in \mathbb{R}^d$ and $\sigma^2 > 0$. Suppose $k(x,y) = \psi(x-y)$, where $\psi \in L^1(\mathbb{R}^d) \cap C_b(\mathbb{R}^d)$ is positive definite and and k is characteristic. Then*

$$\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} = 2\int_{\mathbb{R}^d} \left(1 - e^{-i\langle w, \mu_0 - \mu_1\rangle}\right) \left(\psi^\wedge(w)\right)^2 e^{-\sigma^2\|w\|^2} dw. \tag{34}$$

Proof First of all, note that $\psi \in L^2(\mathbb{R}^d)$ since $\psi \in L^1(\mathbb{R}^d)$ and ψ is bounded. This shows that $\theta_0, \theta_1 \in L^2(\mathbb{R}^d)$. We will use P_0 and P_1 to denote the corresponding Gaussian distributions $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$. By definition we have

$$\begin{aligned}\langle \theta_0, \theta_1\rangle_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} k(x,y)\, dP_0(x)\right)\left(\int_{\mathbb{R}^d} k(z,y)\, dP_1(z)\right) dy \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k(x,y)k(z,y)\, dP_0(x)\, dP_1(z)\right) dy.\end{aligned}$$

Using the fact that ψ is bounded (Wendland, 2005, Theorem 6.2) we get

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \big|k(x,y)k(z,y)\big|\, dy\, dP_0(x)\, dP_1(z) &\leq \psi(0)\iiint_{\mathbb{R}^d} \big|k(x,y)\big|\, dy\, dP_0(x)\, dP_1(z) \\ &= \psi(0)\iiint_{\mathbb{R}^d} \big|\psi(x-y)\big| dy\, dP_0(x) dP_1(z) \\ &= \psi(0)\|\psi\|_{L^1(\mathbb{R}^d)} \iint_{\mathbb{R}^d} dP_0(x) dP_1(z) < \infty.\end{aligned}$$

This allows us to use Tonelli-Fubini theorem (Dudley, 2002, Theorem 4.4.5) and get

$$\begin{aligned}
\langle \theta_0, \theta_1 \rangle_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} k(x,y)k(z,y)dy \right) dP_0(x) dP_1(z) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \psi(y)\psi(y+z-x)dy \right) dP_0(x) dP_1(z) \\
&\overset{(\star)}{=} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \psi(y) \int_{\mathbb{R}^d} e^{-i\langle y+z-x,w\rangle} d\Lambda_\psi(w) dy \right) dP_0(x) dP_1(z) \\
&= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \psi(y) e^{-i\langle y+z-x,w\rangle} d\Lambda_\psi(w) dy \right) dP_0(x) dP_1(z),
\end{aligned}$$

where we used (4) in $(\star)$. Since $\psi \in L^1(\mathbb{R}^d)$ we have

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left| e^{-i\langle y,w\rangle} \psi(y) e^{-i\langle z-x,w\rangle} \right| dy\, d\Lambda_\psi(w) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\psi(y)|\, dy\, d\Lambda_\psi(w) < \infty$$

and thus we can use Tonelli-Fubini theorem to switch the order of integration:

$$\begin{aligned}
\langle \theta_0, \theta_1 \rangle_{L^2(\mathbb{R}^d)} &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \psi(y) e^{-i\langle y+z-x,w\rangle} dy\, d\Lambda_\psi(w) \right) dP_0(x) dP_1(z) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \psi^\wedge(w) e^{-i\langle z-x,w\rangle}\, d\Lambda_\psi(w) \right) dP_0(x) dP_1(z). \qquad (35)
\end{aligned}$$

Next we are going to argue that if both ψ and $\psi^\wedge$ belong to $L^1(\mathbb{R}^d)$ (the latter is true as it follows from Wendland, 2005, Corollary 6.12) then $\psi^\wedge$ is the Radon-Nikodym derivative of Λ_ψ with respect to the Lebesgue measure. To this end, since $\psi^\wedge \in L^1(\mathbb{R}^d)$, Fourier inversion theorem (Wendland, 2005, Corollary 5.24) yields that for all $x \in \mathbb{R}^d$, the following holds:

$$\psi(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle w,x\rangle} \psi^\wedge(w) dw.$$

On the other hand, using (4) and Lemma C.2, we also have

$$\psi(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle w,x\rangle} d\Lambda_\psi(w).$$

These two identities show that for all $x \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} e^{i\langle w,x\rangle} \psi^\wedge(w) dw = \int_{\mathbb{R}^d} e^{i\langle w,x\rangle} d\Lambda_\psi(w). \qquad (36)$$

Note that since k is translation invariant and characteristic, ψ is a strictly positive definite function (Sriperumbudur et al., 2010, Section 3.4) and therefore it follows from (Wendland, 2005, Theorem 6.11) that $\psi^\wedge$ is non-negative (and nonvanishing). Since $\psi^\wedge \in L^1(\mathbb{R}^d)$ we conclude that $\psi^\wedge$ is the Radon-Nikodym derivative of a finite non-negative measure $\mathcal{T}_\psi$ on $\mathbb{R}^d$, which is absolutely continuous with respect to the Lebesgue measure. (36) (after proper normalization) shows that the characteristic functions of measures Λ_ψ and $\mathcal{T}_\psi$ coincide. We finally conclude from (Dudley, 2002, Theorem 9.5.1) that $\Lambda_\psi = \mathcal{T}_\psi$, which means that Λ_ψ is absolutely continuous with respect to the Lebesgue measure and has a density $\psi^\wedge$.

Returning to (35) we can write it as

$$\langle \theta_0, \theta_1 \rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \left(\psi^\wedge(w)\right)^2 e^{-i\langle z-x, w\rangle}\, dw \right) dP_0(x) dP_1(z).$$

We already showed that $\psi \in L^2(\mathbb{R}^d)$. From Plancherel's theorem (Wendland, 2005, Corollary 5.25), we have $\|\psi\|_{L^2(\mathbb{R}^d)} = \|\psi^\wedge\|_{L^2(\mathbb{R}^d)}$ and thus $\psi^\wedge \in L^2(\mathbb{R}^d)$. Another application of Tonelli-Fubini theorem yields

$$\begin{aligned}
\langle \theta_0, \theta_1 \rangle_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \left(\psi^\wedge(w)\right)^2 \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-i\langle z-x, w\rangle} dP_0(x) dP_1(z) \right) dw \\
&= \int_{\mathbb{R}^d} \left(\psi^\wedge(w)\right)^2 e^{i\langle w, \mu_0 - \mu_1\rangle} e^{-\sigma^2 \|w\|_2^2} dw.
\end{aligned}$$

Noticing that the Fourier transform of a real and even function is also even, we conclude that $\psi^\wedge$ is also even. This finishes the proof since $\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} = \|\theta_1\|^2_{L^2(\mathbb{R}^d)} + \|\theta_0\|^2_{L^2(\mathbb{R}^d)} - 2\langle \theta_0, \theta_1 \rangle_{L^2(\mathbb{R}^d)}$. ■

Now we turn to the proof of Theorem 11. We will write $\tilde{\Lambda}_\psi$ to denote a non-negative finite measure, absolutely continuous with respect to the Lebesgue measure with density $(2\pi)^{d/2}(\psi^\wedge)^2$. Then (34) in Lemma 16 can be written as

$$\begin{aligned}
\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} &= \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(1 - e^{i\langle w, \mu_0 - \mu_1\rangle}\right) e^{-\sigma^2 \|w\|_2^2} d\tilde{\Lambda}_\psi(w). \\
&= \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left(1 - \cos(\langle w, \mu_0 - \mu_1\rangle)\right) e^{-\sigma^2 \|w\|_2^2} d\tilde{\Lambda}_\psi(w),
\end{aligned}$$

which is exactly of the form in Lemma 14 but with Λ_ψ replaced by $\tilde{\Lambda}_\psi$. From the proof of Lemma 16, since $\psi^\wedge$ is even and non-vanishing, the corresponding measure $\tilde{\Lambda}_\psi$ is symmetric and $\mathrm{supp}(\tilde{\Lambda}_\psi) = \mathbb{R}^d$. The result therefore follows by carrying out the proof of Proposition 3 verbatim but for replacing Λ_ψ with $\tilde{\Lambda}_\psi$.

5.8 Proof of Theorem 13

The proof will closely follow that of Theorem 8. Let $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ be two d-dimensional Gaussian distributions with mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ and variance $\sigma^2 > 0$. Let θ_0 and θ_1 denote the kernel mean embeddings of $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ respectively.

(A) Deriving a closed form expression for $\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)}$.

