Recovering Distributions from Gaussian RKHS Embeddings

Motonobu Kanagawa
Graduate University for Advanced Studies
kanagawa@ism.ac.jp

Kenji Fukumizu
Institute of Statistical Mathematics
fukumizu@ism.ac.jp

Abstract

Recent advances of kernel methods have yielded a framework for nonparametric statistical inference called RKHS embeddings, in which all probability distributions are represented as elements in a reproducing kernel Hilbert space, namely kernel means. In this paper, we consider the recovery of the information of a distribution from an estimate of the kernel mean, when a Gaussian kernel is used. To this end, we theoretically analyze the properties of a consistent estimator of a kernel mean, which is represented as a weighted sum of feature vectors. First, we prove that the weighted average of a function in a Besov space, whose weights and samples are given by the kernel mean estimator, converges to the expectation of the function. As corollaries, we show that the moments and the probability measures on intervals can be recovered from an estimate of the kernel mean. We also prove that a consistent estimator of the density of a distribution can be defined using a kernel mean estimator. This result confirms that we can in fact completely recover the information of distributions from RKHS embeddings.

1 Introduction

The RKHS embedding approach for nonparametric statistical inference has been developed in the machine learning community as a recent advance of kernel methods (Smola et al., 2007; Sriperumbudur et al., 2010; Song et al., 2013). This approach has been successfully applied to a wide variety of applications ranging from hypothesis testings (Gretton et al., 2012; Gretton et al., 2008) to machine learning problems including state-space modeling (Song et al., 2009; Fukumizu et al., 2011), belief propagation (Song et al., 2010; Song et al., 2011), predictive state representations (Boots et al., 2013), and reinforcement learning (Grünewälder et al., 2012; Nishiyama et al., 2012).

Let k be a positive definite kernel on a measurable space $\mathcal{X}$, and $\mathcal{H}$ be its reproducing kernel Hilbert space (RKHS). In this framework, any probability distribution P on $\mathcal{X}$ is represented as the expectation of feature vector $k(\cdot, x)$ in $\mathcal{H}$

$$m_P := \mathbb{E}_{X \sim P}[k(\cdot, X)] = \int k(\cdot, x) dP(x),$$

which is called the *kernel mean* of P. We can realize statistical inference by directly estimating the kernel mean m_P from data, instead of estimating the target distribution P itself.

In this paper, we deal with RKHS embeddings from another direction: given a good estimate $\hat{m}_P$ of the kernel mean m_P, *recover the information of the underlying distribution* P. This is motivated by the machine learning applications of RKHS embeddings. For example, in the application to state-space modeling (Song et al., 2009; Fukumizu et al., 2011), we often wish to predict future observations. In this case, P corresponds to the predictive distribution on the observation space. However, what we obtain from an algorithm is an estimate of its kernel mean $\hat{m}_P$, which is an element of an RKHS. Hence, to obtain meaningful information of the future observation, we need to extract the information of P from $\hat{m}_P$.

There have been some works on methods for recovering specific statistics of distribution P from kernel mean estimate $\hat{m}_P$. If the goal is to obtain a point estimate of P, a popular method is to represent $\hat{m}_P$ with the point in the original space called pre-image, i.e. $\arg\min_{x \in \mathcal{X}} \|k(\cdot, x) - \hat{m}_P\|_{\mathcal{H}}$ (Song et al., 2009). While having used to many applications to provide point estimation, this method is a heuristic since it is theoretically unclear what kind of information of

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

P the pre-image represents. On the other hand, a method for estimating the density of P from $\hat{m}_P$ has been proposed, assuming that P is a Gaussian mixture (Song et al., 2008; McCalman et al., 2013). This method may cause significant errors, however, if the assumption dose not hold.

Before going into the contributions of this paper, we restrict our considerations to RKHS embeddings using a *Gaussian kernel* on the Euclidian space $\mathbb{R}^d$. The main reasons are (1) it defines a injective mapping from distributions to an RKHS, and therefore can be used for RKHS embeddings (Fukumizu et al., 2004; Sriperumbudur et al., 2010), (2) it is ubiquitous in the applications of RKHS embeddings, and (3) theoretical properties of a Gaussian RKHS have been extensively investigated, e.g. (Steinwart and Christmann, 2008). Note the existing approaches mentioned above also use Gaussian kernels for computational feasibility.

Contributions. Our contributions are threefold. First, we analyze the theoretical properties of a consistent estimator of a kernel mean as a basis for the distribution recovery. In general, a finite sample estimate of a kernel mean takes a form of weighted average $\hat{m}_P = \sum_{i=1}^n w_i k(\cdot, X_i)$, where $X_1, \ldots, X_n$ are samples and $w_1, \ldots, w_n \in \mathbb{R}$ are (possibly negative) weights[1], which appears in all the machine learning applications mentioned above (Song et al., 2013). Assume that $\|\hat{m}_P - m_P\|_{\mathcal{H}} \to 0$ as $n \to \infty$, where $\|\cdot\|_{\mathcal{H}}$ denotes the RKHS norm. Let f be a function in a *Besov space*, which consists of functions with certain degree of smoothness and contains the Gaussian RKHS. Then we prove that the weighted average of f given by $\hat{m}_P = \sum_{i=1}^n w_i k(\cdot, X_i)$ converges to the expectation of f:

$$\sum_{i=1}^n w_i f(X_i) \to \mathbb{E}_{X\sim P}[f(X)].$$

This result is a generalization of the one known for functions in an RKHS (Smola et al., 2007).

Second, using the above result, we prove that certain statistics of P, namely its moments and measures on intervals, can be recovered from $\hat{m}_P$. Note that these quantities are defined as the expectations of polynomial functions or index functions, which are included in Besov spaces under certain assumptions. Hence, we can use the first result to prove that the expectations of these functions, and thus their corresponding quantities, can be consistently estimated from $\hat{m}_P$. Note that this result is not obvious beforehand without the first result, since polynomial and index functions are not included in a Gaussian RKHS (Minh, 2010).

[1]In general, X_i and w_i may depend of the sample size n, but we omit it in this paper for notational brevity.

Third, by employing arguments using a Besov space, which is similar to the first result, we prove that the density of P can be estimated from $\hat{m}_P$ *without any parametric assumptions on* P. We define a nonparametric estimator of the density of P using $\hat{m}_P$, and prove that it converges to the true density as $\hat{m}_P$ converges to m_P. This result shows that we can in fact completely recover the information of P from a consistent kernel mean estimator of the kernel mean m_P.

