Supplementary materials to ”Kernel Bayes’ Rule”

A Proof of Propositions 3 and 4

These propositions can be proved in a similar manner with simple linear algebra. We show the proofs for completeness.

Proof of Proposition 3. We show only the proof for C_{ZW}, as the case of C_{WW} is exactly the same. Let $h = (\widehat{C}_{XX} + \varepsilon_n I)^{-1}\widehat{m}_\Pi^{(\ell)}$, and decompose it as $h = \sum_{i=1}^n \alpha_i k_\mathcal{X}(\cdot, X_i) + h_\perp = \alpha^T \mathbf{k}_X + h_\perp$, where $h_\perp$ is orthogonal to all $k_\mathcal{X}(\cdot, X_i)$. Expansion of $(\widehat{C}_{XX} + \varepsilon_n I)h = \widehat{m}_\Pi^{(\ell)}$ derives $\frac{1}{n}\mathbf{k}_X^T G_X \alpha + \varepsilon_n \mathbf{k}_X^T \alpha + \varepsilon_n h_\perp = \widehat{m}_\Pi^{(\ell)}$. By taking the inner product with $k_\mathcal{X}(\cdot, X_j)$, we have

$$\Big(\frac{1}{n}G_X + \varepsilon_n I_n\Big) G_X \alpha = \widehat{\mathbf{m}}_\Pi.$$

The coefficient $\widehat{\mu}$ in $C_{ZW} = \widehat{C}_{(YX)X} h = \sum_{i=1}^n \widehat{\mu}_i k_\mathcal{X}(\cdot, X_i) \otimes k_\mathcal{Y}(\cdot, Y_i)$ is given by $\widehat{\mu} = G_X \alpha$, and thus

$$\widehat{\mu} = \Big(\frac{1}{n}G_X + \varepsilon_n I_n\Big)^{-1} \widehat{\mathbf{m}}_\Pi.$$

$\square$

Proof of Proposition 4. Let $h = (\widehat{C}_{WW}^2 + \delta_n I)^{-1}\widehat{C}_{WW} k_\mathcal{Y}(\cdot, y)$, and decompose it as $h = \sum_{i=1}^n \alpha_i k_\mathcal{Y}(\cdot, Y_i) + h_\perp = \alpha^T \mathbf{k}_Y + h_\perp$, where $h_\perp$ is orthogonal to all $k_\mathcal{Y}(\cdot, Y_i)$. Expansion of $(\widehat{C}_{WW}^2 + \delta_n I)h = \widehat{C}_{WW} k_\mathcal{Y}(\cdot, y)$ derives $\mathbf{k}_Y^T (\Lambda G_Y)^2 \alpha + \delta_n \mathbf{k}_Y^T \alpha + \delta_n h_\perp = \mathbf{k}_Y^T \Lambda \mathbf{k}_Y(y)$. Taking the inner product with $k_\mathcal{Y}(\cdot, Y_j)$ derives

$$\big((G_Y\Lambda)^2 + \delta_n I_n\big) G_Y \alpha = G_Y \Lambda \mathbf{k}_Y(y).$$

The coefficient w in $\widehat{m}_{Q_\mathcal{X}|y} = \widehat{C}_{ZW} h = \sum_{i=1}^n w_i k_\mathcal{X}(\cdot, X_i)$ is given by $w = \Lambda G_Y \alpha$, and thus

$$w = \Lambda\big((G_Y\Lambda)^2 + \delta_n I_n\big)^{-1} G_Y \Lambda \mathbf{k}_Y(y) = \Lambda G_Y \big((\Lambda G_Y)^2 + \delta_n I_n\big)^{-1} \Lambda \mathbf{k}_Y(y).$$

$\square$

B Derivation of the KBR update rule for nonparametric state-space model

This section gives a more detailed derivation of the update rule for nonparametric state-space model, which we sketched in Section 3.