The condition in (15) ensures that $\psi_\nu \in L^2(\mathbb{R}^d)$ (see Remark A.3). In fact, using a similar argument it can be shown that $\psi_\nu \in L^1(\mathbb{R}^d)$. Also it is easy to verify that $\psi_\nu \in C_b(\mathbb{R}^d)$. Next, under this moment condition we may apply Tonelli-Fubini theorem to compute the Fourier transform of ψ_ν:

$$\psi_\nu^\wedge(w) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle w, x\rangle} \int_0^\infty e^{-t\|x\|_2^2} d\nu(t) dx = \int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t). \tag{37}$$

It is immediate to see that $\psi_\nu^\wedge \in L_1(\mathbb{R}^d)$. Therefore Lemma 16 yields

$$\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} = 2\int_{\mathbb{R}^d} \left(1 - e^{-i\langle w, \mu_0 - \mu_1\rangle}\right) \left(\psi_\nu^\wedge(w)\right)^2 e^{-\sigma^2\|w\|_2^2} dw. \tag{38}$$

Denoting $G(w) := \psi_\nu^\wedge(w) e^{-\frac{\sigma^2\|w\|_2^2}{2}}$ and using a well-known property of the Fourier transform we get

$$(G^2)^\wedge(\tau) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle w,\tau\rangle} G^2(w) dw = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} G^\wedge(x) G^\wedge(\tau - x) dx. \tag{39}$$

Next we compute the Fourier transform of G using (37):

$$G^\wedge(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle w,x\rangle} \left(\int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t)\right) e^{-\frac{\sigma^2\|w\|_2^2}{2}} dw.$$

Using the moment condition on ν we have

$$\begin{aligned}\int_0^\infty \frac{1}{(4\pi t)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{\|x\|_2^2}{4t}} e^{-\frac{\sigma^2\|x\|_2^2}{2}} dx d\nu(t) &= \frac{1}{(2\sigma^2)^{d/2}} \int_0^\infty \left(t + \frac{1}{2\sigma^2}\right)^{-d/2} d\nu(t) \\ &\le \frac{1}{(2\sigma^2)^{d/2}} \int_0^\infty t^{-d/2} d\nu(t) < \infty. \end{aligned} \tag{40}$$

This allows us to use Tonelli-Fubini theorem and write

$$\begin{aligned} G^\wedge(x) &= \frac{1}{(2\pi)^{d/2}} \int_0^\infty \frac{1}{(2t)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle w,x\rangle} e^{-\frac{\|w\|_2^2}{4t}} e^{-\frac{\sigma^2\|w\|_2^2}{2}} dw\, d\nu(t) \\ &= \int_0^\infty \frac{1}{(2t\sigma^2 + 1)^{d/2}} \exp\left(-\frac{t\|x\|_2^2}{2t\sigma^2 + 1}\right) d\nu(t). \end{aligned}$$

Returning to (39) and denoting $\Delta_1 := 2t_1\sigma^2 + 1$, $\Delta_2 := 2t_2\sigma^2 + 1$ we obtain

$$(G^2)^\wedge(\tau) = \int_{\mathbb{R}^d} \int_0^\infty \int_0^\infty \frac{1}{(2\pi\Delta_1\Delta_2)^{d/2}} \exp\left(-\frac{t_1\|x\|_2^2}{\Delta_1} - \frac{t_2\|\tau - x\|_2^2}{\Delta_2}\right) d\nu(t_1)\, d\nu(t_2)\, dx.$$

Using a simple identity

$$a\|x\|_2^2 + b\|x - y\|_2^2 = (a+b)\left\|x - \frac{b}{a+b} y\right\|_2^2 + \frac{ab}{a+b}\|y\|_2^2,$$

which holds for any $x, y \in \mathbb{R}^d$ and $a, b \in \mathbb{R}$ with $a + b \neq 0$, we obtain

$$\begin{aligned}(G^2)^\wedge(\tau) = \int_{\mathbb{R}^d} \int_0^\infty \int_0^\infty \frac{1}{(2\pi\Delta_1\Delta_2)^{d/2}} \exp\left(-\left(\frac{t_1}{\Delta_1} + \frac{t_2}{\Delta_2}\right)\left\|x - \frac{t_2\Delta_1\tau}{t_1\Delta_2 + t_2\Delta_1}\right\|_2^2\right) \\ \times \exp\left(-\frac{t_1 t_2\|\tau\|_2^2}{t_1\Delta_2 + t_2\Delta_1}\right) d\nu(t_1)\, d\nu(t_2)\, dx.\end{aligned}$$

Using an argument similar to (40) we can show that Tonelli-Fubini theorem is applicable to the r.h.s. of the above equation. Therefore, changing the order of integration we get

$$(G^2)^\wedge(\tau) = \int_0^\infty \int_0^\infty \left(\frac{1}{2(t_1\Delta_2 + t_2\Delta_1)}\right)^{d/2} \exp\left(-\frac{t_1 t_2 \|\tau\|_2^2}{t_1\Delta_2 + t_2\Delta_1}\right) d\nu(t_1)\, d\nu(t_2).$$

Noticing that $(G^2)^\wedge(0) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} G^2(w)\, dw$ and returning to (38) we get

$$\begin{aligned}
&\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} = 2(2\pi)^{d/2} \left((G^2)^\wedge(0) - (G^2)^\wedge(\mu_0 - \mu_1)\right) \\
&= 2\int_0^\infty \int_0^\infty \left(\frac{\pi}{t_1\Delta_2 + t_2\Delta_1}\right)^{d/2} \left(1 - \exp\left(-\frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{t_1\Delta_2 + t_2\Delta_1}\right)\right) d\nu(t_1) d\nu(t_2).
\end{aligned}$$

(B) Lower bounding $\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)}$ in terms of $\|\mu_0 - \mu_1\|_2^2$.

Consider

$$\begin{aligned}
&\|\theta_0 - \theta_1\|^2_{L_2(\mathbb{R}^d)} \\
&\geq 2\int_{\delta_0}^{\delta_1} \int_{\delta_0}^{\delta_1} \left(\frac{\pi}{t_1\Delta_2 + t_2\Delta_1}\right)^{d/2} \left(1 - \exp\left(-\frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{t_1\Delta_2 + t_2\Delta_1}\right)\right) d\nu(t_1) d\nu(t_2) \\
&= 2\int_{\delta_0}^{\delta_1} \int_{\delta_0}^{\delta_1} \left(\frac{\pi}{4t_1 t_2\sigma^2 + t_1 + t_2}\right)^{d/2} \left(1 - \exp\left(-\frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{4t_1 t_2\sigma^2 + t_1 + t_2}\right)\right) d\nu(t_1) d\nu(t_2).
\end{aligned}$$

Using the fact that $1 - e^{-x} \geq \frac{x}{2}$ for $0 \leq x \leq 1$, we obtain

$$\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} \geq \int_{\delta_0}^{\delta_1} \int_{\delta_0}^{\delta_1} \left(\frac{\pi}{4t_1 t_2\sigma^2 + t_1 + t_2}\right)^{d/2} \frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{4t_1 t_2\sigma^2 + t_1 + t_2}\, d\nu(t_1)\, d\nu(t_2) \tag{41}$$

whenever

$$\frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{4t_1 t_2\sigma^2 + t_1 + t_2} \leq 1.$$

Note that the expression on the left hand side of the previous inequality is increasing both in t_1 and t_2. This means that for $t_1, t_2 \in [\delta_0, \delta_1]$ we have:

$$\frac{t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{4t_1 t_2\sigma^2 + t_1 + t_2} \leq \frac{\delta_1 \|\mu_0 - \mu_1\|_2^2}{4\delta_1\sigma^2 + 2}$$

and thus (41) holds whenever

$$\delta_1 \|\mu_0 - \mu_1\|_2^2 \leq 4\delta_1\sigma^2 + 2$$

which will be satisfied later. (41) can be rewritten as

$$\begin{aligned}
\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} &\geq \int_{\delta_0}^{\delta_1} \int_{\delta_0}^{\delta_1} \frac{(\pi)^{d/2} t_1 t_2 \|\mu_0 - \mu_1\|_2^2}{(4t_1 t_2\sigma^2 + t_1 + t_2)^{d/2+1}}\, d\nu(t_1)\, d\nu(t_2) \\
&= \int_{\delta_0}^{\delta_1} \left(\frac{\pi}{t_1 + t_2}\right)^{d/2} \int_{\delta_0}^{\delta_1} \frac{\frac{t_1 t_2}{t_1 + t_2} \|\mu_0 - \mu_1\|_2^2}{\left(4\frac{t_1 t_2}{t_1 + t_2}\sigma^2 + 1\right)^{d/2+1}}\, d\nu(t_1)\, d\nu(t_2) \\
&\geq \left(\frac{\pi}{2\delta_1}\right)^{d/2} \int_{\delta_0}^{\delta_1} \int_{\delta_0}^{\delta_1} \frac{S(t_1, t_2) \|\mu_0 - \mu_1\|_2^2}{\left(4S(t_1, t_2)\sigma^2 + 1\right)^{d/2+1}}\, d\nu(t_1)\, d\nu(t_2),
\end{aligned} \tag{42}$$

where $S(t_1, t_2) := \frac{t_1 t_2}{t_1 + t_2}$. Note that $S(t_1, t_2)$ takes values in $[\frac{\delta_0}{2}, \frac{\delta_1}{2}]$ as t_1 and t_2 varies in $[\delta_0, \delta_1]$. We can now repeat part of the proof of Theorem 8 where we showed that the function $t \mapsto \frac{t}{(1+4t\sigma^2)^{(d+2)/2}}$ monotonically increases on $[0, \frac{1}{2d\sigma^2}]$, reaches its global maximum at $t = \frac{1}{2d\sigma^2}$, and then decreases on $[\frac{1}{2d\sigma^2}, \infty)$. Using this fact we have

$$\int_{\delta_0}^{\delta_1}\int_{\delta_0}^{\delta_1} \frac{S(t_1,t_2)}{\left(4S(t_1,t_2)\sigma^2+1\right)^{\frac{d}{2}+1}} d\nu(t_1)d\nu(t_2) \geq \frac{\beta^2}{2} \min\left\{ \frac{\delta_0}{(1+2\delta_0\sigma^2)^{\frac{d}{2}+1}}, \frac{\delta_1}{(1+2\delta_1\sigma^2)^{\frac{d}{2}+1}} \right\}.$$

