Appendices

In this section we prove Theorem 1 and Theorem 2.

A Preliminaries

A.1 Notation

We first introduce some relevant concepts from functional analysis. If $E$ is a Hilbert space we denote by $\langle \cdot, \cdot \rangle_E$ and $\| \cdot \|_E$ its corresponding inner product and norm, respectively. If $E$ and $F$ are two Hilbert spaces, we use $\| \cdot \|$ to denote the operator norm $\|A\| = \sup_{f: \|f\| \leq 1} \|Af\|$, where $A$ is an operator from $E$ to $F$. We denote by $A^*$ the adjoint of $A$.

If $E$ is separable with an orthonormal basis $\{e_k\}_k$, then $\| \cdot \|_1$ and $\| \cdot \|_2$ are the trace norm and Hilbert-Schmidt norm on $E$ and are given by:

$$\|A\|_1 = \sum_k \langle (A^*A)^{\frac{1}{2}} e_k, e_k \rangle$$

$$\|A\|_2 = \|A^*A\|_1,$$

where $A$ is an operator from $E$ to $E$. $\lambda_{\max}(A)$ is used to denote the algebraically largest eigenvalue of $A$. For $f$ in $E$ and $g$ in $F$ we denote by $g \otimes f$ the tensor product viewed as an application from $E$ to $F$ with $(g \otimes f)h = g(f,h)_E$ for all $h$ in $E$. $C^1(\Omega)$ denotes the space of continuously differentiable functions on $\Omega$ and $L^r(\Omega)$ the space of $r$-power Lebesgue-integrable function. Finally for any vector $\beta$ in $\mathbb{R}^{nd}$, we use the notation $\beta_{(a,i)} = \beta_{(a-1)d+i}$ for $a \in [n]$ and $i \in [d]$.

A.2 Operator valued kernels and feature map derivatives

Let $\mathcal{X}$ and $\mathcal{Y}$ be two open subsets of $\mathbb{R}^p$ and $\mathbb{R}^d$. $\mathcal{H}_Y$ is a reproducing kernel Hilbert space of functions $f : \mathcal{Y} \to \mathbb{R}$ with kernel $k_Y$. We denote by $\mathcal{H}$ a vector-valued reproducing kernel Hilbert space of functions $T : x \mapsto T_x$ from $\mathcal{X}$ to $\mathcal{H}_Y$ and we introduce the feature operator $\Gamma : x \mapsto \Gamma_x$ from $\mathcal{X}$ to $\mathcal{L}(\mathcal{H}_Y, \mathcal{H})$ where $\mathcal{L}(\mathcal{H}_Y, \mathcal{H})$ is the set of bounded operators from $\mathcal{H}_Y$ to $\mathcal{H}$. For every $x \in \mathcal{X}$, $\Gamma_x$ is an operator defined from $\mathcal{H}_Y$ to $\mathcal{H}$.

The following reproducing properties will be used extensively:

- Reproducing property of the derivatives of a function in $\mathcal{H}_Y$ (Steinwart et al., 2008, Lemma 4.34): provided that the kernel $k_Y$ is differentiable $m$-times with respect to each coordinate, then all $f \in \mathcal{H}_Y$ are differentiable for every multi-index $\alpha \in \mathbb{N}_0^n$ such that $\alpha \leq m$, and

$$\partial^\alpha f(y) = \langle f, \partial^\alpha k(y, \cdot) \rangle_{\mathcal{H}_Y} \quad \forall y \in \mathcal{Y},$$

where $\partial^\alpha k(y, y') = \frac{\partial^\alpha k(y, y')}{\partial y^\alpha}$. In particular we will use the notation

$$\partial_i k(y, y') = \frac{\partial k(y, y')}{\partial y_i}, \quad \partial_{i+a} k(y, y') = \frac{\partial k(y, y')}{\partial y_i}.$$

- Reproducing property in the vector-valued space $\mathcal{H}$: For any $f \in \mathcal{H}_Y$ and any $T \in \mathcal{H}$ we have the following:

$$\langle T_x, f \rangle_{\mathcal{H}_Y} = \langle T, \Gamma_x f \rangle_{\mathcal{H}}$$

In particular for every $y \in \mathcal{Y}$ we get:

$$\langle T_x, k(y, \cdot) \rangle_{\mathcal{H}_Y} = \langle T, \Gamma_x k(y, \cdot) \rangle_{\mathcal{H}}$$

Using now the reproducing property in $\mathcal{H}_Y$ we get:

$$T(x, y) := T_x(y) = \langle T, \Gamma_x k(y, \cdot) \rangle_{\mathcal{H}}$$
A.3 The conditional infinite dimensional exponential family

Let \( q_0 \) be a base density function of a probability distribution over \( \mathcal{Y} \) and \( \pi \) a probability distribution over \( \mathcal{X} \). \( \pi \) and \( q_0 \) are fixed and are assumed to be supported in the whole spaces \( \mathcal{X} \) and \( \mathcal{Y} \), respectively.

We introduce the following functions \( Z : \mathcal{H}_Y \to \mathbb{R}_+^* \), such that for every \( f \in \mathcal{H}_Y \) we have

\[
Z(f) := \int_Y \exp((f, k(y, .))_{\mathcal{H}_Y}) q_0(dy).
\]

We consider now the following family of operators

\( \mathcal{T} = \{ T \in \mathcal{H} : Z(T_x) < \infty, \forall x \in \mathcal{X} \} \).

This allows to introduce the Kernel Conditional Exponential Family as the family of conditional distributions satisfying

\[
\mathcal{P} = \left\{ p_T(x|y) = q_0(y)^{\mathcal{H}(T, k(y, .))_{\mathcal{H}_Y}} Z(T_x) \right\}.
\]

Given samples \( (X_i, Y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y} \) following a joint distribution \( p_0 \) the goal is to approximate the conditional density function \( p_0(y|x) \) in the case where \( p_0(y|x) \in \mathcal{P} \) (i.e. \( \exists T_0 \in \mathcal{T} \) such that \( p_0(y|x) = p_{T_0}(y|x) \)).

To this end, we introduce the expected conditional score function between two conditional distributions \( p(.|x) \) and \( q(.|x) \) under \( \pi \),

\[
J(p|q) = \frac{1}{2} \int_x \int_y \sum_{i=1}^d \left[ \partial_i \log p(y|x) - \partial_i \log q(y|x) \right]^2 p(dy|x) \pi(dx).
\]

This function has the nice property that \( J(p|q) \geq 0 \) and that \( J(p|q) = 0 \Leftrightarrow q = p \), which makes it a good candidate as a loss function.

The marginal distribution \( p_0(x) \) doesn’t have to match \( \pi(x) \) in general as long as they have the same support. For purpose of simplicity we will assume that \( p_0(x) = \pi(x) \).

A.4 Assumptions

We make the following assumptions:

(A) (well specified) The true conditional density \( p_0(y|x) = p_{T_0}(y|x) \in \mathcal{P} \) for some \( T_0 \in \mathcal{T} \).

(B) \( \mathcal{Y} \) is a non-empty open subset of of the form \( \mathbb{R}^d \) with a piecewise smooth boundary \( \partial \mathcal{Y} := \overline{\mathcal{Y}} \setminus \mathcal{Y} \), where \( \overline{\mathcal{Y}} \) denotes the closure of \( \mathcal{Y} \).

(C) \( k_Y \) is twice continuously differentiable on \( \mathcal{Y} \times \mathcal{Y} \) and \( \partial^\alpha k_Y \) is continuously extensible to \( \overline{\mathcal{Y}} \times \overline{\mathcal{Y}} \) for all \( |\alpha| \leq 2 \).

