Consistent Kernel Mean Estimation
for Functions of Random Variables

Carl-Johann Simon-Gabriel*, **Adam Ścibior***,†, **Ilya Tolstikhin**, **Bernhard Schölkopf**
Department of Empirical Inference, Max Planck Institute for Intelligent Systems
Spemanstraße 38, 72076 Tübingen, Germany
* joint first authors; † also with: Engineering Department, Cambridge University
`cjsimon@, adam.scibior@, ilya@, bs@tuebingen.mpg.de`

Abstract

We provide a theoretical foundation for non-parametric estimation of functions of random variables using kernel mean embeddings. We show that for any continuous function f, consistent estimators of the mean embedding of a random variable X lead to consistent estimators of the mean embedding of $f(X)$. For Matérn kernels and sufficiently smooth functions we also provide rates of convergence.

Our results extend to functions of multiple random variables. If the variables are dependent, we require an estimator of the mean embedding of their joint distribution as a starting point; if they are independent, it is sufficient to have separate estimators of the mean embeddings of their marginal distributions. In either case, our results cover both mean embeddings based on i.i.d. samples as well as "reduced set" expansions in terms of dependent expansion points. The latter serves as a justification for using such expansions to limit memory resources when applying the approach as a basis for probabilistic programming.

1 Introduction

A common task in probabilistic modelling is to compute the distribution of $f(X)$, given a measurable function f and a random variable X. In fact, the earliest instances of this problem date back at least to Poisson (1837). Sometimes this can be done analytically. For example, if f is linear and X is Gaussian, that is $f(x) = ax + b$ and $X \sim \mathcal{N}(\mu; \sigma)$, we have $f(X) \sim \mathcal{N}(a\mu + b; a\sigma)$. There exist various methods for obtaining such analytical expressions (Mathai, 1973), but outside a small subset of distributions and functions the formulae are either not available or too complicated to be practical.

An alternative to the analytical approach is numerical approximation, ideally implemented as a flexible software library. The need for such tools is recognised in the general programming languages community (McKinley, 2016), but no standards were established so far. The main challenge is in finding a good approximate representation for random variables.

Distributions on integers, for example, are usually represented as lists of $(x_i, p(x_i))$ pairs. For real valued distributions, integral transforms (Springer, 1979), mixtures of Gaussians (Milios, 2009), Laguerre polynomials (Williamson, 1989), and Chebyshev polynomials (Korzeń and Jaroszewicz, 2014) were proposed as convenient representations for numerical computation. For strings, probabilistic finite automata are often used. All those approaches have their merits, but they only work with a specific input type.

There is an alternative, based on Monte Carlo sampling (Kalos and Whitlock, 2008), which is to represent X by a (possibly weighted) sample $\{(x_i, w_i)\}_{i=1}^{n}$ (with $w_i \geq 0$). This representation has several advantages: (i) it works for any input type, (ii) the sample size controls the time-accuracy trade-off, and (iii) applying functions to random variables reduces to applying the functions pointwise

to the sample, i.e., $\{(f(x_i), w_i)\}$ represents $f(X)$. Furthermore, expectations of functions of random variables can be estimated as $\mathbb{E}\left[f(X)\right] \approx \sum_i w_i f(x_i) / \sum_i w_i$, sometimes with guarantees for the convergence rate.

The flexibility of this Monte Carlo approach comes at a cost: without further assumptions on the underlying input space $\mathcal{X}$, it is hard to quantify the accuracy of this representation. For instance, given two samples of the same size, $\{(x_i, w_i)\}_{i=1}^n$ and $\{(x_i', w_i')\}_{i=1}^n$, how can we tell which one is a better representation of X? More generally, how could we optimize a representation with predefined sample size?

There exists an alternative to the Monte Carlo approach, called Kernel Mean Embeddings (KME) (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007). It also represents random variables as samples, but additionally defines a notion of similarity between sample points. As a result, (i) it keeps all the advantages of the Monte Carlo scheme, (ii) it includes the Monte Carlo method as a special case, (iii) it overcomes its pitfalls described above, and (iv) it can be tailored to focus on different properties of X, depending on the user's needs and prior assumptions. The KME approach identifies both sample points and distributions with functions in an abstract Hilbert space. Internally the latter are still represented as weighted samples, but the weights can be negative and the straightforward Monte Carlo interpretation is no longer valid. Schölkopf et al. (2015) propose using KMEs as approximate representation of random variables for the purpose of computing their functions. However, they only provide theoretical justification for it in rather idealised settings, which do not meet practical implementation requirements.

In this paper, we build on this work and provide general theoretical guarantees for the proposed estimators. Specifically, we prove statements of the form "if $\{(x_i, w_i)\}_{i=1}^n$ provides a good estimate for the KME of X, then $\{(f(x_i), w_i)\}_{i=1}^n$ provides a good estimate for the KME of $f(X)$". Importantly, our results *do not assume joint independence* of the observations x_i (and weights w_i). This makes them a powerful tool. For instance, imagine we are given data $\{(x_i, w_i)\}_{i=1}^n$ from a random variable X that we need to compress. Then our theorems guarantee that, whatever compression algorithm we use, as long as the compressed representation $\{(x_j', w_j')\}_{j=1}^n$ still provides a good estimate for the KME of X, the pointwise images $\{(f(x_j'), w_j')\}_{j=1}^n$ provide good estimates of the KME of $f(X)$.

In the remainder of this section we first introduce KMEs and discuss their merits. Then we explain why and how we extend the results of Schölkopf et al. (2015). Section 2 contains our main results. In Section 2.1 we show consistency of the relevant estimator in a general setting, and in Section 2.2 we provide finite sample guarantees when Matérn kernels are used. In Section 3 we show how our results apply to functions of multiple variables, both interdependent and independent. Section 4 concludes with a discussion.

1.1 Background on kernel mean embeddings

Let $\mathcal{X}$ be a measurable input space. We use a positive definite bounded and measurable kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to represent random variables $X \sim P$ and weighted samples $\hat{X} := \{(x_i, w_i)\}_{i=1}^n$ as two functions μ_X^k and $\hat{\mu}_X^k$ in the corresponding Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ by defining

$$\mu_X^k := \int k(x, .)\, dP(x) \quad \text{and} \quad \hat{\mu}_X^k := \sum_i w_i k(x_i, .)\,.$$

These are guaranteed to exist, since we assume the kernel is bounded (Smola et al., 2007). When clear from the context, we omit the kernel k in the superscript. μ_X is called the KME of P, but we also refer to it as the KME of X. In this paper we focus on computing functions of random variables. For $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a measurable space, and for a positive definite bounded $k_z : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ we also write

$$\mu_{f(X)}^{k_z} := \int k_z(f(x), .)\, dP(x) \quad \text{and} \quad \hat{\mu}_{f(X)}^{k_z} := \sum_i w_i k_z(f(x_i), .)\,. \tag{1}$$

The advantage of mapping random variables X and samples $\hat{X}$ to functions in the RKHS is that we may now say that $\hat{X}$ is a good approximation for X if the RKHS distance $\|\hat{\mu}_X - \mu_X\|$ is small. This distance depends on the choice of the kernel and different kernels emphasise different information about X. For example if on $\mathcal{X} := [a, b] \subset \mathbb{R}$ we choose $k(x, x') := x \cdot x' + 1$, then

$\mu_X(x) = \mathbb{E}_{X \sim P}[X] x + 1$. Thus any two distributions and/or samples with equal means are mapped to the same function in $\mathcal{H}_k$ so the distance between them is zero. Therefore using this particular k, we keep track only of the mean of the distributions. If instead we prefer to keep track of all first p moments, we may use the kernel $k(x, x') := (x \cdot x' + 1)^p$. And if we do not want to loose any information at all, we should choose k such that μ^k is injective over all probability measures on $\mathcal{X}$. Such kernels are called *characteristic*. For standard spaces, such as $\mathcal{X} = \mathbb{R}^d$, many widely used kernels were proven characteristic, such as Gaussian, Laplacian, and Matérn kernels (Sriperumbudur et al., 2010, 2011).

The Gaussian kernel $k(x, x') := e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$ may serve as another good illustration of the flexibility of this representation. Whatever positive bandwidth $\sigma^2 > 0$, we do not lose any information about distributions, because k is characteristic. Nevertheless, if σ^2 grows, all distributions start looking the same, because their embeddings converge to a constant function 1. If, on the other hand, σ^2 becomes small, distributions look increasingly different and $\hat{\mu}_X$ becomes a function with bumps of height w_i at every x_i. In the limit when σ^2 goes to zero, each point is only similar to itself, so $\hat{\mu}_X$ reduces to the Monte Carlo method. Choosing σ^2 can be interpreted as controlling the degree of smoothing in the approximation.

1.2 Reduced set methods

An attractive feature when using KME estimators is the ability to reduce the number of expansion points (i.e., the size of the weighted sample) in a principled way. Specifically, if $\hat{X}' := \{(x'_j, 1/N)\}_{j=1}^N$ then the objective is to construct $\hat{X} := \{(x_i, w_i)\}_{i=1}^n$ that minimises $\|\hat{\mu}_{X'} - \hat{\mu}_X\|$ with $n < N$. Often the resulting x_i are mutually dependent and the w_i certainly depend on them. The algorithms for constructing such expansions are known as *reduced set methods* and have been studied by the machine learning community (Schölkopf and Smola, 2002, Chapter 18).

