Biometrika, *pp.* 1–23
Printed in Great Britain

Supplementary material for “Supplementary Material for “Universal Robust Regression via Maximum Mean Discrepancy””

BY P. ALQUIER
ESSEC Business School, Asia-Pacific campus, Singapore
alquier@essec.edu

M. GERBER
School of Mathematics, University of Bristol, UK
mathieu.gerber@bristol.ac.uk

1. COMPUTATION OF THE ESTIMATORS

1.1. *Gradient of the loss*

PROPOSITION S1. *Assume that each P_λ has a density p_λ with respect to a measure μ such that $\lambda \mapsto p_\lambda$ is differentiable, and that $\theta \mapsto g(\theta, x)$ is differentiable for any $x \in \mathcal{X}$.*

1. *Assume that there is a function $\hat{b} : \mathcal{Y}^2 \to \mathbb{R}$ such that*

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \hat{b}(y, y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') < \infty$$

and such that, for all (θ, x, x', y, y'),

$$\left| k((x, y), (x', y'))\nabla_\theta p_{g(\theta,x)}(y) p_{g(\theta',x')}(y') \right| \leq \hat{b}(y, y').$$

Then, for all (θ, x, x', y) we have

$$\nabla_\theta \hat{\ell}(\theta, x, x', y)$$
$$= 2\mathbb{E}_{Y \sim P_{g(\theta,x)},\, Y' \sim P_{g(\theta,x')}} \left[\Big(k\big((x, Y), (x', Y')\big) - k\big((x, Y), (x', y)\big) \Big) \nabla_\theta \log p_{g(\theta,x)}(Y) \right].$$

2. *Assume that there exists a function $\tilde{b} : \mathcal{Y}^2 \to \mathbb{R}$ such that*

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \tilde{b}(y, y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') < \infty$$

and such that, for all (θ, x, y, y'),

$$\left| k(y, y')\nabla_\theta [p_{g(\theta,x)}(y) p_{g(\theta',x)}(y')] \right| \leq \tilde{b}(y, y').$$

Then, for all (θ, x, y) we have

$$\nabla_\theta \tilde{\ell}(\theta, x, y) = 2\mathbb{E}_{Y,Y' \overset{\text{iid}}{\sim} P_{g(\theta,x)}} \left[\Big(k_{\mathcal{Y}}(Y, Y') - k_{\mathcal{Y}}(Y, y) \Big) \nabla_\theta \log p_{g(\theta,x)}(Y) \right].$$

Remark S1. We need more assumption to ensure stability and convergence of the stochastic gradient algorithm. See for example Proposition 5.2 in Chérief-Abdellatif and Alquier (2022) (and the references therein), where the authors require the existence of the variance of

$$\hat{L}(\theta, x, x'.U, U', y) := 2\Big(k\big((x,Y),(x',Y')\big) - k\big((x,Y),(x',y)\big)\Big)\nabla_\theta \log p_{g(\theta,x)}(Y)$$

when $U \sim P_{g(\theta,x)}$ and $U' \sim P_{g(\theta,x')}$. However, under Assumption **??**, it boils down to the corresponding assumption on $\nabla_\theta \log p_{g(\theta,x)}(U)$. For example, if there is $v > 0$ such that for any (x,θ), $\mathbb{E}_{U\sim P_{g(\theta,x)}}[\|\nabla_\theta \log p_{g(\theta,x)}(U)\|^2] \leq v$, then

$$\mathrm{Var}(\hat{L}(\theta, x, x'.U, U', y)) \leq 16v, \quad \forall(\theta, x, x', y).$$

Proof. We start by the proof of point 2. By definition,

$$\begin{aligned}\tilde{\ell}(\theta, X_i, Y_i) &= \mathbb{E}_{Y\sim P_{g(\theta,X_i)}, Y'\sim P_{g(\theta,X_i)}}\left[k_{\mathcal{Y}}(Y,Y') - 2k_{\mathcal{Y}}(Y,Y_i)\right] \\ &= \iint \left[k_{\mathcal{Y}}(y,y') - 2k_{\mathcal{Y}}(y,Y_i)\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &= \iint k_{\mathcal{Y}}(y,y') p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') - 2\int k_{\mathcal{Y}}(y,Y_i) p_{g(\theta,X_i)}(y)\mu(\mathrm{d}y),\end{aligned}$$