This paper is organized as follows. We briefly review RKHS embeddings and Besov spaces in Section 2. In Section 3, a convergence theorem for the expectation of a function in a Besov space is presented, and as corollaries we show that moments and measures on intervals can be estimated from a kernel mean estimate. In Section 4, we define a density estimator using a kernel mean estimate and show its convergence result. Proofs are given in Section 5.

2 Preliminaries

2.1 RKHS Embeddings

We first review RKHS embeddings. For details, we refer to the tutorial papers (Smola et al., 2007; Song et al., 2013).

Kernel mean. A measurable kernel $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a mesurable space $\mathcal{X}$ is called *positive definite*, if $\sum_{i=1}^n \sum_{j=1}^n c_i c_j k_{\mathcal{X}}(X_i, X_j) \geq 0$ for any $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{R}$ and $X_1, \ldots, X_n \in \mathcal{X}$. We will use the terminology *kernel* to refer to a function satisfying the positive definiteness. A kernel $k_{\mathcal{X}}$ uniquely defines a *reproducing kernel Hilbert space (RKHS)* $\mathcal{H}_{\mathcal{X}}$ such that the reproducing property $f(x) = \langle f, k_{\mathcal{X}}(\cdot, x)\rangle_{\mathcal{H}_{\mathcal{X}}}$ holds for all $f \in \mathcal{H}_{\mathcal{X}}$ and $x \in \mathcal{X}$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}_{\mathcal{X}}}$ denotes the inner-product of $\mathcal{H}_{\mathcal{X}}$. Let $\|\cdot\|_{\mathcal{H}_{\mathcal{X}}}$ denote the norm of $\mathcal{H}_{\mathcal{X}}$.

In the RKHS embedding approach, we represent any probability distribution P on $\mathcal{X}$ by the expectation of feature vector $k_{\mathcal{X}}(\cdot, x) \in \mathcal{H}_{\mathcal{X}}$:

$$m_P := \mathbb{E}_{X\sim P}[k_{\mathcal{X}}(\cdot, X)] = \int k_{\mathcal{X}}(\cdot, x) dP(x) \in \mathcal{H}_{\mathcal{X}},$$

which is called the *kernel mean* of P. If the kernel is *characteristic*, any probability distribution is uniquely determined by its kernel mean, i.e. $m_P = m_Q \Rightarrow P = Q$ holds for probability distributions P and Q (Fukumizu et al., 2004; Fukumizu et al., 2009; Sriperumbudur et al., 2010). A representative example of characteristic kernels is a Gaussian kernel $k_\gamma(x, x') = \exp(-\|x - x'\|^2/\gamma^2)$ on $\mathcal{X} = \mathbb{R}^d$, where $\gamma > 0$. This paper focuses on RKHS embeddings using a Gaussian kernel with $\mathcal{X} = \mathbb{R}^d$.

Finite sample estimte. In this framework, we aim to directly estimate kernel mean m_P from samples, instead of distribution P itself. If we have i.i.d. samples $X_1, \ldots, X_n$ from P, then m_P is estimated by the empirical mean $\hat{m}_P := \frac{1}{n}\sum_{i=1}^n k_{\mathcal{X}}(\cdot, X_i)$ with the rate $\|\hat{m}_P - m_P\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\frac{1}{2}})$ (Smola et al., 2007). In general, however, the kernel mean of P may be estimated with samples X_i of a distribution different from P, and therefore it has a form of weighted sum of feature vectors

$$\hat{m}_P := \sum_{i=1}^n w_i k_{\mathcal{X}}(\cdot, X_i) \tag{1}$$

with weights[2] $w_1, \ldots, w_n \in \mathbb{R}$ (Song et al., 2013).

For instance, suppose that we are given i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ from a joint distribution $P_{\mathcal{X}\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a measurable space, and that we wish to estimate the kernel mean of the conditional distribution $P := P_{\mathcal{X}|y}$ conditioned on $y \in \mathcal{Y}$. In this case, a consistent estimator of the kernel mean is given with the weights (Song et al., 2009)

$$w_i = ((G_X + n\varepsilon_n I_n)^{-1}\mathbf{k}_Y(y))_i, \tag{2}$$

where $G_X = (k_{\mathcal{X}}(X_i, X_j)) \in \mathbb{R}^{n\times n}$ is a kernel matrix, I_n is the identity, $\varepsilon_n > 0$ is a regularization constant, and $\mathbf{k}_Y(y) = (k_{\mathcal{Y}}(y, Y_1), \ldots, k_{\mathcal{Y}}(y, Y_n))^T \in \mathbb{R}^n$, where $k_{\mathcal{Y}}$ is a kernel on $\mathcal{Y}$. Other examples include the RKHS embedding realization of the sum rule, the chain rule and Bayes' rule (Song et al., 2009; Fukumizu et al., 2011). By combining these estimators, we can realize various applications of RKHS embeddings mentioned in Section 1, in which the estimates also take the form of Eq. (1).

Expectation of RKHS functions. It is known that the expectation of a function in the RKHS can be estimated with a kernel mean estimate (Smola et al., 2007). Let $\hat{m}_P = \sum_{i=1}^n w_i k_{\mathcal{X}}(\cdot, X_i)$ be a consistent estimate of m_P such that $\lim_{n\to\infty} \|\hat{m}_P - m_P\|_{\mathcal{H}_{\mathcal{X}}} = 0$. Then we have

$$\lim_{n\to\infty} \left| \sum_{i=1}^n w f(X_i) - \mathbb{E}_{X\sim P}[f(X)] \right| = 0, \quad \forall f \in \mathcal{H}_{\mathcal{X}}.$$

This is easily shown by $|\sum_i w_i f(X_i) - \mathbb{E}_{X\sim P}[f(X)]| = |\langle f, \hat{m}_P - m_P\rangle_{\mathcal{H}_{\mathcal{X}}}| \leq \|f\|_{\mathcal{H}_{\mathcal{X}}}\|\hat{m}_P - m_P\|_{\mathcal{H}_{\mathcal{X}}}$, since $\langle m_P, f\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_{X\sim P}[f(X)]$ and $\langle \hat{m}_P, f\rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_{i=1}^n w_i f(X_i)$ hold for any $f \in \mathcal{H}_{\mathcal{X}}$ by the reproducing property.