Although reduced set methods provide significant efficiency gains, their application raises certain concerns when it comes to computing functions of random variables. Let P, Q be distributions of X and $f(X)$ respectively. If $x'_j \sim_{i.i.d.} P$, then $f(x'_j) \sim_{i.i.d.} Q$ and so $\hat{\mu}_{f(X')} = \frac{1}{N} \sum_j k(f(x'_j), .)$ reduces to the commonly used $\sqrt{N}$-consistent empirical estimator of $\mu_{f(X)}$ (Smola et al., 2007). Unfortunately, this is not the case after applying reduced set methods, and it is not known under which conditions $\hat{\mu}_{f(X)}$ is a consistent estimator for $\mu_{f(X)}$.

Schölkopf et al. (2015) advocate the use of reduced expansion set methods to save computational resources. They also provide some reasoning why this should be the right thing to do for characteristic kernels, but as they state themselves, their rigorous analysis does not cover practical reduced set methods. Motivated by this and other concerns listed in Section 1.4, we provide a generalised analysis of the estimator $\hat{\mu}_{f(X)}$, where we do not make assumptions on how x_i and w_i were generated.

Before doing that, however, we first illustrate how the need for reduced set methods naturally emerges on a concrete problem.

1.3 Illustration with functions of two random variables

Suppose that we want to estimate $\mu_{f(X,Y)}$ given i.i.d. samples $\hat{X}' = \{x'_i, 1/N\}_{i=1}^N$ and $\hat{Y}' = \{y'_j, 1/N\}_{j=1}^N$ from two independent random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ respectively. Let Q be the distribution of $Z = f(X, Y)$.

The first option is to consider what we will call the *diagonal estimator* $\hat{\mu}_1 := \frac{1}{N} \sum_{i=1}^n k_z(f(x'_i, y'_i), .)$. Since $f(x'_i, y'_i) \sim_{i.i.d.} Q$, $\hat{\mu}_1$ is $\sqrt{N}$-consistent (Smola et al., 2007). Another option is to consider the *U-statistic* estimator $\hat{\mu}_2 := \frac{1}{N^2} \sum_{i,j=1}^N k_z(f(x'_i, y'_j), .)$, which is also known to be $\sqrt{N}$-consistent. Experiments show that $\hat{\mu}_2$ is more accurate and has lower variance than $\hat{\mu}_1$ (see Figure 1). However, the U-statistic estimator $\hat{\mu}_2$ needs $O(n^2)$ memory rather than $O(n)$. For this reason Schölkopf et al. (2015) propose to use a reduced set method both on $\hat{X}'$ and $\hat{Y}'$ to get new samples $\hat{X} = \{x_i, w_i\}_{i=1}^n$ and $\hat{Y} = \{y_j, u_j\}_{j=1}^n$ of size $n \ll N$, and then estimate $\mu_{f(X,Y)}$ using $\hat{\mu}_3 := \sum_{i,j=1}^n w_i u_j k_x(f(x_i, y_j), .)$.

3

We ran experiments on synthetic data to show how accurately $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\mu}_3$ approximate $\mu_{f(X,Y)}$ with growing sample size N. We considered three basic arithmetic operations: multiplication $X \cdot Y$, division X/Y, and exponentiation X^Y, with $X \sim \mathcal{N}(3; 0.5)$ and $Y \sim \mathcal{N}(4; 0.5)$. As the true embedding $\mu_{f(X,Y)}$ is unknown, we approximated it by a U-statistic estimator based on a large sample (125 points). For $\hat{\mu}_3$, we used the simplest possible reduced set method: we randomly sampled subsets of size $n = 0.01 \cdot N$ of the x_i, and optimized the weights w_i and u_i to best approximate $\hat{\mu}_X$ and $\hat{\mu}_Y$. The results are summarised in Figure 1 and corroborate our expectations: (i) all estimators converge, (ii) $\hat{\mu}_2$ converges fastest and has the lowest variance, and (iii) $\hat{\mu}_3$ is worse than $\hat{\mu}_2$, but much better than the diagonal estimator $\hat{\mu}_1$. Note, moreover, that unlike the U-statistic estimator $\hat{\mu}_2$, the reduced set based estimator $\hat{\mu}_3$ can be used with a fixed storage budget even if we perform a sequence of function applications—a situation naturally appearing in the context of probabilistic programming.

Schölkopf et al. (2015) prove the consistency of $\hat{\mu}_3$ only for a rather limited case, when the points of the reduced expansions $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are i.i.d. copies of X and Y, respectively, and the weights $\{(w_i, u_i)\}_{i=1}^n$ are constants. Using our new results we will prove in Section 3.1 the consistency of $\hat{\mu}_3$ under fairly general conditions, even in the case when both expansion points and weights are interdependent random variables.
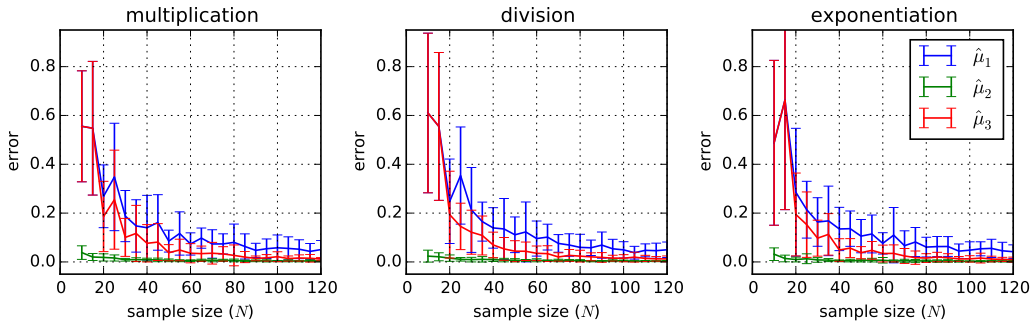

Figure 1: Error of kernel mean estimators for basic arithmetic functions of two variables, $X \cdot Y$, X/Y and X^Y, as a function of sample size N. The U-statistic estimator $\hat{\mu}_2$ works best, closely followed by the proposed estimator $\hat{\mu}_3$, which outperforms the diagonal estimator $\hat{\mu}_1$.

1.4 Other sources of non-i.i.d. samples

Although our discussion above focuses on reduced expansion set methods, there are other popular algorithms that produce KME expansions where the samples are not i.i.d. Here we briefly discuss several examples, emphasising that our selection is not comprehensive. They provide additional motivation for stating convergence guarantees in the most general setting possible.

An important notion in probability theory is that of a conditional distribution, which can also be represented using KME (Song et al., 2009). With this representation the standard laws of probability, such as sum, product, and Bayes' rules, can be stated using KME (Fukumizu et al., 2013). Applying those rules results in KME estimators with strong dependencies between samples and their weights.

Another possibility is that even though i.i.d. samples are available, they may not produce the best estimator. Various approaches, such as kernel herding (Chen et al., 2010; Lacoste-Julien et al., 2015), attempt to produce a better KME estimator by actively generating pseudo-samples that are not i.i.d. from the underlying distribution.

2 Main results

This section contains our main results regarding consistency and finite sample guarantees for the estimator $\hat{\mu}_{f(X)}$ defined in (1). They are based on the convergence of $\hat{\mu}_X$ and avoid simplifying assumptions about its structure.

4

2.1 Consistency

If k_x is c_0-*universal* (see Sriperumbudur et al. (2011)), consistency of $\hat{\mu}_{f(X)}$ can be shown in a rather general setting.

Theorem 1. *Let $\mathcal{X}$ and $\mathcal{Z}$ be compact Hausdorff spaces equipped with their Borel σ-algebras, $f : \mathcal{X} \to \mathcal{Z}$ a continuous function, k_x, k_z continuous kernels on $\mathcal{X}, \mathcal{Z}$ respectively. Assume k_x is c_0-universal and that there exists C such that $\sum_i |w_i| \leq C$ independently of n. The following holds:*

$$\text{If} \quad \hat{\mu}_X^{k_x} \to \mu_X^{k_x} \quad \text{then} \quad \hat{\mu}_{f(X)}^{k_z} \to \mu_{f(X)}^{k_z} \quad \text{as} \quad n \to \infty.$$

Proof. Let P be the distribution of X and $\hat{P}_n = \sum_{i=1}^n w_i \delta_{x_i}$. Define a new kernel on $\mathcal{X}$ by $\widetilde{k}_x(x_1, x_2) := k_z\big(f(x_1), f(x_2)\big)$. $\mathcal{X}$ is compact and $\{\hat{P}_n \,|\, n \in \mathbb{N}\} \cup \{P\}$ is a bounded set (in total variation norm) of finite measures, because $\|\hat{P}_n\|_{TV} = \sum_{i=1}^n |w_i| \leq C$. Furthermore, k_x is continuous and c_0-universal. Using Corollary 52 of Simon-Gabriel and Schölkopf (2016) we conclude that: $\hat{\mu}_X^{k_x} \to \mu_X^{k_x}$ implies that $\hat{P}$ converges weakly to P. Now, k_z and f being continuous, so is $\widetilde{k}_x$. Thus, if $\hat{P}$ converges weakly to P, then $\hat{\mu}_X^{\widetilde{k}_x} \to \mu_X^{\widetilde{k}_x}$ (Simon-Gabriel and Schölkopf, 2016, Theorem 44, Points (1) and (iii)). Overall, $\hat{\mu}_X^{k_x} \to \mu_X^{k_x}$ implies $\hat{\mu}_X^{\widetilde{k}_x} \to \mu_X^{\widetilde{k}_x}$. We conclude the proof by showing that convergence in $\mathcal{H}_{\widetilde{k}_x}$ leads to convergence in $\mathcal{H}_{k_z}$:

$$\left\| \hat{\mu}_{f(X)}^{k_z} - \mu_{f(X)}^{k_z} \right\|_{k_z}^2 = \left\| \hat{\mu}_X^{\widetilde{k}_x} - \mu_X^{\widetilde{k}_x} \right\|_{\widetilde{k}_x}^2 \to 0.$$

For a detailed version of the above, see Appendix A. $\qquad\square$

The continuity assumption is rather unrestrictive. All kernels and functions defined on a discrete space are continuous with respect to the discrete topology, so the theorem applies in this case. For $\mathcal{X} = \mathbb{R}^d$, many kernels used in practice are continuous, including Gaussian, Laplacian, Matérn and other radial kernels. The slightly limiting factor of this theorem is that k_x must be c_0-universal, which often can be tricky to verify. However, most standard kernels—including all radial, non-constant kernels—are c_0-universal (see Sriperumbudur et al., 2011). The assumption that the input domain is compact is satisfied in most applications, since any measurements coming from physical sensors are contained in a bounded range. Finally, the assumption that $\sum_i |w_i| \leq C$ can be enforced, for instance, by applying a suitable regularization in reduced set methods.