so that

$$\begin{aligned}\nabla_\theta \tilde{\ell}(\theta, X_i, Y_i) &= \nabla_\theta \iint k_{\mathcal{Y}}(y,y') p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &\quad - \nabla_\theta \int k_{\mathcal{Y}}(y,Y_i) p_{g(\theta,X_i)}(y)\mu(\mathrm{d}y) \\ &= \iint k_{\mathcal{Y}}(y,y') \nabla_\theta \left[p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\right] \mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &\quad - 2\int k_{\mathcal{Y}}(y,Y_i)\nabla_\theta \left[p_{g(\theta,X_i)}(y)\right]\mu(\mathrm{d}y) \qquad \text{(S1)}\end{aligned}$$

where the inversion of $\int$ and ∇ is jusfified thanks to the existence of the function $\tilde{b}$. Remark that

$$\nabla_\theta \left[p_{g(\theta,X_i)}(y)\right] = \nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}$$

and that

$$\begin{aligned}&\nabla_\theta \left[p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\right] \\ &= \nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y') + \nabla_\theta \left[\log p_{g(\theta,X_i)}(y')\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y').\end{aligned}$$

Plugging this into (S1) gives:

$$\begin{aligned}\nabla_\theta \tilde{\ell}(\theta, X_i, Y_i) &= \iint k_{\mathcal{Y}}(y,y')\nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &\quad + \iint k_{\mathcal{Y}}(y,y')\nabla_\theta \left[\log p_{g(\theta,X_i)}(y')\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y')\mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &\quad - 2\int k_{\mathcal{Y}}(y,Y_i)\nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}\mu(\mathrm{d}y) \\ &= 2\iint k_{\mathcal{Y}}(y,y')\nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}(y) p_{g(\theta,X_i)}(y)\mu(\mathrm{d}y)\mu(\mathrm{d}y') \\ &\quad - 2\sum_{i=1}^{n}\int k_{\mathcal{Y}}(y,Y_i)\nabla_\theta \left[\log p_{g(\theta,X_i)}(y)\right] p_{g(\theta,X_i)}\mu(\mathrm{d}y)\end{aligned}$$

by symmetry, and thus,

$$\nabla_\theta \tilde{\ell}(\theta, X_i, Y_i) = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{Y\sim P_{g(\theta,X_i)},Y'\sim P_{g(\theta,X_i)}}\Big\{\big[k_{\mathcal{Y}}(Y,Y') - k_{\mathcal{Y}}(Y,Y_i)\big]\nabla_\theta\left[\log p_{g(\theta,X_i)}(Y)\right]\Big\}.$$

The proof of point 1, from the expression in (**??**), is exactly similar. □

1.2. *A closer look at the computation of* $\hat{\theta}_n$

Let $k = k_\gamma \otimes k_{\mathcal{Y}}$ with k_γ as in Section **??** and let $L(\theta, x, x', y)$ be a random variable such that $\mathbb{E}[L(\theta, x, x', y)] = \nabla_\Theta \ell(\theta, x, x', y)$, with $\ell(\theta, x, x', y)$ as defined in Section **??**. Then, given n observations $d_n := \{(x_i, y_i)\}_{i=1}^n$ in $\mathcal{Z}$, the random variable

$$H_n\big(\gamma, \theta, d_n\big) := 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_\gamma(x_i, x_j) L(\theta, x_i, x_j, y_j)$$

is such that $\mathbb{E}[H_n\big(\gamma, \theta, d_n\big)] = \nabla_\theta h_n(\gamma, \theta, d_n)$, with $h_n(\gamma, \theta, d_n)$ as defined in (**??**).