2.2 Besov Spaces

Let $\mathcal{X} \subset \mathbb{R}^d$ be a set. Here, we define the Besov space $B^{\alpha}_{2,\infty}(\mathcal{X})$ on $\mathcal{X}$ for any constant $\alpha > 0$. For details of Besov spaces, see (Adams and Fournier, 2003, Chapter 7; DeVore and Lorentz, 1993, Chapter 2). Let $L_p(\mathcal{X})$ be the Lebesgue space of order $p \in [1, \infty]$ with respect to the Lebesgue measure on $\mathcal{X}$ and $\|\cdot\|_{L_p(\mathcal{X})}$ be its norm. Let $r := \lfloor\alpha\rfloor + 1$, where $\lfloor\alpha\rfloor$ is the greatest integer smaller or equal to α. First, for any $h \in [0,\infty)^d \subset \mathbb{R}^d$ and $f \in L_2(\mathcal{X})$, we define a function $\Delta_h^r(f,\cdot) : \mathcal{X} \to \mathbb{R}$ by

$$\Delta_h^r(f,x) = \begin{cases} \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(x+jh) & \text{if } x \in \mathcal{X}_{r,h} \\ 0 & \text{if } x \notin \mathcal{X}_{r,h} \end{cases}$$

where $\mathcal{X}_{r,h} := \{x \in \mathcal{X} : x + sh \in \mathcal{X},\ \forall s \in [0,r]\}$. Then, we define a function $\omega_{r,L_2(\mathcal{X})}(f,\cdot) : [0,\infty) \to [0,\infty)$ by $\omega_{r,L_2(\mathcal{X})}(f,t) := \sup_{\|h\|\leq t} \|\Delta_h^r(f,\cdot)\|_{L_2(\mathcal{X})}, \forall t \geq 0$.

The Besov space $B^{\alpha}_{2,\infty}(\mathcal{X})$ is then defined by

$$B^{\alpha}_{2,\infty}(\mathcal{X}) := \{f \in L_2(\mathcal{X}) : |f|_{B^{\alpha}_{2,\infty}(\mathcal{X})} < \infty\}, \tag{3}$$

where $|f|_{B^{\alpha}_{2,\infty}(\mathcal{X})} := \sup_{t>0}(t^{-\alpha}\omega_{r,L_2(\mathcal{X})}(f,t))$ is the seminorm of $B^{\alpha}_{2,\infty}(\mathcal{X})$. The norm of $B^{\alpha}_{2,\infty}(\mathcal{X})$ is defined by $\|f\|_{B^{\alpha}_{2,\infty}(\mathcal{X})} := \|f\|_{L_2(\mathcal{X})} + |f|_{B^{\alpha}_{2,\infty}(\mathcal{X})}, \forall f \in B^{\alpha}_{2,\infty}(\mathcal{X})$.

Let $W_2^m(\mathbb{R}^d)$ be the Sobolev space with order $m \in \mathbb{N}$, which consists of functions whose (weak) derivatives up to order m exist and are included in $L_2(\mathbb{R}^d)$ (Adams and Fournier, 2003). Importantly, $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ contains $W_2^m(\mathbb{R}^d)$ if $m \geq \alpha$ (Edmunds and Triebel, 1996, pp.26-27 and p.44). Thus, the larger α is, the smoother the functions in $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ are. Relation $W_2^m(\mathbb{R}^d) \subset B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ easily shows that $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ includes functions such as (i) m-times continuously differentiable functions with compact supports and (ii) Gaussian functions $f(x) = A\exp(-B\|x-\mu\|^2)$ for any $A, B > 0$ and $\mu \in \mathbb{R}^d$. Moreover, the relation implies that $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ contains the Gaussian RKHS on $\mathbb{R}^d$ (Steinwart and Christmann, 2008, Theorem 4.48).

3 Main Theorem

Let $k_\gamma(x,x') := \exp(-\|x-x'\|^2/\gamma^2)$ be the Gaussian kernel on $\mathbb{R}^d$ with bandwidth $\gamma > 0$, $\mathcal{H}_\gamma$ be the RKHS of k_γ, and $\langle\cdot,\cdot\rangle_{\mathcal{H}_\gamma}$ and $\|\cdot\|_{\mathcal{H}_\gamma}$ be its inner-product and norm, respectively. Let P be a probability distribution on $\mathbb{R}^d$, $m_P = \mathbb{E}_{X\sim P}[k_\gamma(\cdot, X)]$ be the kernel mean of P, and $\hat{m}_P = \sum_{i=1}^n w_i k_\gamma(\cdot, X_i)$ be its consistent estimate. In Section 2.1, we saw that the expectation $\mathbb{E}_{X\sim P}[f(X)]$ can be estimated by $\sum_{i=1}^n w_i f(X_i) = \langle \hat{m}_P, f\rangle_{\mathcal{H}_\gamma}$ if f belongs to $\mathcal{H}_\gamma$.

In this section, we generalize this to functions in the Besov space $B^{\alpha}_{2,\infty}(\mathbb{R}^d)$, which contains $\mathcal{H}_\gamma$. Namely, we show in Theorem 1 below that $\sum_{i=1}^n w_i f(X_i)$ also converges to $\mathbb{E}_{X\sim P}[f(X)]$ for any $f \in B^{\alpha}_{2,\infty}(\mathbb{R}^d)$.

[2] Note that these weights may take negative values as the example of Eq. (2) shows.

Note that this is not obvious a priori, since we cannot write the estimate in the form of inner-product $\sum_i w_i f(X_i) = \langle \hat{m}_P, f \rangle_{\mathcal{H}_\gamma}$ if f dose not belong to $\mathcal{H}_\gamma$.

Theorem 1. *Let P and Q be probability distributions on $\mathbb{R}^d$. Assume that P and Q have densities that belong to $L_\infty(\mathbb{R}^d)$. Let $\hat{m}_P := \sum_{i=1}^n w_i k_\gamma(\cdot, X_i)$ be a consistent estimate of $m_P = \int k_\gamma(\cdot, x) dP(x)$ such that $\mathbb{E}[\|\hat{m}_P - m_P\|_{\mathcal{H}_\gamma}] = O(n^{-b})$ and $\mathbb{E}[\sum_{i=1}^n w_i^2] = O(n^{-2c})$ for some $0 < b, c \leq 1/2$ as $n \to \infty$, where $X_1, \ldots, X_n$ are i.i.d. samples from Q. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying $f \in B^\alpha_{2,\infty}(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$ for some $\alpha > 0$. Assume that $\alpha b - d(1/2 - c) > 0$. Then we have*

$$\mathbb{E}\left[\left|\sum_{i=1}^n w_i f(X_i) - \mathbb{E}_{X\sim P}[f(X)]\right|\right] = O\left(n^{-\frac{\alpha b - d(1/2-c)}{\alpha + d}}\right) \quad (n \to \infty). \tag{4}$$

The assumption $\alpha b - d(1/2 - c) > 0$ can be satisfied if α is large enough. For example, m-times continuously differentiable functions with compact supports satisfy the assumption if m is large enough. Note that if f belongs to $B^\alpha_{2,\infty}(\mathbb{R}^d)$ for arbitrarily large $\alpha > 0$, the rate (4) becomes $O(n^{-b+\xi})$ for arbitrarily small $\xi > 0$. This shows that the rate can be arbitrarily close to the convergence rate of $\hat{m}_P$ if f is very smooth. Examples of such a case include infinitely continuously differentiable functions with compact supports and Gaussian functions.