(D) For all \( x \in \mathcal{X} \) and all \( i \in [d] \), as \( y \) approaches \( \partial \mathcal{Y} \): \( \| \partial_i k(y, .) \|_{\mathcal{Y}p_0(y)} = o(\|y\|^{1-d}) \)

(E) The operator \( \Gamma \) is uniformly bounded for the operator norm \( \| \Gamma \|_{op} \leq \kappa \) for all \( x \in \mathcal{X} \).

(F) (Integrability) for some \( \epsilon \geq 1 \) and all \( i \in [d] \):

\[
\| \partial_i k(y, .) \|_{\mathcal{Y}} \in L^2(\mathcal{Y}, p_0), \ \| \partial_i^2 k(y, .) \|_{\mathcal{Y}} \in L^1(\mathcal{Y}, p_0), \ \| \partial_i \log q_0(y) \| \in L^1(\mathcal{Y}, p_0).
\]

B Theorems

In this section, we prove the main theorems of the document, by extending the proofs of Sriperumbudur et al., 2017 to the case of the vector-valued RKHS. We provide complete steps for all the proofs, including those that carry over from the earlier work, to make the presentation self-contained; the reader may compare with (Sriperumbudur et al., 2017, Section 8) to see the changes needed in the conditional setting.
B.1 Score Matching

Theorem 3 (Score Matching). Under Assumptions (A) to (F), the following holds:

1. \( J(p_{\mathcal{T}_0}, p_T) < +\infty \) for all \( T \in \mathcal{T} \).
2. For all \( T \in \mathcal{H} \) define
   \[ J(T) = \frac{1}{2} \langle T - T_0, C(T - T_0) \rangle_{\mathcal{H}}, \]
   where
   \[ C := \int_{X \times Y} \sum_{i=1}^d \left[ \Gamma_x \partial_i k(y, \cdot) \right] dx, dy = \mathbb{E}_{\mu_0}[C_{X,Y}], \]
   then \( C \) a trace-class positive operator on \( \mathcal{H} \) and for all \( T \in \mathcal{T} \)
   \( J(T) = J(p_{\mathcal{T}_0}, p_T) \).
3. Alternatively,
   \[ J(T) = \frac{1}{2} \langle T, CT \rangle_{\mathcal{H}} + \langle T, \xi \rangle_{\mathcal{H}} + J(p_{\mathcal{T}_0}, \mu_0), \]
   where
   \[ \mathcal{H} \ni \xi := \int_{X \times Y} \sum_{i=1}^d \Gamma_x \left[ \partial_i \log \mu_0(y) \partial_i k(y, \cdot) + \partial_i^2 k(y, \cdot) \right] dx, dy = \mathbb{E}_{\mu_0}[\xi_{X,Y}], \]
   Moreover, \( T_0 \) satisfies \( CT_0 = -\xi \).
4. For any \( \lambda > 0 \), a unique minimizer \( T_\lambda \) of \( J_\lambda(T) := J(T) + \frac{\lambda}{2} \| T \|^2_{\mathcal{H}} \) over \( \mathcal{H} \) exists and is given by:
   \[ T_\lambda = -(C + \lambda I)^{-1} \xi = (C + \lambda I)^{-1} CT_0. \]

Proof. We prove the results in the same order as stated in the theorem:

1. By the reproducing property of the real valued space \( \mathcal{H}_Y \) we have: \( T(x, y) = \langle T_x, k(y, \cdot) \rangle_{\mathcal{H}_Y} \). Using the reproducing property for the derivatives of real valued functions in an RKHS in Lemma 3, we get
   \[ \partial_i T(x, y) = \partial_i (T_x, k(y, \cdot))_{\mathcal{H}_Y} = \langle T_x, \partial_i k(y, \cdot) \rangle_{\mathcal{H}_Y} \quad \forall i \in [d]. \]
   Finally, using the reproducing property in the vector-valued space \( \mathcal{H} \),
   \[ \partial_i T(x, y) = (T, \Gamma_x \partial_i k(y, \cdot))_{\mathcal{H}}, \quad \forall i \in [d]. \]
   it is easy to see that
   \[ J(p_{\mathcal{T}_0}, p_T) = \frac{1}{2} \int_{X \times Y} \sum_{i=1}^d (T_0 - T, \Gamma_x \partial_i k(y, \cdot))_{\mathcal{H}}^2 \mu_0(dx, dy). \]

By Assumptions (E) and (F),
\[ \| \Gamma_x \partial_i k(y, \cdot) \|_{\mathcal{H}} \leq \| \Gamma_x \|_{\mathcal{O}_2} \| \partial_i k(y, \cdot) \|_{\mathcal{H}_Y} \leq \kappa \sqrt{\| \partial_i \partial_i^d k(y, y) \|_{L^2(\mu_0)}}, \]
and therefore by Cauchy-Schwarz inequality,
\[ J(T) = J(p_{\mathcal{T}_0}, p_T) \leq \frac{1}{2} \| T_0 - T \|_{\mathcal{H}}^2 \int_{X \times Y} \sum_{i=1}^d \| \Gamma_x \partial_i k(y, \cdot) \|_{\mathcal{H}}^2 \mu_0(dx, dy) < +\infty, \]
which means that \( J(T) < \infty \) for all \( T \in \mathcal{T} \).
2. Starting from (8), it is easy to see that:

\[
J(T) = \frac{1}{2} \int_{x \times y} \sum_{i=1}^{d} \langle T_0 - T, \Gamma_x \partial_i k(y, \cdot) \otimes \Gamma_x \partial_i k(y, \cdot) \rangle_H p_0(dx, dy)
= \frac{1}{2} \int_{x \times y} \langle T_0 - T, C_{x,y}(T_0 - T) \rangle_H p_0(dx, dy)
\]

In the first line, we used the fact that \( \langle a, b \rangle_{ HS} = \langle a, b \rangle_H = \langle a, b \otimes ba \rangle_H \) for any \( a \) and \( b \) in a Hilbert space \( H \). By further observing that \( C_{x,y} \) and \( (T_0 - T) \otimes (T_0 - T) \) are Hilbert-Schmidt operators as \( \| C_{x,y} \|_{HS} \leq \kappa^2 \sum_{i=1}^{d} \| \partial_i k(y, \cdot) \|_H \) \( \leq \infty \) by Lemma 1 and \( \|(T_0 - T) \otimes (T_0 - T)\|_H \leq \| (T_0 - T) \|_H^2 \leq \infty \) we get that:

\[
J(T) = \frac{1}{2} \int_{x \times y} \langle (T_0 - T) \otimes (T_0 - T), C_{x,y} \rangle_{HS} p_0(dx, dy)
\]

Using Assumption (F) we have by Lemma 2 that \( C_{x,y} \) is \( p_0 \)-integrable in the Bochner sense (see Retherford, 1978) Definition 1) and that the inner product and integration may be interchanged:

\[
J(T) = \frac{1}{2} \int_{x \times y} \int_{x \times y} C_{x,y} p_0(dx, dy) = \frac{1}{2} \langle (T_0 - T, C(T_0 - T) \rangle_H
\]