2.2 Finite sample guarantees

Theorem 1 guarantees that the estimator $\hat{\mu}_{f(X)}$ converges to $\mu_{f(X)}$ when $\hat{\mu}_X$ converges to μ_X. However, it says nothing about the speed of convergence. In this section we provide a convergence rate when working with Matérn kernels, which are of the form

$$k_x^s(x, x') = \frac{2^{1-s}}{\Gamma(s)} \|x - x'\|_2^{s-d/2} \mathcal{B}_{d/2-s}\left(\|x - x'\|_2\right) , \tag{2}$$

where $\mathcal{B}_\alpha$ is a modified Bessel function of the third kind (also known as Macdonald function) of order α, Γ is the Gamma function and $s > \frac{d}{2}$ is a smoothness parameter. The RKHS induced by k_x^s is the Sobolev space $\mathcal{W}_2^s(\mathbb{R}^d)$ (Wendland, 2004, Theorem 6.13 & Chap.10) containing s-times differentiable functions. The finite-sample bound of Theorem 2 is based on the analysis of Kanagawa et al. (2016), which requires the following assumptions:

Assumptions 1. *Let X be a random variable over $\mathcal{X} = \mathbb{R}^d$ with distribution P and let $\hat{X} = \{(x_i, w_i)\}_{i=1}^n$ be random variables over $\mathcal{X}^n \times \mathbb{R}^n$ with joint distribution S. There exists a probability distribution Q with full support on $\mathbb{R}^d$ and a bounded density, satisfying the following properties:*

(i) P has a bounded density function w.r.t. Q;
(ii) there is a constant $D > 0$ independent of n, such that

$$\mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^n g^2(x_i) \right] \leq D \|g\|_{L^2(Q)}^2 , \qquad \forall g \in L^2(Q) .$$

These assumptions were shown to be fairly general and we refer to Kanagawa et al. (2016, Section 4.1) for various examples where they are met. Next we state the main result of this section.

Theorem 2. *Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Z} = \mathbb{R}^{d'}$, and $f : \mathcal{X} \to \mathcal{Z}$ be an α-times differentiable function ($\alpha \in \mathbb{N}_+$). Take $s_1 > d/2$ and $s_2 > d'$ such that $s_1, s_2/2 \in \mathbb{N}_+$. Let $k_x^{s_1}$ and $k_z^{s_2}$ be Matérn kernels over $\mathcal{X}$ and $\mathcal{Z}$ respectively as defined in (2). Assume $X \sim P$ and $\hat{X} = \{(x_i, w_i)\}_{i=1}^n \sim S$ satisfy 1. Moreover, assume that P and the marginals of $x_1, \ldots x_n$ have a common compact support. Suppose that, for some constants $b > 0$ and $0 < c \le 1/2$:*

(i) $\mathbb{E}_S \left[\|\hat{\mu}_X - \mu_X\|_{k_x^{s_1}}^2 \right] = O(n^{-2b})$;
(ii) $\sum_{i=1}^n w_i^2 = O(n^{-2c})$ *(with probability 1)* .

Let $\theta = \min(\frac{s_2}{2s_1}, \frac{\alpha}{s_1}, 1)$ and assume $\theta b - (1/2 - c)(1 - \theta) > 0$. Then

$$\mathbb{E}_S \left[\left\| \hat{\mu}_{f(X)} - \mu_{f(X)} \right\|_{k_z^{s_2}}^2 \right] = O \left((\log n)^{d'} n^{-2 (\theta b - (1/2 - c)(1 - \theta))} \right). \tag{3}$$

Before we provide a short sketch of the proof, let us briefly comment on this result. As a benchmark, remember that when $x_1, \ldots x_n$ are i.i.d. observations from X and $\hat{X} = \{(x_i, 1/n)\}_{i=1}^n$, we get $\|\hat{\mu}_{f(X)} - \mu_{f(X)}\|^2 = O_P(n^{-1})$, which was recently shown to be a minimax optimal rate (Tolstikhin et al., 2016). How do we compare to this benchmark? In this case we have $b = c = 1/2$ and our rate is defined by θ. If f is smooth enough, say $\alpha > d/2 + 1$, and by setting $s_2 > 2s_1 = 2\alpha$, we recover the $O(n^{-1})$ rate up to an extra $(\log n)^{d'}$ factor.

However, Theorem 2 applies to much more general settings. Importantly, it makes no i.i.d. assumptions on the data points and weights, allowing for complex interdependences. Instead, it asks the convergence of the estimator $\hat{\mu}_X$ to the embedding μ_X to be sufficiently fast. On the downside, the upper bound is affected by the smoothness of f, even in the i.i.d. setting: if $\alpha \ll d/2$ the rate will become slower, as $\theta = \alpha/s_1$. Also, the rate depends both on d and d'. Whether these are artefacts of our proof remains an open question.

Proof. Here we sketch the main ideas of the proof and develop the details in Appendix C. Throughout the proof, C will designate a constant that depends neither on the sample size n nor on the variable R (to be introduced). C may however change from line to line. We start by showing that:

$$\mathbb{E}_S \left[\left\| \hat{\mu}_{f(X)}^{k_z} - \mu_{f(X)}^{k_z} \right\|_{k_z}^2 \right] = (2\pi)^{\frac{d'}{2}} \int_{\mathcal{Z}} \mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz, \tag{4}$$

where h is Matérn kernel over $\mathcal{Z}$ with smoothness parameter $s_2/2$. Second, we upper bound the integrand by roughly imitating the proof idea of Theorem 1 from Kanagawa et al. (2016). This eventually yields:

$$\mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] \le C n^{-2\nu}, \tag{5}$$

where $\nu := \theta b - (1/2 - c)(1 - \theta)$. Unfortunately, this upper bound does not depend on z and can not be integrated over the whole $\mathcal{Z}$ in (4). Denoting B_R the ball of radius R, centred on the origin of $\mathcal{Z}$, we thus decompose the integral in (4) as:

$$\int_{\mathcal{Z}} \mathbb{E} \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz$$
$$= \int_{B_R} \mathbb{E} \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz + \int_{\mathcal{Z} \backslash B_R} \mathbb{E} \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz.$$

On B_R we upper bound the integral by (5) times the ball's volume (which grows like R^d):

$$\int_{B_R} \mathbb{E} \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz \le C R^d n^{-2\nu}. \tag{6}$$

On $\mathcal{X} \backslash B_R$, we upper bound the integral by a value that decreases with R, which is of the form:

$$\int_{\mathcal{Z} \backslash B_R} \mathbb{E} \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz \le C n^{1-2c} (R - C')^{s_2 - 2} e^{-2(R - C')} \tag{7}$$

6

with $C' > 0$ being a constant smaller than R. In essence, this upper bound decreases with R because $[\hat{\mu}^h_{f(X)} - \mu^h_{f(X)}](z)$ decays with the same speed as h when $\|z\|$ grows indefinitely. We are now left with two rates, (6) and (7), which respectively increase and decrease with growing R. We complete the proof by balancing these two terms, which results in setting $R \approx (\log n)^{1/2}$. $\qquad\square$

3 Functions of Multiple Arguments

The previous section applies to functions f of one single variable X. However, we can apply its results to functions of multiple variables if we take the argument X to be a tuple containing multiple values. In this section we discuss how to do it using two input variables from spaces $\mathcal{X}$ and $\mathcal{Y}$, but the results also apply to more inputs. To be precise, our input space changes from $\mathcal{X}$ to $\mathcal{X} \times \mathcal{Y}$, input random variable from X to (X, Y), and the kernel on the input space from k_x to k_{xy}.

To apply our results from Section 2, all we need is a consistent estimator $\hat{\mu}_{(X,Y)}$ of the joint embedding $\mu_{(X,Y)}$. There are different ways to get such an estimator. One way is to sample (x'_i, y'_i) i.i.d. from the joint distribution of (X, Y) and construct the usual empirical estimator, or approximate it using reduced set methods. Alternatively, we may want to construct $\hat{\mu}_{(X,Y)}$ based only on consistent estimators of μ_X and μ_Y. For example, this is how $\hat{\mu}_3$ was defined in Section 1.3. Below we show that this can indeed be done if X and Y are independent.

