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 Hilbert space we denote by $\langle .,. \rangle_E$ and $\|.\|_E$ its corresponding inner product and norm, respectively. If E and F are two Hilbert spaces, we use $\|.\|$ 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 $\|.\|_1$ and $\|.\|_2$ 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 $(g \otimes f)h = g\langle f, h \rangle_E$ for all h in E. $C^1(\Omega)$ denotes the space of continuously differentiable functions on Ω and $L^r(\Omega)$ the space of r-power Lebesgues-integrable function. Finally for any vector β 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 $\mathbb{R}^d$. $\mathcal{H}_\mathcal{Y}$ is a reproducing kernel Hilbert space of functions $f : \mathcal{Y} \to \mathbb{R}$ with kernel $k_\mathcal{Y}$. We denote by $\mathcal{H}$ a vector-valued reproducing kernel Hilbert space of functions $T : x \mapsto T_x$ from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ and we introduce the feature operator $\Gamma : x \mapsto \Gamma_x$ from $\mathcal{X}$ to $\mathcal{L}(\mathcal{H}_\mathcal{Y}, \mathcal{H})$ where $\mathcal{L}(\mathcal{H}_\mathcal{Y}, \mathcal{H})$ is the set of bounded operators from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$. For every $x \in \mathcal{X}$, Γ_x is an operator defined from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{H}$.

The following reproducing properties will be used extensively:

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

$$\partial^\alpha f(y) = \langle f, \partial^\alpha k(y,.) \rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall y \in \mathcal{Y},$$

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

$$\partial_i k(y, y^{'}) = \frac{\partial k(y,y^{'})}{\partial y_i}, \qquad \partial_{i+d} k(y, y^{'}) = \frac{\partial k(y,y^{'})}{\partial y_i^{'}}.$$

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

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

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

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

  Using now the reproducing property in $\mathcal{H}_\mathcal{Y}$ 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 π a probability distribution over $\mathcal{X}$. π 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 $Z : \mathcal{H}_{\mathcal{Y}} \to \mathbb{R}_+^*$, such that for every $f \in \mathcal{H}_{\mathcal{Y}}$ we have

$$Z(f) := \int_{\mathcal{Y}} \exp\left(\langle f, k(y,.)\rangle_{\mathcal{H}_{\mathcal{Y}}}\right) q_0(\mathrm{d}y).$$

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

$$\mathcal{P} = \left\{ p_T(x|y) = q_0(y) \frac{e^{(\langle T, \Gamma_x k(y,\cdot)\rangle_{\mathcal{H}})}}{Z(T_x)} \middle| T \in \mathcal{T} \right\}.$$

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 π,

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

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 distribtion $p_0(x)$ 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**) $k_{\mathcal{Y}}$ is twice continuously differentiable on $\mathcal{Y} \times \mathcal{Y}$ and $\partial^{\alpha,\alpha} k_{\mathcal{Y}}$ 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 $\partial\mathcal{Y}$: $\|\partial_i k(y,\cdot)\|_{\mathcal{Y}} p_0(y|x) = o(\|y\|^{1-d})$

(**E**) The operator Γ is uniformly bounded for the operator norm $\|\Gamma_x\|_{Op} \leq \kappa$ for all $x \in \mathcal{X}$.

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

$$\|\partial_i k(y,\cdot)\|_{\mathcal{Y}} \in L^{2\epsilon}(\mathcal{Y}, p_0),\ \|\partial_i^2 k(y,\cdot)\|_{\mathcal{Y}} \in L^{\epsilon}(\mathcal{Y}, p_0),\ \|\partial_i k(y,\cdot)\|_{\mathcal{Y}} \partial_i \log q_0(y) \in L^{\epsilon}(\mathcal{Y}, p_0).$$

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. *$J(p_{T_0}||p_T) < +\infty$ for all $T \in \mathcal{T}$*
2. *For all $T \in \mathcal{H}$ define*
$$J(T) = \frac{1}{2}\langle T - T_0, C(T - T_0)\rangle_{\mathcal{H}}, \tag{6}$$
where
$$C := \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\sum_{i=1}^{d} \left[\Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)\right]}_{C_{x,y}} p_0(dx,dy) = \mathbb{E}_{p_0}[C_{X,Y}]. \tag{7}$$
then C a trace-class positive operator on $\mathcal{H}$ and for all $T \in \mathcal{T}$ $J(T) = J(p_{T_0}||P_T)$.
3. *Alternatively,*
$$J(T) = \frac{1}{2}\langle T, CT\rangle_{\mathcal{H}} + \langle T, \xi\rangle_{\mathcal{H}} + J(p_{T_0}||q_0).$$
where
$$\mathcal{H} \ni \xi := \int_{\mathcal{X}\times\mathcal{Y}} \underbrace{\sum_{i=1}^{d} \Gamma_x \left[\partial_i \log q_0(y) \partial_i k(y,\cdot) + \partial_i^2 k(y,\cdot)\right]}_{\xi_{x,y}} p_0(dx,dy) = \mathbb{E}_{p_0}[\xi_{X,Y}]$$
Moreover, T_0 satisfies $CT_0 = -\xi$
4. *For any $\lambda > 0$, a unique minimizer T_λ of $J_\lambda(T) := J(T) + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}}$ over $\mathcal{H}$ exists and is given by:*
$$T_\lambda = -(C + \lambda I)^{-1}\xi = (C + \lambda I)^{-1} C T_0.$$

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

1. By the reproducing property of the real valued space $\mathcal{H}_{\mathcal{Y}}$ we have: $T(x,y) = \langle T_x, k(y,\cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}}$. Using the reproducing property for the derivatives of real valued functions in an RKHS in Lemma 3, we get
$$\partial_i T(x,y) = \partial_i \langle T_x, k(y,\cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle T_x, \partial_i k(y,\cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}} \qquad \forall i \in [d].$$
Finally, using the reproducing property in the vector-valued space $\mathcal{H}$,
$$\partial_i T(x,y) = \langle T, \Gamma_x \partial_i k(y,\cdot)\rangle_{\mathcal{H}}, \qquad \forall i \in [d].$$
it is easy to see that
$$J(p_{T_0}||p_T) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \langle T_0 - T, \Gamma_x \partial_i k(y,.)\rangle^2_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y). \tag{8}$$
By Assumptions (**E**) and (**F**),
$$\|\Gamma_x \partial_i k(y,\cdot)\|_{\mathcal{H}} \le \|\Gamma_x\|_{Op} \|\partial_i k(y,\cdot)\|_{\mathcal{H}_{\mathcal{Y}}} \le \kappa \sqrt{\partial_i \partial_{i+d} k(y,y)} \in L^2(p_0),$$
and therefore by Cauchy-Schwarz inequality,
$$J(T) = J(p_{T_0}||p_T) \le \frac{1}{2}\|T_0 - T\|^2_{\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|^2_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y) < +\infty.$$
which means that $J(T) < \infty$ for all $T \in \mathcal{T}$.