By setting $\sigma^2 := \frac{1}{\delta_1 d}$ we ensure that $t = \frac{\delta_1}{2}$ is the global maximum of the function $t \mapsto \frac{t}{\left(1+\frac{4t}{\delta_1 d}\right)^{\frac{d}{2}+1}}$ and thus $\frac{\delta_0}{(1+2\delta_0\sigma^2)^{\frac{d}{2}+1}} \leq \frac{\delta_1}{(1+2\delta_1\sigma^2)^{\frac{d}{2}+1}}$. Combining this with (42) we have

$$\begin{aligned}
\|\theta_0 - \theta_1\|^2_{L^2(\mathbb{R}^d)} &\geq \frac{\beta^2\delta_0}{2\left(1+\frac{2\delta_0}{\delta_1 d}\right)^{\frac{d}{2}+1}} \left(\frac{\pi}{2\delta_1}\right)^{d/2} \|\mu_0 - \mu_1\|_2^2 \\
&\geq \frac{\beta^2\delta_0}{2(1+\frac{2}{d})^{\frac{d}{2}+1}} \left(\frac{\pi}{2\delta_1}\right)^{d/2} \|\mu_0 - \mu_1\|_2^2 \geq \frac{\beta^2\delta_0}{2e}\left(1 - \frac{2}{2+d}\right)\left(\frac{\pi}{2\delta_1}\right)^{d/2} \|\mu_0 - \mu_1\|_2^2,
\end{aligned}$$

where we used an analysis similar to (29).

(C) Application of Theorem B.2.

We finish by repeating all the remaining steps carried out in the proof of Theorem 8 (steps C.1, C.2, and C.3), where we set

$$\sigma^2 := \frac{1}{\delta_1 d} \quad \text{and} \quad c_\nu := \frac{1}{16\delta_1}\left(\log N - \frac{1}{N-1}\right).$$

Acknowledgements

The authors thank the action editor and two anonymous reviewers for their detailed comments, which helped to improve the presentation. The authors would also like to thank David Lopez-Paz, Jonas Peters, Bernhard Schölkopf, and Carl-Johann Simon-Gabriel for useful discussions.

Appendix A. $\sqrt{n}$-consistency of $\mu_k(P_n)$

In the following, we present a general result whose special cases establishes the convergence rate of $n^{-1/2}$ for $\|\mu_k(P_n) - \mu_k(P)\|_{\mathcal{F}}$ when $\mathcal{F} = \mathcal{H}_k$ and $\mathcal{F} = L^2(\mathbb{R}^d)$.

Proposition A.1 *Let $(X_i)_{i=1}^n$ be random samples drawn i.i.d. from P defined on a separable topological space $\mathcal{X}$. Suppose $r : \mathcal{X} \to H$ is continuous and*

$$\sup_{x\in\mathcal{X}} \|r(x)\|_H^2 \leq C_k < \infty, \tag{43}$$

where H is a separable Hilbert space of real-valued functions. Then for any $0 < \delta \leq 1$ with probability at least $1 - \delta$ we have

$$\left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H \leq \sqrt{\frac{C_k}{n}} + \sqrt{\frac{2C_k \log(1/\delta)}{n}}.$$

Proof Note that $r : \mathcal{X} \to H$ is a H-valued measurable function as r is continuous and H is separable (Steinwart and Christmann, 2008, Lemma A.5.18). The condition in (43) ensures that $\int \|r(x)\|_H dQ(x) \leq \sqrt{C_k} < \infty$ for any $Q \in M_+^1(\mathcal{X})$ and therefore $\int r(x)\, dQ(x)$ is well defined as a Bochner integral for any $Q \in M_+^1(\mathcal{X})$ (Diestel and Uhl, 1977, Theorem 2, p.45). By McDiarmid's inequality, it is easy to verify that with probability at least $1 - \delta$,

$$\left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H \leq \mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H + \sqrt{\frac{2C_k \log(1/\delta)}{n}}, \tag{44}$$

where

$$\begin{aligned}
&\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H \leq \sqrt{\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H^2} \\
&= \sqrt{\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) \right\|_H^2 + \left\| \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H^2 - 2\mathbb{E} \left\langle \int_{\mathcal{X}} r(x)\, dP_n(x), \int_{\mathcal{X}} r(x)\, dP(x) \right\rangle_H} \\
&= \sqrt{\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) \right\|_H^2 + \left\| \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H^2 - \frac{2}{n} \sum_{i=1}^{n} \mathbb{E} \left\langle r(X_i), \int_{\mathcal{X}} r(x)\, dP(x) \right\rangle_H}. \qquad (45)
\end{aligned}$$

To simplify the r.h.s. of (45), we make the following observation. Note that for any $g \in H$, $T_g : H \to \mathbb{R}$, $f \mapsto \langle g, f \rangle_H$ is a bounded linear functional on H. Choose $f = \int_{\mathcal{X}} r(y)\, dP(y)$. It follows from (Diestel and Uhl, 1977, Theorem 6, p.47) that

$$\left\langle g, \int_{\mathcal{X}} r(y)\, dP(y) \right\rangle_H = T_g \left(\int_{\mathcal{X}} r(y)\, dP(y) \right) = \int_{\mathcal{X}} T_g(r(y))\, dP(y) = \int_{\mathcal{X}} \langle g, r(y) \rangle_H\, dP(y). \tag{46}$$

Applying (46) to the third term in the r.h.s. of (45) with $g = \int_{\mathcal{X}} r(x)\, dP(x)$, we obtain

$$\mathbb{E} \left\langle r(X_i), \int_{\mathcal{X}} r(x)\, dP(x) \right\rangle_H = \int_{\mathcal{X}} \langle r(x_i), g \rangle_H\, dP(x_i) = \left\langle \int_{\mathcal{X}} r(x_i)\, dP(x_i), g \right\rangle_H = \|g\|_H^2$$

and so (45) reduces to

$$\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) - \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H \leq \sqrt{\mathbb{E} \left\| \int_{\mathcal{X}} r(x)\, dP_n(x) \right\|_H^2 - \left\| \int_{\mathcal{X}} r(x)\, dP(x) \right\|_H^2}. \tag{47}$$

Consider

$$\begin{aligned}\mathbb{E}\left\|\int_{\mathcal{X}} r(x)\,dP_n(x)\right\|_H^2 &= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n r(X_i)\right\|_H^2 = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}\langle r(X_i), r(X_j)\rangle_H \\ &= \frac{1}{n^2}\sum_{i=j}\mathbb{E}\langle r(X_i), r(X_j)\rangle_H + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\langle r(X_i), r(X_j)\rangle_H \\ &= \frac{1}{n}\mathbb{E}_{X\sim P}\|r(X)\|_H^2 + \frac{n-1}{n}\mathbb{E}_{X\sim P, Y\sim P}\langle r(X), r(Y)\rangle_H. \end{aligned} \tag{48}$$

Using (46), the second term in (48) can be equivalently written as

$$\begin{aligned}\mathbb{E}_{X\sim P, Y\sim P}\langle r(X), r(Y)\rangle_H &= \int_{\mathcal{X}}\left(\int_{\mathcal{X}}\langle r(x), r(y)\rangle_H\,dP(y)\right)dP(x) \\ &\overset{(\star)}{=} \int_{\mathcal{X}}\left\langle r(x), \int_{\mathcal{X}} r(y)\,dP(y)\right\rangle_H dP(x) \\ &\overset{(\star)}{=} \left\langle \int_{\mathcal{X}} r(x)\,dP(x), \int_{\mathcal{X}} r(y)\,dP(y)\right\rangle_H x = \left\|\int_{\mathcal{X}} r(x)\,dP(x)\right\|_H^2,\end{aligned}$$

where we invoked (46) in $(\star)$. Combining the above with (48) and using the result in (47) yields

$$\mathbb{E}\left\|\int_{\mathcal{X}} r(x)\,dP_n(x) - \int_{\mathcal{X}} r(x)\,dP(x)\right\|_H \leq \sqrt{\frac{\mathbb{E}_{X\sim P}\|r(X)\|_H^2 - \|\int_{\mathcal{X}} r(x)\,dP(x)\|_H^2}{n}} \leq \sqrt{\frac{C_k}{n}}$$

and the result follows. ∎

Remark A.2 *Suppose H is an RKHS with a reproducing kernel k that is continuous and satisfies $\sup_{x\in\mathcal{X}} k(x,x) < \infty$. Choosing $r(x) = k(\cdot, x)$, $x \in \mathcal{X}$ in Proposition A.1 yields a concentration inequality for $\|\mu_k(P_n) - \mu_k(P)\|_H$ with $C_k := \sup_{x\in\mathcal{X}} k(x,x)$, thereby establishing a convergence rate of $n^{-1/2}$ for $\|\mu_k(P_n) - \mu_k(P)\|_H$. While such a result has already appeared in Smola et al. (2007, Theorem 2), Gretton et al. (2012) and Lopez-Paz et al. (2015), the result derived from Proposition A.1 improves upon them by providing better constants. While all these works including Proposition A.1 are based on McDiarmid's inequality (see (44)), the latter obtains better constants by carefully bounding the expectation term in (44). It is easy to verify that $C_k = 1$ for Gaussian and $C_k = C_M$ for mixture of Gaussian kernels, $C_k = c^{-2\gamma}$ for inverse multiquadrics, and $C_k = 1$ for Matérn kernels.*