3.1 Application to Section 1.3

Following Schölkopf et al. (2015), we consider two independent random variables $X \sim P_x$ and $Y \sim P_y$. Their joint distribution is $P_x \otimes P_y$. Consistent estimators of their embeddings are given by $\hat{\mu}_X = \sum_{i=1}^n w_i k_x(x_i, .)$ and $\hat{\mu}_Y = \sum_{j=1}^n u_j k_y(y_i, .)$. In this section we show that $\hat{\mu}_{f(X,Y)} = \sum_{i,j=1}^n w_i u_j k_z\big(f(x_i, y_j), .\big)$ is a consistent estimator of $\mu_{f(X,Y)}$.

We choose a product kernel $k_{xy}\big((x_1, y_1), (x_2, y_2)\big) = k_x(x_1, x_2)k_y(y_1, y_2)$, so the corresponding RKHS is a tensor product $\mathcal{H}_{k_{xy}} = \mathcal{H}_{k_x} \otimes \mathcal{H}_{k_y}$ (Steinwart and Christmann, 2008, Lemma 4.6) and the mean embedding of the product random variable (X, Y) is a tensor product of their marginal mean embeddings $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$. With consistent estimators for the marginal embeddings we can estimate the joint embedding using their tensor product

$$\hat{\mu}_{(X,Y)} = \hat{\mu}_X \otimes \hat{\mu}_Y = \sum_{i,j=1}^n w_i u_j k_x(x_i, .) \otimes k_y(y_j, .) = \sum_{i,j=1}^n w_i u_j k_{xy}\big((x_i, y_j), (., .)\big).$$

If points are i.i.d. and $w_i = u_i = 1/n$, this reduces to the U-statistic estimator $\hat{\mu}_2$ from Section 1.3.

Lemma 3. *Let $(s_n)_n$ be any positive real sequence converging to zero. Suppose $k_{xy} = k_x k_y$ is a product kernel, $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$, and $\hat{\mu}_{(X,Y)} = \hat{\mu}_X \otimes \hat{\mu}_Y$. Then:*

$$\begin{cases} \|\hat{\mu}_X - \mu_X\|_{k_x} = O(s_n); \\ \|\hat{\mu}_Y - \mu_Y\|_{k_y} = O(s_n) \end{cases} \quad \text{implies} \quad \left\|\hat{\mu}_{(X,Y)} - \mu_{(X,Y)}\right\|_{k_{xy}} = O(s_n).$$

Proof. For a detailed expansion of the first inequality see Appendix B.

$$\left\|\hat{\mu}_{(X,Y)} - \mu_{(X,Y)}\right\|_{k_{xy}} \leq \|\mu_X\|_{k_x} \|\hat{\mu}_Y - \mu_Y\|_{k_y} + \|\mu_Y\|_{k_y} \|\hat{\mu}_X - \mu_X\|_{k_x}$$
$$+ \|\hat{\mu}_X - \mu_X\|_{k_x} \|\hat{\mu}_Y - \mu_Y\|_{k_y} = O(s_n) + O(s_n) + O(s_n^2) = O(s_n). \qquad\square$$

Corollary 4. *If $\hat{\mu}_X \xrightarrow[n\to\infty]{} \mu_X$ and $\hat{\mu}_Y \xrightarrow[n\to\infty]{} \mu_Y$, then $\hat{\mu}_{(X,Y)} \xrightarrow[n\to\infty]{} \mu_{(X,Y)}$.*

Together with the results from Section 2 this lets us reason about estimators resulting from applying functions to multiple independent random variables. Write

$$\hat{\mu}_{XY}^{k_{xy}} = \sum_{i,j=1}^n w_i u_j k_{xy}\big((x_i, y_j), .\big) = \sum_{\ell=1}^{n^2} \omega_\ell k_{xy}(\xi_\ell, .),$$

where ℓ enumerates the (i, j) pairs and $\xi_\ell = (x_i, y_j)$, $\omega_\ell = w_i u_j$. Now if $\hat{\mu}_X^{k_x} \to \mu_X^{k_x}$ and $\hat{\mu}_Y^{k_y} \to \mu_Y^{k_y}$ then $\hat{\mu}_{XY}^{k_{xy}} \to \mu_{(X,Y)}^{k_{xy}}$ (according to Corollary 4) and Theorem 1 shows that $\sum_{i,j=1}^{n} w_i u_j k_z\big(f(x_i, y_j), .\big)$ is consistent as well. Unfortunately, we cannot apply Theorem 2 to get the speed of convergence, because a product of Matérn kernels is not a Matérn kernel any more.

One downside of this overall approach is that the number of expansion points used for the estimation of the joint increases exponentially with the number of arguments of f. This can lead to prohibitively large computational costs, especially if the result of such an operation is used as an input to another function of multiple arguments. To alleviate this problem, we may use reduced expansion set methods before or after applying f, as we did for example in Section 1.2.

To conclude this section, let us summarize the implications of our results for two practical scenarios that should be distinguished.

> ▷ If we have separate samples from two random variables X and Y, then our results justify how to provide an estimate of the mean embedding of $f(X, Y)$ provided that X and Y are *independent*. The samples themselves need not be i.i.d. — we can also work with weighted samples computed, for instance, by a reduced set method.
> ▷ How about *dependent* random variables? For instance, imagine that $Y = -X$, and $f(X, Y) = X + Y$. Clearly, in this case the distribution of $f(X, Y)$ is a delta measure on 0, and there is no way to predict this from separate samples of X and Y. However, it should be stressed that our results (consistency and finite sample bound) apply even to the case where X and Y are dependent. In that case, however, they require a consistent estimator of the joint embedding $\mu_{(X,Y)}$.
> ▷ It is also sufficient to have a reduced set expansion of the embedding of the joint distribution. This setting may sound strange, but it potentially has significant applications. Imagine that one has a large database of user data, sampled from a joint distribution. If we expand the joint's embedding in terms of *synthetic* expansion points using a reduced set construction method, then we can pass on these (weighted) synthetic expansion points to a third party without revealing the original data. Using our results, the third party can nevertheless perform arbitrary continuous functional operations on the joint distribution in a consistent manner.

4 Conclusion and future work

This paper provides a theoretical foundation for using kernel mean embeddings as approximate representations of random variables in scenarios where we need to apply functions to those random variables. We show that for continuous functions f (including all functions on discrete domains), consistency of the mean embedding estimator of a random variable X implies consistency of the mean embedding estimator of $f(X)$. Furthermore, if the kernels are Matérn and the function f is sufficiently smooth, we provide bounds on the convergence rate. Importantly, our results apply beyond i.i.d. samples and cover estimators based on expansions with interdependent points and weights. One interesting future direction is to improve the finite-sample bounds and extend them to general radial and/or translation-invariant kernels.

Our work is motivated by the field of probabilistic programming. Using our theoretical results, kernel mean embeddings can be used to generalize functional operations (which lie at the core of all programming languages) to distributions over data types in a principled manner, by applying the operations to the points or approximate kernel expansions. This is in principle feasible for any data type provided a suitable kernel function can be defined on it. We believe that the approach holds significant potential for future probabilistic programming systems.

References

R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Academic Press, 2003.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Pure and Applied Mathematics. Elsevier Science, 1988.

A. Berlinet and C. Thomas-Agnan. *RKHS in probability and statistics*. Springer, 2004.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *UAI*, 2010.

K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, 2007. Edited by Alan Jeffrey and Daniel Zwillinger.

M. Kalos and P. Whitlock. *Monte Carlo Methods*. Wiley, 2008.

M. Kanagawa, B. K. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. *arXiv:1605.07254 [stat]*, 2016. arXiv: 1605.07254.

Y. Katznelson. *An Introduction to Harmonic Analysis*. Cambridge University Press, 2004.

M. Korzeń and S. Jaroszewicz. PaCAL: A Python package for arithmetic computations with random variables. *Journal of Statistical Software*, 57(10), 2014.

S. Lacoste-Julien, F. Lindsten, and F. Bach. Sequential kernel herding : Frank-Wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, volume 38, pages 544–552, 2015.

A. Mathai. A review of the different techniques used for deriving the exact distributions of multivariate test criteria. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 39–60, 1973.

K. McKinley. Programming the world of uncertain things (keynote). In *ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 1–2, 2016.

D. Milios. *Probability Distributions as Program Variables*. PhD thesis, University of Edinburgh, 2009.

S. Poisson. *Recherches sur la probabilitédes jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. 1837.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015.

C. Scovel, D. Hush, I. Steinwart, and J. Theiler. Radial kernels and their reproducing kernel hilbert spaces. *Journal of Complexity*, 26, 2014.

C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. Technical report, Max Planck Institute for Intelligent Systems, 2016.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*, pages 1–8, 2009.

M. D. Springer. *The Algebra of Random Variables*. Wiley, 1979.

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

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

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

I. Steinwart and C. Scovel. Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

I. Tolstikhin, B. Sriperumbudur, and K. Muandet. Minimax Estimation of Kernel Mean Embeddings. *arXiv:1602.04361 [math, stat]*, 2016.

