

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

Kernel Bayes' Rule

Anonymous Author(s)

Affiliation

Address

email

Abstract

A nonparametric kernel-based method for realizing Bayes' rule is proposed, based on kernel representations of probabilities in reproducing kernel Hilbert spaces. The prior and conditional probabilities are expressed as empirical kernel mean and covariance operators, respectively, and the kernel mean of the posterior distribution is computed in the form of a weighted sample. The kernel Bayes' rule can be applied to a wide variety of Bayesian inference problems: we demonstrate Bayesian computation without likelihood, and filtering with a nonparametric state-space model. A consistency rate for the posterior estimate is established.

1 Introduction

Kernel methods have long provided powerful tools for generalizing linear statistical approaches to nonlinear settings, through an embedding of the sample to a high dimensional feature space, namely a reproducing kernel Hilbert space (RKHS) [16]. The inner product between feature mappings need never be computed explicitly, but is given by a positive definite kernel function, which permits efficient computation without the need to deal explicitly with the feature representation. More recently, the *mean* of the RKHS feature map has been used to represent probability distributions, rather than mapping single points: we will refer to these representations of probability distributions as *kernel means*. With an appropriate choice of kernel, the feature mapping becomes rich enough that its expectation uniquely identifies the distribution: the associated RKHSs are termed *characteristic* [6, 7, 22]. Kernel means in characteristic RKHSs have been applied successfully in a number of statistical tasks, including the two sample problem [9], independence tests [10], and conditional independence tests [8]. An advantage of the kernel approach is that these tests apply immediately to any domain on which kernels may be defined.

We propose a general nonparametric framework for Bayesian inference, expressed entirely in terms of kernel means. The goal of Bayesian inference is to find the posterior of x given observation y ;

$$q(x|y) = \frac{p(y|x)\pi(x)}{q_Y(y)}, \quad q_Y(y) = \int p(y|x)\pi(x)d\mu_{\mathcal{X}}(x), \quad (1)$$

where $\pi(x)$ and $p(y|x)$ are respectively the density function of the prior, and the conditional density or likelihood of y given x . In our framework, the posterior, prior, and likelihood are all expressed as kernel means: the update from prior to posterior is called the Kernel Bayes' Rule (KBR). To implement KBR, the kernel means are learned nonparametrically from training data: the prior and likelihood means are expressed in terms of samples from the prior and joint probabilities, and the posterior as a kernel mean of a weighted sample. The resulting updates are straightforward matrix operations. This leads to the main advantage of the KBR approach: in the absence of a specific parametric model or an analytic form for the prior and likelihood densities, we can still perform Bayesian inference by making sufficient observations on the system. Alternatively, we may have a parametric model, but it might be complex and require time-consuming sampling techniques for inference. By contrast, KBR is simple to implement, and is amenable to well-established approximation techniques which yield an overall computational cost linear in the training sample size [5]. We further

054 establish the rate of consistency of the estimated posterior kernel mean to the true posterior, as a
 055 function of training sample size.

056 The proposed kernel realization of Bayes' rule is an extension of the approach used in [20] for state-
 057 space models. This earlier work applies a heuristic, however, in which the kernel mean of the pre-
 058 vious hidden state and the observation are assumed to combine additively to update the hidden state
 059 estimate. More recently, a method for belief propagation using kernel means was proposed [18, 19]:
 060 unlike the present work, this assumes the prior to be uniform. An alternative to kernel means would
 061 be to use nonparametric density estimates. Classical approaches include finite distribution estimates
 062 on a partitioned domain or kernel density estimation, which perform poorly on high dimensional
 063 data. Alternatively, direct estimates of the density ratio may be used in estimating the conditional
 064 p.d.f. [24]. By contrast with density estimation approaches, KBR makes it easy to compute posterior
 065 expectations (as an RKHS inner product) and to perform conditioning and marginalization, without
 066 requiring numerical integration.

068 2 Kernel expression of Bayes' rule

070 2.1 Positive definite kernel and probabilities

071 We begin with a review some basic concepts and tools concerning statistics on RKHS [1, 3, 6, 7].
 072 Given a set Ω , a (\mathbb{R} -valued) positive definite kernel k on Ω is a symmetric kernel $k : \Omega \times \Omega \rightarrow \mathbb{R}$ such
 073 that $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for arbitrary points x_1, \dots, x_n in Ω and real numbers c_1, \dots, c_n . It is
 074 known [1] that a positive definite kernel on Ω uniquely defines a Hilbert space \mathcal{H} (RKHS) consisting
 075 of functions on Ω , where $\langle f, k(\cdot, x) \rangle = f(x)$ for any $x \in \Omega$ and $f \in \mathcal{H}$ (reproducing property).
 076

077 Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ be measure spaces, and (X, Y) be a random variable on $\mathcal{X} \times$
 078 \mathcal{Y} with probability P . Throughout this paper, it is assumed that positive definite kernels on the
 079 measurable spaces are measurable and bounded, where boundedness is defined as $\sup_{x \in \Omega} k(x, x) <$
 080 ∞ . Let $k_{\mathcal{X}}$ be a positive definite kernel on a measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, with RKHS $\mathcal{H}_{\mathcal{X}}$. The *kernel*
 081 *mean* m_X of X on $\mathcal{H}_{\mathcal{X}}$ is defined by the mean of the $\mathcal{H}_{\mathcal{X}}$ -valued random variable $k_{\mathcal{X}}(\cdot, X)$, namely

$$082 m_X = \int k_{\mathcal{X}}(\cdot, x) dP_X(x). \quad (2)$$

083 For notational simplicity, the dependence on $k_{\mathcal{X}}$ in m_X is not shown. Since the kernel mean depends
 084 only on the distribution of X (and the kernel), it may also be written m_{P_X} ; we will use whichever
 085 of these equivalent notations is clearest in each context. From the reproducing property, we have

$$087 \langle f, m_X \rangle = E[f(X)] \quad (\forall f \in \mathcal{H}_{\mathcal{X}}). \quad (3)$$

088 Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive definite kernels on \mathcal{X} and \mathcal{Y} with respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. The
 089 (uncentered) *covariance operator* $C_{YX} : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}$ is defined by the relation

$$090 \langle g, C_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = E[f(X)g(Y)] \quad (= \langle g \otimes f, m_{(YX)} \rangle_{\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}}) \quad (\forall f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}).$$

091 It should be noted that C_{YX} is identified with the mean $m_{(YX)}$ in the tensor product space $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$,
 092 which is given by the product kernel $k_{\mathcal{Y}} k_{\mathcal{X}}$ [1]. The identification is standard: the tensor product is
 093 isomorphic to the space of linear maps by the correspondence $\psi \otimes \phi \leftrightarrow [h \mapsto \psi(\phi, h)]$. We also
 094 define $C_{XX} : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{X}}$ by $\langle f_2, C_{XX} f_1 \rangle = E[f_2(X)f_1(X)]$ for any $f_1, f_2 \in \mathcal{H}_{\mathcal{X}}$.
 095

096 We next introduce the notion of a characteristic RKHS, which is essential when using kernels to ma-
 097 nipulate probability measures. A bounded measurable positive definite kernel k is called *character-*
 098 *istic* if $E_{X \sim P}[k(\cdot, X)] = E_{X' \sim Q}[k(\cdot, X')]$ implies $P = Q$: probabilities are uniquely determined
 099 by their kernel means [7, 22]. With this property, problems of statistical inference can be cast in
 100 terms of inference on the kernel means. A widely used characteristic kernel on \mathbb{R}^m is the Gaussian
 101 kernel, $\exp(-\|x - y\|^2 / (2\sigma^2))$.

