Supplementary Materials to "Recovering Distributions from Gaussian RKHS Embeddings"

A Kernel mean estimator of conditional distributions

We first review the kernel mean estimator of conditional distributions (Song et al., 2009; Grünewälder et al., 2012). Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces and $P_{\mathcal{X}\mathcal{Y}}$ be a joint distribution on $\mathcal{X} \times \mathcal{Y}$ with density $p(x|y)p(y)$ for $x \in \mathcal{X}, y \in \mathcal{Y}$. Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be bounded kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ be the respective RKHSs. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples from $P_{\mathcal{X}\mathcal{Y}}$ and let $y \in \mathcal{Y}$ be fixed. Then the following is a consistent estimator of the kernel mean of the conditional probability $m_{P_{\mathcal{X}|y}} := \int k_{\mathcal{X}}(\cdot, x) p(x|y) dx$

$$\hat{m}_{P_{\mathcal{X}|y}} = \sum_{i=1}^{n} w_i k_{\mathcal{X}}(\cdot, X_i), \quad w = (G_Y + n\varepsilon_n I_n)^{-1} \mathbf{k}_Y(y) \in \mathbb{R}^n, \tag{A1}$$

where $G_Y = (k_{\mathcal{Y}}(Y_i, Y_j)) \in \mathbb{R}^{n \times n}$ is a Gram matrix, $\varepsilon_n > 0$ is a regularization constant, and $\mathbf{k}_Y(y) = (k_{\mathcal{Y}}(y, Y_i))_{i=1}^n \in \mathbb{R}^n$. Then with $\varepsilon_n = n^{-1/4}$ we have

$$\|\hat{m}_{P_{\mathcal{X}|y}} - m_{P_{\mathcal{X}|y}}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-1/8}) \qquad (n \to \infty)$$

under some smoothness assumption (Song et al., 2010, Theorem 1). We therefore have in this case the value of $b = 1/8$ in Section 3 of the main text.

We aim here to show that $\mathbb{E}[\sum_{i=1}^{n} w_i^2] = O(n^{-c})$ with $c = 1/4$. First, we need the following lemma.

Lemma A1. *The solution to the minimization problem*

$$\min_{\gamma \in \mathbb{R}^n} \|k_{\mathcal{Y}}(\cdot, y) - \sum_{i=1}^{n} \gamma_i k_{\mathcal{Y}}(\cdot, Y_i)\|_{\mathcal{H}_{\mathcal{Y}}}^2$$

is given by $G_Y \gamma_* = \mathbf{k}_Y(y)$, and the minimum value is

$$\min_{\gamma \in \mathbb{R}^n} \|k_{\mathcal{Y}}(\cdot, y) - \sum_{i=1}^{n} \gamma_i k_{\mathcal{Y}}(\cdot, Y_i)\|_{\mathcal{H}_{\mathcal{Y}}}^2 = k_{\mathcal{Y}}(y, y) - \gamma_*^T G_Y \gamma_*$$

Proof. Straightforward calculation. □

Then $c = 1/4$ follows from the following proposition.

Proposition A1. *Let $w \in \mathbb{R}^n$ be given by Eq. (A1). Then we have*

$$nw^T w \leq \frac{k_{\mathcal{Y}}(y,y)}{\varepsilon_n}$$

Proof. From Lemma A1, there is $\gamma \in \mathbb{R}^n$ such that $\mathbf{k}_Y(y) = G_Y\gamma$ and $\gamma^T G_Y \gamma \leq k_{\mathcal{Y}}(y,y)$. We have

$$w = (G_Y + n\varepsilon_n I_n)^{-1} G_Y \gamma,$$

and thus

$$\begin{aligned} nw^{(n)T} w &= n\gamma^T G_Y (G_Y + n\varepsilon_n I_n)^{-2} G_Y \gamma \\ \leq n\gamma^T (G_Y + n\varepsilon_n I_n)^{-1} G_Y \gamma &\leq \frac{1}{\varepsilon_n} \gamma^T G_Y \gamma \leq \frac{k_{\mathcal{Y}}(y,y)}{\varepsilon_n} \end{aligned}$$

where the first inequality in the second line uses $(G_Y + n\varepsilon_n I_n)^{-1} G_Y \leq I_n$ and the second one uses $n(G_Y + n\varepsilon_n)^{-1} \leq (1/\varepsilon_n) I_n$. □

B Numerical Experiments for Corollary 2

We conducted numerical experiments to see the consistency of the estimator for measures on intervals of Corollary 2 in the main text. To this end, we use the kernel mean estimator for conditional distributions (A1) as $\hat{m}_P$ in Corollary 2.