Next, for an integer $M_1 \in \{1, \ldots, (n-1)n/2 - 1\}$ we let

$$\mathcal{S}_{M_1} \subset \mathcal{S} := \{(i,j) : 1 \leq i < j \leq n\}$$

be such that the set $\{k_\gamma(x_i, x_j)\}_{(i,j)\in\mathcal{S}_{M_1}}$ contains the M_1 largest elements of the set $\{k_\gamma(x_i, x_j)\}_{(i,j)\in\mathcal{S}}$, and for an integer $M_2 \in \mathbb{N}$ such that $M_1 + M_2 \leq (n-1)n/2$ we let $\{(I_i, J_i)\}_{i=1}^{M_2}$ be a simple random sample obtained without replacement from the set $\mathcal{S} \setminus \mathcal{S}_{M_1}$. Then, the random variable

$$\begin{aligned} H_n^{(M_1,M_2)}(\gamma, \theta, d_n) := 2 &\sum_{(i,j)\in\mathcal{S}_{M_1}} k_\gamma(x_i, x_j) L(\theta, x_i, x_j, y_j) \\ &+ \frac{(n-1)n - 2M_1}{M_2}\sum_{m=1}^{M_2} k_\gamma(x_{I_m}, x_{J_m}) L(\theta, x_{I_m}, x_{J_m}, y_{J_m}) \end{aligned}$$

is such that $\mathbb{E}[H_n^{(M_1,M_2)}(\gamma, \theta, d_n)] = h_n(\gamma, \theta, d_n)$, and thus

$$\mathbb{E}\Big[\sum_{i=1}^{N} L(\theta, x_i, y_i) + H_n^{(M_1,M_2)}(\gamma, \theta, d_n)\Big] = \nabla_\theta \sum_{i,j=1}^{n} \hat{\ell}(\theta, X_i, X_j, Y_j). \tag{S2}$$

This approach for computing an unbiased estimate of $\nabla_\theta \sum_{i,j=1}^n \hat{\ell}(\theta, X_i, X_j, Y_j)$ involves the construction of the sets $\mathcal{S}$ and $\mathcal{S}_{M_1}$, which requires $\mathcal{O}(n^2)$ operations. However, once these two sets are obtained, obtaining a realization of $G_n(\theta, d_n) := \sum_{i=1}^N L(\theta, x_i, y_i) + H_n^{(M_1,M_2)}(\gamma, \theta, d_n)$ for a given θ can be done in only $\mathcal{O}(n + M_1 + M_2\log(M_2))$ operations using e.g. the simple random sampling without replacement method proposed by Gupta and Bhattacharjee (1984).

For this procedure to work well in practice the parameters M_1 and M_2 must be such that the variance of $G_n(\theta, d_n)$ is small. When a small value for γ is chosen it is often true that $k_\gamma(x_i, x_j) \approx 0$ for most pairs $(i,j) \in \mathcal{S}$. When this happens, taking $M_1 = \mathcal{O}(n)$ and M_2 such that $M_2\log(M_2) = \mathcal{O}(n)$ allows to efficiently compute $\hat{\theta}_n$ using a stochastic gradient algorithm whose cost per iteration is linear in the sample size n. However, the memory requirement the approach we just described is $\mathcal{O}(n^2)$, which limits is applicability to moderate values of n (to n equals to a few thousands, say).

2. PROOF OF LEMMA ??

2.1. *Preliminaries*

We first recall the following result (see Da Prato and Zabczyk, 2014, Proposition 1.6):

LEMMA S1. *Let $\mathcal{A}$ and $\mathcal{B}$ be two Hilbert spaces, $T : \mathcal{A} \to \mathcal{B}$ be a bounded linear operator and Z be a random variable taking values in A and such that $\mathbb{E}\big[\|Z\|_A\big] < \infty$. Then $\mathbb{E}[T(Z)] = T(\mathbb{E}[Z])$.*

We recall that, under Assumption **??**-**??**, for any probability distribution $P \in \mathcal{P}(\mathcal{Z})$ the mean embedding $\mu(P) = \mathbb{E}_{Z\sim P}[k(Z,\cdot)]$ of P is well defined in $\mathcal{H}$, and that $\mu(P)$ has the key property to be such that

$$< f, \mu(P) >_{\mathcal{H}} = < f, \mathbb{E}_{Z\sim P}[k(Z,\cdot)] >_{\mathcal{H}} = \mathbb{E}_{Z\sim P}\big[< f, k(Z,\cdot) >_{\mathcal{H}} \big] = \mathbb{E}_{Z\sim P}[f(Z)], \quad \forall f \in \mathcal{H}$$

where the second equality holds by Lemma S2, noting that for all $f \in \mathcal{H}$ the mapping $g \mapsto < f, g >_{\mathcal{H}}$ is a bounded linear operator on $\mathcal{H}$ while, under Assumption **??**, $\mathbb{E}_{Z\sim P}[\|k(Z,\cdot)\|_{\mathcal{H}}] \leq 1$.