102 Empirical estimates of the kernel mean and covariance operator are straightforward to obtain. Given
 103 an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ with law P , the empirical kernel mean and covariance op-
 104 erator are respectively

$$105 \hat{m}_X^{(n)} = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{X}}(\cdot, X_i), \quad \hat{C}_{YX}^{(n)} = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{Y}}(\cdot, Y_i) \otimes k_{\mathcal{X}}(\cdot, X_i),$$

106 where $\hat{C}_{YX}^{(n)}$ is written in the tensor product form. These are known to be \sqrt{n} -consistent in norm.
 107

2.2 Kernel Bayes' rule

We now derive the kernel mean implementation of Bayes' rule. Let Π be a *prior* distribution on \mathcal{X} with p.d.f. $\pi(x)$. In the following, Q and Q_Y denote the probabilities with p.d.f. $q(x, y) = p(y|x)\pi(x)$ and $q_Y(y)$ in Eq. (1), respectively. Our goal is to obtain an estimator of the kernel mean of posterior $m_{Q_X|Y} = \int k_{\mathcal{X}}(\cdot, x)q(x|y)d\mu_{\mathcal{X}}(x)$. The following theorem is fundamental in manipulating conditional probabilities with positive definite kernels.

Theorem 1 ([6]). *If $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ holds for $g \in \mathcal{H}_{\mathcal{Y}}$, then*

$$C_{XX}E[g(Y)|X = \cdot] = C_{XY}g.$$

If C_{XX} is injective, the above relation can be expressed as

$$E[g(Y)|X = \cdot] = C_{XX}^{-1}C_{XY}g. \quad (4)$$

Using Eq. (4), we can obtain an expression for the kernel mean of Q_Y .

Theorem 2 ([20]). *Assume C_{XX} is injective, and let m_{Π} and m_{Q_Y} be the kernel means of Π in $\mathcal{H}_{\mathcal{X}}$ and Q_Y in $\mathcal{H}_{\mathcal{Y}}$, respectively. If $m_{\Pi} \in \mathcal{R}(C_{XX})$ and $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ for any $g \in \mathcal{H}_{\mathcal{Y}}$, then*

$$m_{Q_Y} = C_{YX}C_{XX}^{-1}m_{\Pi}. \quad (5)$$

As discussed in [20], the operator $C_{YX}C_{XX}^{-1}$ implements forward filtering of the prior π with the conditional density $p(y|x)$, as in Eq. (1). Note, however, that the assumptions $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ and injectivity of C_{XX} may not hold in general; we can easily provide counterexamples. In the following, we nonetheless derive a population expression of Bayes' rule under these strong assumptions, use it as a prototype for an empirical estimator expressed in terms of Gram matrices, and prove its consistency subject to appropriate smoothness conditions on the distributions.

In deriving kernel realization of Bayes' rule, we will also use Theorem 2 to obtain a kernel mean representation of the *joint* probability Q :

$$m_Q = C_{(YX)X}C_{XX}^{-1}m_{\Pi} \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}. \quad (6)$$

In the above equation, $C_{(YX)X}$ is the covariance operator from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ with p.d.f. $\tilde{p}((y, x), x') = p(x, y)\delta_x(x')$, where $\delta_x(x')$ is the point measure at x .

In many applications of Bayesian inference, the probability conditioned on a particular value should be computed. By plugging the point measure at x into Π in Eq. (5), we have a population expression

$$E[k_{\mathcal{Y}}(\cdot, Y)|X = x] = C_{YX}C_{XX}^{-1}k_{\mathcal{X}}(\cdot, x), \quad (7)$$

which was used by [20, 18, 19] as the kernel mean of the conditional probability $p(y|x)$. Let (Z, W) be a random variable on $\mathcal{X} \times \mathcal{Y}$ with law Q . Replacing P by Q and x by y in Eq. (7), we obtain

$$E[k_{\mathcal{X}}(\cdot, Z)|W = y] = C_{ZW}C_{WW}^{-1}k_{\mathcal{Y}}(\cdot, y). \quad (8)$$

This is exactly the kernel mean of the posterior. The next step is to derive the covariance operators in Eq. (8). Recalling that the mean $m_Q = m_{(ZW)} \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ can be identified with the covariance operator $C_{ZW} : \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{H}_{\mathcal{X}}$, and $m_{(WW)} \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$ with C_{WW} , we use Eq. (6) to obtain the operators in Eq. (8), and thus the kernel mean expression of Bayes' rule.

The above argument can be rigorously implemented for empirical estimates of the kernel means and covariances. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an i.i.d. sample with law P , and assume a consistent estimator for m_{Π} given by

$$\hat{m}_{\Pi}^{(\ell)} = \sum_{j=1}^{\ell} \gamma_j k_{\mathcal{X}}(\cdot, U_j),$$

where U_1, \dots, U_{ℓ} is the sample that defines the estimator (which need not be generated by Π), and γ_j are the weights. Negative values are allowed for γ_j . The empirical estimators for C_{ZW} and C_{WW} are identified with $\hat{m}_{(ZW)}$ and $\hat{m}_{(WW)}$, respectively. From Eq. (6), they are given by

$$\hat{m}_Q = \hat{m}_{(ZW)} = \hat{C}_{(YX)X}^{(n)} (\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{m}_{\Pi}^{(\ell)}, \quad \hat{m}_{(WW)} = \hat{C}_{(YY)X}^{(n)} (\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{m}_{\Pi}^{(\ell)},$$

where I is the identity and ε_n is the coefficient of Tikhonov regularization for operator inversion.

The next two propositions express these estimators using Gram matrices. The proofs are simple matrix manipulation and shown in Supplementary material. In the following, G_X and G_Y denote the Gram matrices $(k_{\mathcal{X}}(X_i, X_j))$ and $(k_{\mathcal{Y}}(Y_i, Y_j))$, respectively.

Input: (i) $\{(X_i, Y_i)\}_{i=1}^n$: sample to express P . (ii) $\{(U_j, \gamma_j)\}_{j=1}^\ell$: weighted sample to express the kernel mean of the prior \hat{m}_Π . (iii) ε_n, δ_n : regularization constants.