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

R. Williamson. *Probabilistic Arithmetic*. PhD thesis, University of Queensland, 1989.

A   Detailed Proof of Theorem 1

Proof.

$$\left\| \hat{\mu}_Q^{k_z} - \mu_Q^{k_z} \right\|_{k_z}^2 = \left\| \sum_{i=1}^n w_i k_z(f(x_i), .) - \mathbb{E}\left[k_z(f(X), .)\right] \right\|_{k_z}^2$$

$$= \langle \sum_{i=1}^n w_i k_z(f(x_i), .) - \mathbb{E}\left[k_z(f(X), .)\right], \sum_{j=1}^n w_j k_z(f(x_j), .) - \mathbb{E}\left[k_z(f(X'), .)\right] \rangle$$

$$= \sum_{i,j=1}^n w_i w_j \langle k_z(f(x_i), .), k_z(f(x_j), .) \rangle - 2 \sum_{i=1}^n w_i \mathbb{E}\left[\langle k_z(f(x_i), .), k_z(f(X), .)\rangle\right] + \mathbb{E}\left[\langle k_z(f(X), .), k_z(f(X'), .)\rangle\right]$$

$$= \sum_{i,j=1}^n w_i w_j k_z\big(f(x_i), f(x_j)\big) - 2 \sum_{i=1}^n w_i \mathbb{E}\left[k_z\big(f(x_i), f(X)\big)\right] + \mathbb{E}\left[k_z(f(X), f(X'))\right]$$

$$= \sum_{i,j=1}^n w_i w_j \widetilde{k}_x(x_i, x_j) - 2 \sum_{i=1}^n w_i \mathbb{E}\left[\widetilde{k}_x(x_i, X)\right] + \mathbb{E}\left[\widetilde{k}_x(X, X')\right]$$

$$= \sum_{i,j=1}^n w_i w_j \langle \widetilde{k}_x(x_i, .), \widetilde{k}_x(x_j, .)\rangle - 2 \sum_{i=1}^n w_i \mathbb{E}\left[\langle \widetilde{k}_x(x_i, .), \widetilde{k}_x(X, .)\rangle\right] + \mathbb{E}\left[\langle \widetilde{k}_x(X, .), \widetilde{k}_x(X', .)\rangle\right]$$

$$= \langle \sum_{i=1}^n w_i \widetilde{k}_x(x_i, .) - \mathbb{E}\left[\widetilde{k}_x(X, .)\right], \sum_{j=1}^n w_j \widetilde{k}_x(x_j, .) - \mathbb{E}\left[\widetilde{k}_x(X', .)\right] \rangle$$

$$= \left\| \sum_{i=1}^n w_i \widetilde{k}_x(x_i, .) - \mathbb{E}\left[\widetilde{k}_x(X, .)\right] \right\|_{\widetilde{k}_x}^2 = \left\| \hat{\mu}_X^{\widetilde{k}_x} - \mu_X^{\widetilde{k}_x} \right\|_{\widetilde{k}_x}^2 \xrightarrow[n\to\infty]{} 0. \qquad \square$$

B   Detailed Proof of Lemma 3

Proof.

$$\left\| \hat{\mu}_{XY}^{k_{xy}} - \mu_{XY}^{k_{xy}} \right\|_{k_{xy}} = \left\| \hat{\mu}_X^{k_x} \otimes \hat{\mu}_Y^{k_y} - \mu_X^{k_x} \otimes \mu_Y^{k_y} \right\|_{k_{xy}}$$

$$= \left\| \hat{\mu}_X^{k_x} \otimes \hat{\mu}_Y^{k_y} - \hat{\mu}_X^{k_x} \otimes \mu_Y^{k_y} + \hat{\mu}_X^{k_x} \otimes \mu_Y^{k_y} - \mu_X^{k_x} \otimes \mu_Y^{k_y} \right\|_{k_{xy}}$$

$$= \left\| \hat{\mu}_X^{k_x} \otimes (\hat{\mu}_Y^{k_y} - \mu_Y^{k_y}) + (\hat{\mu}_X^{k_x} - \mu_X^{k_x}) \otimes \mu_Y^{k_y} \right\|_{k_{xy}}$$

$$\leq \left\| \hat{\mu}_X^{k_x} \right\|_{k_x} \left\| \hat{\mu}_Y^{k_y} - \mu_Y^{k_y} \right\|_{k_y} + \left\| \mu_Y^{k_y} \right\|_{k_y} \left\| \hat{\mu}_X^{k_x} - \mu_X^{k_x} \right\|_{k_x}$$

$$= \left\| \mu_X^{k_x} + \hat{\mu}_X^{k_x} - \mu_X^{k_x} \right\|_{k_x} \left\| \hat{\mu}_Y^{k_y} - \mu_Y^{k_y} \right\|_{k_y} + \left\| \mu_Y^{k_y} \right\|_{k_y} \left\| \hat{\mu}_X^{k_x} - \mu_X^{k_x} \right\|_{k_x}$$

$$\leq \left\| \mu_X^{k_x} \right\|_{k_x} \left\| \hat{\mu}_Y^{k_y} - \mu_Y^{k_y} \right\|_{k_y} + \left\| \mu_Y^{k_y} \right\|_{k_y} \left\| \hat{\mu}_X^{k_x} - \mu_X^{k_x} \right\|_{k_x}$$

$$\qquad + \left\| \hat{\mu}_X^{k_x} - \mu_X^{k_x} \right\|_{k_x} \left\| \hat{\mu}_Y^{k_y} - \mu_Y^{k_y} \right\|_{k_y}$$

$$= O(s_n) + O(s_n) + O(s_n^2) = O(s_n + s_n^2). \qquad \square$$

C   Detailed Proof of Theorem 2

C.1   Notations, Reminders and Preliminaries

For any function $\psi \in \mathrm{L}^1(\mathbb{R}^d)$ and any finite (signed or complex regular Borel) measure ν over $\mathbb{R}^d$, we define their convolution as:

$$\nu * \psi(x) := \int \psi(x - x')\, d\nu(x')\,.$$

We define the Fourier and inverse Fourier transforms of ψ and ν as

$$\mathscr{F}\,\psi(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle\omega,x\rangle} \psi(x)\,dx \quad \text{and} \quad \mathscr{F}\,\nu(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle\omega,x\rangle}\,d\nu(x)\,,$$

$$\mathscr{F}^{-1}\,\psi(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{i\langle\omega,x\rangle} \psi(x)\,dx \quad \text{and} \quad \mathscr{F}^{-1}\,\nu(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{i\langle\omega,x\rangle}\,d\nu(x)\,.$$

Fourier transforms are of particular interest when working with translation-invariant kernel because of Bochner's theorem. Here we quote Wendland (2004, Theorem 6.6), but add a useful second sentence, which is immediate to show.

Theorem 5 (Bochner). *A continuous function $\psi : \mathbb{R}^d \to \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite, nonnegative Borel measure ν over $\mathbb{R}^d$. Moreover, ψ is real-valued if and only if ν is symmetric.*

The next theorem, also quoted from Wendland (2004, Corollary 5.25), shows that the Fourier (inverse) transform may be seen as a unitary isomorphism from $L^2(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$.

Theorem 6 (Plancherel). *There exists an isomorphic mapping $T : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ such that:*

    (i) $\|Tf\|_{L^2(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)}$ for all $f \in L^2(\mathbb{R}^d)$.
    (ii) $Tf = \mathscr{F}\,f$ for any $f \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$.
    (iii) $T^{-1}g = \mathscr{F}^{-1}\,g$ for all $g \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$.

The isomorphism is uniquely determined by these properties.

We will call T the Fourier transform over L^2 and note it $\mathscr{F}$.

Remark 7. Combining Plancherel's and Bochner's theorems, we see that, if ψ is a continuous, positive definite (resp. and real-valued) function in $L^2(\mathbb{R}^d)$, then the measure ν from Bochner's theorem is absolutely continuous, and its density is $\mathscr{F}^{-1}\,\psi$. In particular, $\mathscr{F}^{-1}\,\psi$ is real-valued, nonnegative (resp. and symmetric).

Next, our proof of Theorem 2 will need the following result.

Lemma 8. *Let $\mathcal{Z} = \mathbb{R}^{d'}$, $\psi \in L^2(\mathbb{R}^d)$ such that $\mathscr{F}\,\psi \in L^1(\mathbb{R}^d)$. Let k be the translation-invariant kernel $k(z,z') := \psi(z-z')$ and $h(z) := \mathscr{F}^{-1}\sqrt{\mathscr{F}\,\psi}(z)$. Let Z be any random variable on $\mathcal{Z}$, and $\hat{Z} := \{(z_i,w_i)\}_{i=1}^n$. Then:*

$$\left\|\hat{\mu}_Z^k - \mu_Z^k\right\|_k^2 = (2\pi)^{\frac{d'}{2}} \int_{z\in\mathcal{Z}} \left|\hat{\mu}_Z^h - \mu_Z^h\right|^2 dz\,. \tag{8}$$

Proof. (of Lemma 8) For any finite (signed) measure ν over $\mathcal{Z} = \mathbb{R}^{d'}$, we define:

$$\mu_\nu^k := \int k(z,.)\,d\nu(z)\,.$$

Then we have:

$$\begin{aligned}
\left\|\mu_\nu^k\right\|_k^2 &= \int_{z\in\mathbb{R}^d} \int_{z'\in\mathbb{R}^d} \psi(z-z')\,d\nu(z)\,d\nu(z') \\
&= \int_{z\in\mathbb{R}^d} \int_{z'\in\mathbb{R}^d} \left((2\pi)^{-d'/2} \int_{\omega\in\mathbb{R}^d} e^{-i\langle\omega,z-z'\rangle}\,\mathscr{F}^{-1}\,\psi(\omega)\,d\omega\right) d\nu(z)\,d\nu(z') \\
&= \int_{\omega\in\mathbb{R}^d} (2\pi)^{-d'/2} \int_{z\in\mathbb{R}^d} \int_{z'\in\mathbb{R}^d} e^{-i\langle\omega,z-z'\rangle}\,d\nu(z)\,d\nu(z')\,\mathscr{F}^{-1}\,\psi(\omega)\,d\omega \\
&= \int_{\omega\in\mathbb{R}^d} (2\pi)^{d'/2}\,\mathscr{F}\,\nu(\omega)\,\mathscr{F}\,\nu(-\omega)\,\mathscr{F}^{-1}\,\psi(\omega)\,d\omega \\
&= (2\pi)^{d'/2} \int_{\omega\in\mathbb{R}^d} |\mathscr{F}\,\nu(\omega)|^2\,\mathscr{F}^{-1}\,\psi(\omega)\,d\omega
\end{aligned}$$

Second line uses the following: (i) ψ is continuous, because $\mathscr{F}\,\psi \in L^1(\mathbb{R}^d)$ (Riemann-Lebesgue lemma); (ii) Theorem 5 (Bochner) and Remark 7 from the Appendix. Third and fourth line use

Fubini's theorem. Last line uses the fact that $\mathscr{F}\nu(-\omega)$ is the complex conjugate of $\mathscr{F}\nu$ because $\mathscr{F}\psi$ is positive (thus real-valued).