Recall also that the boundedness of $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ (Assumption **??**) implies that $\mathcal{C}_P$ and $\mathcal{C}_{P_X}$ exist, are unique, and that they are bounded, linear operators (see Fukumizu et al., 2004, Section 3).

We then have the following result (also proved in the proof of Corollary 3 in Fukumizu et al., 2004 as well as in Klebanov et al., 2020, Theorem 4.1).

LEMMA S2. *Assume that Assumption **??**-**??** and condition (**??**) hold. Then,* $\mathrm{range}(\mathcal{C}_P^*) \subseteq \mathrm{range}(\mathcal{C}_{P_X})$.

Proof. Let $g \in \mathcal{H}_{\mathcal{Y}}$ and $f \in \mathcal{H}_{\mathcal{X}}$. Then,

$$\begin{aligned} < \mathcal{C}_P^* g, f >_{\mathcal{H}_{\mathcal{X}}} &= < g, \mathcal{C}_P f >_{\mathcal{H}_{\mathcal{Y}}} \\ &= \mathbb{E}_{(X,Y)\sim P}\big[g(Y)f(X)\big] \\ &= \mathbb{E}_{X\sim P_X}\big[\mathbb{E}\big[g(Y)|X\big]f(X)\big] \\ &= < \mathbb{E}_{Y\sim P_{Y|\cdot}}[g(Y)], \mathcal{C}_{P_X} f >_{\mathcal{H}_{\mathcal{X}}} \\ &= < \mathcal{C}_{P_X}\mathbb{E}_{Y\sim P_{Y|\cdot}}[g(Y)], f >_{\mathcal{H}_{\mathcal{X}}} \end{aligned}$$

where the fourth equality holds under **??** and the last one uses the fact that $\mathcal{C}_{P_X}$ is self-adjoint. Since $g \in \mathcal{H}_{\mathcal{Y}}$ and $f \in \mathcal{H}_{\mathcal{X}}$ are arbitrary, it follows that

$$\mathcal{C}_P^* g = \mathcal{C}_{P_X}\mathbb{E}_{Y\sim P_{Y|\cdot}}[g(Y)], \quad \forall g \in \mathcal{H}_{\mathcal{Y}}$$

and the proof of the lemma is complete.

2.2. *Proof of the lemma*

Proof. Let $I : \mathcal{H}_{\mathcal{H}} \to \mathcal{H}_{\mathcal{H}}$ be the identity operator on $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{P}_{\mathrm{Ker}(\mathcal{C}_{P_X})} : \mathcal{H}_{\mathcal{H}} \to \mathcal{H}_{\mathcal{H}}$ be the orthogonal projection on $\mathrm{Ker}(\mathcal{C}_{P_X})$. Recall that $\mathcal{P}_{\mathrm{Ker}(\mathcal{C}_{P_X})}$ is a linear operator such that $\|\mathcal{P}_{\mathrm{Ker}(\mathcal{C}_{P_X})}\|_{\mathrm{o}} = 1$. Therefore, the linear operator $\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_{P_X} = \mathcal{I} - \mathcal{P}_{\mathrm{Ker}(\mathcal{C}_{P_X})}$ is bounded. In addition, by Lemma S2, $\mathrm{range}(\mathcal{C}_P^*) \subseteq \mathrm{range}(\mathcal{C}_{P_X})$ and therefore $\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$ is a bounded linear operator (Arias and Gonzalez, 2009, Theorem 2.3). Hence, recalling that if $A : \mathcal{H}_1 \to \mathcal{H}_2$ is a bounded linear operator between two Hilbert spaces then $\|A^*\|_{\mathcal{H}_2} = \|A\|_{\mathcal{H}_1}$, it follows that $(\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^*)^* : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is a bounded linear operator.