Given the estimate of the kernel mean expression for $p(x_t|\tilde{y}_1, \ldots, \tilde{y}_t)$, the forward filtering with

$$p(y_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t) = \int p(y_{t+1}|x_{t+1}) \int p(x_{t+1}|x_t) p(x_t|\tilde{y}_1, \ldots, \tilde{y}_t) dx_{t+1} dx_t$$

can be realized by the two-times applications of forward filtering procedure similar to Proposition 3. Namely, first the kernel mean of $p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t) = \int p(x_{t+1}|x_t) p(x_t|\tilde{y}_1, \ldots, \tilde{y}_t) dx_t$ can be estimated by

$$\widehat{m}_{x_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \sum_{i=1}^T \beta_i k_\mathcal{X}(\cdot, X_{i+1}), \qquad \text{where} \quad \beta = \big(\tfrac{1}{T}G_X + \varepsilon_T I_T\big)^{-1} G_X \alpha.$$

In the same way, the second step is to compute the kernel mean of $p(y_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t) = \int p(y_{t+1}|x_{t+1}) p(x_{t+1}|\tilde{y}_1, \ldots, \tilde{y}_t) dx_{t+1}$, which is estimated by

$$\widehat{m}_{y_{t+1}|\tilde{y}_1,\ldots,\tilde{y}_t} = \sum_{i=1}^T \gamma_i k_\mathcal{X}(\cdot, Y_i), \qquad \text{where} \quad \gamma = \big(\tfrac{1}{T}G_Y + \varepsilon_T I_T\big)^{-1} G_{X,X_{+1}} \beta.$$

C Rates of consistency

The proof idea for the consistency rates of the KBR estimators is essentially the same as [1, 3], in which the basic techniques are taken from the general theory of regularization [2].

First we give integral expression for the kernel mean and covariance operators. Reacll that the kernel mean m_X of X on $\mathcal{H}_\mathcal{X}$ satisfies

$$\langle f, m_X \rangle = E[f(X)]$$

for any $f \in \mathcal{H}_\mathcal{X}$. Plugging $f = k_\mathcal{X}(\cdot, u)$ into this relation derives

$$m_X(u) = E[k(u, X)] = \int k_\mathcal{X}(u, \tilde{x}) dP_X(\tilde{x}), \tag{15}$$

which shows the explicit functional form of the kernel mean. In a similar manner, the explicit integral expression of the covariance operators C_{YX} and C_{XX} are given by

$$(C_{YX} f)(y) = \int k_\mathcal{Y}(y, \tilde{y}) f(\tilde{x}) dP(\tilde{x}, \tilde{y}), \quad (C_{XX} f)(x) = \int k_\mathcal{X}(x, \tilde{x}) f(\tilde{x}) dP_X(\tilde{x}), \tag{16}$$

respectively. The covariance operators are thus integral operators with integral kernel $k_\mathcal{X}$ or $k_\mathcal{Y}$.

The first preliminary result is a rate of convergence for the mean transition in Theorem 2. In the following $\mathcal{R}(C_{XX}^0)$ means $\mathcal{H}_\mathcal{X}$.

Theorem 6. *Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ for some $\beta \geq 0$, where π and p_X are the p.d.f. of Π and P_X, respectively. Let $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ as $n \to \infty$ for some $0 < \alpha \leq 1/2$. Then, with $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{\alpha}{1+\beta}\}}$, we have*

$$\big\| \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \big)^{-1} \widehat{m}_\Pi^{(n)} - m_{Q_\mathcal{Y}} \big\|_{\mathcal{H}_\mathcal{Y}} = O_p(n^{-\min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}}), \quad (n \to \infty).$$

Proof. Take $\eta \in \mathcal{H}_\mathcal{X}$ such that $\pi/p_X = C_{XX}^\beta \eta$. Then, from Eqs. (15) and (16),

$$m_\Pi = \int k_\mathcal{X}(\cdot, x) \frac{\pi(x)}{p_X(x)} p_X(x) d\mu_\mathcal{X}(x) = C_{XX}^{\beta+1} \eta. \tag{17}$$