Applying this with $\nu = \hat{Q} - Q$, where Q is the distribution of Z and $\hat{Q} := \sum_i w_i \delta_{z_i}$, we get:

$$
\begin{aligned}
\left\|\hat{\mu}_Z^k - \mu_Z^k\right\|_k^2 &= \left\|\mu_{\hat{Q}-Q}^k\right\|_k^2 \\
&= (2\pi)^{d'/2} \int_{\omega \in \mathbb{R}^d} \left|\mathscr{F}[\hat{Q}-Q](\omega)\right|^2 \mathscr{F}\psi(\omega)\, d\omega \\
&= (2\pi)^{d'/2} \int_{\omega \in \mathbb{R}^d} \left|\mathscr{F}[\hat{Q}-Q](\omega)\sqrt{\mathscr{F}\psi(\omega)}\right|^2 d\omega \\
&= (2\pi)^{d'/2} \int_{z \in \mathcal{Z}} \left|\mathscr{F}^{-1}\left[\mathscr{F}[\hat{Q}-Q]\sqrt{\mathscr{F}\psi}\right](z)\right|^2 dz \\
&= (2\pi)^{d'/2} \int_{z \in \mathcal{Z}} \left|[\hat{Q}-Q]*h(z)\right|^2 dz \\
&= (2\pi)^{d'/2} \int_{z \in \mathcal{Z}} \left|\sum_i w_i h(z-z_i) - \int h(z-s)\, dQ(s)\right|^2 dz \\
&= (2\pi)^{d'/2} \int_{z \in \mathcal{Z}} \left|\hat{\mu}_Z^h - \mu_Z^h(z)\right|^2 dz \,.
\end{aligned}
$$

Third line uses the fact that $\mathscr{F}\psi$ is positive (see Appendix, Remark 7). Fourth line uses Plancherel's theorem (see Appendix, Theorem 6). Fifth line uses the fact that the Fourier (inverse) transform of a product equals the convolutional product of the (inverse) Fourier transforms (Katznelson, 2004, Theorem 1.4, and its generalisation to finite measures p.145). □

We now state Theorem 1 from Kanagawa et al. (2016), which serves as basis to our proof. Slightly modifying[1] the notation of Adams and Fournier (2003, Chapter 7), for $0 < \theta < 1$ and $1 \leq q \leq \infty$ we will write $(E_0, E_1)_{\theta,q}$ to denote interpolation spaces, where E_0 and E_1 are Banach spaces that are continuously embedded into some topological Hausdorff vector space $\mathcal{E}$. Following Kanagawa et al. (2016), we also define $(E_0, E_1)_{1,2} := E_1$.

Theorem 9 (Kanagawa et al.)**.** *Let X be a random variable with distribution P and let $\{(x_i, w_i)\}_{i=1}^n$ be random variables with joint distribution S satisfying Assumption 1 (with corresponding distribution Q). Let $\hat{\mu}_X := \sum_i w_i k(x_i, .)$ be an estimator of $\mu_X := \int k(x, .)\, dP(x)$ such that for some constants $b > 0$ and $0 < c \leq 1/2$:*

(i) $\mathbb{E}_S\left[\|\hat{\mu}_X - \mu_X\|_k\right] = O(n^{-b})$,
(ii) $\mathbb{E}_S\left[\sum_i w_i^2\right] = O(n^{-2c})$

as $n \to \infty$. Let θ be a constant such that $0 < \theta \leq 1$.

Then, for any function $g : \mathbb{R}^d \to \mathbb{R}$ in $\left(\mathrm{L}^2(Q), \mathcal{H}_k\right)_{\theta,2}$, there exists a constant C, independent of n, such that:

$$
\mathbb{E}_S\left[\left|\sum_i w_i g(x_i) - \mathbb{E}_{X \sim P}[g(X)]\right|\right] \leq C\, n^{-\theta b + (1/2 - c)(1-\theta)} \,. \tag{9}
$$

In the proof of our finite sample guarantee, we will need the following slightly modified version of this result, where we (a) slightly modify condition (ii) by asking that it holds almost surely, and (b) consider squared norms in Condition (i) and (9).

Theorem 10 (Kanagawa et al.)**.** *Let X be a random variable with distribution P and let $\{(x_i, w_i)\}$ be random variables with joint distribution S satisfying Assumption 1 (with corresponding distribution Q). Let $\hat{\mu}_X := \sum_i w_i k(x_i, .)$ be an estimator of $\mu_X := \int k(x, .)\, dP(x)$ such that for some constants $b > 0$ and $0 < c \leq 1/2$:*

---

[1]Adams and Fournier (2003) introduce interpolation spaces using so-called *J*- and *K-methods*, resulting in two notations $(E_0, E_1)_{\theta,q;J}$ (Definition 7.12) and $(E_0, E_1)_{\theta,q;K}$ (Definition 7.9) respectively. However, it follows from Theorem 7.16 that these two definitions are equivalent if $0 < \theta < 1$ and we simply drop the K and J subindices.

(i) $\mathbb{E}_S\left[\|\hat{\mu}_X - \mu_X\|_k^2\right] = O(n^{-2b})$,

(ii) $\sum_{i=1}^n w_i^2 = O(n^{-2c})$ *(with S-probability 1)* ,

as $n \to \infty$. *Let* θ *be a constant such that* $0 < \theta \le 1$.

Then, for any function $g : \mathbb{R}^d \to \mathbb{R}$ *in* $\left(\mathrm{L}^2(Q), \mathcal{H}_k\right)_{\theta,2}$, *there exists a constant* C, *independent of* n, *such that:*

$$\mathbb{E}_S\left[\left\|\sum_i w_i g(x_i) - \mathbb{E}_{X \sim P}[g(X)]\right\|\right] \le C\, n^{-2\,(\theta b - (1/2 - c)(1-\theta))}. \tag{10}$$

Proof. The proof of this adapted version of Kanagawa et al. (2016, Theorem 1) is almost a copy paste of the original proof, but with the appropriate squares to account for the modified condition (i), and with their f renamed to g here. The only slight non-trivial difference is in their Inequality (20). Replace their triangular inequality by Jensen's inequality to yield:

$$\mathbb{E}_S\left[\left|\sum_{i=1}^n w_i g(x_i) - \mathbb{E}_{X \sim P}[g(X)]\right|^2\right] \le 3\,\mathbb{E}_S\left[\left|\sum_{i=1}^n w_i g(x_i) - \sum_{i=1}^n w_i g_{\lambda_n}(x_i)\right|^2\right]$$

$$+ 3\,\mathbb{E}_S\left[\left|\sum_{i=1}^n w_i g_{\lambda_n}(x_i) - \mathbb{E}_{X \sim P}[g_{\lambda_n}(X)]\right|^2\right]$$

$$+ 3\,\mathbb{E}_S\left[\left|\mathbb{E}_{X \sim P}[g_{\lambda_n}(X)] - \mathbb{E}_{X \sim P}[g(X)]\right|^2\right],$$

where g and g_{λ_n} are the functions that they call f and f_{λ_n}. □

We are now ready to prove Theorem 2.

C.2   Proof of Theorem 2

Proof. This proof is self-contained: the sketch from the main part is not needed. Throughout the proof, C designates constants that depend neither on sample size n nor on radius R (to be introduced). But their value may change from line to line.

Let ψ be such that $k_z^{s_2}(z, z') = \psi(z - z')$. Then $\mathscr{F}\psi(\omega) = (1 + \|\omega\|_2^2)^{-s_2}$ (Wendland, 2004, Chapter 10). Applying Lemma 8 to the Matérn kernel $k_z^{s_2}$ thus yields:

$$\mathbb{E}_S\left[\left\|\hat{\mu}_{f(X)}^{k_z^{s_2}} - \mu_{f(X)}^{k_z^{s_2}}\right\|_{k_z^{s_2}}^2\right] = (2\pi)^{\frac{d'}{2}} \int_{\mathcal{Z}} \mathbb{E}_S\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] dz, \tag{11}$$

where $h = \mathscr{F}^{-1}\sqrt{\mathscr{F} k_z^{s_2}}$ is again a Matérn kernel, but with smoothness parameter $s_2/2 > d'/2$.