To proceed further let $g \in \mathcal{H}_{\mathcal{Y}}$ and $f \in \mathcal{H}_{\mathcal{X}}$. Then,

$$\begin{aligned} < f, \mathcal{C}_{P_X} \mathbb{E}_{Y \sim P_{Y|\cdot}}[g(Y)] >_{\mathcal{H}_{\mathcal{X}}} &= \mathbb{E}_{X \sim P_X}\big[f(X)\mathbb{E}[g(Y)|X]\big] \\ &= \mathbb{E}_{(X,Y) \sim P_X}\big[g(Y)f(X)\big] \\ &=< g, \mathcal{C}_P f >_{\mathcal{H}_{\mathcal{Y}}} \\ &=< \mathcal{C}_P^* g, f >_{\mathcal{H}_{\mathcal{X}}} \end{aligned} \tag{S3}$$

while, on the other hand, recalling that $f' = \mathcal{C}_{P_X}\mathcal{C}_{P_X}^{\dagger} f'$ for all $f' \in \mathrm{range}(\mathcal{C}_{P_X})$, and recalling that $\mathrm{range}(\mathcal{C}_P^*) \subseteq \mathrm{range}(\mathcal{C}_{P_X})$ by Lemma S2,

$$< f, \mathcal{C}_{P_X}\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* g >_{\mathcal{H}_{\mathcal{X}}} =< f, \mathcal{C}_P^* g >_{\mathcal{H}_{\mathcal{X}}} . \tag{S4}$$

Hence, by (S3)-(S4), it follows that

$$< f, \mathcal{C}_{P_X}\big(\mathbb{E}_{Y \sim P_{Y|\cdot}}[g(Y)] - \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* g\big) >_{\mathcal{H}_{\mathcal{X}}}= 0$$

and thus

$$\begin{aligned} &\mathbb{E}_{X \sim P_X}\Big[f(X)\big(\mathbb{E}_{Y \sim P_{Y|\cdot}}[g(Y)] - \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* g\big)(X)\Big] \\ &\qquad =< f, \mathcal{C}_{P_X}\big(\mathbb{E}_{Y \sim P_{Y|\cdot}}[g(Y)] - \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* g\big) >_{\mathcal{H}_{\mathcal{X}}} \\ &\qquad = 0. \end{aligned}$$

Consequently, since $f \in \mathcal{H}_{\mathcal{X}}$ is arbitrary, it follows that, under the assumptions of the lemma,

$$\mathbb{E}_{Y \sim P_{Y|\cdot}}[g(Y)] = \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* g \tag{S5}$$

Remark now that for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we have

$$\begin{aligned} \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* k_{\mathcal{Y}}(y,\cdot)(x) &=< \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* k_{\mathcal{Y}}(y,\cdot), k_{\mathcal{X}}(x,\cdot) >_{\mathcal{H}_{\mathcal{X}}} \\ &=< k_{\mathcal{Y}}(y,\cdot), (\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^*)^* k_{\mathcal{X}}(X,\cdot) >_{\mathcal{H}_{\mathcal{Y}}} \\ &= \big(\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^*\big)^* k_{\mathcal{X}}(x,\cdot)(y) \end{aligned} \tag{S6}$$

where the first equality uses the reproducing property of $k_{\mathcal{X}}$ and the third equality the reproducing property of $k_{\mathcal{Y}}$.

Let $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Then, using (S5) with $g = k_{\mathcal{Y}}(y,\cdot)$ and (S6), we have

$$\begin{aligned} \mu(P_{Y|x})(y) &= \mathbb{E}_{Y \sim P_{Y|x}}\big[k_{\mathcal{Y}}(y,Y)\big] \\ &= \mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^* k_{\mathcal{Y}}(y,\cdot)(x) \\ &= \big(\mathcal{C}_{P_X}^{\dagger}\mathcal{C}_P^*\big)^* k_{\mathcal{X}}(x,\cdot)(y) \end{aligned}$$

and the proof is complete. □

3. PROOF OF THEOREM ??

3.1. *A preliminary result for proving Theorem* ??

LEMMA S3. *Assume that $|k_{\mathcal{X}}| \leq 1$ and let $\mu(\mathrm{d}y)$ be a σ-finite measure on $(\mathcal{Y}, \mathfrak{S}_{\mathcal{Y}})$ and $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be such that*

1. *$f(\cdot, y) \in \mathcal{H}_{\mathcal{X}}$ for all $y \in \mathcal{Y}$,*
2. *The function $\mathcal{Y} \ni y \mapsto f(\cdot, y)$ is Borel measurable,*

3. *The set* $\{f(\cdot,y):\ y\in\mathcal{Y}\}$ *is separable,*
4. $\int_{\mathcal{Y}}\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mu(\mathrm{d}y)<\infty$.