3. From (6) we have \( J(T) = \frac{1}{2} \langle (T, CT)_H - (T, CT_0)_H - \frac{1}{2} (T_0, CT_0)_H \rangle \). Recalling that \( \partial_i T(x, y) = \langle T, \Gamma_x \partial_i k(y, \cdot) \rangle_H \) for all \( i \in [d] \), and using \( \partial_i T_0(x, y) = \partial_i \log p_0(y|x) - \partial_i \log q_0(y|x) \) one gets:

\[
\langle T, CT_0 \rangle_H = \int_{x \times y} \left[ \sum_{i=1}^{d} \partial_i T(x, y) \partial_i T_0(x, y) \right] p_0(dx, dy)
= \int_{x \times y} \left[ \sum_{i=1}^{d} \partial_i T(x, y) \partial_i \log p_0(y|x) \right] p_0(dx, dy) - \int_{x \times y} \left[ \sum_{i=1}^{d} \partial_i T(x, y) \partial_i \log q_0(y|x) \right] p_0(dx, dy)
\]

\[
\langle T, CT_0 \rangle_H = \int_{x \times y} \left[ \sum_{i=1}^{d} \partial_i T(x, y) \partial_i \log p_0(y|x) \right] p_0(dx, dy)
\]

(a) is obtained using the first Green’s identity, where \( \partial Y \) is the boundary of \( Y \) and \( d\tilde{S} \) is the oriented surface element. The first term \( \int_{x} \pi(dx) \int_{\partial Y} p_0(y|x) \nabla_y T(x, y) \cdot d\tilde{S} \) vanishes by Lemma 4, which relies on Assumption (D). The second term can be written as: \( \int_{x \times y} \langle T, \xi_{x,y} \rangle_H p_0(dx, dy) \).

By Assumptions (E) and (F) \( \xi_{x,y} \) is Bochner \( p_0 \)-integrable, therefore:

\[
\int_{x \times y} \langle T, \xi_{x,y} \rangle_H p_0(dx, dy) = \langle T, \int_{x \times y} \xi_{x,y} p_0(dx, dy) \rangle_H = \langle T, \xi \rangle_H.
\]

Hence \( \langle T, CT_0 \rangle_H = -\langle T, \xi \rangle_H \) and \( \xi = -CT_0 \). Moreover, one can clearly see that:

\[
\langle T_0, CT_0 \rangle_H = \int_{x \times y} \left[ \sum_{i=1}^{d} \partial_i T_0(x, y) \right]^2 p_0(dx, dy) = J(p_{T_0} || q_0).
\]

And the result follows.

4. For \( \lambda > 0 \), \( (C + \lambda I)^{-\frac{1}{2}} \) is invertible as \( C \) is a symmetric trace-class operator. Moreover, \( (C + \lambda I)^{-\frac{1}{2}} \) is well defined and one can easily see that:

\[
J_{\lambda}(T) = \frac{1}{2} \| (C + \lambda I)^{-\frac{1}{2}} T + (C + \lambda I)^{-\frac{1}{2}} \xi \|_H^2 - \frac{1}{2} \| \xi, (C + \lambda I)^{-1} \xi \|_H + c_0
\]

with \( c_0 = J(p_{T_0} || q_0) \). \( J_{\lambda}(T) \) is minimized if and only if \( (C + \lambda I)^{-\frac{1}{2}} T = (C + \lambda I)^{-\frac{1}{2}} \xi \) and therefore \( T = (C + \lambda I)^{-1} \xi \) is the unique minimizer of \( J_{\lambda}(T) \).

\[
\square
\]
B.2 Estimator of $T_0$

Given samples $(X_a, Y_a)_{a=1}^n$ drawn i.i.d. from $p_0$ and $\lambda > 0$, we define the empirical score function as

$$\hat{J}(T) := \frac{1}{2} \langle T, \hat{C}T \rangle_{\mathcal{H}} + \langle T, \hat{\xi} \rangle_{\mathcal{H}} + J(p_{\hat{p}_n} \| q_0).$$

where:

$$\hat{C} := \frac{1}{n} \sum_{a=1}^n \sum_{i=1}^d \Gamma_{X_a} \partial_i k(Y_a, \cdot) \odot \Gamma_{X_a} \partial_i k(Y_a, \cdot)$$

$$\hat{\xi} := \frac{1}{n} \sum_{a=1}^n \sum_{i=1}^d \Gamma_{X_a} \left[ \partial_i \log q_0(Y_a) \partial_i k(Y_a, \cdot) + \partial_i^2 k(Y_a, \cdot) \right].$$

are the empirical estimators of $C$ and $\xi$ respectively.

**Theorem 4 (Estimator of $T_0$).** For and any $\lambda > 0$, we have the following:

1. The unique minimizer $T_{\lambda, n}$ of $J_\lambda(T) := \hat{J}(T) + \frac{\lambda}{2} \| T \|^2_{\mathcal{H}}$ over $\mathcal{H}$ exists and is given by

$$T_{\lambda, n} = -(\hat{C} + \lambda I)^{-1} \hat{\xi}.$$

2. Moreover, $T_{\lambda, n}$ is of the form

$$T_{\lambda, n} = \frac{1}{\lambda} \hat{\xi} + \sum_{b=1}^n \sum_{i=1}^d \beta_{(b-1)d+i+1} \Gamma_{X_b} \partial_i k(Y_b, \cdot),$$

where $(\beta_b)$ are obtained by solving the following linear system:

$$(G + n \lambda I) \beta = \frac{h}{\lambda}$$

with:

$$(G)_{(a-1)d+i,(b-1)d+j} = \langle \Gamma_{X_a} \partial_i k(Y_a, \cdot), \Gamma_{X_b} \partial_j k(Y_b, \cdot) \rangle_{\mathcal{H}}.$$

and:

$$(h)_{(a-1)d+i} = \langle \xi, \Gamma_{X_a} \partial_i k(Y_a, \cdot) \rangle_{\mathcal{H}}.$$

**Proof.** 1. The same proof as in Theorem 3 holds with $C$ and $\xi$ replaced by $\hat{C}$ and $\hat{\xi}$.

2. We will use the general representer theorem stated in Lemma 6. We have that:

$$T_{\lambda, n} = \arg \inf_{T \in \mathcal{H}} \frac{1}{2} \langle T, \hat{C}T \rangle_{\mathcal{H}} + \langle T, \hat{\xi} \rangle_{\mathcal{H}} + \frac{\lambda}{2} \| T \|^2_{\mathcal{H}}$$

$$= \arg \inf_{T \in \mathcal{H}} \frac{1}{2} \sum_{a=1}^n \sum_{i=1}^d \langle T, \Gamma_{X_a} \partial_i k(Y_a, \cdot) \rangle_{\mathcal{H}}^2 + \langle T, \hat{\xi} \rangle_{\mathcal{H}} + \frac{\lambda}{2} \| T \|^2_{\mathcal{H}}$$

$$= \arg \inf_{T \in \mathcal{H}} V(T, \phi_1, ..., T, \phi_{nd+1})_{\mathcal{H}} + \frac{\lambda}{2} \| T \|^2_{\mathcal{H}}.$$