2. Starting from (8), it is easy to see that:

$$
\begin{aligned}
J(T) &= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} \langle T_0 - T, \Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)(T_0 - T)\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y) \\
&= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \langle T_0 - T, C_{x,y}(T_0 - T)\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y)
\end{aligned}
$$

.

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

$$
J(T) = \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Y}} \langle (T_0 - T)\otimes(T_0 - T), C_{x,y}\rangle_{HS} p_0(\mathrm{d}x, \mathrm{d}y)
$$

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:

$$
J(T) = \frac{1}{2}\left\langle (T_0 - T)\otimes(T_0 - T), \int_{\mathcal{X}}\int_{\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y)\right\rangle_{HS} = \frac{1}{2}\langle T_0 - T, C(T_0 - T)\rangle_{\mathcal{H}}
$$

3. From (6) we have $J(T) = \frac{1}{2}\langle T, CT\rangle_{\mathcal{H}} - \langle T, CT_0\rangle_{\mathcal{H}} + \frac{1}{2}\langle T_0, CT_0\rangle_{\mathcal{H}}$. Recalling that: $\partial_i T(x,y) = \langle T, \Gamma_x \partial_i k(y,\cdot)\rangle_{\mathcal{H}}$ 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:

$$
\begin{aligned}
\langle T, CT_0\rangle_{\mathcal{H}} &= \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i T_0(x,y)\Big] p_0(\mathrm{d}x, \mathrm{d}y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i \log p_0(y|x)\Big] p_0(\mathrm{d}x)\mathrm{d}y - \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i T(x,y)\partial_i \log q_0(y|x)\Big] p_0(\mathrm{d}x, \mathrm{d}y) \\
&\overset{(a)}{=} \int_{\mathcal{X}} p_0(\mathrm{d}x)\int_{\partial\mathcal{Y}} p_0(y|x)\nabla_y T(x,y).\vec{dS} - \int_{\mathcal{X}\times\mathcal{Y}} \Big[\sum_{i=1}^{d} \partial_i^2 T(x,y) + \partial_i T(x,y)\partial_i \log q_0(y|x)\Big] p_0(\mathrm{d}x, \mathrm{d}y).
\end{aligned}
$$

(a) is obtained using the first Green's identity, where $\partial\mathcal{Y}$ is the boundary of $\mathcal{Y}$ and $\vec{dS}$ is the oriented surface element. The first term $\int_{\mathcal{X}} \pi(\mathrm{d}x)\int_{\partial\mathcal{Y}} p_0(y|x)\nabla_y T(x,y).\vec{dS}$ vanishes by Lemma 4, which relies on Assumption **(D)**. The second term can be written as: $\int_{\mathcal{X}\times\mathcal{Y}} \langle T, \xi_{x,y}\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y)$.
By Assumptions **(E)** and **(F)** $\xi_{x,y}$ is Bochner p_0-integrable, therefore:

$$
\int_{\mathcal{X}\times\mathcal{Y}} \langle T, \xi_{x,y}\rangle_{\mathcal{H}} p_0(\mathrm{d}x, \mathrm{d}y) = \left\langle T, \int_{\mathcal{X}\times\mathcal{Y}} \xi_{x,y} p_0(\mathrm{d}x, \mathrm{d}y)\right\rangle_{\mathcal{H}} = \langle T, \xi\rangle_{\mathcal{H}}.
$$

Hence $\langle T, CT_0\rangle_{\mathcal{H}} = -\langle T, \xi\rangle_{\mathcal{H}}$ and $\xi = -CT_0$. Moreover, one can clearly see that:

$$
\langle T_0, CT_0\rangle_{\mathcal{H}} = \int_{\mathcal{X}\times\mathcal{Y}} \sum_{i=1}^{d} (\partial_i T_0(x,y))^2 p_0(\mathrm{d}x, \mathrm{d}y) = J(p_{T_0}||q_0).
$$

And the result follows.

4. For $\lambda > 0$, $(C + \lambda I)$ 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:

$$
J_\lambda(T) = \frac{1}{2}\|(C + \lambda I)^{\frac{1}{2}} T + (C + \lambda I)^{-\frac{1}{2}}\xi\|_{\mathcal{H}}^2 - \frac{1}{2}\langle \xi, (C + \lambda I)^{-1}\xi\rangle_{\mathcal{H}} + c_0
$$

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 $J_\lambda(T)$.

□

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

$$\hat{J}(T) := \frac{1}{2}\langle T, \hat{C}T\rangle_{\mathcal{H}} + \langle T, \hat{\xi}\rangle_{\mathcal{H}} + J(p_{T_0}||q_0).$$

where:

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

are the empirical estimators of C and ξ respectively.