Then, $\int_{\mathcal{Y}} f(\cdot,y)\mu(\mathrm{d}y)\in\mathcal{H}_{\mathcal{X}}$.

Proof. Since the set $\{f(\cdot,y):\ y\in\mathcal{Y}\}$ is separable and the mapping $y\mapsto f(\cdot,y)$ is Borel measurable the function $y\mapsto f(\cdot,y)$ is strongly measurable. Therefore, there exist (Cohn, 2013, Proposition E.2) a sequence $\big(\{E_{i,n}\}_{i=1}^n\big)_{n\geq 1}$ and a sequence $\big(\{f_{i,n}\}_{i=1}^n\big)_{n\geq 1}$ such that

1. $E_{i,n}\in\mathfrak{S}_{\mathcal{Y}}$ and $f_{i,n}\in\mathcal{H}_{\mathcal{X}}$ for all $n\geq i\geq 1$,
2. $\lim_{n\to 0}\|\sum_{i=1}^n\mathbb{1}_{E_{i,n}}(y)f_{i,n}-f(y,\cdot)\|_{\mathcal{H}_{\mathcal{X}}}=0$ for all $y\in\mathcal{Y}$,
3. $\|\sum_{i=1}^n\mathbb{1}_{E_{i,n}}(y)f_{i,n}\|_{\mathcal{H}_{\mathcal{X}}}\leq\|f(y,\cdot)\|_{\mathcal{H}_{\mathcal{X}}}$ for all $n\geq 1$ and all $y\in\mathcal{Y}$.

For every $n\geq 1$ let $f_n:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be defined by

$$f_n(x,y)=\sum_{i=1}^n\mathbb{1}_{E_{i,n}}(y)f_{i,n}(x),\quad (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Under the assumptions of the lemma we have $\int_{\mathcal{Y}}\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mu(\mathrm{d}y)<\infty$, and thus,

$$\int_{\mathcal{Y}}\|f_n(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mu(\mathrm{d}y)\leq\int_{\mathcal{Y}}\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mu(\mathrm{d}y)<\infty,\quad\forall n\geq 1,$$

showing that, for all $n\geq 1$, the simple function $y\mapsto f_n(\cdot,y)$ is Bochner integrable. Consequently, for all $n\geq 1$ the function

$$\tilde{f}_n:=\int_{\mathcal{Y}}f_n(\cdot,y)\mu(\mathrm{d}y)=\sum_{i=1}^n\Big(\int_{E_{i,n}}\mu(\mathrm{d}y)\Big)f_{i,n}$$

is well-defined. Notice that $\tilde{f}_n\in\mathcal{H}_{\mathcal{X}}$ for all $n\geq 1$.

To proceed further remark that

$$|f_n(x,y)|\leq\|f_n(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\leq\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}},\quad\forall(x,y)\in\mathcal{X}\times\mathcal{Y}$$

where the first inequality holds since $|k_{\mathcal{X}}|\leq 1$ by assumption while the second inequality hods by the third aforementioned properties of $\big(\{E_{i,n}\}_{i=1}^n\big)_{n\geq 1}$ and $\big(\{f_{i,n}\}_{i=1}^n\big)_{n\geq 1}$.

By assumption, $\int_{\mathcal{Y}}\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mathrm{d}y<\infty$ and thus, by the dominated converge theorem, and using the fact that the convergence in $\|\cdot\|_{\mathcal{H}_{\mathcal{X}}}$ norm implies the point-wise convergence,

$$\lim_{n\to\infty}\tilde{f}_n(s)=\int_{\mathcal{Y}}f(s,y)\mathrm{d}y,\quad\forall s\in\mathcal{X}.\tag{S7}$$

Therefore, recalling that $\tilde{f}_n\in\mathcal{H}_{\mathcal{X}}$ for all $n\geq 1$, to complete the proof it remains to show that the sequence $(\tilde{f}_n)_{n\geq 1}$ is Cauchy w.r.t. the $\|\cdot\|_{\mathcal{H}_{\mathcal{X}}}$ norm.