Computation:

1. Compute Gram matrices $G_X = (k_X(X_i, X_j))$, $G_Y = (k_Y(Y_i, Y_j))$, and a vector $\hat{\mathbf{m}}_\Pi = (\sum_{j=1}^\ell \gamma_j k_X(X_i, U_j))_{i=1}^n \in \mathbb{R}^n$.
2. Compute $\hat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1} \hat{\mathbf{m}}_\Pi$.
3. Compute $R_{X|Y} = \Lambda G_Y ((\Lambda G_Y)^2 + \delta_n I_n)^{-1} \Lambda$, where $\Lambda = \text{Diag}(\hat{\mu})$.

Output: $n \times n$ matrix $R_{X|Y}$.

Given conditioning value y , the kernel mean of the posterior $q(x|y)$ is estimated by the weighted sample $\{(X_i, w_i)\}_{i=1}^n$ with $w = R_{X|Y} \mathbf{k}_Y(y)$, where $\mathbf{k}_Y(y) = (k_Y(Y_i, y))_{i=1}^n$.

Figure 1: Kernel Bayes' Rule Algorithm

Proposition 3. *The Gram matrix expressions of \hat{C}_{ZW} and \hat{C}_{WW} are given by*

$\hat{C}_{ZW} = \sum_{i=1}^n \hat{\mu}_i k_X(\cdot, X_i) \otimes k_Y(\cdot, Y_i)$ and $\hat{C}_{WW} = \sum_{i=1}^n \hat{\mu}_i k_Y(\cdot, Y_i) \otimes k_Y(\cdot, Y_i)$, respectively, where the common coefficient $\hat{\mu} \in \mathbb{R}^n$ is

$$\hat{\mu} = n(G_X + n\varepsilon_n I_n)^{-1} \hat{\mathbf{m}}_\Pi, \quad \hat{\mathbf{m}}_{\Pi, i} = \hat{m}_\Pi(X_i) = \sum_{j=1}^\ell \gamma_j k_X(X_i, U_j). \quad (9)$$

Prop. 3 implies that the probabilities Q and Q_Y are estimated by the weighted samples $\{(X_i, Y_i), \hat{\mu}_i\}_{i=1}^n$ and $\{(Y_i, \hat{\mu}_i)\}_{i=1}^n$, respectively, with common weights. Since the weights $\hat{\mu}_i$ may be negative, we use another type of Tikhonov regularization in computing Eq. (8),

$$\hat{m}_{Q_{x|y}} := \hat{C}_{ZW} (\hat{C}_{WW}^2 + \delta_n I)^{-1} \hat{C}_{WW} k_Y(\cdot, y). \quad (10)$$

Proposition 4. *For any $y \in \mathcal{Y}$, the Gram matrix expression of $\hat{m}_{Q_{x|y}}$ is given by*

$$\hat{m}_{Q_{x|y}} = \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y), \quad R_{X|Y} := \Lambda G_Y ((\Lambda G_Y)^2 + \delta_n I_n)^{-1} \Lambda, \quad (11)$$

where $\Lambda = \text{Diag}(\hat{\mu})$ is a diagonal matrix with elements $\hat{\mu}_i$ given by Eq. (9), $\mathbf{k}_X = (k_X(\cdot, X_1), \dots, k_X(\cdot, X_n))^T \in \mathcal{H}_X^n$, and $\mathbf{k}_Y = (k_Y(\cdot, Y_1), \dots, k_Y(\cdot, Y_n))^T \in \mathcal{H}_Y^n$.

We call Eqs.(10) or (11) the *kernel Bayes' rule* (KBR): i.e., the expression of Bayes' rule entirely in terms of kernel means. The algorithm to implement KBR is summarized in Fig. 1. If our aim is to estimate $E[f(Z)|W = y]$, that is, the expectation of a function $f \in \mathcal{H}_X$ with respect to the posterior, then based on Eq. (3) an estimator is given by

$$\langle f, \hat{m}_{Q_{x|y}} \rangle_{\mathcal{H}_X} = \mathbf{f}_X^T R_{X|Y} \mathbf{k}_Y(y), \quad (12)$$

where $\mathbf{f}_X = (f(X_1), \dots, f(X_n))^T \in \mathbb{R}^n$. In using a weighted sample to represent the posterior, KBR has some similarity to Monte Carlo methods such as importance sampling and sequential Monte Carlo ([4]). The KBR method, however, does not generate samples from the posterior, but updates the weights of a sample via matrix operations. We will provide experimental comparisons between KBR and sampling methods in Sec. 4.1.

2.3 Consistency of KBR estimator

We now demonstrate the consistency of the KBR estimator in Eq. (12). We show only the best rate that can be derived under the assumptions, and leave more detailed discussions and proofs to the Supplementary material. We assume that the sample size $\ell = \ell_n$ for the prior goes to infinity as the sample size n for the likelihood goes to infinity, and that $\hat{m}_\Pi^{(\ell_n)}$ is n^α -consistent. In the theoretical results, we assume all Hilbert spaces are separable. In the following, $\mathcal{R}(A)$ denotes the range of A .

Theorem 5. *Let $f \in \mathcal{H}_X$, (Z, W) be a random vector on $\mathcal{X} \times \mathcal{Y}$ such that its law is Q with p.d.f. $p(y|x)\pi(x)$, and $\hat{m}_\Pi^{(\ell_n)}$ be an estimator of m_Π such that $\|\hat{m}_\Pi^{(\ell_n)} - m_\Pi\|_{\mathcal{H}_X} = O_p(n^{-\alpha})$ as $n \rightarrow \infty$ for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$, where p_X is the p.d.f. of P_X , and $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^2)$. For $\varepsilon_n = n^{-\frac{2}{3}\alpha}$ and $\delta_n = n^{-\frac{8}{27}\alpha}$, we have for any $y \in \mathcal{Y}$*

$$\mathbf{f}_X^T R_{X|Y} \mathbf{k}_Y(y) - E[f(Z)|W = y] = O_p(n^{-\frac{8}{27}\alpha}), \quad (n \rightarrow \infty),$$

where $\mathbf{f}_X^T R_{X|Y} \mathbf{k}_Y(y)$ is the estimator of $E[f(Z)|W = y]$ given by Eq. (12).

The condition $\pi/p_X \in \mathcal{R}(C_{XX}^{1/2})$ requires the prior to be smooth. If $\ell_n = n$, and if $\widehat{m}_\Pi^{(n)}$ is a direct empirical kernel mean with an i.i.d. sample of size n from Π , typically $\alpha = 1/2$ and the theorem implies $n^{4/27}$ -consistency. While this might seem to be a slow rate, in practice the convergence may be much faster than the above theoretical guarantee.