Note that the assumption that $X_1, \ldots, X_n$ are i.i.d. samples from some Q is natural. For example, suppose that $\hat{m}_P$ is given by the conditional embedding estimator Eq. (2) (Song et al., 2009). In this case, Q corresponds to the marginal distribution on $\mathcal{X}$ of the joint distribution $P_{\mathcal{XY}}$ that generates training samples $\{(X_i, Y_i)\}_{i=1}^n$. Other kernel mean estimators also satisfy the assumption (Song et al., 2013).

The conditions $\mathbb{E}[\|\hat{m}_P - m_P\|_{\mathcal{H}_\gamma}] = O(n^{-b})$ and $\mathbb{E}[\sum_{i=1}^n w_i^2] = O(n^{-2c})$ depend on the distributions P and Q and the way the estimator $\hat{m}_P$ is defined. For example, if $P = Q$ and the weights are uniform $w_i = 1/n$, then we have $b = c = 1/2$. We can also show that if the estimator is the conditional embedding Eq. (2), then $b = 1/8, c = 1/4$ (see the supplementary materials).

3.1 Polynomial Functions - Estimation of Moments

As a corollary of Theorem 1, we show that the expectation of a polynomial function, and thus the moments of P, can be estimated from $\hat{m}_P$, under the assumption that *the supports of P and Q are bounded*. Note, however, that we cannot directly apply Theorem 1 to polynomial functions since they do not satisfy $f \in B^\alpha_{2,\infty}(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. Here, we first show that the condition can be weakened to

$$f \in B^\alpha_{2,\infty}(B_R) \cap L_\infty(B_R), \tag{5}$$

where $B_R = \{x \in \mathbb{R}^d : \|x\| < R\}$ is an open ball with radius $R > 0$ that contains the supports of P and Q.

To this end, we use Stein's extension theorem (Stein, 1970, pp.180-192; Adams and Fournier, 2003, p.154 and p.230). Let $\mathcal{X} \subset \mathbb{R}^d$ be a set with *minimally smooth boundary* (Stein, 1970, p.189). Stein's extension theorem guarantees that for any $f \in B^\alpha_{2,\infty}(\mathcal{X})$, there exists $\mathfrak{E}(f) \in B^\alpha_{2,\infty}(\mathbb{R}^d)$ satisfying $\mathfrak{E}(f)(x) = f(x)$ for all $x \in \mathcal{X}$. Likewise, the theorem guarantees that for any $f \in L_\infty(\mathcal{X})$, there exists $\mathfrak{E}(f) \in L_\infty(\mathbb{R}^d)$ satisfying the same property. Extended function $\mathfrak{E}(f)$ is defined in a way independent of the function space on $\mathcal{X}$ to which f belongs (Stein, 1970, p.191).

Since B_R has minimally smooth boundary (Stein, 1970, p.189), Stein's extension theorem guarantees that for f satisfying Eq. (5), there exists $\mathfrak{E}(f) : \mathbb{R}^d \to \mathbb{R}$ such that $\mathfrak{E}(f) \in L_\infty(\mathbb{R}^d) \cap B^\alpha_{2,\infty}(\mathbb{R}^d)$ and $\mathfrak{E}(f)(x) = f(x), \quad \forall x \in B_R$. Then, applying Theorem 1 to $\mathfrak{E}(f)$, we obtain the rate (4) for $\mathbb{E}[|\sum_{i=1}^n w_i \mathfrak{E}(f)(X_i) - E_P[\mathfrak{E}(f)(X)]|]$. Since B_R contains the supports of P and Q, we have $\mathbb{E}[|\sum_{i=1}^n w_i \mathfrak{E}(f)(X_i) - E_P[\mathfrak{E}(f)(X)]|] = \mathbb{E}[|\sum_{i=1}^n w_i f(X_i) - E_P[f(X)]|]$. Thus, it turns out that the obtained rate is for $\mathbb{E}[|\sum_{i=1}^n w_i f(X_i) - E_P[f(X)]|]$.

Note that if f is polynomial, f satisfies Eq. (5) for arbitrarily large $\alpha > 0$. Thus, Theorem 1 combined with the above arguments yields the following corollary.

Corollary 1. *Assume the same conditions for P, Q, and $\hat{m}_P$ as in Theorem 1. Assume also that the supports of P and Q are bounded. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a polynomial function. Then for arbitrary small $\xi > 0$, we have*

$$\mathbb{E}\left[\left|\sum_{i=1}^n w_i f(X_i) - E_P[f(X)]\right|\right] = O\left(n^{-b+\xi}\right) \ (n \to \infty).$$

Note that the rate of Corollary 1 does not depend on the order of polynomial: this is mainly because of the assumption that the supports of P and Q are bounded, under which the corollary is derived.

We can use Corollary 1 to show that the moments about the mean of P can be estimated from $\hat{m}_P$. Let $d = 1$. Corollary 1 indicates the convergence in expectation of $\sum_{i=1}^n w_i X_i^k$ to the raw moment $\mathbb{E}_{X\sim P}[X^k]$, where $k \in \mathbb{N}$. Then, an estimate of the moment about the mean of order $\ell \in$

$\mathbb{N}$ can be defined by $\sum_{i=1}^n w_i(X_i - \sum_{j=1}^n w_j X_j)^\ell = \sum_{k=0}^\ell \binom{\ell}{k}(\sum_{i=1}^n w_i X_i^{\ell-k})(\sum_{j=1}^n w_j X_j)^k$, and its consistency in probability follows.

3.2 Index Functions - Estimation of Measures on Rectangles

Next, we show that the probability measure $P(\Omega)$ on any interval $\Omega \subset \mathbb{R}^d$ may be estimated from $\hat{m}_P$ as a corollary of Theorem 1. Here, we define an interval in $\mathbb{R}^d$ as $\Omega := [a_1, b_1] \times \cdots [a_d, b_d]$, where $-\infty < a_i < b_i < +\infty, i = 1, \ldots, d$. Note that we may use the estimated $P(\Omega)$ for making credible intervals from $\hat{m}_P$ in Bayesian inference applications (Fukumizu et al., 2011).