To this aim remark that, since

$$\big\|f_n(\cdot,y)-f(\cdot,y)\big\|_{\mathcal{H}_{\mathcal{X}}}\leq 2\big\|f(\cdot,y)\big\|_{\mathcal{H}_{\mathcal{X}}},\quad\forall n\geq 1$$

while, by assumption, $\int_{\mathcal{Y}}\|f(\cdot,y)\|_{\mathcal{H}_{\mathcal{X}}}\mathrm{d}y<\infty$, the dominated convergence theorem implies that

$$\lim_{n\to\infty}\int_{\mathcal{Y}}\big\|f_n(\cdot,y)-f(\cdot,y)\big\|_{\mathcal{H}_{\mathcal{X}}}\mathrm{d}y=0.\tag{S8}$$

On the other hand, for every $n > m \geq 1$ we have

$$\begin{aligned} \|\tilde{f}_n - \tilde{f}_m\|_{\mathcal{H}_\mathcal{X}} &= \Big\| \int_\mathcal{Y} \{f_n(\cdot, y) - f_m(\cdot, y)\}\mu(\mathrm{d}y) \Big\|_{\mathcal{H}_\mathcal{X}} \\ &\leq \int_\mathcal{Y} \big\|f_n(\cdot, y) - f_m(\cdot, y)\big\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) \qquad \text{(S9)} \\ &\leq \int_\mathcal{Y} \big\|f_n(\cdot, y) - f(\cdot, y)\big\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) + \int_\mathcal{Y} \big\|f_m(\cdot, y) - f(\cdot, y)\big\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) \end{aligned}$$

where the first inequality holds by Cohn (2013, Proposition E.5), since a shown above the function $y \mapsto f_n(\cdot, y)$ is Bochner integrable. Together, (S8) and (S9) show that the sequence $(\tilde{f}_n)_{n\geq 1}$ is indeed Cauchy w.r.t. the $\|\cdot\|_{\mathcal{H}_\mathcal{X}}$ norm, and the proof of the lemma is complete. □

3.2. *Proof of Theorem* ??

Proof. Let $g \in \mathcal{H}_\mathcal{Y}$ so that $g = \sum_{i=1}^{\infty} a_i k_\mathcal{Y}(y_i, \cdot)$ for a sequence $(y_i)_{i\geq 1}$ in $\mathcal{Y}$ and a sequence $(a_i)_{i\geq 1}$ in $\mathbb{R}$. For all $n \geq 1$ let $g_n = \sum_{i=1}^{n} a_i k_\mathcal{Y}(y_i, \cdot)$ and $f_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be defined by $f_n(x, y) = g_n(y)p(y|x)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We first show that, for all $n \geq 1$, the function f_n verifies the assumptions of Lemma S3.

By Conditions ?? and ?? of the theorem, it readily follows that f_n verifies Conditions 1 and 3 of Lemma S3, for all $n \geq 1$. To show that this is also the case for Condition 2 of Lemma S3 let $\mathcal{B}\big(\mathcal{H}_\mathcal{X}\big)$ be the Borel σ-algebra on $\mathcal{H}_\mathcal{X}$. Let $n \geq 1$ and assume first that $\mathcal{H}_\mathcal{X}$ contains the non-zero constant functions so that the function $y \mapsto g_n(y)$ is $\mathcal{B}(\mathcal{H}_\mathcal{X})$-measurable. Then, since by assumption the function $y \mapsto p(y|\cdot)$ is $\mathcal{B}(\mathcal{H}_\mathcal{X})$-measurable and since the product of two Borel measurable functions is a Borel measurable function, it follows that the function $\mathcal{Y} \ni y \mapsto f_n(\cdot, y)$ is $\mathcal{B}(\mathcal{H}_\mathcal{X})$-measurable, as required. Assume now that $\mathcal{H}_\mathcal{X}$ does not contain the non-zero constant functions. Let $\tilde{H}_\mathcal{X}$ be the RKHS on $\mathcal{X}$ having $k_\mathcal{X} + 1$ as reproducing kernel so that, as shown above, the function $\mathcal{Y} \ni y \mapsto f_n(\cdot, y)$ is $\mathcal{B}\big(\tilde{\mathcal{H}}_\mathcal{X}\big)$-measurable. Consequently,

$$\big\{y \in \mathcal{Y} : \, f_n(\cdot|y) \in A\big\} \in \mathfrak{S}_\mathcal{Y}, \quad \forall A \in \mathcal{B}\big(\tilde{\mathcal{H}}_\mathcal{X}\big). \qquad \text{(S10)}$$