3 Bayesian inference with Kernel Bayes' Rule

In Bayesian inference, tasks of interest include finding properties of the posterior (MAP value, moments), and computing the expectation of a function under the posterior. We now demonstrate the use of the kernel mean obtained via KBR in solving these problems.

First, we have already seen from Theorem 5 that we may obtain a consistent estimator under the posterior for the expectation of some $f \in \mathcal{H}_X$. This covers a wide class of functions when characteristic kernels are used (see also experiments in Sec. 4.1).

Next, regarding a point estimate of x , [20] proposes to use the preimage $\widehat{x} = \arg \min_x \|k_X(\cdot, x) - \mathbf{k}_X^T R_{X|Y} \mathbf{k}_Y(y)\|_{\mathcal{H}_X}^2$, which represents the posterior mean most effectively by one point. We use this approach in the present paper where point estimates are considered. In the case of the Gaussian kernel, a fixed point method can be used to sequentially optimize x [13].

In KBR the prior and likelihood are expressed in terms of samples. Thus unlike many methods for Bayesian inference, exact knowledge on their densities is not needed, once samples are obtained. The following are typical situations where the KBR approach is advantageous:

- The relation among variables is difficult to realize with a simple parametric model, however we can obtain samples of the variables (e.g. nonparametric state-space model in Sec. 3).
- The p.d.f of the prior and/or likelihood is hard to obtain explicitly, but sampling is possible: (a) In population genetics, branching processes are used for the likelihood to model the split of species, for which the explicit density is hard to obtain. Approximate Bayesian Computation (ABC) is a popular sampling method in these situations [25, 12, 17]. (b) In nonparametric Bayesian inference (e.g. [14]), the prior is typically given in the form of a process without a density. The KBR approach can give alternative ways of Bayesian computation for these problems. We will show some experimental comparisons between KBR approach and ABC in Sec. 4.2.
- If a standard sampling method such as MCMC or sequential MC is applicable, the computation given y may be time consuming, and real-time applications may not be feasible. Using KBR, the expectation of the posterior given y is obtained simply by the inner product as in Eq. (12), once $\mathbf{f}_X^T R_{X|Y}$ has been computed.

The KBR approach nonetheless has a weakness common to other nonparametric methods: if a new data point appears far from the training sample, the reliability of the output will be low. Thus, we need sufficient diversity in training sample to reliably estimate the posterior.

In KBR computation, Gram matrix inversion is necessary, which would cost $O(n^3)$ for sample size n if attempted directly. Substantial cost reductions can be achieved by low rank matrix approximations such as the incomplete Cholesky decomposition [5], which approximates a Gram matrix in the form of $\Gamma \Gamma^T$ with $n \times r$ matrix Γ . Computing Γ costs $O(nr^2)$, and with the Woodbury identity, the KBR can be approximately computed with cost $O(nr^2)$.

Kernel choice or model selection is key to the effectiveness of KBR, as in other kernel methods. KBR involves three model parameters: the kernel (or its parameters), and the regularization parameters ε_n and δ_n . The strategy for parameter selection depends on how the posterior is to be used in the inference problem. If it is applied in a supervised setting, we can use standard cross-validation (CV). A more general approach requires constructing a related supervised problem. Suppose the prior is given by the marginal P_X of P . The posterior density $q(x|y)$ averaged with P_Y is then equal to the marginal density p_X . We are then able to compare the discrepancy of the kernel mean of P_X and the average of the estimators $\widehat{Q}_{\mathcal{X}|y=Y_i}$ over Y_i . This leads to application of K -fold CV approach. Namely, for a partition of $\{1, \dots, n\}$ into K disjoint subsets $\{T_a\}_{a=1}^K$, let $\widehat{m}_{Q_{\mathcal{X}|y}}^{[-a]}$ be the kernel mean of posterior estimated with data $\{(X_i, Y_i)\}_{i \notin T_a}$, and the prior mean $\widehat{m}_X^{[-a]}$ with data $\{X_i\}_{i \notin T_a}$. We use $\sum_{a=1}^K \left\| \frac{1}{|T_a|} \sum_{j \in T_a} \widehat{m}_{Q_{\mathcal{X}|y=Y_j}}^{[-a]} - \widehat{m}_X^{[a]} \right\|_{\mathcal{H}_X}^2$ for CV, where $\widehat{m}_X^{[a]} = \frac{1}{|T_a|} \sum_{j \in T_a} k_X(\cdot, X_j)$.

Application to nonparametric state-space model. Consider the state-space model,

$$p(X, Y) = \pi(X_1) \prod_{t=1}^T p(Y_t | X_t) \prod_{t=1}^{T-1} q(X_{t+1} | X_t),$$

where Y_t is observable and X_t is a hidden state. We do not assume the conditional probabilities $p(Y_t | X_t)$ and $q(X_{t+1} | X_t)$ to be known explicitly, nor do we estimate them with simple parametric models. Rather, we assume a sample $(X_1, Y_1), \dots, (X_{T+1}, Y_{T+1})$ is given for both the observable and hidden variables in the training phase. This problem has already been considered in [20], but we give a more principled approach based on KBR. The conditional probability for the transition $q(x_{t+1} | x_t)$ and observation process $p(y | x)$ are represented by the covariance operators as computed with the training sample; $\widehat{C}_{X, X_{+1}} = \frac{1}{T} \sum_{i=1}^T k_{\mathcal{X}}(\cdot, X_i) \otimes k_{\mathcal{X}}(\cdot, X_{i+1})$, $\widehat{C}_{XY} = \frac{1}{T} \sum_{i=1}^T k_{\mathcal{X}}(\cdot, X_i) \otimes k_{\mathcal{Y}}(\cdot, Y_i)$, and \widehat{C}_{YX} and \widehat{C}_{XX} are defined similarly. Note that though the data are not i.i.d., consistency is achieved by the mixing property of the Markov model.