Where $V(\phi_1, ..., \phi_{nd+1}) := \frac{1}{2} \sum_{a=1}^n \sum_{i=1}^d \theta_{(a-1)d+i+1}^2 + \theta_{nd+1}$ is a convex differentiable function and $\phi_{(a-1)d+i} := \Gamma_{X_a} \partial_i k(Y_a, \cdot)$ where $a \in [n], i \in [d]$ and $\phi_{nd+1} = \hat{\xi}$. Therefore, it follows from Lemma 6 that:

$$T_{\lambda, n} = \delta \hat{\xi} + \sum_{a=1}^n \sum_{i=1}^d \beta_{(a-1)d+i+1} \phi_{(a-1)d+i}.$$
where $\delta$ and $\beta$ satisfy:

$$\lambda(\beta, \delta) + \nabla V(K(\beta, \delta)) = 0$$

with $K = \begin{pmatrix} G & h \\ h^T & \|\delta\|_H^2 \end{pmatrix}$. The gradient $\nabla V$ of $V$ is given by $\nabla V(z, t) = (\frac{1}{2}z, 1)$. The above equation reduces then to $\lambda\delta + 1 = 0$ and $\lambda\beta + \frac{1}{2}G\beta + \frac{1}{2}h = 0$ which yields $\delta = -\frac{1}{\lambda}$ and $(\frac{1}{2}G + \lambda I)\beta = \frac{1}{\lambda} h$.

\[ \square \]

### B.3 Consistency and convergence

**Theorem 5 (Consistency and convergence rates for $T_{\lambda,n}$).** Let $\gamma > 0$ be a positive number and define $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{2}) \in (\frac{1}{2}, \frac{1}{2})$, under Assumptions (A) to (F):

1. if $T_0 \in \overline{R(C)}$ then $\|T_{\lambda,n} - T_0\| \to 0$ when $\lambda \sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.
2. if $T_0 \in R(C^n)$ for some $\gamma > 0$ then $\|T_{\lambda,n} - T_0\| = O_p(\sqrt{n}^{-1+\gamma})$ for $\lambda = n^{-\alpha}$

**Proof.** Recalling that $T_{\lambda,n} = -(\hat{C} + \lambda I)^{-1}\hat{\xi}$ We consider the following decomposition:

$$T_{\lambda,n} - T_\lambda = -(\hat{C} + \lambda I)^{-1}(\xi + (\hat{C} + \lambda I)T_\lambda) = -(\hat{C} + \lambda I)^{-1}(\hat{C} + \lambda I)T_\lambda + C(T_0 - T_\lambda)$$

$$= (\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) - (\hat{C} + \lambda I)^{-1}(\hat{\xi} - \xi) + (\hat{C} + \lambda)^{-1}(C - \hat{C})T_0.$$

We used the fact that $\lambda T_\lambda = C(T_0 - T_\lambda)$ in (\star). Define now

$$S_1 := \| (\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) \|_H$$

$$S_2 := \| (\hat{C} + \lambda I)^{-1}(\hat{\xi} - \xi) \|_H$$

$$S_3 := \| (\hat{C} + \lambda I)^{-1}(C - \hat{C})T_0 \|_H$$

$$A_0(\lambda) := \| T_{\lambda,n} - T_0 \|_H.$$

It comes then:

$$\| T_\lambda - T_0 \|_H \leq \| T_{\lambda,n} - T_\lambda \|_H + \| T_\lambda - T_0 \|_H$$

$$\leq S_1 + S_2 + S_3 + A_0(\lambda).$$

Using Lemma 10 we can bound $S_1$, $S_2$ and $S_3$. Note that $C_{x,y}$ as defined in (7) is a positive, self-adjoint trace-class operator by Lemma 1, we therefore have:

$$\| C_{x,y} \|_H^2 = \sum_{i,j=1}^d (\Gamma_x \partial_i k(y, \cdot), \Gamma_x \partial_j k(y, \cdot))_H^2 \leq \sum_{i,j=1}^d \| \Gamma_x \partial_i k(y, \cdot) \|_H^2 \sum_{i,j=1}^d \| \Gamma_x \partial_j k(y, \cdot) \|_H^2$$

$$\leq \sum_{i=1}^d \| \Gamma_x \partial_i k(y, \cdot) \|_H^2 \leq a \sum_{i=1}^d \| \Gamma_x \partial_i k(y, \cdot) \|_H^4 \leq d \sum_{i=1}^d \| \partial_i k(y, \cdot) \|_{H_y}^4.$$

The last inequality is obtained using Assumption (E). Using now Assumption (F) for $\epsilon = 2$ one can get:

$$\int_{X \times Y} \| C_{x,y} \|_H^2 \rho_0(dx, dy) \leq d \sum_{i=1}^d \int_{X \times Y} \| \partial_i k(y, \cdot) \|_{H_y}^4 \rho_0(dx, dy) < +\infty.$$  

16
Lemma 10 can then be applied to get the following inequalities:

\[ S_1 \leq \| (\hat{C} + \lambda I)^{-1} (C - \hat{C})(T_{\lambda} - T_0) \|_H = O_{p_0}\left( \frac{A(\lambda)}{\lambda \sqrt{n}} \right) \]

\[ S_3 \leq \| (\hat{C} + \lambda I)^{-1} (C - \hat{C})T_0 \| = O_{p_0}\left( \frac{1}{\lambda \sqrt{n}} \right) \]

\[ \| (C + \lambda I)^{-1} \| \leq \frac{1}{\lambda} \]

To bound \( S_2 \) we need to show that \( \| \xi - \xi \|_H = O_{p_0}(n^{-\frac{1}{2}}) \). The same argument as in Sriperumbudur et al., 2017 holds:

\[
E_{p_0} \| \xi - \xi \|_H^2 = \frac{1}{n} \int_{X \times Y} \| \xi_{x,y} \|_H^2 p_0(dx, dy) - \| \xi \|^2
\]

\[ \leq \frac{1}{n} \int_{X \times Y} \| \xi_{x,y} \|_H^2 p_0(dx, dy) \]

By Assumption (F) for \( \epsilon = 2 \) we have that \( \int_{X \times Y} \| \xi_{x,y} \|_H^2 p_0(dx, dy) < \infty \). One can therefore apply Chebyshev inequality to get the results. It comes that:

\[ S_2 \leq \| (\hat{C} + \lambda I)^{-1} \| \| \xi - \xi \|_H = O_{p_0}\left( \frac{1}{\lambda \sqrt{n}} \right) \]

Using the bounds on \( S_1, S_2 \) and \( S_3 \) we get:

\[ \| T_{\lambda,n} - T_0 \|_H = O_{p_0}\left( \frac{1}{\lambda \sqrt{n}} + \frac{A_0(\lambda)}{\lambda \sqrt{n}} \right) + A_0(\lambda) \]

(9)

1. By Lemma 9 we have \( A_0(\lambda) \to 0 \) as \( \lambda \to 0 \) if \( T_0 \in \mathcal{R}(C) \). Therefore it follows from (9) that \( \| T_{\lambda,n} - T_0 \| \to 0 \) as \( \lambda \to 0, \lambda \sqrt{n} \to \infty \) and \( n \to \infty \).

2. We have by Lemma 9 that if \( T_0 \in \mathcal{R}(C^\gamma) \) for \( \gamma > 0 \) then:

   \[ \lambda \leq \max \{ 1, \| C \|^{-1} \} \| C^{-\gamma} T_0 \|_H \lambda^{-\min\{1, \gamma\}}. \]

   The result follows by choosing \( \lambda = n^{-\max\{ \frac{1}{2}, \frac{1}{1+\gamma} \}} = n^{-\alpha}. \)

\[ \square \]

We denote by \( KL(p_{T_0}\|p_T) \) the expected KL divergence between \( p_{T_0} \) and \( p_T \) under the marginal \( p_0(x) \).

**Theorem 6 (Consistency and convergence rates for \( p_{T_{\lambda,n}} \)).** Assuming Assumptions (A) to (F), and \( \| k \|_\infty := \sup_{y \in Y} k(y, y) < \infty \) and that \( p_{T_0}(y|x) \) is supported on \( Y \) for all \( x \in X \) then the following holds:

1. \( KL(p_{T_0}\|p_{T_{\lambda,n}}) \to 0 \) as \( \lambda \sqrt{n} \to \infty \), \( \lambda \to 0 \) and \( n \to \infty \).