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

1. *The unique minimizer $T_{\lambda,n}$ of $\hat{J}_\lambda(T) := \hat{J}(T) + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}}$ 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*

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

where (β_b) are obtained by solving the following linear system:

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

with:

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

and:

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

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

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

$$\begin{aligned}
T_{\lambda,n} &= \operatorname*{arginf}_{T\in\mathcal{H}} \frac{1}{2}\langle T\hat{C}T\rangle_{\mathcal{H}} + \langle T, \hat{\xi}\rangle_{\mathcal{H}} + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}} \\
&= \operatorname*{arginf}_{T\in\mathcal{H}} \frac{1}{2}\sum_{a=1}^{n}\sum_{i=1}^{d}\langle T, \Gamma_{X_a}\partial_i k(Y_a,.)\rangle^2_{\mathcal{H}} + \langle T, \hat{\xi}\rangle_{\mathcal{H}} + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}} \\
&= \operatorname*{arginf}_{T\in\mathcal{H}} V(\langle T, \phi_1\rangle_{\mathcal{H}}, ..., \langle T, \phi_{nd+1}\rangle_{\mathcal{H}}) + \frac{\lambda}{2}\|T\|^2_{\mathcal{H}}.
\end{aligned}$$

Where $V(\theta_1, ..., \theta_{nd+1}) := \frac{1}{2n}\sum_{a=1}^{n}\sum_{i=1}^{d}\theta^2_{(a-1)d+i} + \theta_{nd+1}$ is a convex differentiable function and $\phi_{(a-1)d+i} := \Gamma_{X_a}\partial_i k(Y_a,.)$ where $a \in [n], i \in [d]$ and $\phi_{nd+1} = \hat{\xi}$. Therefore, it follows from Lemma 6 that:

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

where δ and β satisfy:

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

with $K = \begin{pmatrix} G & h \\ h^T & \|\hat{\xi}\|^2_{\mathcal{H}} \end{pmatrix}$.

The gradient ∇V of V is given by $\nabla V(z,t) = (\frac{1}{n}z, 1)$. The above equation reduces then to $\lambda\delta + 1 = 0$ and $\lambda\beta + \frac{1}{n}G\beta + \frac{\delta}{n}h = 0$ which yields $\delta = -\frac{1}{\lambda}$ and $(\frac{1}{n}G + \lambda I)\beta = \frac{1}{n\lambda}h$.

□

B.3 Consistency and convergence

Theorem 5 (**Consistency and convergence rates for** $T_{\lambda,n}$)**.** *Let $\gamma > 0$ be a positive number and define $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$, under Assumptions (**A**) to (**F**):*

1. *if $T_0 \in \overline{\mathcal{R}(C)}$ then $\|T_{n,\lambda} - T_0\| \to 0$ when $\lambda\sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.*
2. *if $T_0 \in \mathcal{R}(C^\gamma)$ for some $\gamma > 0$ then $\|T_{n,\lambda} - T_0\| = \mathcal{O}_{p_0}(n^{-\frac{1}{2}+\alpha})$ for $\lambda = n^{-\alpha}$*

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

$$\begin{aligned}
T_{\lambda,n} - T_\lambda &= -(\hat{C} + \lambda I)^{-1}(\hat{\xi} + (\hat{C} + \lambda I)T_\lambda) \overset{(*)}{=} -(\hat{C} + \lambda I)^{-1}(\hat{\xi} + \hat{C}T_\lambda + C(T_0 - T_\lambda)) \\
&= (\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) - (\hat{C} + \lambda I)^{-1}(\hat{\xi} + \hat{C}T_0) \\
&= (\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0) - (\hat{C} + \lambda I)^{-1}(\hat{\xi} - \xi) + (\hat{C} + \lambda)^{-1}(C - \hat{C})T_0.
\end{aligned}$$

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

$$\begin{aligned}
S_1 &:= \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})(T_\lambda - T_0)\|_{\mathcal{H}} \\
S_2 &:= \|(\hat{C} + \lambda I)^{-1}(\hat{\xi} - \xi)\|_{\mathcal{H}} \\
S_3 &:= \|(\hat{C} + \lambda I)^{-1}(C - \hat{C})T_0\|_{\mathcal{H}} \\
\mathcal{A}_0(\lambda) &:= \|T_{\lambda,n} - T_0\|_{\mathcal{H}}.
\end{aligned}$$

it comes then:

$$\begin{aligned}
\|T_\lambda - T_0\|_{\mathcal{H}} &\leq \|T_{\lambda,n} - T_\lambda\|_{\mathcal{H}} + \|T_\lambda - T_0\|_{\mathcal{H}} \\
&\leq S_1 + S_2 + S_2 + \mathcal{A}_0(\lambda),
\end{aligned}$$

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:

$$\begin{aligned}
\|C_{x,y}\|^2_{HS} &= \sum_{i,j=1}^{d} \langle \Gamma_x \partial_i k(y,\cdot), \Gamma_x \partial_j k(y,\cdot)\rangle^2_{\mathcal{H}} \leq \sum_{i,j=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|^2_{\mathcal{H}} \|\Gamma_x \partial_j k(y,\cdot)\|^2_{\mathcal{H}} \\
&\leq (\sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|^2_{\mathcal{H}})^2 \leq d \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|^4_{\mathcal{H}} \leq d\kappa^4 \sum_{i=1}^{d} \|\partial_i k(y,\cdot)\|^4_{\mathcal{H}_\mathcal{Y}}.
\end{aligned}$$

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

$$\int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|^2_{HS} p_0(\mathrm{d}x, \mathrm{d}y) \leq d\kappa^4 \sum_{i=1}^{d} \int_{\mathcal{X}\times\mathcal{Y}} \|\partial_i k(y,\cdot)\|^4_{\mathcal{H}_\mathcal{Y}} p_0(\mathrm{d}x, \mathrm{d}y) < +\infty.$$

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

$$
\begin{aligned}
S_1 &\leq \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})(T_\lambda - T_0)\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{\mathcal{A}(\lambda)}{\lambda\sqrt{n}}) \\
S_3 &\leq \|(\hat{C}+\lambda I)^{-1}\| \|(C-\hat{C})T_0\| = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}})\|_{\mathcal{H}} \\
&\|(C+\lambda I)^{-1}\| \leq \frac{1}{\lambda}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}_{p_0} \|\hat{\xi}-\xi\|^2_{\mathcal{H}} &= \frac{1}{n}\left(\int_{\mathcal{X}\times\mathcal{Y}} \|\xi_{x,y}\|^2_{\mathcal{H}} p_0(\mathrm{d}x,\mathrm{d}y) - \|\xi\|^2\right) \\
&\leq \frac{1}{n}\int_{\mathcal{X}\times\mathcal{Y}} \|\xi_{x,y}\|^2_{\mathcal{H}} p_0(\mathrm{d}x,\mathrm{d}y)
\end{aligned}
$$