Remark A.3 *Assuming $\mathcal{X} = \mathbb{R}^d$, $H = L^2(\mathbb{R}^d)$ and $r(x) = k(\cdot, x)$, $x \in \mathbb{R}^d$, where k is a continuous positive definite kernel on $\mathbb{R}^d$, Proposition A.1 establishes a convergence rate of $n^{-1/2}$ for $\|\mu_k(P_n) - \mu_k(P)\|_{L^2(\mathbb{R}^d)}$ under the condition that $\sup_{x\in\mathbb{R}^d}\|k(x,\cdot)\|_{L^2(\mathbb{R}^d)}^2 < \infty$. If k is translation invariant on $\mathbb{R}^d$, i.e., $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ where $\psi \in C(\mathbb{R}^d)$ is positive definite, then $\psi \in L^2(\mathbb{R}^d)$ ensures that $\sup_{x\in\mathbb{R}^d}\|k(x,\cdot)\|_{L^2(\mathbb{R}^d)}^2 = \sup_{x\in\mathbb{R}^d}\|\psi(x-\cdot)\|_{L^2(\mathbb{R}^d)}^2 = \|\psi\|_{L^2(\mathbb{R}^d)}^2$ and therefore Propositions A.1 holds with $C_k := \|\psi\|_{L^2(\mathbb{R}^d)}^2$. On the other hand, for radial kernels on $\mathbb{R}^d$, i.e., kernels of the form in (5), the condition in (43) is ensured if*

$$\int_0^\infty t^{-d/2} d\nu(t) < \infty \tag{49}$$

since

$$\begin{aligned}\sup_{x\in\mathbb{R}^d}\|k(x,\cdot)\|^2_{L^2(\mathbb{R}^d)} &= \sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\int_0^\infty e^{-t\|x-y\|^2}d\nu(t)\right)^2 dy\\ &\overset{(\dagger)}{\leq} \nu([0,\infty))\sup_{x\in\mathbb{R}^d}\int_{\mathbb{R}^d}\int_0^\infty e^{-2t\|x-y\|^2}d\nu(t)dy\\ &\overset{(\ddagger)}{=} \nu([0,\infty))\sup_{x\in\mathbb{R}^d}\int_0^\infty\int_{\mathbb{R}^d} e^{-2t\|x-y\|^2}dy\,d\nu(t) = \frac{\nu([0,\infty))}{(2/\pi)^{d/2}}\int_0^\infty\frac{d\nu(t)}{t^{d/2}},\end{aligned}$$

where we used Jensen's inequality in (†) *and Fubini's theorem in* (‡). *Therefore the bound in Proposition A.1 holds with* $C_k := \nu([0,\infty))\int_0^\infty\left(\frac{\pi}{2t}\right)^{d/2}d\nu(t)$. *(49) is satisfied by Gaussian, mixture of Gaussian and Matérn (see Sriperumbudur, 2016, Equation 6.17) kernels. For inverse multiquadrics, while (49) holds for* $\gamma > d/2$ *since* $\nu = c^{-2\gamma}Gamma(\gamma, c^2)$ *(see Wendland, 2005, Theorem 7.15), in fact the condition in (43) holds for* $\gamma > d/4$ *(see Lemma C.4).*

Appendix B. Minimax Lower Bounds and Le Cam's Method

Let Θ be a set of parameters (or functions) containing the element θ which we want to estimate. Assume there is a class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures on $\mathbb{R}^d$ indexed by Θ. Suppose $d\colon \Theta\times\Theta\to[0,\infty)$ is a metric on Θ. Le Cam's method provides a lower bound on the minimax probability, $\inf_{\hat\theta_n}\sup_{\theta\in\Theta}P^n_\theta(d(\hat\theta_n,\theta)\geq s)$ for $s>0$, where the infimum is taken over all possible estimators $\hat\theta_n\colon\mathbb{R}^d\to\Theta$ that are constructed from an i.i.d. sample $(X_i)_{i=1}^n$ drawn from P_θ. The following two results which we used throughout this work are based on Le Cam's method and they provide a lower bound on the minimax probability. The first one follows from Theorem 2.2 and Equation (2.9) of Tsybakov (2008). It requires a construction of two sufficiently distant elements of the set Θ corresponding to the probability distributions similar in the Kullback-Leibler (KL) divergence sense, where the KL divergence between two distributions P and Q with P absolutely continuous w.r.t. Q is defined as $\mathrm{KL}(P\|Q) = \int\log\frac{dP}{dQ}\,dP$.

Theorem B.1 (Lower bound based on two hypotheses) *Assume* Θ *contains* θ_0 *and* θ_1 *such that* $d(\theta_0,\theta_1)\geq 2s$ *and* $\mathrm{KL}(P^n_{\theta_0}\|P^n_{\theta_1})\leq\alpha$ *for some* $s>0$ *and* $0<\alpha<\infty$. *Then*

$$\inf_{\hat\theta_n}\sup_{\theta\in\Theta}P^n_\theta\left\{d(\hat\theta_n,\theta)\geq s\right\}\geq\max\left(\frac{1}{4}e^{-\alpha},\frac{1-\sqrt{\alpha/2}}{2}\right).$$

Note that the second condition of the theorem bounds the distance between the n-fold product distributions by a constant independent of n. Recalling the chain rule of the KL-divergence, which states that $\mathrm{KL}(P^n_{\theta_0}\|P^n_{\theta_1}) = n\cdot\mathrm{KL}(P_{\theta_0}\|P_{\theta_1})$, we can see that this condition is rather restrictive and requires the marginal distributions to satisfy $\mathrm{KL}(P_{\theta_0}\|P_{\theta_1}) = O(n^{-1})$. This condition is slightly relaxed in the following result, which follows from Theorem 2.5 of Tsybakov (2008).

Theorem B.2 (Lower bound based on many hypotheses) *Assume $M \geq 2$ and suppose that there exist $\theta_0, \ldots, \theta_M \in \Theta$ such that (i) $d(\theta_i, \theta_j) \geq 2s > 0$, $\forall\, 0 \leq i < j \leq M$; (ii) P_{θ_j} is absolutely continuous w. r. t. P_{θ_0} for all $j = 1, \ldots, M$, and $\frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P^n_{\theta_j} \| P^n_{\theta_0}) \leq \alpha \log M$ with $0 < \alpha < 1/8$. Then*

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P^n_\theta \left\{ d(\hat{\theta}_n, \theta) \geq s \right\} \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \right) > 0.$$

The above result is commonly used with M tending to infinity as $n \to \infty$. In this case the second condition on the KL-divergence indeed becomes less restrictive than the one of Theorem B.1, since the upper bound $\alpha \log M$ may now grow with the sample size n. At the same time, Theorem B.2 still provides a lower bound on the minimax probability independent of n, since $\sqrt{M}/(1+\sqrt{M})$ and $\log M$ can be lower bounded by $1/2$ and $\log 2$ respectively.

Appendix C. Technical Lemmas

The following technical results are used to prove the main results of Sections 3 and 4.

Lemma C.1 (Theorem 5.18, Wendland, 2005) *For any $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$ the following holds:*

$$\left[\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-\mu\|_2^2}{2\sigma^2}} \right]^\wedge (w) = \frac{1}{(2\pi)^{d/2}} \exp\left(-i\langle \mu, w \rangle - \frac{\sigma^2 \|w\|_2^2}{2} \right), \quad w \in \mathbb{R}^d.$$

Lemma C.2 *Let $\psi \colon \mathbb{R}^d \to \mathbb{R}$ be a symmetric and positive definite function. Let Λ_ψ be the corresponding finite non-negative Borel measure from (4). Then Λ_ψ is symmetric, i.e., $\Lambda_\psi(A) = \Lambda_\psi(-A)$ for all $A \subset \mathbb{R}^d$.*

Proof From the definition of Λ_ψ we know that it is finite, non-negative, and

$$\psi(x) = \int_{\mathbb{R}^d} e^{-i\langle w, x \rangle} \Lambda_\psi(dw) = \int_{\mathbb{R}^d} \cos(\langle w, x \rangle) \Lambda_\psi(dw) - i \cdot \int_{\mathbb{R}^d} \sin(\langle w, x \rangle) \Lambda_\psi(dw).$$

Since $\psi(-x) = \psi(x)$ for all $x \in \mathbb{R}^d$, we get $\int_{\mathbb{R}^d} \sin(\langle w, x \rangle) \Lambda_\psi(dw) = 0$. Note that $\psi(-x)$ is by definition a characteristic function of measure Λ_ψ, and we have just proved that it is real-valued. It is known (Bogachev, 2007, Corollary 3.8.7) that in this case the measure Λ_ψ is invariant under the mapping $x \to -x$. ∎

Lemma C.3 *Assume $X, Y \subseteq \mathbb{R}^d$. If $f \colon X \times Y \to \mathbb{R}$ is a continuous function and Y is a compact set then $g(x) := \inf_{y \in Y} f(x, y)$ is continuous.*

Proof First, the map $g \colon X \to \mathbb{R}$ is well defined since $f_x(y) := f(x, y)$ is a continuous function for any $x \in X$ and thus f_x achieves its infimum since Y is a compact set. We will prove that the map $g \colon X \to \mathbb{R}$ is continuous by showing that $g^{-1}(-\infty, a)$ and $g^{-1}(a, \infty)$ are open sets for all $a \in \mathbb{R}$ (Dudley, 2002, Corollary 2.2.7 (a)).