Let χ_Ω be the index function of Ω defined by

$$\chi_\Omega(x) = \begin{cases} 1 & (x \in \Omega) \\ 0 & (x \notin \Omega) \end{cases}. \tag{6}$$

Note that have $\mathbb{E}_{X\sim P}[\chi_\Omega(X)] = P(\Omega)$. Observing that $\sum_{i=1}^n w_i \chi_\Omega(X_i) = \sum_{X_i \in \Omega} w_i$, we define an estimator of $P(\Omega)$ by the sum of weights of points in Ω, i.e. $\sum_{X_i \in \Omega} w_i$. We can show that χ_Ω is included in $B^\alpha_{2,\infty}(\mathbb{R}^d)$ for any α with $0 < \alpha < 1/2$. Thus, we can apply Theorem 1 to χ_Ω and derive the following corollary.

Corollary 2. *Assume the same conditions for P, Q, and $\hat{m}_P$ as in Theorem 1. Let $\Omega := [a_1, b_1] \times \cdots [a_d, b_d]$. Assume that $b - d(1-2c) > 0$. Then for arbitrary small $\xi > 0$, we have*

$$\mathbb{E}\left[\left|\sum_{X_i \in \Omega} w_i - P(\Omega)\right|\right] = O\left(n^{-\frac{b-d(1-2c)}{1+2d}+\xi}\right).$$

We need the assumption $b - d(1-2c) > 0$ in Corollary 2 to guarantee the consistency of the estimator. For example, if $P = Q$ and $w_i = 1/n$ we have $b = c = 1/2$, and thus the assumption is satisfied. Note, however, that the assumption is strong: for example, if $\hat{m}_P$ is given by the estimator for conditional distributions Eq. 2, we have $b = 1/8$ and $c = 1/4$ and therefore the assumption is not satisfied.

Note that the only assumption of Corollary 2 for P is that its density is bounded. Thus, the density can be arbitrarily complicated and even discontinuous, and we therefore need such a strong condition on $\hat{m}_P$ for consistency. In other words, we may obtain better bounds by assuming additional conditions on P, e.g. smoothness of the density. In fact, numerical experiments (reported in the supplements) show that $\sum_{X_i \in \Omega} w_i$ converges to $P(\Omega)$ for the estimator for conditional distributions Eq. (2). Investigation for better bounds remains as a topic for a future research.

4 Recovery of the Density

Assume that P has a density function p. In this section, we show that we can estimate the density $p(x_0)$ at any fixed point $x_0 \in \mathbb{R}^d$ from $\hat{m}_P$, by defining its nonparametric estimator using $\hat{m}_P$. Let δ_{x_0} be the Dirac delta function at x_0. Then we have $p(x_0) = \int \delta_{x_0}(x) p(x) dx = \mathbb{E}_{X\sim P}[\delta_{x_0}(X)]$. Thus, intuitively, if we can define an estimator for the expectation of δ_{x_0} using $\hat{m}_P$, this would be an estimator of $p(x_0)$. Theorem 1 cannot be used in this case, however, since the delta function is included neither in Gaussian RKHS $\mathcal{H}_\gamma$ nor in Besov space $B^\alpha_{2,\infty}(\mathbb{R}^d)$.

Here, we introduce a new (smoothing) kernel for approximating the delta function, as for usual kernel density estimation. For brevity, we also use a Gaussian kernel as a smoothing kernel[3]

$$J_h(x - x_0) := \frac{1}{\pi^{d/2} h^d} \exp(-\|x - x_0\|^2 / h^2),$$

where $h > 0$. Let $J_{x_0,h} := J_h(\cdot - x_0)$. Then $J_{x_0,h}$ is included in $B^\alpha_{2,\infty}(\mathbb{R}^d)$ for arbitrarily large $\alpha > 0$, as shown in Section 2.2.

Thus, we can apply Theorem 1 to J_h for fixed h. A consistent estimator of $\mathbb{E}_{X\sim P}[J_{x_0,h}](X)$ is thus given by

$$\sum_{i=1}^n w_i J_{x_0,h}(X_i). \tag{7}$$

Note that this is not obvious without Theorem 1, since $J_{x_0,h}$ is not included in Gaussian RKHS $\mathcal{H}_\gamma$ if $h < \gamma$. On the other hand, we have

$$\lim_{h \to 0} \mathbb{E}_{X\sim P}[J_{x_0,h}(X)] = p(x_0).$$

By this argument, we expect that Eq. (7) converges to $p(x_0)$ if we take $h := h_n \to 0$ as $n \to \infty$. Theorem 2 below shows that Eq. (7) is in fact a consistent density estimator.

Theorem 2. *Assume the same conditions for P, Q, and $\hat{m}_P$ as in Theorem 1. Assume also that density p of P is Lipschitz. Then with $h_n = n^{-\frac{2b}{3d+2}+\xi}$ for an arbitrarily small $\xi > 0$, we have for all $x_0 \in \mathbb{R}^d$*

$$\mathbb{E}\left[\left|\sum_{i=1}^n w_i J_{h_n}(X_i - x_0) - p(x_0)\right|\right] = O\left(n^{-\frac{2b}{3d+2}+\xi}\right) \ (n \to \infty). \tag{8}$$

Note that the reason why α does not appear in the resulting rate (8) is that we take $\alpha \to \infty$ in the proof:

[3] We can also use any kernel that belongs to $B^\alpha_{2,\infty}(\mathbb{R}^d)$ for arbitrarily large $\alpha > 0$, as the proof of Theorem 2 only depends on this property in regard to the smoothing kernel.

recall that we have $J_h \in B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ for arbitrarily large $\alpha > 0$.

Theorem 2 shows that we can in fact completely recover the information of distribution P from an estimate of the corresponding kernel mean m_P. The assumptions Theorem 2 imposes on density p is boundedness and Lipschitz continuity. Comparing with (Song et al., 2008; McCalman et al., 2013), in which they assume that p is a Gaussian mixture, consistency of the estimator (7) can be guaranteed for a wider class of probability distributions.

For instance, if the kernel mean estimators for conditional distributions Eq. (2) or Bayes' rule (Fukumizu et al., 2011) are used, then their respective densities can be estimated with Eq. (7). Note that if $P = Q$ and $w_i = 1/n$, Eq. (7) exactly corresponds to the usual kernel density estimator. In this case, we have $b = c = 1/2$, and thus the rate becomes $O(n^{-\frac{1}{2+3d}})$. On the other hand, the minimax optimal rate for this setting is $O(n^{-\frac{1}{2+d}})$ (Stone, 1980), so the rates of Theorem 2 may be improved.

Eq. (7) may also be useful in practice. For example, in the application to state-space modeling (Song et al., 2009; Fukumizu et al., 2011), the kernel means of posterior distributions on hidden state are estimated, and their densities can be estimated by Eq. (7). If the posterior distributions are highly multimodal, we can use the estimated densities for MAP estimation of the hidden state, as in (McCalman et al., 2013).