For simplicity, we focus on the filtering problem, but smoothing and prediction can be done similarly. In filtering, we wish to estimate the current hidden state x_t , given observations $\tilde{y}_1, \dots, \tilde{y}_t$. The sequential estimate of $p(x_t | \tilde{y}_1, \dots, \tilde{y}_t)$ can be derived using KBR (we give only a sketch below; see Supplementary material for the detailed derivation). Suppose we already have an estimator of the kernel mean of $p(x_t | \tilde{y}_1, \dots, \tilde{y}_t)$ in the form

$$\widehat{m}_{x_t | \tilde{y}_1, \dots, \tilde{y}_t} = \sum_{i=1}^T \alpha_i^{(t)} k_{\mathcal{X}}(\cdot, X_i),$$

where $\alpha_i^{(t)} = \alpha_i^{(t)}(\tilde{y}_1, \dots, \tilde{y}_t)$ are the coefficients at time t . By applying Theorem 2 twice, the kernel mean of $p(y_{t+1} | \tilde{y}_1, \dots, \tilde{y}_t)$ is estimated by $\widehat{m}_{y_{t+1} | \tilde{y}_1, \dots, \tilde{y}_t} = \sum_{i=1}^T \widehat{\mu}_i^{(t+1)} k_{\mathcal{Y}}(\cdot, Y_i)$, where

$$\widehat{\mu}^{(t+1)} = \left(\frac{1}{T} G_X + \varepsilon_T I_T\right)^{-1} G_{X, X_{+1}} \left(\frac{1}{T} G_X + \varepsilon_T I_T\right)^{-1} G_X \alpha^{(t)}. \quad (13)$$

Here $G_{X_{+1}X}$ is the ‘‘transfer’’ matrix defined by $(G_{X_{+1}X})_{ij} = k_{\mathcal{X}}(X_{i+1}, X_j)$. With the notation $\Lambda^{(t+1)} = \text{Diag}(\widehat{\mu}_1^{(t+1)}, \dots, \widehat{\mu}_T^{(t+1)})$, kernel Bayes’ rule yields

$$\alpha^{(t+1)} = \Lambda^{(t+1)} G_Y \left((\Lambda^{(t+1)} G_Y)^2 + \delta_T I_T \right)^{-1} \Lambda^{(t+1)} \mathbf{k}_Y(\tilde{y}_{t+1}). \quad (14)$$

Eqs. (13) and (14) describe the update rule of $\alpha^{(t)}(\tilde{y}_1, \dots, \tilde{y}_t)$. By contrast with [20], where the estimates of the previous hidden state and observation are assumed to combine additively, the above derivation is based only on applying KBR. In sequential filtering, a substantial reduction of computational cost can be achieved by low rank approximations for the matrices of a training phase: given rank r , the computation costs only $O(Tr^2)$ for each step in filtering.

Bayesian computation without likelihood. When the likelihood and/or prior is not obtained in an analytic form but sampling is possible, the ABC approach [25, 12, 17] is popular for Bayesian computation. The ABC *rejection method* generates a sample from $q(X | Y = y)$ as follows: (1) generate X_t from the prior Π , (2) generate Y_t from $p(y | X_t)$, (3) if $D(y, Y_t) < \rho$, accept X_t ; otherwise reject, (4) go to (1). In Step (3), D is a distance on \mathcal{X} , and ρ is the tolerance to acceptance.

In the exactly the same situation as the above, the KBR approach gives the following method: (i) generate X_1, \dots, X_n from the prior Π , (ii) generate a sample Y_t from $p(y | X_t)$ ($t = 1, \dots, n$), (iii) compute Gram matrices G_X and G_Y with $(X_1, Y_1), \dots, (X_n, Y_n)$, and $R_{X|Y} \mathbf{k}_Y(y)$.

The distribution of a sample given by ABC approaches the true posterior if $\rho \rightarrow 0$, while the empirical posterior estimate of KBR converges to the true one as $n \rightarrow \infty$. The efficiency of ABC, however, can be arbitrarily low for a small ρ , since X_t is then rarely accepted in Step (3). Finally, ABC generates a sample, which allows any statistic of the posterior to be approximated. In the case of KBR, certain statistics of the posterior (such as confidence intervals) can be harder to obtain, since consistency is guaranteed only for expectations of RKHS functions. In Sec. 4.2, we provide experimental comparisons addressing the trade-off between computational time and accuracy for ABC and KBR.

4 Experiments

4.1 Nonparametric inference of posterior

First we compare KBR and the standard kernel density estimation (KDE). Let $\{(X_i, Y_i)\}_{i=1}^n$ be an i.i.d. sample from P on $\mathbb{R}^d \times \mathbb{R}^r$. With p.d.f. $K(x)$ on \mathbb{R}^d and $H(y)$ on \mathbb{R}^r , the conditional

324 p.d.f. $p(y|x)$ is estimated by $\hat{p}(y|x) = \sum_{j=1}^n K_{h_X}(x - X_j)H_{h_Y}(y - Y_j) / \sum_{j=1}^n K_{h_X}(x - X_j)$,
 325 where $K_{h_X}(x) = h_X^{-d}K(x/h_X)$ and $H_{h_Y}(x) = h_Y^{-r}H(y/h_Y)$. Given an i.i.d. sample $\{U_j\}_{j=1}^\ell$
 326 from the prior Π , the posterior $q(x|y)$ is represented by the weighted sample (U_i, w_i) with $w_i =$
 327 $\hat{p}(y|U_i) / \sum_{j=1}^\ell \hat{p}(y|U_j)$ as importance weight (IW).
 328

329 We compare the estimates of $\int xq(x|y)dx$ obtained by KBR and KDE + IW, using Gaussian kernels
 330 for both the methods. Note that with Gaussian kernel, the function $f(x) = x$ does not belong to
 331 $\mathcal{H}_{\mathcal{X}}$, and the consistency of the KBR method is not rigorously guaranteed (*c.f.* Theorem 5). Gaussian
 332 kernels, however, are known to be able to approximate any continuous function on a compact subset
 333 with arbitrary accuracy [23]. We can thus expect that the posterior mean can be estimated effectively.
 334

335 In the experiments, the dimensionality was given by
 336 $r = d$ ranging from 2 to 64. The distribution P of
 337 (X, Y) was $N((0, 1_d)^T, V)$ with V randomly generated
 338 for each run. The prior Π was $P_X = N(0, V_{XX}/2)$,
 339 where V_{XX} is the X -component of V . The sample sizes
 340 were $n = \ell = 200$. The bandwidth parameter h_X, h_Y
 341 in KDE were set $h_X = h_Y$ and chosen by two ways,
 342 the least square cross-validation [15] and the best mean
 343 performance, over the set $\{2 * i \mid i = 1, \dots, 10\}$. For
 344 the KBR, we used two methods to choose the deviation
 345 parameter in Gaussian kernel: the median over the
 346 pairwise distances in the data [10] and the 10-fold CV
 347 described in Sec. 3. Fig. 2 shows the MSE of the estimates
 348 over 1000 random points $y \sim N(0, V_{YY})$. While the accuracy of the both methods decrease
 349 for larger dimensionality, the KBR significantly outperforms the KDE+IW.

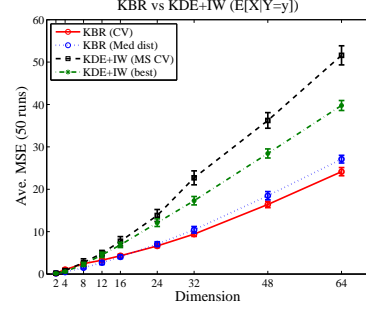


Figure 2: KBR v.s. KDE+IW.