2. If \( T_0 \in \mathcal{R}(C^\gamma) \) for some \( \gamma > 0 \) then by defining \( \alpha = \max\{ \frac{1}{2(1+\gamma)}, \frac{1}{2} \} \in (\frac{1}{2}, \frac{1}{2}) \), and choosing \( \lambda = n^{-\alpha} \) we have that \( KL(p_0\|p_{T_{\lambda,n}}) = O_{p_0}(n^{-1+2\alpha}) \)

**Proof.** By Lemma 8, we have that \( T = \mathcal{H} \) and we can assume without loss of generality that \( T_0 \in \mathcal{R}(C) \). Using Lemma 7 (also see van der Vaart et al., 2008 Lemma 3.1), one can see that for a given \( x \):

\[
KL(p_{T_0}(y|x)\|p_{T_{\lambda,n}}(y|x)) \leq \| T_0(x) - T_{\lambda,n}(x) \|_\infty^2 \exp \| T_0(x) - T_{\lambda,n}(x) \|_\infty (1 + \| T_0(x) - T_{\lambda,n}(x) \|_\infty)
\]

(10)

Moreover, using Assumption (E) and the fact that \( \| k \|_\infty < \infty \) one can see that
\[ |T_0(x, y) - T_{\lambda,n}(x, y)|_{\mathcal{H}} = \langle T_0 - T_{\lambda,n}, \Gamma_x k(y, \cdot) \rangle_{\mathcal{H}} \]
\[ \leq \|T_0 - T_{\lambda,n}\|_{\mathcal{H}} \|\Gamma_x k(y, \cdot)\|_{\mathcal{H}} \]
which gives after taking the supremum:
\[ \|T_0(x) - T_{\lambda,n}(x)\|_{\infty} \leq \kappa \|k\|_{\infty} \|T_0 - T_{\lambda,n}\|_{\mathcal{H}} \quad (11) \]
for all \( x \in \mathcal{X} \). Using (11) in (10) and taking the expectation with respect to \( x \), one can conclude using Theorem 5.

\[ \Box \]

C Auxiliary results

**Lemma 1.** Under Assumptions (C), (E) and (F) we have that:

1. \( C_{x,y} \) is a trace-class positive and symmetric operator for all \((x, y) \in \mathcal{X} \times \mathcal{Y}\)
2. \( C_{x,y} \) is Bochner-integrable for all \((x, y) \in \mathcal{X} \times \mathcal{Y}\)
3. \( C \) is a trace-class positive and symmetric operator

**Proof.** Recall that \( C = \int_{\mathcal{X} \times \mathcal{Y}} C_{x,y} p_0(dx, dy) \) where \( C_{x,y} = \sum_{i=1}^d \Gamma_x \partial_i k(y, \cdot) \otimes \Gamma_x \partial_i k(y, \cdot) \) is a positive self-adjoint operator. The trace norm of \( C_{x,y} \) satisfies:

\[ \|C_{x,y}\|_1 \leq \sum_{i=1}^d \|\Gamma_x \partial_i k(y, \cdot) \otimes \Gamma_x \partial_i k(y, \cdot)\|_1 \]
\[ = \sum_{i=1}^d \|\Gamma_x \partial_i k(y, \cdot)\|_{\mathcal{H}}^2 \leq \sum_{i=1}^d \|\Gamma_x \partial_i k(y, \cdot)\|_{\mathcal{H}}^2 \]
\[ \overset{(a)}{\leq} \kappa^2 \sum_{i=1}^d \|\partial_i k(y, \cdot)\|_{\mathcal{H}}^2 < \infty. \]

(a) comes from Assumption (E). This implies that \( C_{x,y} \) is trace-class. Moreover, by Assumption (F) for \( \epsilon = 1 : \|\partial_i k(y, \cdot)\|_{\mathcal{H}} \in L^2(\mathcal{Y}, p_0) \) which leads to:

\[ \int_{\mathcal{X} \times \mathcal{Y}} \|C_{x,y}\|_1 p_0(dx, dy) < \infty. \]

This means that \( C_{x,y} \) is \( p_0 \)-integrable in the Bochner sense (Retherford, 1978, Definition 1 and Theorem 2) and its integral \( C \) is trace-class with:

\[ \|C\|_1 = \left\| \int_{\mathcal{X} \times \mathcal{Y}} C_{x,y} p_0(dx, dy) \right\|_1 \leq \int_{\mathcal{X} \times \mathcal{Y}} \|C_{x,y}\|_1 p_0(dx, dy) < \infty. \]

\[ \Box \]

**Lemma 2.** Let \( \mathcal{X} \) be a topological space endowed with a probability distribution \( \mathbb{P} \). Let \( B \) be a separable Banach space. Define \( R \) to be an \( B \)-valued measurable function on \( \mathcal{X} \) in the Bochner sense (Retherford, 1978 Definition 1), satisfying \( \int_{\mathcal{X}} \|R(x)\|_B d\mathbb{P}(x) < \infty \), then \( R \) is \( \mathbb{P} \)-integrable in the Bochner sense (Retherford, 1978 Definition 1, Theorem 6) and for any continuous linear operator \( T \) from \( B \) to another Banach space \( \mathcal{A} \), then \( TR \) is also \( \mathbb{P} \)-integrable in the Bochner sense and:

\[ \int_{\mathcal{X}} TR(x) d\mathbb{P}(x) = T \int_{\mathcal{X}} R(x) d\mathbb{P}(x) \]

18
For a proof of this result see Retherford, 1978, Definition 1, Theorem 6 and 7.

**Lemma 3 (RKHS of differentiable kernels (Steinwart et al., 2008 Chap 4.4, Corollary 4.36)).** Let \( X \subset \mathbb{R}^d \) be an open subset, \( m \geq 0 \), and \( k \) be an \( m \)-times continuously differentiable kernel on \( X \) with RKHS \( \mathcal{H} \). Then every function \( f \in \mathcal{H} \) is \( m \)-times continuously differentiable, and for \( \alpha \in \mathbb{N}^d_0 \) with \( |\alpha| \leq m \) we have:

\[
|\partial^\alpha f(x)| \leq \|f\|_\mathcal{H}^2 (\partial^\alpha k(x,x))^{\frac{1}{2}}
\]

\[
\partial^\alpha f(x) = (f, \partial^\alpha k(x,\cdot))_\mathcal{H}
\]

A proof of this result can be found in Steinwart et al., 2008 (Chap 4.4, Corollary 4.36)

**Lemma 4.** Under Assumptions (B) to (D) we have the following:

\[
\int_X \pi(dx) \int_{\partial \mathcal{Y}} p_0(y|x) \nabla_T \mathcal{Y}(x,y) \tilde{dS} = 0 \quad \forall T \in T
\]

where \( \partial \mathcal{Y} \) is the boundary of \( \mathcal{Y} \) and \( \tilde{dS} \) is an oriented surface element of \( \partial \mathcal{Y} \).