Now we will show that $g^{-1}(-\infty, a)$ is open for any $a \in \mathbb{R}$. It suffices to show that for any $x \in g^{-1}(-\infty, a)$ there is an open neighborhood U_x of x which also belongs to $g^{-1}(-\infty, a)$. The set $g^{-1}(-\infty, a)$ consists of elements $x \in X$ for which $g(x) < a$. In other words, it consists of such elements $x \in X$ for which there is corresponding $y_x \in Y$ satisfying $f(x, y_x) < a$. Take any $x \in g^{-1}(-\infty, a)$. Since f is continuous, $f^{-1}(-\infty, a)$ is open and contains (x, y_x). Moreover $f^{-1}(-\infty, a)$ contains $U_x \times V_y$, where U_x and V_y are open sets with $x \in U_x$ and $y_x \in V_y$. Now suppose $x' \in U_x$. Then for any $y \in V_y$ we have $f(x', y) < a$. In particular, $f(x', y_x) < a$, which means that $g(x') < a$ and $x' \in g^{-1}(-\infty, a)$. This shows that $g^{-1}(-\infty, a)$ is open.

Next we will show that $g^{-1}(a, \infty)$ is also an open set for any $a \in \mathbb{R}$. Assume this is not the case. Then there is $x \in g^{-1}(a, \infty)$ such that for any neighborhood U_x of x there is a point $x' \in U_x$ such that $x' \notin g^{-1}(a, \infty)$. This means that for any such x' there is $y_{x'}$ satisfying $f(x', y_{x'}) \leq a$. Using this we can construct a sequence $\{x_n, y_n\}$ from $X \times Y$, such that $x_n \notin g^{-1}(a, \infty)$ for every n, $\lim_{n\to\infty} x_n = x$ and for any n it holds that $f(x_n, y_n) \leq a$. Since Y is compact we conclude that $\{y_n\}$ has a converging subsequence $\{y_{n(k)}\}$ (Dudley, 2002, Theorem 2.3.1) with limit $y^* \in Y$. We just showed that there is a sequence $\{x_{n(k)}, y_{n(k)}\}$ in $X \times Y$, which converges to (x, y^*), such that $\lim_{k\to\infty} f(x_{n(k)}, y_{n(k)}) \leq a$. Since f is continuous, this also means that $\lim_{k\to\infty} f(x_{n(k)}, y_{n(k)}) = f(x, y^*) \leq a$. This means that $\inf_{y\in Y} f(x, y) \leq f(x, y^*) \leq a$. In other words, this shows that $x \notin g^{-1}(a, \infty)$ leading to a contradiction and therefore $g^{-1}(a, \infty)$ is open. ■

Lemma C.4 (L^2 norm of inverse multiquadrics kernels) *For any $c > 0$ and $\gamma > \frac{d}{4}$,*

$$\int_{\mathbb{R}^d} (c^2 + \|x\|_2^2)^{-2\gamma}\, dx = c^{d-4\gamma}\pi^{d/2}\frac{\Gamma(2\gamma - \frac{d}{2})}{\Gamma(2\gamma)}.$$

Proof

$$\begin{aligned}
\int_{\mathbb{R}^d} (c^2 + \|x\|_2^2)^{-2\gamma} dx &= c^{-4\gamma} \int_{\mathbb{R}^d} \left(1 + \left\|\frac{x}{c}\right\|_2^2\right)^{-2\gamma} dx = c^{d-4\gamma} \int_{\mathbb{R}^d} \left(1 + \|x\|_2^2\right)^{-2\gamma} dx \\
&= c^{d-4\gamma} \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_0^\infty \left(1 + r^2\right)^{-2\gamma} r^{d-1} dr \\
&= c^{d-4\gamma} \frac{\pi^{d/2}}{\Gamma(d/2)} \int_0^\infty (1 + x)^{-2\gamma} x^{d/2-1} dx \\
&= c^{d-4\gamma} \frac{\pi^{d/2}}{\Gamma(d/2)} \frac{\Gamma(d/2)\Gamma(2\gamma - d/2)}{\Gamma(2\gamma)},
\end{aligned}$$

where last identity can be found in (Gradshteyn and Ryzhik, 2000, 3.194.3). ■

Appendix D. Bounds on Constants for Various Radial Kernels

In this appendix, we present bounds on the constants that appear in Corollaries 2, 10 and Theorems 8, 13. for various radial kernels.

D.1 α in Corollary 2

In Corollary 2, we assumed that there exist $0 < t_1 < \infty$ and $\alpha > 0$ such that $\nu([t_1, \infty)) \geq \alpha$. In the following, we present the values of t_1 and α for various radial kernels.

(i) Gaussian kernel: $\nu = \delta_{\frac{1}{2\eta^2}}$ and so for any $t_1 < \frac{1}{2\eta^2}$, we obtain $\alpha = 1$.

(ii) Mixture of Gaussians: $\nu = \sum_{i=1}^{M} \beta_i \delta_{\frac{1}{2\eta_i^2}}$ and so $\alpha = C_M$ for any $t_1 < \frac{1}{2\eta_1^2}$.

(iii) Inverse multiquadric kernel: It follows from (Wendland, 2005, Theorem 7.15) that

$$k(x, y) = \int_0^\infty e^{-t\|x-y\|_2^2} \frac{t^{\gamma-1} e^{-c^2 t}}{\Gamma(\gamma)} dt, \tag{50}$$

and so

$$\nu = c^{-2\gamma} \mathrm{Gamma}(\gamma, c^2) \tag{51}$$

where the density of a Gamma distribution with parameters $a, b > 0$ is defined as

$$\mathrm{Gamma}(t; a, b) = \frac{b^a}{\Gamma(a)} t^{a-1} e^{-tb}, \ t \geq 0.$$

Therefore choosing t_1 to be the median of ν, we obtain $\alpha = \frac{c^{-2\gamma}}{2}$.

(iv) Matérn kernel: We know from (Wendland, 2005, Theorem 6.13) that Matérn kernel is related to the Fourier transform of the inverse multiquadric kernel as

$$\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle v, w\rangle} \left(c^2 + \|v\|_2^2\right)^{-\tau} dv = \frac{2^{1-\tau}}{\Gamma(\tau)} K_{d/2-\tau}(c\|w\|_2) \left(\frac{\|w\|_2}{c}\right)^{\tau-d/2},$$

where $c > 0$ and $\tau > d/2$. Using this together with the representation (50) of an inverse multiquadrics kernel we obtain the following identity, which already appeared in (Sriperumbudur, 2016, Equation (72)):

$$\begin{aligned}
k(x, y) &= \frac{\Gamma(\tau) c^{2\tau-d} 2^{d/2}}{\Gamma(\tau - d/2)} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle v, x-y\rangle} \left[\int_0^\infty e^{-t\|v\|_2^2} \frac{t^{\tau-1} e^{-c^2 t}}{\Gamma(\tau)} dt \right] dv \\
&\overset{(\star)}{=} \frac{c^{2\tau-d} 2^{d/2}}{\Gamma(\tau - d/2)} \int_0^\infty t^{\tau-1} e^{-c^2 t} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle v, x-y\rangle} e^{-t\|v\|_2^2} dv \, dt \\
&= \frac{c^{2\tau-d} 2^{d/2}}{\Gamma(\tau - d/2)} \int_0^\infty t^{\tau-1} e^{-c^2 t} \frac{1}{(2t)^{d/2}} e^{-\frac{\|x-y\|_2^2}{4t}} dt \\
&= \frac{c^{2\tau-d}}{\Gamma(\tau - d/2)} \int_0^\infty t^{\tau-d/2-1} e^{-c^2 t} e^{-\frac{\|x-y\|_2^2}{4t}} dt,
\end{aligned}$$

where we invoked Tonelli-Fubini theorem (Dudley, 2002, Theorem 4.4.5) in $(\star)$ since $\left(c^2 + \|\cdot\|_2^2\right)^{-\tau} \in L_1(\mathbb{R}^d)$ for $\tau > d/2$. After change of variables we finally obtain

$$k(x, y) = \frac{1}{\Gamma\left(\tau - \frac{d}{2}\right)} \left(\frac{c^2}{4}\right)^{\tau - \frac{d}{2}} \int_0^\infty e^{-t\|x-y\|^2} t^{d/2-\tau-1} e^{-\frac{c^2}{4t}} dt,$$

which shows that Matérn kernel is a particular instance of radial kernels with

$$\nu = \text{InvGamma}\left(\tau - \frac{d}{2}, \frac{c^2}{4}\right),$$

where the density of an inverse-Gamma distribution with parameters $a, b > 0$ has the form

$$\text{InvGamma}(t; a, b) = \frac{b^a}{\Gamma(a)} t^{-a-1} e^{-b/t}.$$

Therefore $\alpha = \frac{1}{2}$ for the choice of t_1 to be the median of ν.

D.2 $\frac{\beta t_0}{t_1}$ in Theorem 8

In Theorem 8, we assumed that there exist $0 < t_0 \leq t_1 < \infty$ and $0 < \beta < \infty$ such that $\nu([t_0, t_1]) \geq \beta$. Define $B_k := \frac{\beta t_0}{t_1}$. In the following, we present the values of B_k for various radial kernels.

(i) Gaussian kernel: Choose $t_0 = t_1 = \frac{1}{2\eta^2}$ so that $\beta = 1$ and $B_k = 1$.

(ii) Mixture of Gaussians: Set $t_0 = \frac{1}{2\eta_1^2}$, $t_1 = \frac{1}{2\eta_M^2}$ so that $\beta = C_M$ implying $B_k = \frac{C_M \eta_M^2}{\eta_1^2}$.