5 Proofs

In the following, $L_p(\nu)$ for arbitrary measure ν and $p \in (0, \infty]$ denotes the Banach space consisting of p-integrable functions with respect to ν. We will use the following inequity in our proofs, which holds for arbitrary $f \in B^{\alpha}_{2,\infty}(\mathcal{X})$:

$$\omega_{r,L_2(\mathcal{X})}(f,t) \leq |f|_{B^{\alpha}_{2,\infty}(\mathcal{X})} t^{\alpha}, \qquad t > 0\ , \tag{9}$$

where $r = \lfloor \alpha \rfloor + 1$.

5.1 Proof of Theorem 1

Our strategy in the proof of Theorem 1 is to approximate the function in the Besov space by a sequence of functions in the RKHS. A recent study on learning theory has yielded bounds for errors when approximating a Besov function with certain RKHS functions and for their associated RKHS norms (Eberts and Steinwart, 2013, Theorem 2.2., Theorem 2.3). Some of the inequalities derived in our proof use these results. They are reviewed in the supplementary materials.

Proof. Let $\gamma_n = n^{-\beta}\gamma$ for some constant $\beta > 0$ (a concrete value of β is determined in the end of the proof). Let $\mathcal{H}_{\gamma_n}$ denote the RKHS of the Gaussian kernel k_{γ_n}. At first, we show the inequalities which will be used in the proof. Note that assumption $f \in B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ implies $f \in L_2(\mathbb{R}^d)$.

By $\gamma_n \leq \gamma$, we have the following inequality (Steinwart and Christmann, 2008, Proposition 4.46):

$$\|\hat{m}_P - m_P\|_{\mathcal{H}_{\gamma_n}} \leq \left(\frac{\gamma}{\gamma_n}\right)^{d/2} \|\hat{m}_P - m_P\|_{\mathcal{H}_{\gamma}} \tag{10}$$

We define $k_\gamma : \mathbb{R}^d \to \mathbb{R}$ by $k_\gamma(x) = \exp(-\|x\|^2/\gamma^2)$ for $\gamma > 0$. Let $r = \lfloor \alpha \rfloor + 1$ and define $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ by

$$K_\gamma(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2\pi}\right)^{\frac{d}{2}} k_{j\gamma/\sqrt{2}}(x). \tag{11}$$

Let $f_n : \mathbb{R}^d \to \mathbb{R}$ be the convolution of K_{γ_n} and f

$$f_n(x) := (K_{\gamma_n} * f)(x) := \int_{\mathbb{R}^d} K_{\gamma_n}(x-t) f(t) dt\ , \quad x \in \mathbb{R}^d.$$

Then by $f \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, the following inequalities hold by (Eberts and Steinwart, 2013, Theorem 2.2.) and Eq. (9):

$$\begin{aligned} & \|f_n - f\|_{L_2(P)} \\ \leq\ & \left(C_{r,1} \|g_1\|_{L_\infty(\mathbb{R}^d)}\right)^{1/2} \omega_{r,L_2(\mathbb{R}^d)}(f, \gamma_n/2) \\ \leq\ & A\gamma_n^{\alpha}, \end{aligned} \tag{12}$$

$$\begin{aligned} & \|f_n - f\|_{L_2(Q)} \\ \leq\ & \left(C_{r,2} \|g_2\|_{L_\infty(\mathbb{R}^d)}\right)^{1/2} \omega_{r,L_2(\mathbb{R}^d)}(f, \gamma_n/2) \\ \leq\ & B\gamma_n^{\alpha}, \end{aligned} \tag{13}$$

where g_1 and g_2 denotes the Lebesgue densities of P and Q, respectively, $C_{r,1}$ and $C_{r,2}$ are constants only depending on r, and A and B are constants independent of γ_n.

By $f \in L_2(\mathbb{R}^d)$, (Eberts and Steinwart, 2013, Theorem 2.3.) yields $f_n \in \mathcal{H}_{\gamma_n}$ and

$$\|f_n\|_{\mathcal{H}_{\gamma_n}} \leq C\gamma_n^{-d/2}, \tag{14}$$

where C is a constant independent of γ_n.

We are now ready to prove the assertion. The triangle

inequality yields the following inequality:

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^n w_i f(X_i) - \mathbb{E}_{X\sim P}[f(X)]\right|\right] \\
\leq\ &\mathbb{E}\left[\left|\sum_{i=1}^n w_i f(X_i) - \sum_{i=1}^n w_i f_n(X_i)\right|\right] \quad (15)\\
+\ &\mathbb{E}\left[\left|\sum_{i=1}^n w_i f_n(X_i) - \mathbb{E}_{X\sim P}[f_n(X)]\right|\right] \quad (16)\\
+\ &\left|\mathbb{E}_{X\sim P}[f_n(X)] - \mathbb{E}_{X\sim P}[f(X)]\right|. \quad (17)
\end{aligned}$$

We first derive a rate of convergence for the first term Eq. (15):

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^n w_i f(X_i) - \sum_{i=1}^n w_i f_n(X_i)\right|\right] \\
=\ &\mathbb{E}\left[\left|\sum_{i=1}^n w_i (f(X_i) - f_n(X_i))\right|\right] \\
\leq\ &\mathbb{E}\left[\left(\sum_{i=1}^n w_i^2\right)^{1/2}\left(\sum_{i=1}^n (f(X_i)-f_n(X_i))^2\right)^{1/2}\right] \\
\leq\ &\left(\mathbb{E}\left[\sum_{i=1}^n w_i^2\right]\right)^{1/2} \\
&\left(\mathbb{E}\left[n\left(\frac{1}{n}\sum_{i=1}^n (f(X_i)-f_n(X_i))^2\right)\right]\right)^{1/2} \\
=\ &\left(\mathbb{E}\left[\sum_{i=1}^n (w_i)^2\right]\right)^{1/2} n^{1/2}\|f-f_n\|_{L_2(Q)},
\end{aligned}$$

where we used the Cauchy-Schwartz inequality in the first two inequalities. Note that since the weights $w_1,\ldots,w_n$ depend on the random variables $X_1,\ldots,X_n$, the term $\left(\sum_{i=1}^n w_i^2\right)^{1/2}$ in the third line is not independent of the term $\left(\sum_{i=1}^n (f(X_i)-f_n(X_i))^2\right)^{1/2}$. By the assumption $\mathbb{E}\left[\sum_{i=1}^n (w_i)^2\right] = O(n^{-2c})$, Eq. (13), and $\gamma_n = n^{-\beta}\gamma$, the rate of the first term is $O\left(n^{-c+1/2-\alpha\beta}\right)$.