**Proof.** First let’s prove that \( \|\nabla_T \mathcal{Y}(x,y)\| \|p_0(y|x)\| = o(\|y\|^{1-d}) \) for all \( x \in X \). Where the norm used is the euclidian norm in \( \mathbb{R}^d \). Using the reproducing property and Cauchy-Schwarz inequality one can see that:

\[
\|\nabla_T \mathcal{Y}(x,y)\|^2 = \sum_{i=1}^d (\partial_i T(x,y))^2 = \sum_{i=1}^d (T_x, \partial_i k(y,\cdot))^2
\]

\[
\leq \|T_x\|^2 (\sum_{i=1}^d \|\partial_i k(y,\cdot)\|^2)
\]

By Assumption (D), one can see that \( \sqrt{\sum_{i=1}^d \|\partial_i k(y,\cdot)\|^2} p_0(y|x) = O(\|y\|^{1-d}) \), therefore it comes that \( \|\nabla_T \mathcal{Y}(x,y)\| p_0(y|x) = O(\|y\|^{1-d}) \). Using Lemma 5 one gets that \( \int_{\partial \mathcal{Y}} p_0(y|x) \nabla_T \mathcal{Y}(x,y) \tilde{dS} = 0 \) for all \( x \in X \) which leads to the result.

**Lemma 5.** Let \( \Omega \) be an open set in \( \mathbb{R}^d \) with piece-wise smooth boundary \( \partial \Omega \). Let \( u \) be a real valued function defined over \( \Omega \) and \( v : \mathbb{R}^d \rightarrow \mathbb{R}^d \) a vector valued function. We assume that \( u \) and \( v \) are measurable and that \( \|v(x)\| u(x) = O(\|x\|^{1-d}) \). Then the following surface integral is null:

\[
\int_{\partial \Omega} u(x)v(x) \tilde{dS} = 0
\]

where \( \tilde{dS} \) is an element of the surface \( \partial \Omega \).

More details on this result can be found in Pietzsch, 1994

**Lemma 6 (Generalized representer theorem).** Let \( \mathcal{H} \) be a vector-valued Hilbert space and let \( (\phi_i)_{i=1}^m \in \mathcal{H}^m \). Suppose \( J : \mathcal{H} \rightarrow \mathbb{R} \) is such that \( J(T) = V((T, \phi_1)_H, ..., (T, \phi_m)_H) \) for \( T \in \mathcal{H} \), where \( V : \mathbb{R}^m \rightarrow \mathbb{R} \) is a convex and gâteaux-differentiable function. Define:

\[
T_\lambda = \arg \inf_{T \in \mathcal{H}} J(T) + \frac{\lambda}{2} \|T\|^2_\mathcal{H}
\]

where \( \lambda > 0 \). Then there exists \( (\alpha_i)_{i=1}^m \in \mathbb{R}^m \) such that \( T_\lambda = \sum_{i=1}^m \alpha_i \phi_i \) where \( \alpha := (\alpha_1, ..., \alpha_m) \) satisfies the following equation:

\[
(\lambda I + (\nabla V) \circ K)\alpha = 0,
\]

with \( (K)_{i,j} = (\phi_i, \phi_j)_\mathcal{H}, 8 \in [m], j \in [m] \).
Lemma 3.1

Let $A : \mathcal{H} \to \mathbb{R}^m$, $T \mapsto \langle (T, \phi_i)_\mathcal{H} \rangle_{i=1}^m$. Then $T_\lambda = \arg\inf_{T \in \mathcal{H}} V(A(T)) + \frac{\lambda}{2} \|T\|_H^2$. Taking the gâteaux-differential at $T$, the optimality condition yields:

$$0 = A^* \nabla V(A(T_\lambda)) + \lambda T_\lambda \Leftrightarrow A^* \left( - \frac{1}{\lambda} \nabla V(A(T_\lambda)) \right) = T_\lambda$$

$$\Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^* \alpha, \alpha = - \frac{1}{\lambda} \nabla V(A(T_\lambda))$$

$$\Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^* \alpha, \alpha = - \frac{1}{\lambda} \nabla (AA^* \alpha)$$

where $A^* : \mathbb{R}^m \to \mathcal{H}$ is the adjoint of $A$ which can be obtained as follows. Note that:

$$\langle \forall T \in \mathcal{H} \rangle (\forall \alpha \in \mathbb{R}^m) \langle AT, \alpha \rangle = \sum_{i=1}^m \alpha_i (T, \phi_i)_\mathcal{H} = \langle T, \sum_{i=1}^m \alpha_i \phi_i \rangle_\mathcal{H}$$

thus $A^* \alpha = \sum_{i=1}^m \alpha_i \phi_i$. Therefore $AA^* \alpha = \sum_{i=1}^m \alpha_j \langle \phi_j, \phi_i \rangle_\mathcal{H}$ and hence $AA^* = K$. \hfill \Box

Lemma 7 (Bound on KL divergence between $p_f$ and $p_y$ (van der Vaart et al., 2008 Lemma 3.1)). Assume that $\|k\|_\infty < \infty$ and let $f$ and $g$ in $\mathcal{H}_Y$ such that $Z(f)$ and $Z(g)$ are finite, then: $KL(p_f || p_g) \leq \|f - g\|_\infty \exp \|f - g\|_\infty (1 + \|f - g\|_\infty)$

Lemma 8 (see Lemma 14 in Sriperumbudur et al., 2014). Suppose $\sup_{y \in \mathcal{Y}} k(y, y) < \infty$ and $supp(q_0) = \mathcal{Y}$. Then $T = \mathcal{H}$ and for any $T_0$ there exists $\tilde{T}_0 \in \mathcal{R}(C)$ such that $p_{\tilde{T}_0} = p_{T_0}$.

Proof. Since $\|k\|_\infty < \infty$ then $Z(T_0) \leq \exp \|T_0\|_H \|k\|_\infty < \infty$ for all $T \in \mathcal{H}$, therefore $T = \mathcal{H}$. Moreover, since $\supp(p_{T_0})(y|x) = \mathcal{Y}$ for all $x \in \mathcal{X}$, this implies that the null space of $C \mathcal{N}(C)$ can either be the set of functions $T(x, y) = m(x)$ or $\{0\}$. Indeed, for $T \in \mathcal{N}(C)$ we have $\langle T, CT \rangle = 0$ which leads to $\int_{\mathcal{X} \times \mathcal{Y}} \nabla_y T(x, y)^T \tilde{p}_0(dx, dy) = 0$ which means that $p_{\tilde{T}_0}$-almost surely, $T_0(y) = m(x)$ a constant function of $y$ if the set of constant functions belong to $\mathcal{H}_Y$, or $T_0(x, y) = 0$ otherwise. Let $\tilde{T}_0$ be the orthogonal projection of $T_0$ onto $\mathcal{R}(C) = \mathcal{N}(C)^\perp$ then $\tilde{T}_0$ can be written in the form $\tilde{T}_0(x, y) = m(x) + \tilde{T}_0(x, y)$. It comes that $\int_{\mathcal{Y}} \exp \tilde{T}_0(x, y) q_0(dy) = \exp m(x) \int_{\mathcal{Y}} \exp \tilde{T}_0(x, y) q_0(dy)$ almost surely in $x$. And we finally get $p_{\tilde{T}_0}$-almost surely:

$$p_{\tilde{T}_0}(y|x) = \frac{\exp \tilde{T}_0(x, y)}{Z(\tilde{T}_0(x))} = \frac{\exp \tilde{T}_0(x, y) + m(x)}{\exp m(x)Z(\tilde{T}_0(x))} = p_{\tilde{T}_0}(y|x)$$

\hfill \Box

Lemma 9 (Proposition A.3 in Sriperumbudur et al., 2014). Let $C$ be a bounded, positive self-adjoint compact operator on a separable Hilbert space $\mathcal{H}$. For $\lambda > 0$ and $T \in \mathcal{H}$, define $T_\lambda := (C + \lambda I)^{-1}CT$ and $\mathcal{A}_\theta(\lambda) := \|C^\theta(T_\lambda - T)\|_H$ for $\theta \geq 0$. Then the following hold.