(iii) Inverse multiquadric kernel: From (51), we have $\nu = c^{-2\gamma}\text{Gamma}(\gamma, c^2)$. Therefore

$$\begin{aligned}
\nu\left(\left[\frac{\gamma}{2c^2}, \frac{\gamma}{c^2}\right]\right) &= \frac{1}{\Gamma(\gamma)} \int_{\gamma/(2c^2)}^{\gamma/c^2} t^{\gamma-1} e^{-tc^2} dt \\
&\geq \begin{cases} \frac{1}{\Gamma(\gamma)} \left(\frac{\gamma}{2c^2}\right)^{\gamma-1} \exp\left(-\frac{\gamma}{c^2} c^2\right) \frac{\gamma}{2c^2}, & \text{for } \gamma \geq 1; \\ \frac{1}{\Gamma(\gamma)} \left(\frac{\gamma}{c^2}\right)^{\gamma-1} \exp\left(-\frac{\gamma}{c^2} c^2\right) \frac{\gamma}{2c^2}, & \text{for } \gamma \in (0, 1). \end{cases}
\end{aligned}$$

Therefore with $t_0 = \frac{\gamma}{2c^2}$, $t_1 = \frac{\gamma}{c^2}$ and $\beta = \begin{cases} \frac{c^{-2\gamma}}{\Gamma(\gamma)} \left(\frac{\gamma}{2e}\right)^{\gamma}, & \text{for } \gamma \geq 1; \\ \frac{c^{-2\gamma}}{2\Gamma(\gamma)} \left(\frac{\gamma}{e}\right)^{\gamma}, & \text{for } \gamma \in (0, 1) \end{cases}$, we obtain

$$B_k = \begin{cases} \frac{c^{-2\gamma}}{2\Gamma(\gamma)} \left(\frac{\gamma}{2e}\right)^{\gamma}, & \text{for } \gamma \geq 1; \\ \frac{c^{-2\gamma}}{4\Gamma(\gamma)} \left(\frac{\gamma}{e}\right)^{\gamma}, & \text{for } \gamma \in (0, 1) \end{cases}.$$

(iv) Matérn kernel: It is easy to check that if $X \sim \text{Gamma}(a, b)$ and $Y \sim \text{InvGamma}(a, b)$ for $a, b > 0$ then for any $0 < x \leq y < \infty$ the following holds:

$$\mathbb{P}\{x \leq X \leq y\} = \mathbb{P}\{1/y \leq Y \leq 1/x\}.$$

This means, the above calculations for inverse multiquadrics can be used to obtain the following for the Matérn kernel:

$$B_k = \begin{cases} \frac{1}{2\Gamma(\tau - \frac{d}{2})} \left(\frac{2\tau - d}{4e}\right)^{\tau - \frac{d}{2}}, & \text{for } \tau - \frac{d}{2} \geq 1; \\ \frac{1}{4\Gamma(\tau - \frac{d}{2})} \left(\frac{2\tau - d}{2e}\right)^{\tau - \frac{d}{2}}, & \text{for } \tau - \frac{d}{2} \in (0, 1) \end{cases}.$$

D.3 $\beta^2\delta_1^{-d/2}$ in Corollary 10

In Corollary 10, we assumed that there exist $0 < \delta_0 \leq \delta_1 < \infty$ and $0 < \beta < \infty$ such that $\nu([\delta_0, \delta_1]) \geq \beta$. Define $A_k := \beta^2\delta_1^{-d/2}$. Based on the analysis carried out in Appendix D.2, in the following, we present the values of A_k for various radial kernels.

(i) Gaussian kernel: Choose $\delta_0 = \delta_1 = \frac{1}{2\eta^2}$ so that $\beta = 1$ and $A_k = (2\eta^2)^{d/2}$.

(ii) Mixture of Gaussians: Set $\delta_0 = \frac{1}{2\eta_1^2}$, $\delta_1 = \frac{1}{2\eta_M^2}$ so that $\beta = C_M$ implying $A_k = C_M^2(2\eta_M^2)^{d/2}$.

(iii) Inverse multiquadric kernels: Choosing $\delta_0 = t_0$ and $\delta_1 = t_1$ as in Appendix D.2, we obtain

$$A_k = \begin{cases} \frac{c^{d-4\gamma}}{\Gamma^2(\gamma)} \frac{\gamma^{2\gamma-\frac{d}{2}}}{(2e)^{2\gamma}}, & \text{for } \gamma \geq 1; \\ \frac{c^{d-4\gamma}}{4\Gamma^2(\gamma)} \frac{\gamma^{2\gamma-\frac{d}{2}}}{e^{2\gamma}}, & \text{for } \gamma \in (0,1) \end{cases}.$$

(iv) Matérn kernel: Define $\tilde{\gamma} := \tau - \frac{d}{2}$ and $\tilde{c} := \frac{c}{2}$. Choosing $\delta_0 = \frac{\tilde{c}^2}{\tilde{\gamma}}$ and $\delta_1 = \frac{2\tilde{c}^2}{\tilde{\gamma}}$, we obtain

$$\beta = \begin{cases} \frac{1}{\Gamma(\tilde{\gamma})} \left(\frac{\tilde{\gamma}}{2e}\right)^{\tilde{\gamma}}, & \text{for } \tilde{\gamma} \geq 1; \\ \frac{1}{2\Gamma(\tilde{\gamma})} \left(\frac{\tilde{\gamma}}{e}\right)^{\tilde{\gamma}}, & \text{for } \tilde{\gamma} \in (0,1) \end{cases},$$

using the analysis in Appendix D.2. Therefore,

$$A_k = \begin{cases} \frac{c^{-d}e^{-2\tilde{\gamma}}}{\Gamma^2(\tilde{\gamma})} \left(\frac{\tilde{\gamma}}{2}\right)^{2\tilde{\gamma}+\frac{d}{2}}, & \text{for } \tilde{\gamma} \geq 1; \\ \frac{c^{-d}e^{-2\tilde{\gamma}}}{\Gamma^2(\tilde{\gamma})} \frac{\tilde{\gamma}^{2\tilde{\gamma}+\frac{d}{2}}}{2^{2+\frac{d}{2}}}, & \text{for } \tilde{\gamma} \in (0,1) \end{cases}.$$

D.4 $\beta^2\delta_0\delta_1^{-\frac{d+2}{2}}$ in Theorem 13

In Theorem 13, we assumed that there exist $0 < \delta_0 \leq \delta_1 < \infty$ and $0 < \beta < \infty$ such that $\nu([\delta_0, \delta_1]) \geq \beta$. Define $B_k := \beta^2\delta_0\delta_1^{-\frac{d+2}{2}}$. Based on the analysis carried out in Appendix D.2, in the following, we present the values of B_k for various radial kernels.

(i) Gaussian kernel: Choose $\delta_0 = \delta_1 = \frac{1}{2\eta^2}$ so that $\beta = 1$ and $B_k = (2\eta^2)^{d/2}$.

(ii) Mixture of Gaussians: Set $\delta_0 = \frac{1}{2\eta_1^2}$, $\delta_1 = \frac{1}{2\eta_M^2}$ so that $\beta = C_M$ implying $B_k = \frac{C_M^2 2^{d/2}\eta_M^{d+2}}{\eta_1^2}$.

(iii) Inverse multiquadric kernels: Choosing $\delta_0 = t_0$ and $\delta_1 = t_1$ as in Appendix D.2, we obtain

$$B_k = \begin{cases} \frac{c^{d-4\gamma}}{2\Gamma^2(\gamma)} \frac{\gamma^{2\gamma-\frac{d}{2}}}{(2e)^{2\gamma}}, & \text{for } \gamma \geq 1; \\ \frac{c^{d-4\gamma}}{8\Gamma^2(\gamma)} \frac{\gamma^{2\gamma-\frac{d}{2}}}{e^{2\gamma}}, & \text{for } \gamma \in (0,1) \end{cases}.$$

(iv) Matérn kernel: With the choice of δ_0 and δ_1 as in Appendix D.3, we obtain

$$B_k = \begin{cases} \frac{c^{-d}e^{-2\tilde{\gamma}}}{2\Gamma^2(\tilde{\gamma})}\left(\frac{\tilde{\gamma}}{2}\right)^{2\tilde{\gamma}+\frac{d}{2}}, & \text{for } \tilde{\gamma} \geq 1; \\ \frac{c^{-d}e^{-2\tilde{\gamma}}}{\Gamma^2(\tilde{\gamma})}\frac{\tilde{\gamma}^{2\tilde{\gamma}+\frac{d}{2}}}{2^{3+\frac{d}{2}}}, & \text{for } \tilde{\gamma} \in (0,1) \end{cases}.$$

Appendix E. Alternate Proof of Theorem 8

In Theorem 8 we presented a minimax lower bound for radial kernels based on an appropriate construction of d-dimensional Gaussian distributions. By a clever choice of the variance σ^2, which decays to zero as $d \to \infty$, we obtained a lower bound of the order $\Omega(n^{-1/2})$ independent of d. This result was based on the direct analysis and special properties of radial kernels. In this appendix we will show that we can recover almost the same result using only Proposition 3, which holds for any translation invariant kernel. As we will see, this leads to slightly worse constant factors and an additional lower bound on the sample size n in terms of the properties of distribution ν, which specifies the kernel. Essentially we will repeat the main steps of the proof of Theorem 8. However, we will use Proposition 3 instead of direct computations (based on the form of radial kernels) to lower bound the RKHS distance between embeddings of Gaussian distributions with the Euclidean distance between their mean vectors.