We next show a convergence rate for the second term Eq. (16):

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^n w_i f_n(X_i) - \mathbb{E}_{X\sim P}[f_n(X)]\right|\right] \\
=\ &\mathbb{E}\left[\langle \hat{m}_P - m_P, f_n\rangle_{\mathcal{H}_{\gamma_n}}\right] \\
\leq\ &\mathbb{E}\left[\|\hat{m}_P - m_P\|_{\mathcal{H}_{\gamma_n}}\right]\|f_n\|_{\mathcal{H}_{\gamma_n}} \\
\leq\ &\left(\frac{\gamma}{\gamma_n}\right)^{\frac{d}{2}}\mathbb{E}\left[\|\hat{m}_P - m_P\|_{\mathcal{H}_{\gamma}}\right]\|f_n\|_{\mathcal{H}_{\gamma_n}},
\end{aligned}$$

where the equality follows from $f_n \in \mathcal{H}_{\gamma_n}$, and the second inequality follows from Eq. (10). By the assumption $\mathbb{E}\left[\|\hat{m}_P - m_P\|_{\mathcal{H}_\gamma}\right] = O(n^{-b})$, $\gamma_n = n^{-\beta}\gamma$, and Eq. (14), the rate of the second term is $O(n^{-b+\beta d})$.

The third term Eq. (16) is bounded as

$$\begin{aligned}
\left|\mathbb{E}_{X\sim P}[f_n(X)] - \mathbb{E}_{X\sim P}[f(X)]\right| &\leq \|f_n - f\|_{L_1(P)} \\
&\leq \|f_n - f\|_{L_2(P)}\ .
\end{aligned}$$

By Eq. (12) and $\gamma_n = n^{-\beta}\gamma$, the rate of the third term is $O\left(n^{-\alpha\beta}\right)$, which is faster than that of the first term.

Balancing the first and second term yields $\beta = \frac{b-c+1/2}{\alpha+d}$. The assertion is obtained by substituting this into the above terms. □

5.2 Proof of Corollary 2

Proof. Let $\mathcal{F}(\chi_\Omega)$ denote the Fourier transform of χ_Ω. It can be easily shown that the function $(1+\|\cdot\|^2)^{\alpha/2}\mathcal{F}(\chi_\Omega)(\cdot)$ belongs to $L_2(\mathbb{R}^d)$ for any α satisfying $0 < \alpha < 1/2$. Therefore χ_Ω is included in the fractional order Sobolev space $W_2^\alpha(\mathbb{R}^d)$ (Adams and Fournier, 2003, p.252). Since $W_2^\alpha(\mathbb{R}^d) \subset B_{2,\infty}^\alpha(\mathbb{R}^d)$ holds (Edmunds and Triebel, 1996, pp.26-27, p.44), we have $\chi_\Omega \subset B_{2,\infty}^\alpha(\mathbb{R}^d)$.

For arbitrary constant α satisfying $0 < \alpha < 1/2$ and $\alpha b - d(1/2-c) > 0$, Theorem 1 then yields the rate of $O\left(n^{-\frac{\alpha b - d(1/2-c)}{\alpha+d}}\right)$ for the lhs of the assertion. Let $\alpha = 1/2-\zeta$, where $0 < \zeta < 1/2$. Then by the assumption $b - d(1-2c) > 0$ we have $\alpha b - d(1/2-c) > 0$ for sufficiently small ζ. It is not hard to check that $\frac{\alpha - d(1/2-c)}{\alpha+d}$ is monotonically decreasing as a function of ζ. Therefore in the limit of $\zeta \to 0$ we have the supremum value $\frac{b-d(1-2c)}{1+2d}$ over $\zeta \in (0,1/2)$. Since we can take an arbitrarily small value for ζ, the assertion of the corollary follows. □

5.3 Proof of Theorem 2

First, we need the following lemmas (their proofs can be found in the supplementary materials).

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz function. Then there exists a constant $M > 0$ such that for all $x_0 \in \mathbb{R}^d$ and $h > 0$ we have*

$$\left|\int J_h(x - x_0) f(x) dx - f(x_0)\right| \leq Mh\ . \quad (18)$$

Lemma 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying $f \in B_{2,\infty}^\alpha(\mathbb{R}^d)$ for some $\alpha > 0$. Then for any $h > 0$, we have*

$$|f(\cdot/h)|_{B_{2,\infty}^\alpha(\mathbb{R}^d)} = h^{-\alpha+d/2}|f|_{B_{2,\infty}^\alpha(\mathbb{R}^d)}\ . \quad (19)$$

We are now ready to prove Theorem 2.

Proof. Let $\gamma_n = n^{-\beta}\gamma$ and $h_n = n^{-\tau}$ for some constants $\beta, \tau > 0$ (concrete values for β and τ will be determined later).

Let $\alpha > 0$ be an arbitrary positive constant. We define $J_{h_n,x_0} := h_n^{-d} J_{1,x_0}(\cdot/h_n)$. Since $J_{1,x_0} \in B^{\alpha}_{2,\infty}(\mathbb{R}^d)$ holds, we then have by Lemma 2

$$|J_{h_n,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} = h_n^{-\alpha-d/2}|J_{1,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)}. \quad (20)$$

Let $r := \lfloor \alpha \rfloor + 1$ and define the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ by Eq. (11). Then by (Eberts and Steinwart, 2013, Theorem 2.2.) and Eqs. (9)(20) we have

$$\begin{aligned} &\|K_{\gamma_n} * J_{h_n,x_0} - J_{h_n,x_0}\|_{L_2(P)} \\ &\leq C_1 \omega_{r,L_2(\mathbb{R}^d)}(J_{h_n,x_0}, \gamma_n/2) \\ &\leq C_1' |J_{h_n,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} \gamma_n^{\alpha} \\ &\leq C_1' |J_{1,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} h_n^{-\alpha-d/2} \gamma_n^{\alpha} , \end{aligned} \quad (21)$$

$$\begin{aligned} &\|K_{\gamma_n} * J_{h_n,x_0} - J_{h_n,x_0}\|_{L_2(Q)} \\ &\leq C_2 \omega_{r,L_2(\mathbb{R}^d)}(J_{h_n,x_0}, \gamma_n/2) \\ &\leq C_2' |J_{h_n,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} \gamma_n^{\alpha} \\ &\leq C_2' |J_{1,x_0}|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} h_n^{-\alpha-d/2} \gamma_n^{\alpha}, \end{aligned} \quad (22)$$

where C_1, C_1', C_2, and C_2' are constants independent of h_n and γ_n.