By Assumption (F) for $\epsilon = 2$ we have that $\int_{\mathcal{X}\times\mathcal{Y}} \|\xi_{x,y}\|^2_{\mathcal{H}} p_0(\mathrm{d}x,\mathrm{d}y) < \infty$. One can therefore apply Chebychev inequality to get the results. It comes that:

$$
S_2 \leq \|(\hat{C}+\lambda I)^{-1}\| \|\hat{\xi}-\xi\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}})
$$

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

$$
\|T_{\lambda,n} - T_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}(\frac{1}{\lambda\sqrt{n}} + \frac{\mathcal{A}_0(\lambda)}{\lambda\sqrt{n}}) + \mathcal{A}_0(\lambda) \tag{9}
$$

1. By Lemma 9 we have $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$ if $T_0 \in \overline{\mathcal{R}(C)}$. 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:

$$
\mathcal{A}_0(\lambda) \leq \max\{1, \|C\|^{\gamma-1}\}\|C^{-\gamma}T_0\|_{\mathcal{H}}\lambda^{\min\{1,\gamma\}}.
$$

The result follows by choosing $\lambda = n^{-\max\{\frac{1}{4},\frac{1}{2(\gamma+1)}\}} = n^{-\alpha}$.

□

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 $\|k\|_\infty := \sup_{y\in\mathcal{Y}} k(y,y) < \infty$ and that $p_{T_0}(y|x)$ is supported on $\mathcal{Y}$ for all $x \in \mathcal{X}$ 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 $\alpha = \max(\frac{1}{2(\gamma+1)}, \frac{1}{4}) \in (\frac{1}{4}, \frac{1}{2})$, and choosing $\lambda = n^{-\alpha}$ we have that $KL(p_0||p_{T_{n,\lambda}}) = \mathcal{O}_{p_0}(n^{-1+2\alpha})$*

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

$$
KL(p_{T_0}(Y|x)||p_{T_{\lambda,n}}(Y|x)) \leq \|T_0(x) - T_{\lambda,n}(x)\|^2_\infty \exp \|T_0(x) - T_{\lambda,n}(x)\|_\infty (1 + \|T_0(x) - T_{\lambda,n}(x)\|_\infty) \tag{10}
$$

Moreover, using Assumption (E) and the fact that $\|k\|_\infty < \infty$ one can see that

$$
\begin{aligned}
|T_0(x,y) - T_{\lambda,n}(x,y)|_{\mathcal{H}_{\mathcal{Y}}} &= \langle T_0 - T_{\lambda,n}, \Gamma_x k(y,\cdot)\rangle_{\mathcal{H}} \\
&\leq \|T_0 - T_{\lambda,n}\|_{\mathcal{H}} \|\Gamma_x k(y,\cdot)\|_{\mathcal{H}}
\end{aligned}
$$

which gives after taking the supremum:

$$
\|T_0(x) - T_{\lambda,n}(x)\|_{\infty} \leq \kappa \|k\|_{\infty} \|T_0 - T_{\lambda,n}\|_{\mathcal{H}} \tag{11}
$$

for all $x \in \mathcal{X}$. Using (11) in (10) and taking the expectation with respect to x, one can conclude using Theorem 5.

□

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 $C = \int_{\mathcal{X}\times\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y)$ 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:

$$
\begin{aligned}
\|C_{x,y}\|_1 &\leq \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot) \otimes \Gamma_x \partial_i k(y,\cdot)\|_1 \\
&= \sum_{i=1}^{d} \|\Gamma_x \partial_i k(y,\cdot)\|_{\mathcal{H}}^2 \leq \sum_{i=1}^{d} \|\Gamma_x\|_{Op}^2 \|\partial_i k(y,\cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^2 \\
&\overset{(a)}{\leq} \kappa^2 \sum_{i=1}^{d} \|\partial_i k(y,\cdot)\|_{\mathcal{H}_{\mathcal{Y}}}^2 < \infty.
\end{aligned}
$$

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

$$
\int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|_1 p_0(\mathrm{d}x, \mathrm{d}y) < \infty.
$$

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:

$$
\|C\|_1 = \left\| \int_{\mathcal{X}\times\mathcal{Y}} C_{x,y} p_0(\mathrm{d}x, \mathrm{d}y) \right\|_1 \leq \int_{\mathcal{X}\times\mathcal{Y}} \|C_{x,y}\|_1 p_0(\mathrm{d}x, \mathrm{d}y) < \infty.
$$

□

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 A, then TR is also $\mathbb{P}$-integrable in the Bochner sense and:*

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

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

Lemma 3 (**RKHS of differentiable kernels (Steinwart et al., 2008 Chap 4.4, Corollary 4.36)**). *Let $\mathcal{X} \in \mathbb{R}^d$ be an open subset, $m \geq 0$, and k be an m-times continuously differentiable kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then every function $f \in \mathcal{H}$ is m-times continuously differentiable, and for $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$ we have:*

$$|\partial^\alpha f(x)| \leq \|f\|_{\mathcal{H}}^2 (\partial^{\alpha,\alpha} k(x,x))^{\frac{1}{2}}$$
$$\partial^\alpha f(x) = \langle f, \partial^\alpha k(x,\cdot)\rangle_{\mathcal{H}}$$

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:*

$$\int_{\mathcal{X}} \pi(\mathrm{d}x) \int_{\partial\mathcal{Y}} p_0(y|x) \nabla_y T(x,y).\vec{dS} = 0 \qquad\qquad \forall T \in \mathcal{T}$$

where $\partial\mathcal{Y}$ is the boundary of $\mathcal{Y}$ and $\vec{dS}$ is an oriented surface element of $\partial\mathcal{Y}$.