Step 1: Applying Theorem 10

We now want to upper bound the integrand by using Theorem 10. To do so, let $\mathcal{K}$ be the common compact support of P and marginals of $x_1, \ldots, x_n$. Now, rewrite the integrand as:

$$\mathbb{E}_S\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] = \mathbb{E}_S\left[\left(\sum_i w_i h\big(f(x_i) - z\big) - \mathbb{E}_{X \sim P}\big[h\big(f(X) - z\big)\big]\right)^2\right]$$

$$= \mathbb{E}_S\left[\left(\sum_i w_i h\big(f(x_i) - z\big)\varphi_{\mathcal{K}}(x_i) - \mathbb{E}_{X \sim P}\big[h\big(f(X) - z\big)\varphi_{\mathcal{K}}(X)\big]\right)^2\right]$$

$$= \mathbb{E}_S\left[\left(\sum_i w_i g_z(x_i) - \mathbb{E}_{X \sim P}[g_z(X)]\right)^2\right], \tag{12}$$

13

where $\varphi_{\mathcal{K}}$ is any smooth function ≤ 1, with compact support, that equals 1 on a neighbourhood of $\mathcal{K}$ and where $g_z(x) := h(f(x) - z)\varphi_{\mathcal{K}}(x)$.

To apply Theorem 10, we need to prove the existence of $0 < \theta \leq 1$ such that $g_z \in \left(L^2(Q), \mathcal{H}_{k_x^{s_1}}\right)_{\theta,2}$ for each $z \in \mathcal{Z}$. We will prove this fact in two steps: (a) first we show that $g_z \in \left(L^2(\mathbb{R}^d), \mathcal{H}_{k_x^{s_1}}\right)_{\theta,2}$ for each $z \in \mathcal{Z}$ and certain choice of θ and (b) we argue that $\left(L^2(\mathbb{R}^d), \mathcal{H}_{k_x^{s_1}}\right)_{\theta,2}$ is continuously embedded in $\left(L^2(Q), \mathcal{H}_{k_x^{s_1}}\right)_{\theta,2}$.

Step (a): Note that $g_z \in \mathcal{W}_2^{\min(\alpha, s_2/2)}(\mathbb{R}^d)$ because f is α-times differentiable, $h \in \mathcal{W}_2^{s_2/2}(\mathbb{R}^{d'})$ (thus g_z is $\min(\alpha, s_2/2)$-times differentiable in the distributional sense), and g_z has compact support (thus meets the integrability conditions of Sobolev spaces). As $k_x^{s_1}$ is a Matérn kernel with smoothness parameter s_1, its associated RKHS $\mathcal{H}_{k_x^{s_1}}$ is the Sobolev space $\mathcal{W}_2^{s_1}(\mathbb{R}^d)$ (Wendland, 2004, Chapter 10). Now, if $s_1 \leq \min(\alpha, s_2/2)$, then $g_z \in \mathcal{W}_2^{s_1}(\mathbb{R}^d) = \mathcal{H}_{k_x^{s_1}} = \left(L^2(\mathbb{R}^d), \mathcal{W}_2^{s_1}(\mathbb{R}^d)\right)_{1,2}$ and step (a) holds for $\theta = 1$. Thus for the rest of this step, we assume $s_1 > \min(\alpha, s_2/2)$. It is known that $\mathcal{W}_2^s(\mathbb{R}^d) = B_{2,2}^s(\mathbb{R}^d)$ for $0 < s < \infty$ (Adams and Fournier, 2003, Page 255), where $B_{2,2}^s(\mathbb{R}^d)$ is the Besov space of smoothness s. It is also known that $B_{2,2}^s(\mathbb{R}^d) = \left(L^2(\mathbb{R}^d), \mathcal{W}_2^m(\mathbb{R}^d)\right)_{s/m,2}$ for any integer $m > s$ (Adams and Fournier, 2003, Page 230). Applying this to $\mathcal{W}_2^{\min(\alpha, s_2/2)}(\mathbb{R}^d)$ and denoting $s' = \min(\alpha, s_2/2)$ we get

$$g_z \in \mathcal{W}_2^{s'}(\mathbb{R}^d) = \left(L^2(\mathbb{R}^d), \mathcal{W}_2^{s_1}(\mathbb{R}^d)\right)_{s'/s_1,2} = \left(L^2(\mathbb{R}^d), \mathcal{H}_{k_x^{s_1}}\right)_{s'/s_1,2}, \qquad \forall z \in \mathcal{Z} .$$

Thus, whatever s_1, step (a) is always satisfied with $\theta := \min(\frac{\alpha}{s_1}, \frac{s_2}{2s_1}, 1) \leq 1$.

Step (b): If $\theta = 1$, then $\left(L^2(\mathbb{R}^d), \mathcal{H}_{k_x^{s_1}}\right)_{1,2} = \mathcal{H}_{k_x^{s_1}} = \left(L^2(Q), \mathcal{H}_{k_x^{s_1}}\right)_{1,2}$. Now assume $\theta < 1$. Note that $L^2(\mathbb{R}^d)$ is continuously embedded in $L^2(Q)$, because we assumed that Q has a bounded density. Thus Theorem V.1.12 of Bennett and Sharpley (1988) applies and gives the desired inclusion.

Now we apply Theorem 10, which yields a constant C_z independent of n such that:

$$\mathbb{E}_S\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] \leq C_z n^{-2\nu} ,$$

with $\nu := \theta b - (1/2 - c)(1 - \theta)$.

We now prove that the constants C_z are uniformly bounded. From Equations (18-19) of Kanagawa et al. (2016), it appears that $C_z = C \left\|T^{-\theta/2} g_z\right\|_{L^2(Q)}$, where C is a constant independent of z and $T^{-\theta/2}$ is defined as follows. Let T be the operator from $L^2(Q)$ to $L^2(Q)$ defined by

$$Tf := \int k_x(x,.) f(x) \, dQ(x).$$

It is continuous, compact and self-adjoint. Denoting $(e_i)_i$ an orthonormal basis of eigenfunctions in $L^2(Q)$ with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq 0$, let $T^{\theta/2}$ be the operator from $L^2(Q)$ to $L^2(Q)$ defined by:

$$T^{\theta/2} f := \sum_{i=1}^{\infty} \mu_i^{\theta/2} \langle e_i, f\rangle_{L^2(Q)} e_i .$$

Using Scovel et al. (2014, Corollary 4.9.i) together with Steinwart and Christmann (2008, Theorem 4.26.i) we conclude that $T^{\theta/2}$ is injective. Thus $\mu_i > 0$ for all i. Thus, if $\theta = 1$, Lemma 6.4 of Steinwart and Scovel (2012) shows that the range of $T^{\theta/2}$ is $[\mathcal{H}_k]_\sim$, the image of the canonical embedding of $\mathcal{H}_k$ into $L^2(Q)$. And as Q has full support, we may identify $[\mathcal{H}_k]_\sim$ and $\mathcal{H}_k = \left(L^2(Q), \mathcal{H}_k\right)_{\theta,2}$. Now, if $\theta < 1$, Theorem 4.6 of Steinwart and Scovel (2012) shows that the range of $T^{\theta/2}$ is $\left(L^2(Q), \mathcal{H}_k\right)_{\theta,2}$.

Thus the inverse operator $T^{-\theta/2}$ is well-defined, goes from $\left(L^2(Q), \mathcal{H}_k\right)_{\theta,2}$ to $L^2(Q)$ and can be written in the following form:

$$T^{-\theta/2} f := \sum_{i=1}^{\infty} \mu_i^{-\theta/2} \langle e_i, f\rangle_{L^2(Q)} e_i . \tag{13}$$

Using this, we get:

$$|C_z| = C \left\| T^{-\theta/2} g_z \right\|_{\mathrm{L}^2(Q)}$$

$$= C \left\| \sum_{i=1}^{\infty} \mu_i^{-\theta/2} \langle e_i, h(f(\cdot) - z)\varphi_{\mathcal{K}}(\cdot) \rangle_{\mathrm{L}^2(Q)} e_i \right\|_{\mathrm{L}^2(Q)}$$

$$\leq C \max_{z \in \mathcal{Z}} |h(z)| \left\| \sum_{i=1}^{\infty} \mu_i^{-\theta/2} \langle e_i, \varphi_{\mathcal{K}} \rangle_{\mathrm{L}^2(Q)} e_i \right\|_{\mathrm{L}^2(Q)}$$

$$= C \max_{z \in \mathcal{Z}} |h(z)| \left\| T^{-\theta/2} \varphi_{\mathcal{K}} \right\|_{\mathrm{L}^2(Q)},$$

which is a constant independent of z. Hereby, we used the fact that $\varphi_{\mathcal{K}} \in \left(\mathrm{L}^2(Q), \mathcal{H}_k \right)_{\theta,2}$, because it is infinitely smooth and has compact support. Thus we just proved that

$$\mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] \leq C n^{-2\nu} . \tag{14}$$

Step 2: Splitting the integral in two parts

However, now that this upper bound does not depend on z anymore, we cannot integrate over all $\mathcal{Z}$ ($= \mathbb{R}^{d'}$). Thus we now decompose the integral in (11) as:

$$\int_{\mathcal{Z}} \mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz$$

$$= \int_{B_R} \mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz + \int_{\mathcal{Z} \setminus B_R} \mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz , \tag{15}$$

where B_R denotes the ball of radius R, centred on the origin of $\mathcal{Z} = \mathbb{R}^{d'}$. We will upper bound each term by a function depending on R, and eventually make R depend on the sample size so as to balance both upper bounds.