Theorem E.1 *Let $\mathcal{P}$ be the set of distributions over $\mathbb{R}^d$ whose densities are continuously infinitely differentiable and k be radial on $\mathbb{R}^d$, i.e.,*

$$k(x,y) = \int_0^\infty e^{-t\|x-y\|_2^2}\, d\nu(t),$$

where $\nu \in M_+^b([0,\infty))$ such that $\operatorname{supp}(\nu) \neq \{0\}$. Assume that there exist $0 < t_0 \leq t_1 < \infty$ and $0 < \beta < \infty$ such that $\nu([t_0,t_1]) \geq \beta$. Suppose $n \geq 24\frac{t_1 Z_\nu}{\beta t_0}$ where $Z_\nu := \nu([0,\infty))$. Then

$$\inf_{\hat{\theta}_n} \sup_{P\in\mathcal{P}} P^n \left\{ \|\hat{\theta}_n - \mu_k(P)\|_{\mathcal{H}_k} \geq \frac{1}{50}\sqrt{\frac{1}{2n}\cdot\frac{\beta t_0}{t_1 e}\left(1-\frac{2}{2+d}\right)} \right\} \geq \frac{1}{5}.$$

Proof We apply Proposition 3 to the radial kernel k. In order to do so, we need to lower bound the quantity appearing in r.h.s. of Condition (7), which we do as follows. We already saw in the proof of Theorem 8 that in our case Λ_ψ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and has the following density:

$$\lambda_\psi(w) = \int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t), \quad w \in \mathbb{R}^d.$$

Therefore the r.h.s. of (7) reduces to

$$\begin{aligned} &\frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w) \\ &= \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 \cos\left(\langle a, w\rangle\right) \left(\int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|w\|_2^2}{4t}} d\nu(t)\right) dw \end{aligned}$$

$$= \frac{2}{(2\pi)^{d/2}} \int_0^\infty \frac{1}{(2t)^{d/2}} \underbrace{\int_{\mathbb{R}^d} e^{-\frac{1}{2}\left(2\sigma^2+\frac{1}{2t}\right)\|w\|_2^2} \langle e_z, w\rangle^2 e^{-i\langle a,w\rangle} dw}_{\clubsuit} \, d\nu(t), \tag{52}$$

where we used Euler's formula and Tonelli-Fubini theorem in the last equality. Denoting $\delta := 2\sigma^2 + \frac{1}{2t}$, we have

$$\begin{aligned}
\clubsuit &= \int_{\mathbb{R}^d} \exp\left(-\frac{\delta}{2}\sum_{\ell=1}^d w_\ell^2\right)\left(\sum_{j=1}^d (e_z)_j^2 w_j^2 + \sum_{j\neq\ell}(e_z)_j(e_z)_\ell w_j w_\ell\right)\exp\left(-i\sum_{\ell=1}^d a_\ell w_\ell\right) dw \\
&= \bigstar + \spadesuit,
\end{aligned}$$

where

$$\bigstar := \sum_{j=1}^d \int_{\mathbb{R}^d} e^{-\frac{1}{2}\delta\|w\|_2^2} (e_z)_j^2 w_j^2 e^{-i\langle a,w\rangle} dw$$

and

$$\spadesuit := \sum_{j\neq l}^d \int_{\mathbb{R}^d} e^{-\frac{1}{2}\delta\|w\|_2^2} (e_z)_j (e_z)_l w_j w_l e^{-i\langle a,w\rangle} dw.$$

Note that

$$\begin{aligned}
(e_z)_j^2 \int_{\mathbb{R}^d} e^{-\frac{1}{2}\delta\|w\|_2^2} w_j^2 e^{-i\langle a,w\rangle} dw &= (e_z)_j^2 \left(\prod_{\ell\neq j}\int_{-\infty}^\infty e^{-\frac{\delta}{2}w_\ell^2} e^{-ia_\ell w_\ell} dw_\ell\right) \\
&\qquad \times \left(\int_{-\infty}^\infty e^{-\frac{\delta}{2}w_j^2} w_j^2 e^{-ia_j w_j} dw_j\right) \\
&= (e_z)_j^2 \left(\prod_{\ell\neq j}\sqrt{\frac{2\pi}{\delta}} e^{-\frac{a_\ell^2}{2\delta}}\right) \cdot \left(\int_{-\infty}^\infty e^{-\frac{\delta}{2}w_j^2} w_j^2 e^{-ia_j w_j} dw_j\right),
\end{aligned}$$

where we used Lemma C.1. It follows from (Folland, 1999, Theorem 8.22(d)) that if $g = x^2 f \in L^1(\mathbb{R})$, then $f^\wedge$ is twice differentiable and

$$g^\wedge(y) = -\frac{\partial^2 f^\wedge(y)}{\partial^2 y},$$

which together with Lemma C.1 shows that

$$\int_{-\infty}^\infty e^{-\frac{\delta}{2}w_j^2} w_j^2 e^{-ia_j w_j} dw_j = \frac{1}{\delta}\sqrt{\frac{2\pi}{\delta}} e^{-\frac{a_j^2}{2\delta}}\left(1 - \frac{a_j^2}{\delta}\right).$$

Therefore, we get

$$\begin{aligned}
(e_z)_j^2 \int_{\mathbb{R}^d} e^{-\frac{1}{2}\delta\|w\|_2^2} w_j^2 e^{-i\langle a,w\rangle} dw &= (e_z)_j^2 \left(\prod_{\ell\neq j}\sqrt{\frac{2\pi}{\delta}} e^{-\frac{a_\ell^2}{2\delta}}\right)\sqrt{\frac{2\pi}{\delta}}\frac{1}{\delta} e^{-\frac{a_j^2}{2\delta}}\left(1 - \frac{a_j^2}{\delta}\right) \\
&= \frac{(e_z)_j^2}{\delta}\left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}}\left(1 - \frac{a_j^2}{\delta}\right).
\end{aligned}$$

Summing over $j = 1, \ldots, d$ we get

$$\bigstar = \sum_{j=1}^{d} (e_z)_j^2 \int_{\mathbb{R}^d} e^{-\frac{1}{2}\delta\|w\|_2^2} w_j^2 e^{-i\langle a, w\rangle} dw = \frac{1}{\delta}\left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}} - \sum_{j=1}^{d} \frac{(e_z)_j^2 a_j^2}{\delta^2}\left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}}.$$

Next, for any $j \neq \ell$ we compute

$$\begin{aligned}
&\int_{\mathbb{R}^d} \exp\left(-\frac{\delta}{2}\sum_{\ell=1}^{d} w_\ell^2\right)(e_z)_j (e_z)_\ell w_j w_\ell \exp\left(-i\sum_{\ell=1}^{d} a_\ell w_\ell\right) dw \\
&= (e_z)_j (e_z)_\ell \left(\prod_{q\notin\{j,\ell\}} \int_{-\infty}^{\infty} e^{-\frac{\delta}{2}w_q^2} e^{-ia_q w_q} dw_q\right)\left(\prod_{q\in\{j,\ell\}} \int_{-\infty}^{\infty} e^{-\frac{\delta}{2}w_q^2} w_q e^{-ia_q w_q} dw_q\right) \\
&= (e_z)_j (e_z)_\ell \left(\prod_{q\notin\{j,\ell\}} \sqrt{\frac{2\pi}{\delta}} e^{-\frac{a_q^2}{2\delta}}\right)\left(\prod_{q\in\{j,\ell\}} \sqrt{\frac{2\pi}{\delta}} \frac{ia_q}{\delta} e^{-\frac{a_q^2}{2\delta}}\right) \\
&= (e_z)_j (e_z)_\ell \left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}} \left(-\frac{a_j a_\ell}{\delta^2}\right).
\end{aligned}$$

Summing over $j \neq \ell$ we get

$$\spadesuit = -\frac{\langle e_z, a\rangle^2}{\delta^2}\left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}} + \sum_{j=1}^{d} \frac{(e_z)_j^2 a_j^2}{\delta^2}\left(\frac{2\pi}{\delta}\right)^{d/2} e^{-\frac{\|a\|_2^2}{2\delta}}.$$

Returning to (52), we get

$$\begin{aligned}
&\frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w) \\
&= \frac{2}{(2\pi)^{d/2}} \int_0^\infty \frac{1}{(2t)^{d/2}} e^{-\frac{\|a\|_2^2}{2\delta}} \left(\frac{2\pi}{\delta}\right)^{d/2} \frac{1}{\delta}\left(1 - \frac{\langle e_z, a\rangle^2}{\delta}\right) d\nu(t) \\
&= 4\int_0^\infty \exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t + 1}\right) \frac{t}{(4\sigma^2 t + 1)^{1+d/2}}\left(1 - \frac{2t\langle e_z, a\rangle^2}{4\sigma^2 t + 1}\right) d\nu(t).
\end{aligned}$$

In order to apply Proposition 3 we need to lower bound the following value, appearing in Condition (7):

$$\begin{aligned}
\Delta(a) &:= \min_{z\in\mathbb{R}^d\setminus\{0\}} \frac{2}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\sigma^2\|w\|_2^2} \langle e_z, w\rangle^2 \cos\left(\langle a, w\rangle\right) d\Lambda_\psi(w) \\
&= 4\int_0^\infty \exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t + 1}\right) \frac{t}{(4\sigma^2 t + 1)^{1+d/2}}\left(1 - \frac{2t\|a\|_2^2}{4\sigma^2 t + 1}\right) d\nu(t).
\end{aligned}$$

Next we will separately treat two different cases.