First we show the rate of the estimation error:

$$\big\| \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \big)^{-1} \widehat{m}_\Pi^{(n)} - C_{YX} \big(C_{XX} + \varepsilon_n I \big)^{-1} m_\Pi \big\|_{\mathcal{H}_\mathcal{Y}} = O_p\big(n^{-\alpha} \varepsilon_n^{-1/2}\big), \tag{18}$$

as $n \to \infty$. By using the fact that $B^{-1} - A^{-1} = B^{-1}(A - B)A^{-1}$ holds for any invertible operators A and B, the left hand side of Eq. (18) is upper bounded by

$$\begin{aligned} &\big\| \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \big)^{-1} \big(\widehat{m}_\Pi^{(n)} - m_\Pi \big) \big\|_{\mathcal{H}_\mathcal{Y}} + \big\| \big(\widehat{C}_{YX}^{(n)} - C_{YX} \big) \big(C_{XX} + \varepsilon_n I \big)^{-1} m_\Pi \big\|_{\mathcal{H}_\mathcal{Y}} \\ &\qquad + \big\| \widehat{C}_{YX}^{(n)} \big(\widehat{C}_{XX}^{(n)} + \varepsilon_n I \big)^{-1} \big(C_{XX} - \widehat{C}_{XX}^{(n)} \big) \big(C_{XX} + \varepsilon_n I \big)^{-1} m_\Pi \big\|_{\mathcal{H}_\mathcal{Y}}. \end{aligned}$$

By the decomposition $\widehat{C}_{YX}^{(n)} = \widehat{C}_{YY}^{(n)1/2} \widehat{W}_{YX}^{(n)} \widehat{C}_{XX}^{(n)1/2}$ with $\|\widehat{W}_{YX}^{(n)}\| \leq 1$ (see [2]), the first term is of $O_p(n^{-\alpha} \varepsilon_n^{-1/2})$. From Eq. (17), the second and third terms are of the order $O_p(n^{-1/2})$ and $O_p(n^{-1/2} \varepsilon_n^{-1/2})$, respectively, by $\|(C_{XX} + \varepsilon_n I)^{-1} C_{XX}\| \leq 1$. This means Eq. (18).

Next, we show

$$\big\| C_{YX} \big(C_{XX} + \varepsilon_n I \big)^{-1} m_\Pi - m_{Q_\mathcal{Y}} \big\|_{\mathcal{H}_\mathcal{Y}} = O(\varepsilon_n^{\min\{(1+2\beta)/2, 1\}}) \qquad (n \to \infty). \tag{19}$$

Let $C_{YX} = C_{YY}^{1/2} W_{YX} C_{XX}^{1/2}$ be the decomposition with $\|W_{YX}\| \leq 1$. It follows from the relation

$$m_{Q_\mathcal{Y}} = \int \int k(\cdot, y) \frac{\pi(x)}{p_X(x)} p(x, y) d\mu_\mathcal{X}(x) d\mu_\mathcal{Y}(y) = C_{YX} C_{XX}^\beta \eta$$

that the left hand side of Eq. (19) is upper bounded by

$$\|C_{YY}^{1/2} W_{YX}\| \, \| \big(C_{XX} + \varepsilon_n I \big)^{-1} C_{XX}^{(2\beta+3)/2} \eta - C_{XX}^{(2\beta+1)/2} \eta \|_{\mathcal{H}_\mathcal{X}}.$$

By the eigendecomposition $C_{XX} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot \rangle$, where $\{\phi_i\}$ are the unit eigenvectors and $\{\lambda_i\}$ are the corresponding eigenvalues, the expansion

$$\big\|\big(C_{XX}+\varepsilon_n I\big)^{-1} C_{XX}^{(2\beta+3)/2}\eta - C_{XX}^{(2\beta+1)/2}\eta\big\|_{\mathcal{H}_\mathcal{X}}^2 = \sum_i \left(\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n}\right)^2 \langle \eta, \phi_i\rangle^2$$

holds. If $0 \le \beta < 1/2$, we have $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n} = \frac{\lambda_i^{(2\beta+1)/2}}{(\lambda_i+\varepsilon_n)^{(2\beta+1)/2}} \frac{\varepsilon_n^{(1-2\beta)/2}}{(\lambda_i+\varepsilon_n)^{(1-2\beta)/2}} \varepsilon_n^{(2\beta+1)/2} \le \varepsilon_n^{(2\beta+1)/2}$. If $\beta \ge 1/2$, then $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i+\varepsilon_n} \le \|C_{XX}\|\varepsilon_n$. The dominated convergence theorem shows that the the above sum converges to zero as $\varepsilon_n \to 0$ of the order $O(\varepsilon_n^{\min\{2\beta+1,2\}})$.