349 4.2 Bayesian computation without likelihood

350 We compare KBR and ABC in terms of the estimation
 351 accuracy and computational time. To compute the
 352 estimation accuracy rigorously, Gaussian distributions
 353 are used for the true prior and likelihood. The sam-
 354 ples are taken from the same model as in Sec. 4.1, and
 355 $\int xq(x|y)dx$ is evaluated at 10 different points of y . We
 356 performed 10 runs with different covariance.
 357

358 For ABC, we used only the rejection method; while
 359 there are more advanced sampling schemes [12, 17], im-
 360 plementation is not straightforward. Various parameters
 361 for the acceptance are used, and the accuracy and com-
 362 putational time are shown in Fig.3 together with total
 363 sizes of generated samples. For the KBR method, the sample sizes n of the likelihood and prior are
 364 varied. The regularization parameters are given by $\varepsilon_n = 0.01/n$ and $\delta_n = 2\varepsilon_n$. In KBR, Gaussian
 365 kernels are used and the incomplete Cholesky decomposition is employed. The results indicate that
 366 KBR achieves more accurate results than ABC in the same computational time.

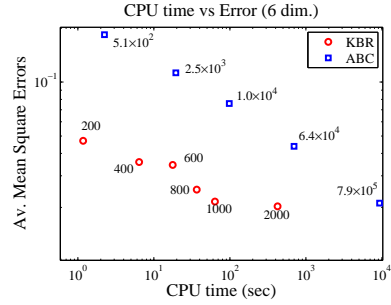


Figure 3: Estimation accuracy and computational time with KBR and ABC.

367 4.3 Filtering problems

368 The KBR filter proposed in Sec. 3 is applied. Alternative strategies for state-space models with
 369 complex dynamics involve the extended Kalman filter (EKF) and unscented Kalman filter (UKF,
 370 [11]). There are some works on nonparametric state-space model or HMM which use nonparametric
 371 estimation of conditional p.d.f. such as KDE or partitions [27, 26] and, more recently, kernel method
 372 [20, 21]. In the following, the KBR method is compared with linear and nonlinear Kalman filters.
 373

374 KBR has the regularization parameters ε_T, δ_T , and kernel parameters for $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ (*e.g.*, the de-
 375 viation parameter for Gaussian kernel). The validation approach is applied for selecting them by
 376 dividing the training sample into two. To reduce the search space, we set $\delta_T = 2\varepsilon_T$ and use the
 377 Gaussian kernel deviation $\beta\sigma_{\mathcal{X}}$ and $\beta\sigma_{\mathcal{Y}}$, where $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{Y}}$ are the median of pairwise distances
 among the training samples ([10]), leaving only two parameters β and ε_T to be tuned.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

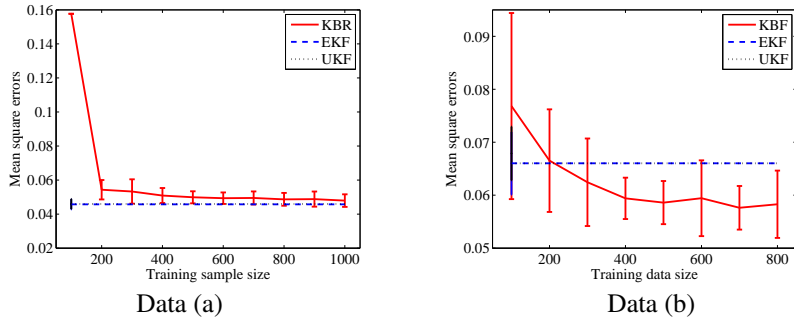


Figure 4: Comparisons with the KBR Filter and EKF. (Average MSEs over 30 runs.)

	KBR (Gauss)	KBR (Tr)	Kalman (9 dim.)	Kalman (Quat.)
$\sigma^2 = 10^{-4}$	0.210 ± 0.015	0.146 ± 0.003	1.980 ± 0.083	0.557 ± 0.023
$\sigma^2 = 10^{-3}$	0.222 ± 0.009	0.210 ± 0.008	1.935 ± 0.064	0.541 ± 0.022

Table 1: Average MSE of camera angle estimates (10 runs).

We first use two synthetic data sets with KBR, EKF, and UKF, assuming that EKF and UKF *know* the exact dynamics. The dynamics has a hidden state $X_t = (u_t, v_t)^T \in \mathbb{R}^2$, and is given by

$$(u_{t+1}, v_{t+1}) = (1 + b \sin(M\theta_{t+1}))(\cos \theta_{t+1}, \sin \theta_{t+1}) + Z_t, \quad \theta_{t+1} = \theta_t + \eta \pmod{2\pi},$$

where $Z_t \sim N(0, \sigma_h^2 I_2)$ is independent noise. Note that the dynamics of (u_t, v_t) is nonlinear even for $b = 0$. The observation Y_t follows $Y_t = X_t + W_t$, where $W_t \sim N(0, \sigma_o^2 I)$. The two dynamics are defined as follows: (a) (noisy rotation) $\eta = 0.3$, $b = 0$, $\sigma_h = \sigma_o = 0.2$, (b) (noisy oscillatory rotation) $\eta = 0.4$, $b = 0.4$, $M = 8$, $\sigma_h = \sigma_o = 0.2$. The results are shown in Fig. 4. In all the cases, EKF and UKF show unrecognizably small difference. The dynamics in (a) has weak nonlinearity, and KBR shows slightly worse MSE than EKF and UKF. For dataset (b) of strong nonlinearity, KBR outperforms for $T \geq 200$ the nonlinear Kalman filters, which know the true dynamics.

Next, we applied the KBR filter to the camera rotation problem used in [20]¹, where the angle of a camera is the hidden variable and the movie frames of a room taken by the camera are observed. As in [20], we are given 3600 frames of 20×20 RGB pixels ($Y_t \in [0, 1]^{1200}$), where the first 1800 frames are used for training, and the second half are used for test. We make the data noisy by adding Gaussian noise $N(0, \sigma^2)$ to Y_t . Our experiments cover two settings. In the first, we assume we do not know the hidden state X_t is included in $SO(3)$, but is a general 3×3 matrix. In this case, we use the Kalman filter by estimating the relations under a linear assumption, and the KBR filter with Gaussian kernels for X_t and Y_t . In the second setting, we exploit the fact $X_t \in SO(3)$: for the Kalman filter, X_t is represented by a quaternion, and for the KBR filter the kernel $k(A, B) = \text{Tr}[AB^T]$ is used for X_t . Table 1 shows the Frobenius norms between the estimated matrix and the true one. The KBR filter significantly outperforms the Kalman filter, since KBR has the advantage in extracting the complex nonlinear dependence of the observation on the hidden state.