By (Eberts and Steinwart, 2013, Theorem 2.3.) and Eq. (20), we have $K_{\gamma_n,r} * J_{h_n,x_0} \in \mathcal{H}_{\gamma_n}$ and

$$\begin{aligned} &\|K_{\gamma_n} * J_{h_n,x_0}\|_{\mathcal{H}_{\gamma_n}} \\ \leq\ & C_3 \|J_{h_n,x_0}\|_{L_2(\mathbb{R}^d)} \gamma_n^{-d/2} \\ =\ & C_3 \|J_{1,x_0}\|_{L_2(\mathbb{R}^d)} h_n^{-d/2} \gamma_n^{-d/2}, \end{aligned} \quad (23)$$

where C_3 is a constant independent of h_n and γ_n.

Similar arguments with the proof of Theorem 1 yields the following inequality:

$$\begin{aligned} &\mathbb{E}\left[\left|\sum_{i=1}^n w_i J_{h_n}(X_i - x_0) - \mathbb{E}_{X\sim P}[J_{h_n}(X - x_0)]\right|\right] \\ \leq\ & \left(\mathbb{E}\left[\sum_{i=1}^n w_i^2\right]\right)^{1/2} n^{1/2} \\ & \|K_{\gamma_n} * J_{h_n,x_0} - J_{h_n,x_0}\|_{L_2(Q)} \\ +\ & \left(\frac{\gamma}{\gamma_n}\right)^{\frac{d}{2}} \mathbb{E}\left[\|\hat{m}_P - m_P\|_{\mathcal{H}_\gamma}\right] \|K_{\gamma_n} * J_{h_n,x_0}\|_{\mathcal{H}_{\gamma_n}} \\ +\ & \|K_{\gamma_n} * J_{h_n,x_0} - J_{h_n,x_0}\|_{L_2(P)} . \end{aligned}$$

By Eq. (22) and the assumption $\mathbb{E}\left[\sum_{i=1}^n w_i^2\right] = O(n^{-c})$, the rate of the first term is $O(n^{-c+1/2-\alpha\beta+\alpha(\tau+d/2)})$. For the second term, Eq. (23) and the assumption $\mathbb{E}\left[\|\hat{m}_P - m_P\|_{\mathcal{H}_\gamma}\right] = O(n^{-b})$ yields the rate of $O(n^{-b+\beta d+d\tau/2})$. By Eq. (21), the rate of the third term is $O(n^{-\alpha\beta+\alpha(\tau+d/2)})$, which is faster than that of the first term. Balancing the first and second terms yields $\beta = \frac{-c+1/2+\alpha\tau+b}{\alpha+d}$. Substituting this into the above terms, the overall rate is then given by

$$O\left(n^{-\frac{\alpha(b-3\tau d/2)-d(1/2-c)-d^2\tau/2}{\alpha+d}}\right) . \quad (24)$$

Note that we can take an arbitrarily large constant for α . We therefore have for arbitrarily small $\zeta > 0$

$$\begin{aligned} &\mathbb{E}\left[\left|\sum_{i=1}^n w_i J_{h_n}(X_i - x_0) - \mathbb{E}_{X\sim P}[J_{h_n}(X - x_0)]\right|\right] \\ &= O\left(n^{-b+3\tau d/2+\zeta}\right) . \end{aligned} \quad (25)$$

On the other hand, since

$$\mathbb{E}_{X\sim P}[J_{h_n}(X - x_0)] = \int J_{h_n}(x - x_0) p(x) dx$$

holds, Lemma 1 and the Lipschitz continuity of p yield

$$|\mathbb{E}_{X\sim P}[J_{h_n}(X - x_0)] - p(x_0)| \leq M h_n . \quad (26)$$

By balancing Eqs. (25) and (26) we have $\tau = \frac{2b}{3d+2} - \frac{2\zeta}{3d+2}$. We therefore have $E\left[\left|\sum_{i=1}^n w_i J_{h_n}(X_i - x_0) - p(x_0)\right|\right] = O\left(n^{-\frac{2b}{3d+2}+\frac{2\zeta}{3d+2}}\right)$, and letting $\xi := \frac{2\zeta}{3d+2}$ yields the assertion of the theorem. □

6 Conclusions

In this paper, we discussed methodology for recovering the information of distributions from estimates of their corresponding kernel means. To this end, we theoretically analyzed the properties of a consistent estimator of a kernel mean in a Gaussian RKHS, and proved that the expectations of functions in a Besov space can be consistently estimated with the kernel mean estimator. Using this result and an argument similar to its proof, we show that moments, probability measures on intervals, and the density can be recovered from an estimate of a kernel mean. This work will serve as a theoretical basis for developing practical applications of RKHS embeddings.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported in part by JSPS Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

References

Adams, R. A. and Fournier, J. J. F. (2003). *Sobolev Spaces*. Academic Press, New York, 2nd ed. edition.

Boots, B., Gretton, A., and Gordon., G. J. (2013). Hilbert space embeddings of predictive state representations. In *UAI*.

Eberts, M. and Steinwart, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.*, 7:1–42.

Edmunds, D. E. and Triebel, H. (1996). *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge.

Fukumizu, K., Bach, F., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*, 5:73–99.

Fukumizu, K., Bach, F., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.*, 37(4):1871–1905.

Fukumizu, K., Song, L., and Gretton, A. (2011). Kernel Bayes' rule. In *NIPS*, pages 1737–1745.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *JMLR*, 13:723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schoelkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *NIPS*.

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. (2012). Modeling transition dynamics in MDPs with RKHS embeddings. In *ICML*.

McCalman, L., O'Callaghan, S., and Ramos, F. (2013). Multi-modal estimation with kernel embeddings for learning motion models. In *ICRA*.

Minh, H. Q. (2010). Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338.

Nishiyama, Y., Boularias, A., Gretton, A., and Fukumizu, K. (2012). Hilbert space embeddings of POMDPs. In *UAI*.

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

Song, L., Fukumizu, K., and Gretton, A. (2013). Kernel embeddings of conditional distributions. *IEEE Signal Processing Magazine*, 30(4):98–111.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. (2011). Kernel belief propagation. In *AISTATS*.

Song, L., Gretton, A., and Guestrin, C. (2010). Nonparametric tree graphical models. In *AISTATS*.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, pages 961–968.

Song, L., Zhang, X., Smola, A., Gretton, A., and Schölkopf, B. (2008). Tailoring density estimation via reproducing kernel moment matching. In *ICML*, pages 992–999.

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

Stein, E. M. (1970). *Singular integrals and differentiability properties of functions*. Princeton University Press, Princeton, NJ.

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

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360.