From Eqs. (18) and (19), the optimal order of ε_n and the optimal rate of consistency are given as claimed. $\square$

The following theorem shows the consistency rate of the estimator used in the conditioning step Eq. (8).

Theorem 7. *Let f be a function in $\mathcal{H}_\mathcal{X}$, and (Z,W) be a random variable taking value in $\mathcal{X}\times\mathcal{Y}$. Assume that $E[f(Z)|W=\cdot] \in \mathcal{R}(C_{WW}^\nu)$ for some $\nu \ge 0$, and $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be compact operators, which may not be positive definite, such that $\|\widehat{C}_{WZ}^{(n)} - C_{WZ}\| = O_p(n^{-\gamma})$ and $\|\widehat{C}_{WW}^{(n)} - C_{WW}\| = O_p(n^{-\gamma})$ for some $\gamma > 0$. Then, for $\delta_n = n^{-\max\{\frac{4}{9}\gamma, \frac{4}{2\nu+5}\gamma\}}$ and any $y \in \mathcal{Y}$, we have as $n\to\infty$*

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)} f - E[f(X)|W=\cdot]\big\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\min\{\frac{4}{9}\gamma, \frac{2\nu}{2\nu+5}\gamma\}}).$$

Proof. Let $\eta \in \mathcal{H}_\mathcal{X}$ such that $E[f(Z)|W=\cdot] = C_{WW}^\nu \eta$. First we show

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)} f - C_{WW}(C_{WW}^2+\delta_n I)^{-1} C_{WZ} f\big\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\gamma}\delta_n^{-5/4}). \tag{20}$$

The left hand side of Eq. (20) is upper bounded by

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}(\widehat{C}_{WZ}^{(n)}-C_{WZ})f\big\|_{\mathcal{H}_\mathcal{Y}} + \big\|(\widehat{C}_{WW}^{(n)}-C_{WW})(C_{WW}^2+\delta_n I)^{-1}C_{WZ}f\big\|_{\mathcal{H}_\mathcal{Y}} \\ + \big\|\widehat{C}_{WW}^{(n)}((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\big((\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2\big)\big(C_{WW}^2+\delta_n I\big)^{-1}C_{WZ}f\big\|_{\mathcal{H}_\mathcal{Y}}.$$

Let $\widehat{C}_{WW}^{(n)} = \sum_i \lambda_i \phi_i \langle \phi_i, \cdot\rangle$ be the eigendecomposition, where $\{\phi_i\}$ is the unit eigenvectors and $\{\lambda_i\}$ is the corresponding eigenvalues. From $\big|\lambda_i/(\lambda_i^2+\delta_n)\big| = 1/|\lambda_i + \delta_n/\lambda_i| \le 1/(2\sqrt{|\lambda_i|}\sqrt{\delta_n/|\lambda_i|}) = 1/(2\sqrt{\delta_n})$, we have $\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\| \le 1/(2\sqrt{\delta_n})$, and thus the first term of the above bound is of $O_p(n^{-\gamma}\delta_n^{-1/2})$. A similar argument by the eigendecomposition of C_{WW} combined with the decomposition $C_{WZ} = C_{WW}^{1/2} U_{WZ} C_{ZZ}^{1/2}$ with $\|U_{WZ}\| \le 1$ shows that the second term is of $O_p(n^{-\gamma}\delta_n^{-3/4})$. From the fact $\|(\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2\| \le \|\widehat{C}_{WW}^{(n)}(\widehat{C}_{WW}^{(n)} - C_{WW})\| + \|(\widehat{C}_{WW}^{(n)} - C_{WW})C_{WW}\| = O_p(n^{-\gamma})$, the third term is of $O_p(n^{-\gamma}\delta_n^{-5/4})$. This implies Eq. (20).