Proof. First let's prove that $\|\nabla_y T(x,y)\| p_0(y|x) = o(\|y\|^{1-d})$ for all $x \in \mathcal{X}$. Where the norm used is the euclidian norm in $\mathbb{R}^d$. Using the reproducing property and Cauchy-Schwarz inequality one can see that:

$$\begin{aligned}\|\nabla_y T(x,y)\|^2 &= \sum_{i=1}^d (\partial_i T(x,y))^2 = \sum_{i=1}^d \langle T_x, \partial_i k(y,.)\rangle^2 \\ &\leq \|T_x\|^2 \Big(\sum_{i=1}^d \|\partial_i k(y,.)\|^2\Big)\end{aligned}$$

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

□

Lemma 5. *Let Ω be an open set in $\mathbb{R}^d$ with piece-wise smooth boundary $\partial\Omega$. Let u be a real valued function defined over Ω and $v : \mathbb{R}^d \to \mathbb{R}^d$ 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:*

$$\int_{\partial\Omega} u(x)v(x).\mathrm{d}\vec{S} = 0$$

where $\mathrm{d}\vec{S}$ 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 $J : \mathcal{H} \to \mathbb{R}$ is such that $J(T) = V(\langle T, \phi_1\rangle_{\mathcal{H}}, ..., \langle T, \phi_m\rangle_{\mathcal{H}})$ for $T \in \mathcal{H}$, where $V : \mathbb{R}^m \to \mathbb{R}$ is a convex and gâteaux-differentiable function. Define:*

$$T_\lambda = \operatorname*{arginf}_{T\in\mathcal{H}} J(T) + \frac{\lambda}{2}\|T\|_{\mathcal{H}}^2$$

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 $(K)_{i,j} = \langle \phi_i, \phi_j\rangle_{\mathcal{H}}, ß \in [m], j \in [m]$

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

$$\begin{aligned} 0 = A^*\nabla V(AT_\lambda) + \lambda T_\lambda \Leftrightarrow A^*\Big(-\frac{1}{\lambda}\nabla V(AT_\lambda)\Big) = T_\lambda \\ \Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^*\alpha, \alpha = -\frac{1}{\lambda}\nabla V(AT_\lambda) \\ \Leftrightarrow (\exists \alpha \in \mathbb{R}^m) T_\lambda = A^*\alpha, \alpha = -\frac{1}{\lambda}\nabla V(AA^*\alpha) \end{aligned}$$

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

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

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

Lemma 7 (Bound on KL divergence between p_f and p_g (van der Vaart et al., 2008 Lemma 3.1))**.** *Assume that $\|k\|_\infty < \infty$ and let f and g in $\mathcal{H}_\mathcal{Y}$ such that $Z(f)$ and $Z(g)$ are finite, then: $KL(p_f\|q_g) \le \|f-g\|_\infty^2 \exp\|f-g\|_\infty(1+\|f-g\|_\infty)$*

Lemma 8 (see Lemma 14 in Sriperumbudur et al., 2017)**.** *Suppose* $\sup_{y\in\mathcal{Y}} k(y,y) < \infty$ *and supp*$(q_0) = \mathcal{Y}$*. Then $\mathcal{T} = \mathcal{H}$ and for any T_0 there exists $\widetilde{T}_O \in \overline{\mathcal{R}(C)}$ such that $p_{\widetilde{T}_0} = p_0$.*

Proof. Since $\|k\|_\infty < \infty$ then $Z(T_x) \le \exp\|T_x\|\|k\|_\infty < \infty$ for all $T \in \mathcal{H}$, therefore $\mathcal{T} = \mathcal{H}$. Moreover, since $\operatorname{supp}(p_{T_0})(y|x) = \mathcal{Y}$ 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 $\int_{\mathcal{X}\times\mathcal{Y}} \|\nabla_y T\|_2^2 p_0(\mathrm{d}x,\mathrm{d}y) = 0$ which means that p_0-almost surely, $T_x(y) = m(x)$ a constant function of y if the set of constant functions belong to $\mathcal{H}_\mathcal{Y}$, or $T_x(y) = 0$ otherwise. Let $\widetilde{T}_0$ be the orthogonal projection of T_0 onto $\overline{\mathcal{R}(C)} = \mathcal{N}(C)^\perp$ then T_0 can be written in the form $T_0(x,y) = m(x) + \widetilde{T}_0(x,y)$. It comes that $\int_\mathcal{Y} \exp T_0(x,y) q_0(\mathrm{d}y) = \exp m(x) \int_\mathcal{Y} \exp \widetilde{T}_0(x,y) q_0(\mathrm{d}y)$ almost surely in x. And we finally get p_0-almost surely:

$$p_{T_0}(y|x) = \frac{\exp T_0(x,y)}{Z(T_0(x))} = \frac{\exp T_0(x,y) + m(x)}{\exp m(x) Z(T_0(x))} = p_{T_0}(y|x)$$

☐

Lemma 9 (Proposition A.3 in Sriperumbudur et al., 2017)**.** *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 $\mathcal{A}_\theta(\lambda) := \|C^\theta(T_\lambda - T)\|_{\mathcal{H}}$ for $\theta \ge 0$. Then the following hold.*

1. *For any $\theta > 0$, $\mathcal{A}_\theta(\lambda) \to 0$ as $\lambda \to 0$ and if $T \in \overline{\mathcal{R}(C)}$, then $\mathcal{A}_0(\lambda) \to 0$ as $\lambda \to 0$.*
2. *If $T \in \mathcal{R}(C^\beta)$ for $\beta \ge 0$ and $\beta + \theta > 0$, then*

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

Proof. 1. Since C is bounded, compact and positive self-adjoint, Hilbert-Shmidt and $\mathcal{H}$ is a separable Hilbert space then C admits an Eigen-decomposition of the form $C = \sum_l \alpha_l\phi_l\langle\phi_l\rangle_{\mathcal{H}}$ where $(\alpha_l)_{l\in\mathbb{N}}$ are positive eigenvalues and $(\phi_l)_{l\in\mathbb{N}}$ are the corresponding unit eigenvectors that form an ONB for $\mathcal{R}(C)$. Let $\theta = 0$. Since $T \in \overline{\mathcal{R}(C)}$,

$$\begin{aligned} \mathcal{A}_0^2(\lambda) = \|(C+\lambda I)^{-1}CT - T\|_{\mathcal{H}}^2 &= \Big\| \sum_i \frac{\alpha_i}{\alpha_i+\lambda}\langle T, \phi_i\rangle_{\mathcal{H}}\phi_i - \sum_i \langle T, \phi_i\rangle_{\mathcal{H}}\phi_i \Big\|_{\mathcal{H}}^2 \\ &= \Big\| \sum_i \frac{\lambda}{\alpha_i+\lambda}\langle T, \phi_i\rangle_{\mathcal{H}}\phi_i \Big\|_{\mathcal{H}}^2 = \sum_i \Big(\frac{\lambda}{\alpha_i+\lambda}\Big)^2 \langle T, \phi_i\rangle_{\mathcal{H}}^2 \to 0 \text{ as } \lambda \to 0 \end{aligned}$$

by the dominated convergence theorem. For any $\theta > 0$, we have:

$$\mathcal{A}_0^2(\lambda) = \|C^\theta(C+\lambda I)^{-1}CT - C^\theta T\|_{\mathcal{H}}^2 = \Big\|\sum_i \frac{\alpha_i}{\alpha_i+\lambda}\langle T,\phi_i\rangle_{\mathcal{H}}\phi_i - \sum_i \langle T,\phi_i\rangle_{\mathcal{H}}\phi_i\Big\|_{\mathcal{H}}^2.$$