1. For any $\theta > 0$, $\mathcal{A}_\theta(\lambda) \to 0$ as $\lambda \to 0$ and if $T \in \mathcal{R}(C)$, then $\mathcal{A}_\theta(\lambda) \to 0$ as $\lambda \to 0$.

2. If $T \in \mathcal{R}(C^\beta)$ for $\beta \geq 0$ and $\beta + \theta > 0$, then

$$\mathcal{A}_\theta(\lambda) \leq \max\{1, \|C\|^{\beta + \theta - 1}\} \lambda^{\min\{1, \beta + \theta\}} \|C^{-\beta}T\|_H$$

Proof. 1. Since $C$ is bounded, compact and positive self-adjoint, Hilbert-Schmidt and $\mathcal{H}$ is a separable Hilbert space then $C$ admits an Eigen-decomposition of the form $C = \sum_i \alpha_i \phi_i \phi_i^\mathcal{H}$ where $(\alpha_i)_{i \in \mathbb{N}}$ are positive eigenvalues and $(\phi_i)_{i \in \mathbb{N}}$ are the corresponding unit eigenvectors that form an ONB for $\mathcal{R}(C)$. Let $\theta = 0$. Since $T \in \mathcal{R}(C)$,

$$\mathcal{A}_0^2(\lambda) = \|(C + \lambda I)^{-1}CT - T\|_H^2 = \left\| \sum_i \frac{\alpha_i}{\alpha_i + \lambda} (T, \phi_i)_\mathcal{H} \phi_i - \sum_i (T, \phi_i)_\mathcal{H} \phi_i \right\|_H^2$$

$$= \left\| \sum_i \frac{\lambda}{\alpha_i + \lambda} (T, \phi_i)_\mathcal{H} \phi_i \right\|_H^2 = \sum_i \left( \frac{\lambda}{\alpha_i + \lambda} \right)^2 (T, \phi_i)_\mathcal{H} \to 0 \text{ as } \lambda \to 0$$

20
by the dominated convergence theorem. For any $\theta > 0$, we have:
\[
A^2(\lambda) = \|C^\theta(C + \lambda I)^{-1}CT - C^\theta T\|_H^2 = \left\| \sum_{i} \frac{\alpha_i}{\alpha_i + \lambda} (T, \phi_i)_H \phi_i - \sum_{i} (T, \phi_i)_H \phi_i \right\|_H^2.
\]
Let $T = T_R + T_N$ where $T_R \in \mathcal{R}(C^\theta)$, $T_N \in \mathcal{R}(C^\theta)^\perp$ if $0 < \theta \leq 1$ and $T_N \in \mathcal{R}(C^\theta)^\perp$ if $\theta \geq 1$. Then
\[
A^2(\lambda) = \|C^\theta(C + \lambda I)^{-1}CT - C^\theta T\|_H^2 = \|C^\theta(C + \lambda I)^{-1}CT_R - C^\theta T_R\|_H^2
\]
\[
= \left\| \sum_{i} \frac{\alpha_i^{1+\theta}}{\alpha_i + \lambda} (T_R, \phi_i)_H \phi_i - \sum_{i} \alpha_i^\theta (T_R, \phi_i)_H \phi_i \right\|_H^2
\]
\[
= \left\| \sum_{i} \frac{\alpha_i^\theta}{\alpha_i + \lambda} (T_R, \phi_i)_H \phi_i \right\|_H^2 = \sum_{i} \left( \frac{\alpha_i^\theta}{\alpha_i + \lambda} \right)^2 (T_R, \phi_i)_H^2 \to 0 \text{ as } \lambda \to 0
\]

2. If $T \in \mathcal{R}(C^\theta)$, then there exists $g \in \mathcal{H}$ such that $T = C^\beta g$. This yields
\[
A^2(\lambda) = \|C^\theta(C + \lambda I)^{-1}CT - C^\theta T\|_H^2 = \|C^\theta(C + \lambda I)^{-1}C^1C^\beta g - C^\beta C^\theta g\|_H^2
\]
\[
= \left\| \sum_{i} \frac{\lambda \alpha_i^\theta}{\alpha_i + \lambda} (g, \phi_i)_H \phi_i \right\|_H^2 = \sum_{i} \left( \frac{\lambda \alpha_i^\theta}{\alpha_i + \lambda} \right)^2 (g, \phi_i)_H^2.
\]

Suppose $0 < \beta + \theta < 1$. Then
\[
\frac{\lambda \alpha_i^\theta}{\alpha_i + \lambda} = \left( \alpha_i \right) \frac{\lambda}{\alpha_i + \lambda} \left( 1 - \frac{\theta}{\alpha_i + \lambda} \right) \leq \lambda^{\beta + \theta}
\]
On the other hand, for $\beta + \theta \geq 1$, we have:
\[
\frac{\lambda \alpha_i^\theta}{\alpha_i + \lambda} = \left( \alpha_i \right) \frac{\lambda \alpha_i^{\beta + \theta - 1}}{\alpha_i + \lambda} \leq \lambda^{\beta + \theta - 1}.\alpha
\]

Using the above bounds yields the result.

\[\square\]

Lemma 10 (Proposition A.4 in Sriperumbudur et al., 2017). Let $\mathcal{X}$ be a topological space, $\mathcal{H}$ be a separable Hilbert space and $\mathcal{L}_+^2(\mathcal{H})$ be the space of positive, self-adjoint Hilbert-Schmidt operators on $\mathcal{H}$. Define $R := \int_{\mathcal{X}} r(x) dP(x)$ and $\hat{R} := \frac{1}{N} \sum_{n=1}^{N} r(X_n)$ where $P \in \mathcal{M}_1^+(\mathcal{X})$ is a positive measure with finite mean, $(X_n)_{n=1}^{N} \sim P$ and $r$ is an $\mathcal{L}_+^2(\mathcal{H})$-valued measurable function on $\mathcal{X}$ satisfying $\int_{\mathcal{X}} \|r(x)\|_{\mathcal{L}_+^2(\mathcal{H})} dP(x) < \infty$. Define $g_\lambda := (R + \lambda I)^{-1}Rg$ for $g \in \mathcal{H}$, $\lambda > 0$ and $A_\lambda(\lambda) := \|g_\lambda - g\|_{\mathcal{H}}$. Let $\alpha \geq 0$ and $\theta \geq 0$. Then the following hold:

1. $\|\hat{R} - R\|_{\mathcal{H}} = O_p\left( \frac{\lambda^{\alpha \theta}}{\sqrt{N}} \right)$
2. $\|R^\alpha (R + \lambda I)^{-\theta}\| \leq \lambda^{\alpha - \theta}$.
3. $\|\hat{R}^\alpha (\hat{R} + \lambda I)^{-\theta}\| \leq \lambda^{\alpha - \theta}$.
4. $\|(R + \lambda I)^{-\theta}(\hat{R} - R)\| = O_p\left( \frac{1}{\sqrt{\lambda N}} \right)$.