From $E[f(Z)|W=\cdot] = C_{WW}^\nu\eta$ and $C_{WZ}f = C_{WW}E[f(Z)|W=\cdot] = C_{WW}^{\nu+1}\eta$, the convergence rate

$$\big\|C_{WW}(C_{WW}^2+\delta_n I)^{-1}C_{WZ}f - E[f(Z)|W=\cdot]\big\|_{\mathcal{H}_\mathcal{Y}} = O(\delta_n^{\min\{1,\frac{\nu}{2}\}}). \tag{21}$$

can be proved by the same way as Eq. (19).

Combination of Eqs.(20) and (21) proves the assertion. $\square$

It is possible to extend the covariance operator C_{WW} to the one defined on $L^2(Q_W)$ by

$$\tilde{C}_{WW}\phi = \int k_\mathcal{Y}(y,w)\phi(w)dQ_W(w), \qquad (\phi \in L^2(Q_W)). \tag{22}$$

The following theorem shows the consistency rate on average. Here $\mathcal{R}(\tilde{C}_{WW}^0)$ means $L^2(Q_W)$.

Theorem 8. *Let f be a function in $\mathcal{H}_\mathcal{X}$, and (Z,W) be a random variable taking values in $\mathcal{X}\times\mathcal{Y}$ with distribution Q. Assume that $E[f(Z)|W=\cdot]\in\mathcal{R}(\tilde{C}_{WW}^{\nu})\cap\mathcal{H}_\mathcal{Y}$ for some $\nu>0$, and $\widehat{C}_{WZ}^{(n)}:\mathcal{H}_\mathcal{X}\to\mathcal{H}_\mathcal{Y}$ and $\widehat{C}_{WW}^{(n)}:\mathcal{H}_\mathcal{Y}\to\mathcal{H}_\mathcal{Y}$ be compact operators, which may not be positive definite, such that $\|\widehat{C}_{WZ}^{(n)}-C_{WZ}\|=O_p(n^{-\gamma})$ and $\|\widehat{C}_{WW}^{(n)}-C_{WW}\|=O_p(n^{-\gamma})$ for some $\gamma>0$. Then, for $\delta_n=n^{-\max\{\frac{1}{2}\gamma,\frac{2}{\nu+2}\gamma\}}$, we have as $n\to\infty$*

$$\big\|\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)}f-E[f(X)|W=\cdot]\big\|_{L^2(Q_W)}=O_p(n^{-\min\{\frac{1}{2}\gamma,\frac{\nu}{\nu+2}\gamma\}}),$$

where Q_W is the marginal distribution of W.

Proof. Note that for $h,g\in\mathcal{H}_\mathcal{Y}$ we have $(h,g)_{L^2(Q_W)}=E[h(W)g(W)]=\langle h,C_{WW}g\rangle_{\mathcal{H}_\mathcal{Y}}$. It follows that the left hand side of the assertion is equal to

$$\big\|C_{WW}^{1/2}\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)}f-C_{WW}^{1/2}E[f(Z)|W=\cdot]\big\|_{\mathcal{H}_\mathcal{Y}}.$$

First, by the similar argument to the proof of Eq. (20), it is easy to show that the rate of the estimation error is given by

$$\begin{aligned}\big\|C_{WW}^{1/2}\big\{\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{WW}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)}f-C_{WW}(C_{WW}^2+\delta_n I)^{-1}C_{WZ}f\big\}\big\|_{\mathcal{H}_\mathcal{Y}}\\ =O_p(n^{-\gamma}\delta_n^{-1}).\end{aligned}$$

It suffices then to prove

$$\big\|C_{WW}(C_{WW}^2+\delta_n I)^{-1}C_{WZ}f-E[f(Z)|W=\cdot]\big\|_{L^2(Q_W)}=O(\delta_n^{\min\{1,\frac{\nu}{2}\}}).$$

Let $\xi\in L^2(Q_W)$ such that $E[f(Z)|W=\cdot]=\tilde{C}_{WW}^{\nu}\xi$. In a similar way to Theorem 1, $\tilde{C}_{WW}E[f(Z)|W]=\tilde{C}_{WZ}f$ holds, where $\tilde{C}_{WZ}$ is the extension of C_{WZ}, and thus $C_{WZ}f=\tilde{C}_{WW}^{\nu+1}\xi$. The left hand side of the above equation is equal to

$$\big\|\tilde{C}_{WW}(\tilde{C}_{WW}^2+\delta_n I)^{-1}\tilde{C}_{WW}^{\nu+1}\xi-\tilde{C}_{WW}^{\nu}\xi\big\|_{L^2(Q_W)}.$$

By the eigendecomposition of $\tilde{C}_{WW}$ in $L^2(Q_W)$, a similar argument to the proof of Eq. (21) shows the assertion. □

Combining the above theorems, we have the following consistency of KBR.