Let $T = T_R + T_N$ where $T_R \in \overline{\mathcal{R}(C^\theta)}$, $T_N \in \overline{\mathcal{R}(C^\theta)}^\perp$ if $0 < \theta \le 1$ and $T_N \in \overline{\mathcal{R}(C)}^\perp$ if $\theta \ge 1$. Then

$$\begin{aligned}\mathcal{A}_0^2(\lambda) &= \|C^\theta(C+\lambda I)^{-1}CT - C^\theta T\|_{\mathcal{H}}^2 = \|C^\theta(C+\lambda I)^{-1}CT_R - C^\theta T_R\|_{\mathcal{H}}^2\\ &= \Big\|\sum_i \frac{\alpha_i^{1+\theta}}{\alpha_i+\lambda}\langle T_R,\phi_i\rangle_{\mathcal{H}}\phi_i - \sum_i \alpha_i^\theta\langle T_R,\phi_i\rangle_{\mathcal{H}}\phi_i\Big\|_{\mathcal{H}}^2\\ &= \Big\|\sum_i \frac{\lambda\alpha_i^{\theta}}{\alpha_i+\lambda}\langle T_R,\phi_i\rangle_{\mathcal{H}}\phi_i\Big\|_{\mathcal{H}}^2 = \sum_i\Big(\frac{\lambda\alpha_i^\theta}{\alpha_i+\lambda}\Big)^2\langle T_R,\phi_i\rangle_{\mathcal{H}}^2 \to 0 \text{ as } \lambda\to 0\end{aligned}$$

2. If $T \in \mathcal{R}(C^\beta)$, then there exists $g \in \mathcal{H}$ such that $T = C^\beta g$. This yields

$$\begin{aligned}\mathcal{A}_0^2(\lambda) &= \|C^\theta(C+\lambda I)^{-1}CT - C^\theta T\|_{\mathcal{H}}^2 = \|C^\theta(C+\lambda I)^{-1}C^{1+\beta}g - C^{\beta+\theta}g\|_{\mathcal{H}}^2\\ &= \Big\|\sum_i \frac{\lambda\alpha_i^{\beta+\theta}}{\alpha_i+\lambda}\langle g,\phi_i\rangle_{\mathcal{H}}\phi_i\Big\|_{\mathcal{H}}^2 = \sum_i\Big(\frac{\lambda\alpha_i^{\beta+\theta}}{\alpha_i+\lambda}\Big)^2\langle g,\phi_i\rangle_{\mathcal{H}}^2\end{aligned}$$

Suppose $0 < \beta+\theta < 1$. Then

$$\frac{\lambda\alpha_i^{\beta+\theta}}{\alpha_i+\lambda} = \Big(\frac{\alpha_i}{\alpha_i+\lambda}\Big)^{\beta+\theta}\Big(\frac{\lambda}{\alpha_i+\lambda}\Big)^{1-\beta-\theta}\lambda^{\beta+\theta} \le \lambda^{\beta+\theta}$$

On the other hand, for $\beta+\theta \ge 1$, we have:

$$\frac{\lambda\alpha_i^{\beta+\theta}}{\alpha_i+\lambda} = \Big(\frac{\alpha_i}{\alpha_i+\lambda}\Big)\alpha_i^{\beta+\theta-1}\lambda \le \|\|^{\beta+\theta-1}\lambda.$$

Using the above bounds yields the result.

□

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 $R := \int_{\mathcal{X}} r(x)d\mathbb{P}(x)$ and $\hat{R} := \frac{1}{n}\sum_{a=1}^m r(X_a)$ where $\mathbb{P} \in M_+^1(\mathcal{X})$ is a positive measure with finite mean, $(X_a)_{a=1}^m \sim \mathbb{P}$ and r is an $\mathcal{L}_2^+(\mathcal{H})$-valued measurable function on $\mathcal{X}$ satisfying $\int_{\mathcal{X}}\|r(x)\|_{HS}^2 d\mathbb{P}(x) < \infty$. Define $g_\lambda := (R+\lambda I)^{-1}Rg$ for $g \in \mathcal{H}$, $\lambda > 0$ and $\mathcal{A}_0(\lambda) := \|g_\lambda - g\|_{\mathcal{H}}$. Let $\alpha \ge 0$ and $\theta \ge 0$. Then the following hold:*

1. $\|(\hat{R}-R)(g_\lambda - g)\|_{\mathcal{H}} = O_{\mathbb{P}}(\frac{\mathcal{A}_0(\lambda)}{\sqrt{m}})$
2. $\|R^\alpha(R+\lambda I)^{-\theta}\| \le \lambda^{\alpha-\theta}$.
3. $\|\hat{R}^\alpha(\hat{R}+\lambda I)^{-\theta}\| \le \lambda^{\alpha-\theta}$.
4. $\|(R+\lambda I)^{-\theta}(\hat{R}-R)\| = O_{\mathbb{P}}(\frac{1}{\sqrt{m}\lambda^{2\theta}})$.