5 Conclusion

We have proposed a general, novel framework for implementing Bayesian inference, where the prior, likelihood, and posterior are expressed as kernel means in reproducing kernel Hilbert spaces. The model is expressed in terms of a set of training samples, and inference consists of a small number of straightforward matrix operations. Our approach is well suited to cases where simple parametric models or an analytic forms of density are not available, but samples are easily obtained. We have addressed two applications: Bayesian inference without likelihood, and sequential filtering with nonparametric state-space model. Future applications could include inference in nonparametric Bayesian models, and Bayesian reinforcement learning.

¹Due to some difference in noise model, the results here are not directly comparable with those of [20].

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

References

- [1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, 1950.
- [2] C.R. Baker. Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.*, 186:273–289, 1973.
- [3] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publisher, 2004.
- [4] A. Doucet, N. De Freitas, and N.J. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.
- [5] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *JMLR*, 2:243–264, 2001.
- [6] K. Fukumizu, F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *JMLR*, 5:73–99, 2004.
- [7] K. Fukumizu, F.R. Bach, and M.I. Jordan. Kernel dimension reduction in regression. *Anna. Stat.*, 37(4):1871–1905, 2009.
- [8] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in NIPS 20*, pages 489–496. MIT Press, 2008.
- [9] A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in NIPS 19*, pages 513–520. MIT Press, 2007.
- [10] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in NIPS 20*, pages 585–592. MIT Press, 2008.
- [11] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proc. AeroSense: The 11th Intern. Symp. Aerospace/Defence Sensing, Simulation and Controls*, 1997.
- [12] P. Marjoram, Jo. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *PNAS*, 100(26):15324–15328, 2003.
- [13] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Ra’tsch. Kernel pca and de-noising in feature spaces. In *Advances in NIPS 11*, pages 536–542. MIT Press, 1999.
- [14] P. Müller and F.A. Quintana. Nonparametric bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.
- [15] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian J. Statistics*, 9(2):pp. 65–78, 1982.
- [16] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [17] S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential monte carlo without likelihoods. *PNAS*, 104(6):1760–1765, 2007.
- [18] L. Song, A. Gretton., and C. Guestrin. Nonparametric tree graphical models via kernel embeddings. In *AISTATS 2010*, pages 765–772, 2010.
- [19] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *AISTATS 2011*.
- [20] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. *Proc ICML2009*, pages 961–968. 2009.
- [21] L. Song and S. M. Siddiqi and G. Gordon and A. Smola. Hilbert Space Embeddings of Hidden Markov Models. *Proc. ICML2010*, 991–998. 2010.
- [22] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.
- [23] I. Steinwart. On the Influence of the Kernel on the Consistency of Support Vector Machines. *JMLR*, 2:67–93, 2001.
- [24] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiyu, and D. Okanojara. Conditional density estimation via least-squares density ratio estimation. In *AISTATS 2010*, pages 781–788, 2010.
- [25] S. Tavaré, D.J. Balding, R.C. Griffithis, and P. Donnelly. Inferring coalescence times from dna sequece data. *Genetics*, 145:505–518, 1997.
- [26] S. Thrun, J. Langford, and D. Fox. Monte carlo hidden markov models: Learning non-parametric models of partially observable stochastic processes. In *ICML 1999*, pages 415–424, 1999.
- [27] V. Monbet , P. Ailliot, and P.F. Marteau. l^1 -convergence of smoothing densities in non-parametric state space models. *Statistical Inference for Stochastic Processes*, 11:311–325, 2008.

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

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

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

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

□

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

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

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

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

□

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 $p(x_t | \tilde{y}_1, \dots, \tilde{y}_t)$, the forward filtering with

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

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

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

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

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

540 C Rates of consistency

541
542 The proof idea for the consistency rates of the KBR estimators is essentially the same as [1, 3], in
543 which the basic techniques are taken from the general theory of regularization [2].

544 First we give integral expression for the kernel mean and covariance operators. Recall that the kernel
545 mean m_X of X on \mathcal{H}_X satisfies

$$546 \langle f, m_X \rangle = E[f(X)]$$

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

$$548 m_X(u) = E[k(u, X)] = \int k_{\mathcal{X}}(u, \tilde{x}) dP_X(\tilde{x}), \quad (15)$$

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

$$551 (C_{YX}f)(y) = \int k_Y(y, \tilde{y}) f(\tilde{x}) dP(\tilde{x}, \tilde{y}), \quad (C_{XX}f)(x) = \int k_{\mathcal{X}}(x, \tilde{x}) f(\tilde{x}) dP_X(\tilde{x}), \quad (16)$$

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

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

555 **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
556 and P_X , respectively. Let $\hat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\hat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_X} = O_p(n^{-\alpha})$ as
557 $n \rightarrow \infty$ for some $0 < \alpha \leq 1/2$. Then, with $\varepsilon_n = n^{-\max\{\frac{2}{3}\alpha, \frac{\alpha}{1+\beta}\}}$, we have

$$558 \|\hat{C}_{YX}^{(n)}(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{m}_\Pi^{(n)} - m_{Q_Y}\|_{\mathcal{H}_Y} = O_p(n^{-\min\{\frac{2}{3}\alpha, \frac{2\beta+1}{2\beta+2}\alpha\}}), \quad (n \rightarrow \infty).$$

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

$$560 m_\Pi = \int k_{\mathcal{X}}(\cdot, x) \frac{\pi(x)}{p_X(x)} p_X(x) d\mu_X(x) = C_{XX}^{\beta+1} \eta. \quad (17)$$

561 First we show the rate of the estimation error:

$$562 \|\hat{C}_{YX}^{(n)}(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \hat{m}_\Pi^{(n)} - C_{YX}(C_{XX} + \varepsilon_n I)^{-1} m_\Pi\|_{\mathcal{H}_Y} = O_p(n^{-\alpha} \varepsilon_n^{-1/2}), \quad (18)$$

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

$$565 \|\hat{C}_{YX}^{(n)}(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}(\hat{m}_\Pi^{(n)} - m_\Pi)\|_{\mathcal{H}_Y} + \|(\hat{C}_{YX}^{(n)} - C_{YX})(C_{XX} + \varepsilon_n I)^{-1} m_\Pi\|_{\mathcal{H}_Y}$$

$$566 + \|\hat{C}_{YX}^{(n)}(\hat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1}(C_{XX} - \hat{C}_{XX}^{(n)})(C_{XX} + \varepsilon_n I)^{-1} m_\Pi\|_{\mathcal{H}_Y}.$$

567 By the decomposition $\hat{C}_{YX}^{(n)} = \hat{C}_{YY}^{(n)1/2} \hat{W}_{YX}^{(n)} \hat{C}_{XX}^{(n)1/2}$ with $\|\hat{W}_{YX}^{(n)}\| \leq 1$ (see [2]), the first term
568 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
569 $O_p(n^{-1/2} \varepsilon_n^{-1/2})$, respectively, by $\|(C_{XX} + \varepsilon_n I)^{-1} C_{XX}\| \leq 1$. This means Eq. (18).