Case 1: $d > 2$. Note that the function $\rho(t) = t(4\sigma^2 t + 1)^{-(d+2)/2}$ is positive and bounded on $[0, \infty)$ for any $d > 0$. Thus, we can define a non-negative and finite measure

$\tilde{\tau}$, absolutely continuous with respect to ν with density $\rho(t)$. If we denote $Z_\tau := \int_0^\infty 1\, d\tilde{\tau}(t)$ and write τ for the normalized version of $\tilde{\tau}$, then we can rewrite

$$\begin{aligned}\Delta(a) &= 4\int_0^\infty \exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\left(1-\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right) d\tilde{\tau}(t)\\ &= 4\,Z_\tau \mathbb{E}_{t\sim\tau}\left[\exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\left(1-\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\right]\\ &= 4\,Z_\tau \mathbb{E}_{t\sim\tau}\left[\exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\right] - 4\,Z_\tau \mathbb{E}_{t\sim\tau}\left[\exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right].\end{aligned}$$

Note that for $d>2$, $\mathbb{E}_{t\sim\tau}[|t|]$ is finite, since in this case $t\mapsto \frac{t^2}{(4\sigma^2 t+1)^{(d+2)/2}}$ is bounded and ν is a finite measure. Denote $\mu_\tau := \mathbb{E}_{t\sim\tau}[t]$ and note that $t\mapsto \exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)$ is a convex function on $[0,\infty)$. Thus, for $d>2$ we can use Jensen's inequality to get

$$\mathbb{E}_{t\sim\tau}\left[\exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\right] \geq \exp\left(-\frac{1}{2}\frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1}\right).$$

Also note that

$$-4\,Z_\tau \mathbb{E}_{t\sim\tau}\left[\exp\left(-\frac{1}{2}\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right)\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right] \geq 4\,Z_\tau \mathbb{E}_{t\sim\tau}\left[-\frac{2t\|a\|_2^2}{4\sigma^2 t+1}\right] \geq -4\,Z_\tau \frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1},$$

where we used inequality $e^{-x}\leq 1$, which holds for $x\geq 0$, together with Jensen's inequality and the fact that $t\mapsto -\frac{2t\|a\|_2^2}{4\sigma^2 t+1}$ is concave on $[0,\infty)$. Summarizing, we have

$$\begin{aligned}\Delta(a) &\geq 4Z_\tau\left(\exp\left(-\frac{1}{2}\frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1}\right) - \frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1}\right)\\ &\geq 4Z_\tau\left(1-\frac{1}{2}\frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1} - \frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1}\right)\\ &= 2Z_\tau\left(2-3\frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1}\right),\end{aligned}$$

where we used a simple inequality $e^x\geq 1+x$. If the following condition is satisfied:

$$\frac{2\mu_\tau\|a\|_2^2}{4\sigma^2\mu_\tau+1} \leq \frac{1}{3}, \tag{53}$$

then we get

$$\Delta(a) \geq 2Z_\tau = 2\int_0^\infty \frac{t}{(4\sigma^2 t+1)^{1+d/2}} d\nu(t). \tag{54}$$

Together with Proposition 3 this leads to the following lower bound, which holds for any $\mu_0,\mu_1\in\mathbb{R}^d$ and $\sigma^2>0$ satisfying (53) with $a:=\mu_0-\mu_1$:

$$\|\theta_0-\theta_1\|_{\mathcal{H}_k}^2 \geq \int_0^\infty \frac{t\|\mu_0-\mu_1\|_2^2}{(4\sigma^2 t+1)^{1+d/2}} d\nu(t),$$

where θ_0 and θ_1 are KME's of Gaussian measures $G(\mu_0, \sigma^2 I)$ and $G(\mu_1, \sigma^2 I)$ respectively. Note that this lower bound is identical to the one in (27), which we obtained using direct analysis for the radial kernels. However, condition (26) is now replaced with the stronger one in (53). We can now repeat the proof of Theorem 8 starting from inequality (27) and making sure that condition (53) is satisfied when we choose constants appearing in definitions of μ_0, μ_1 and σ^2.

In order to check condition (53) we need to upper bound the expectation μ_τ. It is easily seen that for $d > 2$, $t \mapsto \frac{t^2}{(4\sigma^2 t+1)^{(d+2)/2}}$ achieves its maximum on $[0, \infty)$ for $t^* = \frac{1}{\sigma^2(d-2)}$. Using this fact, denoting $Z_\nu = \int_0^\infty 1 \, d\nu(t)$, and setting $\sigma^2 = \frac{1}{2t_1 d}$ we get

$$\begin{aligned}
\mu_\tau &= \frac{1}{Z_\tau} \int_0^\infty \frac{t^2}{(4\sigma^2 t + 1)^{1+d/2}} d\nu(t) \leq \frac{Z_\nu}{Z_\tau} \frac{(t^*)^2}{(4\sigma^2 t^* + 1)^{1+d/2}} \\
&= \frac{4t_1^2 Z_\nu}{Z_\tau} \frac{d^2}{(d-2)^2} \frac{1}{(\frac{4}{d-2} + 1)^{1+d/2}} \\
&= \frac{4t_1^2 Z_\nu}{Z_\tau} \left(1 + \frac{2}{d-2}\right)^2 \left(\left(1 - \frac{4}{d+2}\right)^{(d+2)/4}\right)^2 \leq \frac{4t_1^2 Z_\nu}{Z_\tau e^2} \left(1 + \frac{2}{d-2}\right)^2.
\end{aligned}$$

We may finally use (29) to get

$$Z_\tau \geq \frac{\beta t_0}{e} \left(1 - \frac{2}{d+2}\right),$$

which leads to the following upper bound on the expectation μ_τ:

$$\mu_\tau \leq \frac{4t_1^2 Z_\nu}{\beta t_0 e} \frac{d(d+2)}{(d-2)^2}.$$

This upper bound shows that the condition (53) is satisfied if the following holds:

$$t_1 \|a\|_2^2 \leq \frac{2}{3} t_1 \sigma^2 + \frac{\beta t_0 e}{24 t_1 Z_\nu} \frac{(d-2)^2}{d(d+2)}. \tag{55}$$

We conclude the proof by repeating the remaining steps of the proof of Theorem 8 and replacing condition (26) on the value $\|\mu_0 - \mu_1\|_2^2$ with (55) specified to $a = \mu_0 - \mu_1$.

Case 2: $d \leq 2$. We can use a simple inequality $e^{-x/2}(1-x) \geq 1 - 3x/2$ which holds for any x and get the following lower bound:

$$\Delta(a) \geq 4 \int_0^\infty \frac{t}{(4\sigma^2 t + 1)^{1+d/2}} \left(1 - \frac{3t\|a\|_2^2}{4\sigma^2 t + 1}\right) d\nu(t).$$

Assuming $\|a\|_2^2 \leq \sigma^2$ we further get

$$\left(1 - \frac{3t\|a\|_2^2}{4\sigma^2 t + 1}\right) \geq \left(1 - \frac{3t\sigma^2}{4\sigma^2 t + 1}\right) \geq \left(1 - \frac{3t\sigma^2}{4\sigma^2 t}\right) = \frac{1}{4}$$

and as a consequence, we also get

$$\Delta(a) \geq \int_0^\infty \frac{t}{(4\sigma^2 t + 1)^{1+d/2}} d\nu(t),$$

which coincides with (54) up to an additional factor of 2. We can now repeat all the steps for the previous case, and it is also easy to check that in this case $\|\mu_0 - \mu_1\|_2^2 \leq \sigma^2$ will be indeed satisfied. This concludes the proof. ■

Remark E.2 *This result should be compared to Theorem 8, which was based on the direct analysis for radial kernels. We see that apart from an extra factor* 2 *appearing under the square root in the lower bound, Theorem E.1 also requires a superfluous condition on the minimal sample size* n*, which depends on properties of* ν*. For instance, for Gaussian kernel with* ν *concentrated on a single point* $\frac{1}{2\eta^2}$ *for some* $\eta^2 > 0$*, the result holds as long as* $n \geq 24$*, because in this case we can take* $t_0 = t_1 = \frac{1}{2\eta^2}$ *and* $\beta = 1$*. However, other choices of* ν *may lead to quite restrictive lower bounds on* n*.*

Remark E.3 *Conceptually, the main difference between the proofs of Theorems 8 and E.1 lies in the way we lower bound the RKHS distance between embeddings of Gaussian measures with the Euclidean distance between their mean vectors. In Theorem 8 we derived a closed-form expression for the RKHS distance in* (25) *and then lower bounded it directly using the properties specific to its form. On the other hand, in Theorem E.1 we resorted to the lower bound of Lemma 3, which holds for any translation invariant kernel and hence is less tight.*

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

V. I. Bogachev. *Measure Theory*, volume 1. Springer, 2007.

J. Diestel and J. J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

N. Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley, 2000.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.*, 14:3753–3783, 2013.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, San Diego, USA, 2000.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520, Cambridge, MA, 2007. MIT Press.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

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

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, 2008.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015.

K. Muandet, B. Sriperumbudur, K. Fukumizu, A. Gretton, and B. Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 2016. To appear.

A. Ramdas, S. Reddi, B. Poczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI Conference on Artificial Intelligence*, 2015.

I. J. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811–841, 1938.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

B. K. Sriperumbudur. Mixture density estimation via Hilbert space embedding of measures. In *Proceedings of International Symposium on Information Theory*, pages 1027–1030, 2011.

B. K. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

Z. Szabó, A. Gretton, B. Póczos, and B. K. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 948–957. JMLR Workshop and Conference Proceedings, 2015.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, NY, 2008.

R. Vert and J-P. Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

V. Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.