Let $d = 1$. We generated i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ as $X_i | Y_i \sim \mathbb{N}(Y_i, 1), Y_i \sim$ Uniform$[0,1]$, where $\mathbb{N}(Y_i, 1)$ is the normal distribution with mean Y_i and variance 1 and Uniform$[0,1]$ is the uniform distribution on $[0,1]$. We used the Gaussian kernel $k_\gamma(x,x') = \exp(-\|x - x'\|^2/\gamma)$ as $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ with parameter fixed as $\gamma = 1$. We set the regularization constant as $\varepsilon_n = n^{-1/4}$. The conditional distribution $P_{\mathcal{X}|y}$ for which the kernel mean is estimated is $X|y \sim \mathbb{N}(y, 1)$ with $y = 0.5$. In this experiment, the intervals $[a,b]$ are set to (1) $[a,b] = [0.5, \infty]$ and (2) $[a,b] = [-1, 1]$. The performance was evaluated by the absolute error in Corollary 2, $\left|\sum_{X_i \in [a,b]} w_i - P([a,b])\right|$, where $P([a,b]) = P_{\mathcal{X}|y}([a,b])$ with $y = 0.5$ in this setting. We run experiments for each sample size for 20 times and averaged the results.

The results are shown in Table 1. The results empirically show the consistency of the estimator of Corollary 2 for both cases (1) (2). The convergence rates are roughly $O(n^{-0.45})$ for (1) and $O(n^{-0.4})$ for (2). Thus, these results suggest that the rates in Corollary 2 may be further improved.

Table 1: Result of Numerical Experiments. The performance was evaluated by the absolute errors $\left|\sum_{X_i \in [a,b]} w_i - P([a,b])\right|$

sample size	(1) $[a,b] = [0.5, \infty]$	(2) $[a,b] = [-1, 1]$
2^4	0.0958 ± 0.0521	0.0869 ± 0.0630
2^5	0.0752 ± 0.0521	0.0730 ± 0.0490
2^6	0.0533 ± 0.0466	0.0584 ± 0.0450
2^7	0.0417 ± 0.0315	0.0306 ± 0.0199
2^8	0.0264 ± 0.0245	0.0334 ± 0.0249
2^9	0.0313 ± 0.0206	0.0228 ± 0.0226
2^{10}	0.0141 ± 0.0121	0.0192 ± 0.0154
2^{11}	0.0094 ± 0.0074	0.0124 ± 0.0092
2^{12}	0.0086 ± 0.0072	0.0097 ± 0.0067

C Theorems from (Eberts and Steinwart, 2013)

We review the theorems from (Eberts and Steinwart, 2013) which are used in our proof. Let $k_\gamma : \mathbb{R}^d \to \mathbb{R}$ be a function on $\mathbb{R}^d$ defined by $k_\gamma = \exp\left(-\frac{\|x\|^2}{\gamma^2}\right)$ and $k_\gamma(x, x') := k_\gamma(x - x')$ be the Gaussian kernel with the band-width parameter $\gamma > 0$. Let $\mathcal{H}_\gamma$ be the RKHS defined by the kernel k_γ, and $\langle \cdot, \cdot \rangle_{\mathcal{H}_\gamma}$ and $\| \cdot \|_{\mathcal{H}_\gamma}$ be its inner-product and norm, respectively.