Theorem 9. *Let f be a function in $\mathcal{H}_\mathcal{X}$, (Z,W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)}-m_\Pi\|_{\mathcal{H}_\mathcal{X}}=O_p(n^{-\alpha})$ ($n\to\infty$) for some $0<\alpha\le 1/2$. Assume that $\pi/p_X\in\mathcal{R}(C_{XX}^{\beta})$ with $\beta\ge 0$, and $E[f(Z)|W=\cdot]\in\mathcal{R}(C_{WW}^{\nu})$ for some $\nu\ge 0$. For the regularization constants $\varepsilon_n=n^{-\max\{\frac{2}{3}\alpha,\frac{1}{1+\beta}\alpha\}}$ and $\delta_n=n^{-\max\{\frac{4}{9}\gamma,\frac{4}{2\nu+5}\gamma\}}$, where $\gamma=\min\{\frac{2}{3}\alpha,\frac{2\beta+1}{2\beta+2}\alpha\}$, we have for any $y\in\mathcal{Y}$*

$$\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y)-E[f(Z)|W=y]=O_p(n^{-\min\{\frac{4}{9}\gamma,\frac{2\nu}{2\nu+5}\gamma\}}),\quad(n\to\infty),$$

where $\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y)$ is the estimator of $E[f(Z)|W=y]$ given by Eq. (11).

Proof. By applying Theorem 6 to $Y=(Y,X)$ and $Y=(Y,Y)$, we see that both of $\|\widehat{C}_{ZW}^{(n)}-C_{ZW}\|$ and $\|\widehat{C}_{WW}^{(n)}-C_{WW}\|$ are of $O_p(n^{-\gamma})$. Since

$$\begin{aligned}&\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(y)-E[f(Z)|W=y]\\ &\qquad=\langle k_\mathcal{Y}(\cdot,y),\widehat{C}_{WW}^{(n)}\big((\widehat{C}_{YY}^{(n)})^2+\delta_n I\big)^{-1}\widehat{C}_{WZ}^{(n)}f-E[f(Z)|W=\cdot]\rangle_{\mathcal{H}_\mathcal{Y}},\end{aligned}$$

combination of Theorems 6 and 7 proves the theorem. □

The next theorem shows the rate on average w.r.t. Q_W. The proof is similar to the above theorem, and omitted.

Theorem 10. *Let f be a function in $\mathcal{H}_\mathcal{X}$, (Z,W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ ($n \to \infty$) for some $0 < \alpha \le 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ with $\beta \ge 0$, and $E[f(Z)|W = \cdot] \in \mathcal{R}(\tilde{C}_{WW}^\nu) \cap \mathcal{H}_\mathcal{Y}$ for some $\nu > 0$. For the regularization constants $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{1}{1+\beta}\alpha\}}$ and $\delta_n = n^{-\max\{\frac{1}{2}\gamma, \frac{2}{\nu+2}\gamma\}}$, where $\gamma = \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$, we have*

$$\left\|\mathbf{f}_X^T R_{X|Y}\mathbf{k}_Y(W) - E[f(Z)|W]\right\|_{L^2(Q_W)} = O_p(n^{-\min\{\frac{1}{2}\gamma, \frac{\nu}{\nu+2}\gamma\}}), \quad (n \to \infty).$$

We have also the consistency of estimator for the kernel mean of posterior, if we make stronger assumptions. First, we formulate the mean of the conditional probability $q(x|y)$ in terms of operators. Let (Z,W) be a random variable with distribution Q. Assume that for any $f \in \mathcal{H}_\mathcal{X}$ the conditional mean $E[f(Z)|W = \cdot]$ is included in $\mathcal{H}_\mathcal{Y}$. We have a linear operator S defined by

$$S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}, \qquad f \mapsto E[f(Z)|W = \cdot].$$