Proof. 1. Not that for any $f \in \mathcal{H}$,

$$\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_{\mathcal{H}}^2 = \mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_{\mathcal{H}}^2 + \|Rf\|_{\mathcal{H}}^2 - 2\,\mathbb{E}_{\mathbb{P}}\langle\hat{R}f, Rf\rangle_{\mathcal{H}}$$

where $\mathbb{E}_{\mathbb{P}}\langle\hat{R}f,Rf\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{a=1}^n\mathbb{E}_{\mathbb{P}}\langle r(X_a)f, Rf)\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{a=1}^n\mathbb{E}_{\mathbb{P}}\langle r(X_a), f\otimes Rf\rangle_{HS}$. Since $\int_{\mathcal{X}}\|r(x)\|_{HS}^2 d\mathbb{P}(x) < \infty$, $r(x)$ is $\mathbb{P}$-integrable in the Bochner sense (see Retherford, 1978), and therefore it follows $\mathbb{E}_{\mathbb{P}}\langle r(X_a), f\otimes Rf\rangle_{HS} = \langle\int_{\mathcal{X}} r(x)d\mathbb{P}(x), f\otimes Rf\rangle_{HS} = \|Rf\|_{HS}^2$. Therefore,

$$\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_{\mathcal{H}}^2 = \mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_{\mathcal{H}}^2 - \|Rf\|_{\mathcal{H}}^2$$

where

$$\mathbb{E}_{\mathbb{P}}\|\frac{1}{m}\sum_{a=1}^{m} r(X_a)f\|_{\mathcal{H}}^2 = \frac{1}{m^2}\sum_{a,b=1}^{m}\mathbb{E}_{\mathbb{P}}\langle r(X_A)f, r(X_b)f\rangle_{\mathcal{H}}.$$

Splitting the sum into two parts (one with $a = b$ and the other with $a \neq b$), it is easy to verify that $\mathbb{E}_{\mathbb{P}}\|\hat{R}f\|_{\mathcal{H}}^2 = \frac{1}{m}\int_{\mathcal{X}}\|r(x)f\|_{\mathcal{H}}^2 d\mathbb{P}(x) + \frac{m-1}{m}\|Rf\|_{\mathcal{H}}^2$, therefore yielding

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\|(\hat{R}-R)f\|_{\mathcal{H}}^2 = \frac{1}{m}\Big(\int_{\mathcal{X}}\|r(x)f\|_{\mathcal{H}}^2 d\mathbb{P}(x) - \|Rf\|_{\mathcal{H}}^2\Big) &\leq \frac{1}{m}\int_{\mathcal{X}}\|r(x)f\|_{\mathcal{H}}^2 d\mathbb{P}(x)) \\ &\leq \frac{\|f\|_{\mathcal{H}}^2}{m}\int_{\mathcal{X}}\|r(x)\|_{HS}^2 d\mathbb{P}(x)\end{aligned}$$

Using $f = g_\lambda - g$, an application of Chebyshev's inequality yields the result.

2. $\|R^\alpha(R+\lambda I)^{-\theta}\| = \sup_i \frac{\gamma_i^\alpha}{(\gamma_i+\lambda)^\theta} = \sup_i\left[(\frac{\gamma_i}{\gamma_i+\lambda})^\alpha\frac{1}{(\gamma_i+\lambda)^{\theta-\alpha}}\right] \leq \sup_i \frac{1}{(\gamma_i+\lambda)^{\theta-\alpha}} \leq \lambda^{\alpha-\theta}$, where $(\gamma_i)_{i\in n}$ are the eigenvalues of R.

3. Same as above, after replacing $(\gamma_i)_{i\in\mathbb{N}}$ by the eigenvalues of $\hat{R}$

4. Since $\|(R+\lambda I)^{-\theta}(\hat{R}-R)\| \leq \|(R+\lambda I)^{-\theta}(\hat{R}-R)\|_{HS}^2$, consider $\mathbb{E}_{\mathbb{P}}\|(R+\lambda I)^{-\theta}(\hat{R}-R)\|_{HS}^2$, which using the technique in the proof of (1), can be shown to be bounded as

$$\mathbb{E}_{\mathbb{P}}\|(R+\lambda I)^{-\theta}(\hat{R}-R)\|_{HS}^2 \leq \frac{1}{m}\int_{\mathcal{X}}\|(R+\lambda I)^{-\theta}r(x)\|_{HS}^2 d\mathbb{P}(x) \tag{12}$$

Note that

$$\begin{aligned}\|(R+\lambda I)^{-\theta}r(x)\|_{HS}^2 &= \langle R+\lambda I)^{-\theta}r(x), R+\lambda I)^{-\theta}r(x)\rangle_{HS} \\ &= \|(R+\lambda I)^{-2\theta}\|Tr(r(x)r(x)) = \|(R+\lambda I)^{-2\theta}\|\|r(x)\|_{HS}^2 \\ &\leq \lambda^{-2\theta}\|r(x)\|_{HS}^2 \end{aligned}\tag{13}$$

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

$$\mathbb{E}_{\mathbb{P}}\|(R+\lambda I)^{-\theta}r(x)\|_{HS}^2 \leq \frac{1}{m\lambda^{2\theta}}\int_{\mathcal{X}}\|(R+\lambda I)^{-\theta}r(x)\|_{HS}^2 d\mathbb{P}(x)$$

The result follows by an application of Chebyshev's inequality.

□

D 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:

$$\mathcal{J}(r||s) := \frac{1}{2}\int_{\mathcal{X}} r(x)\|\nabla_x \log r(x) - \nabla_x \log s(x)\|^2 dx \tag{14}$$

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:

$$J(p_0|q) = \int_{\mathcal{X}}\mathcal{J}(p_0(.|x)q(.|x))\pi(x)dx \tag{15}$$

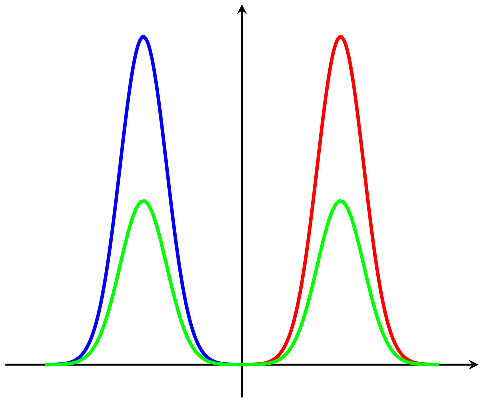

Figure 3: A Failure case for the expected conditional score-matching. Here a conditional density of the form $p_0(y|x) = p_A(y)H(x)+(1-H(x))p_B(y)$ is considered, where p_A and p_B are supported on two disjoint sets $A \subset \mathbb{R}^*_-$ and $B \subset \mathbb{R}^*_+$ and H denotes the Heaviside step function. The red curve and blue curve represent $p_0(y|x>0) = p_A$ and and $p_0(y|x<=0) = p_B$ 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 α_A and α_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

Hyvärinen, Aapo and Peter Dayan (2005). "Estimation of non-normalized statistical models by score matching." In: *Journal of Machine Learning Research* 6, pp. 695–709.