570 Next, we show

$$571 \|C_{YX}(C_{XX} + \varepsilon_n I)^{-1} m_\Pi - m_{Q_Y}\|_{\mathcal{H}_Y} = O(\varepsilon_n^{\min\{(1+2\beta)/2, 1\}}) \quad (n \rightarrow \infty). \quad (19)$$

572 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

$$573 m_{Q_Y} = \int \int k(\cdot, y) \frac{\pi(x)}{p_X(x)} p(x, y) d\mu_X(x) d\mu_Y(y) = C_{YX} C_{XX}^\beta \eta$$

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

$$575 \|C_{YY}^{1/2} W_{YX}\| \|(C_{XX} + \varepsilon_n I)^{-1} C_{XX}^{(2\beta+3)/2} \eta - C_{XX}^{(2\beta+1)/2} \eta\|_{\mathcal{H}_X}.$$

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

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

holds. If $0 \leq \beta < 1/2$, we have $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} = \frac{\lambda_i^{(2\beta+1)/2}}{(\lambda_i + \varepsilon_n)^{(2\beta+1)/2}} \frac{\varepsilon_n^{(1-2\beta)/2}}{(\lambda_i + \varepsilon_n)^{(1-2\beta)/2}} \varepsilon_n^{(2\beta+1)/2} \leq \varepsilon_n^{(2\beta+1)/2}$. If $\beta \geq 1/2$, then $\frac{\varepsilon_n \lambda_i^{(2\beta+1)/2}}{\lambda_i + \varepsilon_n} \leq \|C_{XX}\| \varepsilon_n$. The dominated convergence theorem shows that the above sum converges to zero as $\varepsilon_n \rightarrow 0$ 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}_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 $\nu \geq 0$, and $\widehat{C}_{WZ}^{(n)} : \mathcal{H}_X \rightarrow \mathcal{H}_Y$ and $\widehat{C}_{WW}^{(n)} : \mathcal{H}_Y \rightarrow \mathcal{H}_Y$ 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 $\delta_n = n^{-\max\{\frac{4}{5}\gamma, \frac{4}{2\nu+5}\gamma\}}$ and any $y \in \mathcal{Y}$, we have as $n \rightarrow \infty$*

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

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

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

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

$$\begin{aligned} & \|\widehat{C}_{WW}^{(n)} ((\widehat{C}_{WW}^{(n)})^2 + \delta_n I)^{-1} (\widehat{C}_{WZ}^{(n)} - C_{WZ}) f\|_{\mathcal{H}_Y} + \|(\widehat{C}_{WW}^{(n)} - C_{WW}) (C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f\|_{\mathcal{H}_Y} \\ & + \|\widehat{C}_{WW}^{(n)} ((\widehat{C}_{WW}^{(n)})^2 + \delta_n I)^{-1} ((\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2) (C_{WW}^2 + \delta_n I)^{-1} C_{WZ} f\|_{\mathcal{H}_Y}. \end{aligned}$$

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 $|\lambda_i / (\lambda_i^2 + \delta_n)| = 1 / |\lambda_i + \delta_n / \lambda_i| \leq 1 / (2\sqrt{|\lambda_i|} \sqrt{\delta_n / |\lambda_i|}) = 1 / (2\sqrt{\delta_n})$, we have $\|\widehat{C}_{WW}^{(n)} ((\widehat{C}_{WW}^{(n)})^2 + \delta_n I)^{-1}\| \leq 1 / (2\sqrt{\delta_n})$, 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 $\|U_{WZ}\| \leq 1$ shows that the second term is of $O_p(n^{-\gamma} \delta_n^{-3/4})$. From the fact $\|(\widehat{C}_{WW}^{(n)})^2 - C_{WW}^2\| \leq \|\widehat{C}_{WW}^{(n)} (\widehat{C}_{WW}^{(n)} - C_{WW})\| + \|(\widehat{C}_{WW}^{(n)} - C_{WW}) C_{WW}\| = O_p(n^{-\gamma})$, 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

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

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

$$\tilde{C}_{WW} \phi = \int k_Y(y, w) \phi(w) dQ_W(w), \quad (\phi \in L^2(Q_W)). \quad (22)$$

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

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

where Q_W is the marginal distribution of W .

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

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

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

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

It suffices then to prove

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

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, $\tilde{C}_{WW} E[f(Z)|W = \cdot] = \tilde{C}_{WZ} f$ 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

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

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

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

Theorem 9. Let f be a function in \mathcal{H}_X , (Z, W) be a random variable that has the distribution Q with p.d.f. $p(y|x)\pi(x)$, and $\hat{m}_\Pi^{(n)}$ be an estimator of m_Π such that $\|\hat{m}_\Pi^{(n)} - m_\Pi\|_{\mathcal{H}_X} = O_p(n^{-\alpha})$ ($n \rightarrow \infty$) for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^\beta)$ with $\beta \geq 0$, and $E[f(Z)|W = \cdot] \in \mathcal{R}(C_{WW}^\nu)$ for some $\nu \geq 0$. 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 $y \in \mathcal{Y}$

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

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 $\|\hat{C}_{ZW}^{(n)} - C_{ZW}\|$ and $\|\hat{C}_{WW}^{(n)} - C_{WW}\|$ are of $O_p(n^{-\gamma})$. Since

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

combination of Theorems 6 and 7 proves the theorem. \square

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 $\|\widehat{m}_{\Pi}^{(n)} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\alpha})$ ($n \rightarrow \infty$) for some $0 < \alpha \leq 1/2$. Assume that $\pi/p_X \in \mathcal{R}(C_{XX}^{\beta})$ with $\beta \geq 0$, and $E[f(Z)|W = \cdot] \in \mathcal{R}(\widetilde{C}_{WW}^{\nu}) \cap \mathcal{H}_{\mathcal{Y}}$ 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

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

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

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

If we further assume that S is bounded, the adjoint operator $S^* : \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{H}_{\mathcal{X}}$ 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 $A_W : \mathcal{H}_{\mathcal{Y}} \rightarrow L^2(Q_W)$ 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 $\eta_f \in \mathcal{H}_{\mathcal{Y}}$ with $Sf = C_{WW}^{\nu} \eta_f$, and the linear map

$$C_{WW}^{-\nu} S : \mathcal{H}_{\mathcal{X}} \rightarrow \mathcal{H}_{\mathcal{Y}}, \quad f \mapsto \eta_f$$

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 $\|\widehat{m}_{\Pi}^{(n)} - m_{\Pi}\|_{\mathcal{H}_{\mathcal{X}}} = O_p(n^{-\alpha})$ ($n \rightarrow \infty$) for some $0 < \alpha \leq 1/2$. Assume (S) above, and $\pi/p_X \in \mathcal{R}(C_{XX}^{\beta})$ with some $\beta \geq 0$. 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

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

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

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

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

The assertion is thus obtained if

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

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

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

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

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

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.