We define, for $r \in \mathbb{N}$ and $\gamma > 0$, the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ by

$$K_\gamma(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \pi}\right)^{\frac{d}{2}} k_{j\gamma/\sqrt{2}}(x) \quad . \tag{C2}$$

Theorem C1 (Theorem 2.2 in (Eberts and Steinwart, 2013))**.** *Let q be a constant such that $q \in [1, \infty)$. Let P be probability distribution $\mathbb{R}^d$. Assume that P has a density function satisfying $g \in L_2(\mathbb{R}^d)$ for some $p \in [1, \infty]$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $f \in L_q(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$. Then, for $r \in \mathbb{N}$, $\gamma > 0$, and $s \geq 1$ with $1 = \frac{1}{s} + \frac{1}{p}$, we have*

$$\|K_\gamma * f - f\|_{L_q(P)}^q \leq C_{r,q} \|g\|_{L_2(\mathbb{R}^d)} \omega_{r, L_{qs}(\mathbb{R}^d)}^q (f, \gamma/2) \ ,$$

where $C_{r,q}$ is a constant depending only on r and q.

Theorem C2 (Theorem 2.3 in (Eberts and Steinwart, 2013))**.** *Let $f \in L_2(\mathbb{R}^d)$, $\mathcal{H}_\gamma$ be the RKHS of the Gaussian kernel k_γ over $\mathcal{X} \subset \mathbb{R}^d$ with $\gamma > 0$. Then we have $K_{\gamma,r} * f \in \mathcal{H}_\gamma$ with*

$$\|K_\gamma * f\|_{\mathcal{H}_\gamma} \leq (\gamma\sqrt{\pi})^{-\frac{d}{2}} (2^r - 1) \|f\|_{L_2(\mathbb{R}^d)} \ .$$

D Proof of Lemma 1.

Proof. Since f is Lipschitz, there exists a constant C such that $|f(x) - f(y)| \leq C\|x-y\|$ holds for all $x, y \in \mathbb{R}^d$. We therefore have

$$\begin{aligned} & \left|\int J_h(x-x_0) f(x) dx - f(x_0)\right| \\ = & \left|\int J_h(x-x_0)(f(x)-f(x_0)) dx\right| \\ \leq & C \int |J_h(x-x_0)| \|x-x_0\| dx \\ = & C \int \left|\frac{1}{h^d} J_1((x-x_0)/h)\right| \|x-x_0\| dx \\ \leq & Ch \int |J_1(u)| \|u\| du =: Mh. \end{aligned}$$

□

E Proof of Lemma 2.

Proof. First, we have for all $v \in [0,\infty)^d$

$$\begin{aligned} \Delta_v^r(f(\cdot/h), x) &= \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} f\left(\frac{x+jv}{h}\right) \\ &= \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} f\left(\frac{x}{h} + \frac{jv}{h}\right) \\ &= \Delta_{v/h}^r(f, x/h) \ . \end{aligned}$$

Then, we have for all $t > 0$

$$\begin{aligned} \omega_{r,L_2(\mathbb{R}^d)}(f(\cdot/h), t) &= \sup_{\|v\|_2 \leq t} \|\Delta_v^r(f(\cdot/h), \cdot)\|_{L_2(\mathbb{R}^d)} \\ &= \sup_{\|v\|_2 \leq t} \|\Delta_{v/h}^r(f, \cdot/h)\|_{L_2(\mathbb{R}^d)} \\ &= \sup_{\|v\| \leq t/h} \|\Delta_v^r(f, \cdot/h)\|_{L_2(\mathbb{R}^d)} \\ &= h^{d/2} \sup_{\|v\| \leq t/h} \|\Delta_v^r(f, \cdot)\|_{L_2(\mathbb{R}^d)} \ . \end{aligned}$$

Thus, we have

$$\begin{aligned}
|f(\cdot/h)|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)} &= \sup_{t>0}\Big(t^{-\alpha}\omega_{r,L_2(\mathbb{R}^d)}(f(\cdot/h),t)\Big) \\
&= h^{d/2}\sup_{t>0}\left(t^{-\alpha}\sup_{\|v\|\le t/h}\|\Delta^r_v(f,\cdot)\|_{L_2(\mathbb{R}^d)}\right) \\
&= h^{d/2}\sup_{t>0}\left(t^{-\alpha}h^{-\alpha}\sup_{\|v\|\le t}\|\Delta^r_v(f,\cdot)\|_{L_2(\mathbb{R}^d)}\right) \\
&= h^{-\alpha+d/2}|f|_{B^{\alpha}_{2,\infty}(\mathbb{R}^d)}
\end{aligned}$$

□

References

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

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012). Conditional mean embeddings as regressors. In *ICML*.

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.