Proof. 1. Note that for any $f \in \mathcal{H}$,
\[
E_p\|\hat{R} - R\|_{\mathcal{H}}^2 = \|\hat{R}f\|_{\mathcal{H}}^2 + \|Rf\|_{\mathcal{H}}^2 - 2E_p\langle \hat{R}f, Rf \rangle_{\mathcal{H}}
\]
where $E_p\langle \hat{R}f, Rf \rangle_{\mathcal{H}} = \frac{1}{N} \sum_{n=1}^{N} E_p\langle r(X_n), f \rangle_{\mathcal{H}} = \frac{1}{N} \sum_{n=1}^{N} E_p\langle r(X_n), f \rangle_{HS} \|Rf\|_{HS}. \|Rf\|_{HS.}$. Since $\int_{\mathcal{X}} \|r(x)\|_{\mathcal{L}_+^2(\mathcal{H})} dP(x) < \infty$, $r(x)$ is $P$-integrable in the Bochner sense (see Retherford, 1978), and therefore it follows $E_p\langle r(X_n), f \rangle_{HS} = \langle \int_{\mathcal{X}} r(x) dP(x), f \rangle_{HS} = \|Rf\|_{HS}$. Therefore,
\[
E_p\|\hat{R} - R\|_{\mathcal{H}}^2 = E_p\|\hat{R}f\|_{\mathcal{H}}^2 - \|Rf\|_{\mathcal{H}}^2
\]
where

\[
\mathbb{E}_\phi \left[ \frac{1}{m} \sum_{a=1}^{m} r(X_a) f \right] = \mathbb{E}_\phi \left[ \frac{1}{m} \sum_{a,b=1}^{m} \mathbb{E}_\phi (r(X_a) f, r(X_b) f) \right].
\]

Splitting the sum into two parts (one with \(a = b\) and the other with \(a \neq b\)), it is easy to verify that

\[
\mathbb{E}_\phi \| \hat{R} f \|^2_{\mathcal{H}} = \frac{1}{m} \int_X \| r(x) f \|^2_{\mathcal{D}} + \frac{m-1}{m} \| R f \|^2_{\mathcal{H}},
\]

therefore yielding

\[
\mathbb{E}_\phi \| (\hat{R} - R) f \|^2_{\mathcal{H}} = \frac{1}{m} \left( \int_X \| r(x) f \|^2_{\mathcal{D}} \right) \leq \frac{1}{m} \int_X \| r(x) f \|^2_{\mathcal{H}} \leq \frac{m}{m-1} \| R f \|^2_{\mathcal{H}}.
\]

Using Chebyshev’s inequality, yields the result.

1. Since \(\mathbb{E}_\phi \| R f \|^2_{\mathcal{H}} = \mathbb{E}_\phi \| \hat{R} f \|^2_{\mathcal{H}}\), the inequality follows from (3). Using (12), can be shown to be bounded as

\[
\mathbb{E}_\phi \| (\hat{R} - R) f \|^2_{\mathcal{H}} \leq \frac{1}{m} \int_X \| (R + \lambda I) - \theta f(x) \|^2_{\mathcal{H}} d\mathbb{P}(x)
\]

Note that

\[
\| (R + \lambda I) - \theta f(x) \|^2_{\mathcal{H}} = (R + \lambda I) - \theta f(x), R + \lambda f(x) \|^2_{\mathcal{H}} = \| (R + \lambda I) - \theta f(x) \|^2_{\mathcal{H}} \leq \lambda - \theta \| f(x) \|^2_{\mathcal{H}}
\]

where the inequality follows from (3). Using (12) and (13), we obtain

\[
\mathbb{E}_\phi \| (R + \lambda I) - \theta f(x) \|^2_{\mathcal{H}} \leq \frac{1}{m\lambda \theta} \int_X \| (R + \lambda I) - \theta f(x) \|^2_{\mathcal{H}} d\mathbb{P}(x)
\]

The result follows by applying Chebyshev’s inequality.

\[\square\]

\section{Failure case for the score-matching approach}

We first recall the expressions of the score and expected conditional score for convenience. If \(r\) and \(s\) are two densities that are differentiable and positive, then the score objective as introduced in Hyvärinen et al., 2005 is given by:

\[
\mathcal{J}(r|s) := \frac{1}{2} \int_X r(x) \| \nabla_s \log r(x) - \nabla_s \log s(x) \|^2 dx
\]

If \(p_0(y|x)\) and \(q(y|x)\) are two conditional densities, then the expected conditional score under some marginal distribution \(\pi(x)\) is given by:

\[
J(p_0|q) = \int_X \mathcal{J}(p_0(.|x)q(.|x))\pi(x) dx
\]
Figure 3: A failure case for the expected conditional score-matching. Here a conditional density of the form $p_0(y|x) = p_A(y)H(x) + (1 - H(x))p_B(y)$ is considered, where $p_A$ and $p_B$ are supported on two disjoint sets $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^n$ and $H$ denotes the Heaviside step function. The red curve and blue curve represent $p_0(y|x > 0) = p_A$ and $p_0(y|x \leq 0) = p_B$ respectively, while the green curve represent the mixture $q(y) = \frac{1}{2}(p_A(y) + p_B(y))$. This is a case where the expected conditional score fails to separate the two conditional distributions $p_0(y|x)$ and $q(y)$.

The positivity condition of the target density $r$ is crucial to get a well-behaved divergence between $r$ and $s$ in (14). When this condition fails, the score becomes degenerate. For instance, if $r$ is supported on two disjoint sets $A$ and $B$ of $\mathcal{X}$ it can be written in the form:

$$r(x) = \alpha_A p_A(x) + \alpha_B p_B(x)$$

where $\alpha_A$ and $\alpha_B$ are non-negative and sum to 1, and $p_A$ and $p_B$ are two distributions supported on $A$ and $B$ respectively. In this case, any mixture $s(x) = \beta_A p_A(x) + \beta_B p_B(x)$ satisfies $J(r||s) = 0$.

Similarly, for the conditional expected score in (15) to be well behaved, the conditional density $p_0(y|x)$ needs to be positive on $\mathcal{Y}$ for all $x$ in $\mathcal{X}$. When this condition fails to hold, the same degeneracy happens. Indeed, as shown in Figure 3, consider $p_0$ of the form:

$$p_0(y|x) = p_A(y)H(x) + (1 - H(x))p_B(y)$$

where $p_A$ and $p_B$ are supported on two disjoint sets $A$ and $B$ respectively and $H$ denotes the Heaviside step function. For this choice of $p_0$ any mixture $q(y) = \beta_A p_A(y) + \beta_B p_B(y)$ of $p_A$ and $p_B$ satisfies $J(p_0||q) = 0$. This is because their scores match exactly: $\nabla_y \log p_0(y|x) = \nabla_y \log q(y)$ whenever $p_0(y|x) > 0$. Note that in this case $q$ doesn’t depend on $x$, which means that this approach might learn a model where $x$ and $y$ are independent while a simple investigation of the joint samples $(X_i, Y_i)$ would suggest the opposite.

E Additional experimental results

Additional experimental results are shown in Figure 4 on the Red Wine and Parkinsons datasets.

Experimental results on the synthetic grid dataset are shown in Figure 5 in the case where an isotropic RBF kernel is used.

References


23
Figure 4: Scatter plot of 2-d slices of red wine and parkinsons data sets, the dimensions are $(x_6, x_7)$ for red wine and $(x_{15}, x_{16})$ for parkinsons. The black points represent 1000 data points from the data sets. In red, 1000 samples from each of the three models KEF, KCEF and NADE.
Figure 5: Experimental comparison of proposed method KCEF and other methods (LSCDE and NADE) on synthetic grid dataset. log-likelihood per dimension vs dimension, $N = 2000$. The log-likelihood is evaluated on a separate test set of size 2000.