On B_R we upper bound the integral by Rate (14) times the ball's volume (which grows like $R^{d'}$):

$$\int_{B_R} \mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right] dz \leq C R^{d'} n^{-2\nu} . \tag{16}$$

On $\mathcal{Z} \setminus B_R$ we upper bound the integral by a value that decreases with R. The intuition is that, according to (12), the integrand is the expectation of sums of Matérn functions, which are all centred on a compact domain. Thus it should decay exponentially with z outside of a sufficiently large ball. Next we turn to the formal argument.

Let us define $\|f\|_{\mathcal{K}} := \max_{x \in \mathcal{X}} \|f(x)\varphi_{\mathcal{K}}(x)\|$, which is finite because $f\varphi_{\mathcal{K}}$ is an α-times differentiable (thus continuous) function with compact support. Now, Matérn kernels are radial kernels, meaning that there exists a function $\tilde{h}$ over $\mathbb{R}$ such that $h(x) = \tilde{h}(\|x\|)$ (Tolstikhin et al., 2016, page 5). Moreover $\tilde{h}$ is strictly positive and decreasing. Using (12) we may write

$$\mathbb{E}_S \left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z) \right)^2 \right]$$

$$= \mathbb{E}_S \left[\left(\sum_i w_i h\big(f(x_i) - z\big)\varphi_{\mathcal{K}}(x_i) - \mathbb{E}_{X \sim P} \left[h\big(f(X) - z\big)\varphi_{\mathcal{K}}(X) \right] \right)^2 \right]$$

$$\leq \mathbb{E}_S \left[\left(\sum_i w_i h\big(f(x_i) - z\big)\varphi_{\mathcal{K}}(x_i) \right)^2 + \left(\mathbb{E}_{X \sim P} \left[h\big(f(X) - z\big)\varphi_{\mathcal{K}}(X) \right] \right)^2 \right]$$

$$\overset{(\dagger)}{\leq} \tilde{h}(\|z\| - \|f\|_{\mathcal{K}})^2 \mathbb{E}_S \left[\left(\left(\sum_i w_i \right)^2 + 1 \right) \right] ,$$

15

where we assumed $R > \|f\|_{\mathcal{K}}$ and used the fact that $\tilde{h}$ is a decreasing function in (†). Using Cauchy-Schwarz and applying hypothesis (ii), we get:

$$\mathbb{E}\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] \leq \tilde{h}(\|z\| - \|f\|_{\mathcal{K}})^2 \underset{S}{\mathbb{E}}\left[\left(n\left(\sum_i w_i^2\right) + 1\right)\right]$$

$$\leq C\, n^{1-2c}\, \tilde{h}(\|z\| - \|f\|_{\mathcal{K}})^2\,.$$

Let $\mathcal{S}_{d'}$ be the surface area of the unit sphere in $\mathbb{R}^{d'}$. We have:

$$\int_{\mathcal{Z}\setminus B_R} \mathbb{E}\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] dz \leq Cn^{1-2c} \int_{\mathcal{Z}\setminus B_R} \tilde{h}(\|z\| - \|f\|_{\mathcal{K}})^2\, dz$$

$$\overset{(\dagger)}{=} Cn^{1-2c} \int_{r=R-\|f\|_{\mathcal{K}}}^{+\infty} \tilde{h}(r)^2 \mathcal{S}_{d'}(r + \|f\|_{\mathcal{K}})^{d'-1}\, dr$$

$$\leq Cn^{1-2c}2^{d'-1} \int_{r=R-\|f\|_{\mathcal{K}}}^{+\infty} \tilde{h}(r)^2 \mathcal{S}_{d'} r^{d'-1}\, dr,$$

$$\text{(for } R \geq 2\,\|f\|_{\mathcal{K}}) \tag{17}$$

where (†) switches to radial coordinates. From Lemma 5.13 of Wendland (2004) we get, for any $r > 0$:

$$|\tilde{h}(r)| \leq Cr^{s_2/2 - d'/2}\sqrt{\frac{2\pi}{r}}e^{-r}e^{|d'/2 - s_2/2|^2/(2r)}.$$

Recalling that $s_2 > d'$ by assumption we have

$$\int_{\mathcal{Z}\setminus B_R} \mathbb{E}\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] dz$$

$$\leq Cn^{1-2c} \int_{r=R-\|f\|_{\mathcal{K}}}^{+\infty} r^{s_2-2}e^{-2r}e^{(s_2-d')^2/(4r)}\, dr$$

$$\leq Cn^{1-2c}e^{\frac{(s_2-d')^2}{4(R-\|f\|_{\mathcal{K}})}} \int_{r=R-\|f\|_{\mathcal{K}}}^{+\infty} r^{s_2-2}e^{-2r}\, dr.$$

Now, $s_2/2$ being by assumption a strictly positive integer, s_2-2 is an integer. Thus, using (Gradshteyn and Ryzhik, 2007, 2.321.2)

$$\int_{r=R-\|f\|_{\mathcal{K}}}^{+\infty} r^{s_2-2}e^{-2r}\, dr = e^{-2(R-\|f\|_{\mathcal{K}})}\left(\sum_{k=0}^{s_2-2} \frac{k!\binom{s_2-2}{k}}{2^{k+1}}(R-\|f\|_{\mathcal{K}})^{s_2-2-k}\right)$$

we continue by writing

$$\int_{\mathcal{Z}\setminus B_R} \mathbb{E}\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] dz$$

$$\leq Cn^{1-2c}e^{\frac{(s_2-d')^2}{4(R-\|f\|_{\mathcal{K}})}}e^{-2(R-\|f\|_{\mathcal{K}})}(R-\|f\|_{\mathcal{K}})^{s_2-2}$$

$$\text{(for } R \geq \|f\|_{\mathcal{K}} + 1) \tag{18}$$

$$\leq Cn^{1-2c}(R-\|f\|_{\mathcal{K}})^{s_2-2}e^{-2(R-\|f\|_{\mathcal{K}})} \tag{19}$$

$$\text{(for } R \geq \|f\|_{\mathcal{K}} + \frac{(s_2-d')^2}{4})\,. \tag{20}$$

Step 3: Choosing R to balance the terms

Compiling Equations (15), (16) and (19), we get:

$$\int_{\mathcal{Z}} \underset{S}{\mathbb{E}}\left[\left([\hat{\mu}_{f(X)}^h - \mu_{f(X)}^h](z)\right)^2\right] dz \leq CR^{d'}n^{-2\nu} + Cn^{1-2c}(R-\|f\|_{\mathcal{K}})^{s_2-2}e^{-2(R-\|f\|_{\mathcal{K}})}.$$

We now let R depend on the sample size n so that both rates be (almost) balanced. Ideally, defining $\gamma := \nu + 1/2 - c \geq 0$ and taking the log of these rates, we would thus solve

$$d' \log R = 2\gamma \log n + (s_2 - 2) \log(R - \|f\|_{\mathcal{K}}) - 2(R - \|f\|_{\mathcal{K}}), \tag{21}$$

and stick the solution R_s back into either of the two rates. Instead, we will upper bound R_s and stick the upper bound into the first rate, $R^d n^{-2\nu}$, which is the one increasing with R. More precisely, we will now show that for large enough n we can upper bound R_s essentially with $2\gamma \log(n) + \|f\|_{\mathcal{K}}$, which also satisfies conditions (17), (18) and (20). This will complete the proof.

Note that $s_2 > d' \geq 1$ and $s_2 \in \mathbb{N}_+$. First assume $s_2 = 2$. Then it is easy to check that (21) has a unique solution R_s satisfying $R_s \leq \gamma \log n + \|f\|_{\mathcal{K}}$ as long as $n \geq \exp\left(\frac{1 - \|f\|_{\mathcal{K}}}{\gamma}\right)$.

Next, assume $s_2 > 2$. Then for n large enough (21) has exactly 2 solutions and the larger of which will be denoted R_s. We now replace the right hand side of (21) with a lower bound $d' \log(R - \|f\|_{\mathcal{K}})$:

$$(d' - s_2 + 2) \log(R - \|f\|_{\mathcal{K}}) = 2\gamma \log n - 2(R - \|f\|_{\mathcal{K}}), \tag{22}$$

Clearly, (22) has one (if $d' - s_2 + 2 \geq 0$) or two (if $d' - s_2 + 2 < 0$) solutions, and in both cases the larger one, R_s^*, satisfies $R_s^* \geq R_s$. If $d' - s_2 + 2 \geq 0$ then, for $n \geq e^{1/\gamma}$, $R_s^* \leq \|f\|_{\mathcal{K}} + \gamma \log n$, because

$$(d' - s_2 + 2) \log(\gamma \log n) \geq 0$$

Finally, if $d' - s_2 + 2 < 0$ then the smaller solution of (22) decreases to $\|f\|_{\mathcal{K}}$ and the larger one R_s^* tends to infinity with growing n. Evaluating both sides of (22) for $R = \|f\|_{\mathcal{K}} + 2\gamma \log n$ we notice that

$$(d' - s_2 + 2) \log(2\gamma \log n) \geq -2\gamma \log n$$

for n large enough, as $\log n$ increases faster than $\log \log n$. This shows that $R_s^* \leq \|f\|_{\mathcal{K}} + 2\gamma \log n$.

Thus there exists a constant C independent of n, such that, for any $n \geq 1$:

$$\mathbb{E}\left[\left\|[\hat{\mu}_{f(X)}^{k_z} - \mu_{f(X)}^{k_z}]\right\|_{k_z}^2\right] \leq C(\log n)^{d'} n^{-2\nu} = O\left((\log n)^{d'} n^{-2\nu}\right). \qquad \square$$