If we further assume that S is bounded, the adjoint operator $S^* : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{X}$ satisfies

$$\langle S^* k_\mathcal{Y}(\cdot, y), f\rangle_{\mathcal{H}_\mathcal{X}} = \langle k_\mathcal{Y}(\cdot, y), Sf\rangle_{\mathcal{H}_\mathcal{Y}} = E[f(Z)|W = y]$$

for any $y \in \mathcal{Y}$, and thus $S^* k_\mathcal{Y}(\cdot, y)$ is equal to the kernel mean of conditional probability distribution of Z given $W = y$.

We make the following further assumptions:
Assumption (S)

1. The canonical map $A_W : \mathcal{H}_\mathcal{Y} \to L^2(Q_W)$ is injective, that is, C_{WW} is injective.
2. There exists $\nu > 0$ such that for any $f \in \mathcal{H}_\mathcal{X}$ there is $\eta_f \in \mathcal{H}_\mathcal{X}$ with $Sf = C_{WW}^\nu \eta_f$, and the linear map
$$C_{WW}^{-\nu} S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}, \qquad f \mapsto \eta_f$$
is bounded.

Theorem 11. *Let (Z,W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\widehat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\widehat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\alpha})$ ($n \to \infty$) for some $0 < \alpha \le 1/2$. Assume (S) above, and $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ with some $\beta \ge 0$. For the regularization constants $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{1}{1+\beta}\alpha\}}$ and $\delta_n = n^{-\max\{\frac{4}{9}\gamma, \frac{4}{2\nu+5}\gamma\}}$, where $\gamma = \min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}$, we have*

$$\left\|\mathbf{k}_X^T R_{X|Y}\mathbf{k}_Y(y) - m_{Q_\mathcal{X}|y}\right\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\min\{\frac{4}{9}\gamma, \frac{2\nu}{2\nu+5}\gamma\}}),$$

as $n \to \infty$, where $m_{Q_\mathcal{X}|y}$ is the kernel mean of the posterior given y.

Proof. First, in a similar manner to the proof of Eq. (20), we have

$$\left\|\widehat{C}_{ZW}^{(n)}\left((\widehat{C}_{WW}^{(n)})^2 + \delta_n I\right)^{-1}\widehat{C}_{WW}^{(n)} k_\mathcal{Y}(\cdot, y) - C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_\mathcal{Y}(\cdot, y)\right\|_{\mathcal{H}_\mathcal{X}} = O_p(n^{-\gamma}\delta_n^{-5/4}).$$

The assertion is thus obtained if

$$\left\|C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} k_\mathcal{Y}(\cdot, y) - S^* k_\mathcal{Y}(\cdot, y)\right\|_{\mathcal{H}_\mathcal{X}} = O(\delta_n^{\min\{1, \frac{\nu}{2}\}}) \tag{23}$$

is proved. The left hand side of Eq. (23) is upper-bounded by

$$\left\|C_{ZW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} - S^*\right\| \|k_\mathcal{Y}(\cdot, y)\|_{\mathcal{H}_\mathcal{Y}} = \left\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} - S\right\| \|k_\mathcal{Y}(\cdot, y)\|_{\mathcal{H}_\mathcal{Y}}.$$

It follows from Theorem 1 that $C_{WZ} = C_{WW}S$, and thus $\|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WZ} - S\| = \|C_{WW}(C_{WW}^2 + \delta_n I)^{-1} C_{WW} S - S\| \le \delta_n \|(C_{WW}^2 + \delta_n I)^{-1} C_{WW}^\nu\| \|C_{WW}^{-\nu} S\|$. The eigendecomposition of C_{WW} together with the inequality $\frac{\delta_n \lambda^\nu}{\lambda^2 + \delta_n} \le \delta_n^{\min\{1, \nu/2\}}$ ($\lambda \ge 0$) completes the proof. $\square$

References

[1] A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[2] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 2000.

[3] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximation. *Constructive Approximation*, 26:153–172, 2007.