Pietzsch, W. (1994). "D. E. Bourne. P. C. Kendall. Vector analysis and cartesian tensors. Third edition Chapman and Hall. London-Glasgow-New York-Tokyo-Melbourne-Madras 1992, Seitenzahl: 304, Zahl der Abbildungen und Tabellen: 115. ISBN: 0-412-42750-8." In: *Crystal Research and Technology* 29.1, pp. 50–50. URL: http://dx.doi.org/10.1002/crat.2170290115.

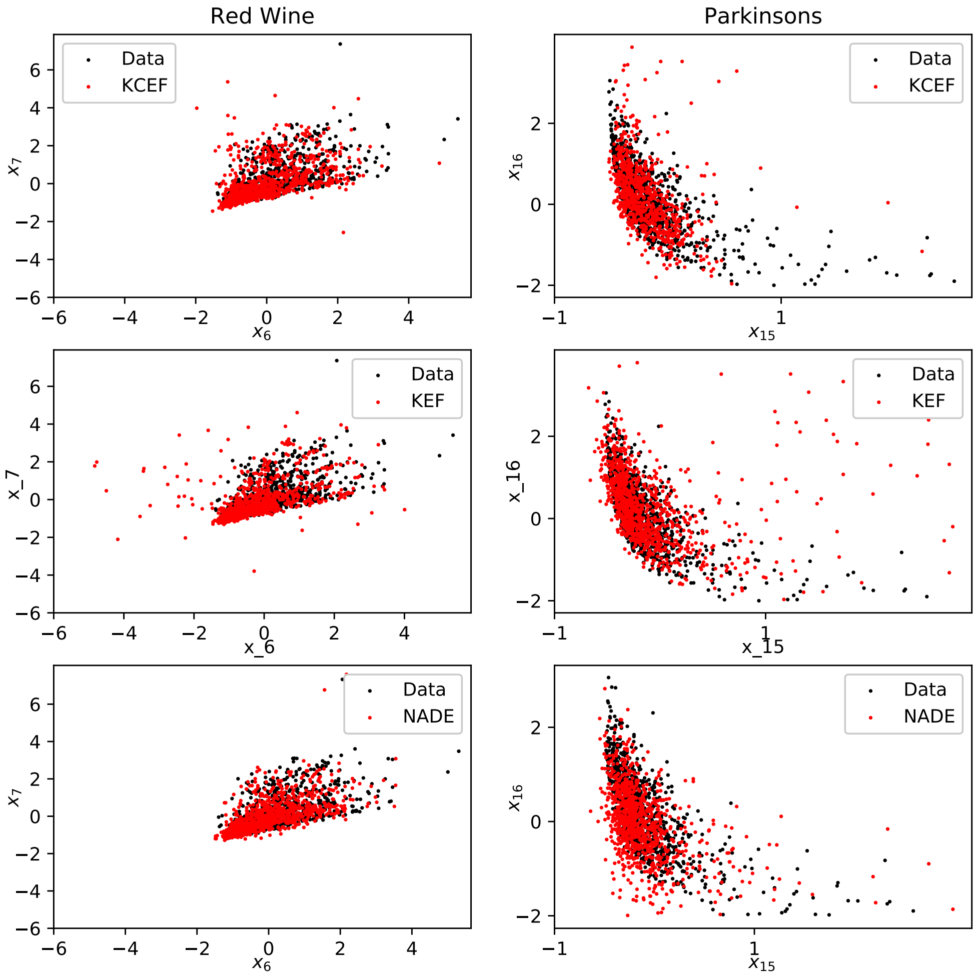

Figure 4: Scatter plot of 2-d slices of *red wine* and *parkinsons* data sets, the dimensions are (x_6, x_7) for *red wine* and (x_{15}, x_{16}) for *parkinsons*. The black points represent 1000 data points from the data sets. In red, 1000 samples from each of the three models KEF, KCEF and NADE.

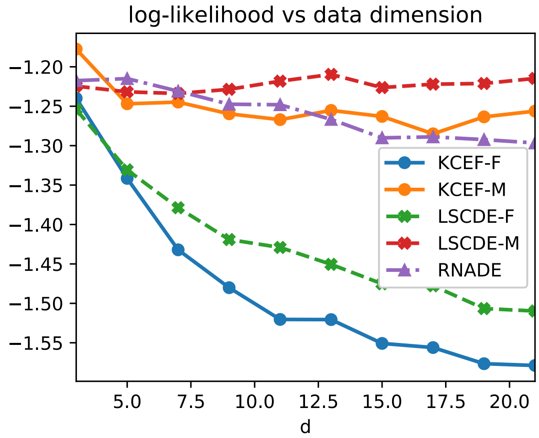

Figure 5: Experimental comparison of proposed method KCEF and other methods (LSCDE and NADE) on synthetic *grid* dataset. log-likelihood per dimension vs dimension, $N = 2000$. The log-likelihood is evaluated on a separate test set of size 2000.

Retherford, J. R. (1978). "Review: J. Diestel and J. J. Uhl, Jr., Vector measures." In: *Bull. Amer. Math. Soc.* 84.4, pp. 681–685. URL: http://projecteuclid.org/euclid.bams/1183540941.

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2017). "Density Estimation in Infinite Dimensional Exponential Families." In: *Journal of Machine Learning Research* 18.57, pp. 1–59. URL: http://jmlr.org/papers/v18/16-011.html.

Steinwart, Ingo and Andreas Christmann (2008). *Support Vector Machines*. 1st. Springer Publishing Company, Incorporated.

van der Vaart, A. W. and J. H. van Zanten (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." In: eprint: 0806.3024. URL: https://arxiv.org/abs/0806.3024.