Recalling that $\tilde{\mathcal{H}}_\mathcal{X} = \big\{f + c, \, f \in \mathcal{H}_\mathcal{X}, \, c \in \mathbb{R}\big\}$ and that $\|f\|_{\tilde{\mathcal{H}}_\mathcal{X}} = \|f\|_{\mathcal{H}_\mathcal{X}}$ for all $f \in \mathcal{H}_\mathcal{X}$ (Paulsen and Raghupathi, 2016, Theorem 5.1), it follows that $\mathcal{B}(\mathcal{H}_\mathcal{X}) \subset \mathcal{B}(\tilde{\mathcal{H}}_\mathcal{X})$ which, together with (S10), implies that

$$\big\{y \in \mathcal{Y} : \, f_n(\cdot|y) \in A\big\} \in \mathfrak{S}_\mathcal{Y}, \quad \forall A \in \mathcal{B}\big(\mathcal{H}_\mathcal{X}\big).$$

This shows that the function $\mathcal{Y} \ni y \mapsto f_n(\cdot, y)$ is $\mathcal{B}(\mathcal{H}_\mathcal{X})$-measurable, and thus, for all $n \geq 1$, f_n satisfies Condition 2 of Lemma S3.

Lastly, using the fact that $|k_\mathcal{Y}| \leq 1$ and Condition ?? of the theorem, for all $n \geq 1$ we have

$$\begin{aligned} \int_\mathcal{Y} \|f_n(\cdot, y)\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) &\leq \big(\sup_{y\in\mathcal{Y}} |g_n(y)|\big) \int_\mathcal{Y} \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) \\ &\leq \|g_n\|_{\mathcal{H}_\mathcal{Y}} \int_\mathcal{Y} \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) \\ &< \infty \end{aligned}$$

and thus, for all $n \geq 1$, f_n verifies Condition 4 of Lemma S3, which concludes to show that, for all $n \geq 1$, f_n verifies all the assumptions of Lemma S3.

Therefore, by Lemma S3, the function $\tilde{f}_n := \int_\mathcal{Y} f_n(\cdot, y)\mu(\mathrm{d}y)$ exists and belongs to $\mathcal{H}_\mathcal{X}$, for all $n \geq 1$. In addition, for all $m > n \geq 1$ we have (see Cohn, 2013, Proposition E.5, for the first inequality)

$$\begin{aligned}\|\tilde{f}_n - \tilde{f}_m\|_{\mathcal{H}_\mathcal{X}} &= \Big\| \int_\mathcal{Y} (g_n - g_m)(y) p(y|\cdot)\mu(\mathrm{d}y) \Big\|_{\mathcal{H}_\mathcal{X}} \\ &\leq \int_\mathcal{Y} |g_n(y) - g_m(y)|\, \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y) \\ &\leq \sup_{y\in\mathcal{Y}} |g_n(y) - g_m(y)| \int_\mathcal{Y} \big\|p(y|\cdot)\big\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y)\end{aligned}$$

where, since $|k_\mathcal{Y}| \leq 1$ by assumption,

$$\limsup_{n\to\infty} \sup_{m>n} \sup_{y\in\mathcal{Y}} |g_n(y) - g_m(y)| \leq \limsup_{n\to\infty} \sup_{m>n} \|g_n - g_m\|_{\mathcal{H}_\mathcal{Y}} = 0. \tag{S11}$$

Consequently, the sequence $(\tilde{f}_n)_{n\geq 1}$ is Cauchy w.r.t. the $\|\cdot\|_{\mathcal{H}_\mathcal{X}}$ norm and therefore converges point-wise to a function $\tilde{f} \in \mathcal{H}_\mathcal{X}$. Thus, to complete the proof it remains to show that

$$\lim_{n\to\infty} \tilde{f}_n(x) = \mathbb{E}_{Y\sim P_{Y|X=x}}[g(Y)], \quad \forall x \in \mathcal{X}.$$

Since for every $n \geq 1$ and $x \in \mathcal{X}$ we have

$$\big|\tilde{f}_n(x) - \mathbb{E}_{Y\sim P_{Y|X=x}}[g(Y)]\big| \leq \int_\mathcal{Y} |g_n(y) - g(y)|\, p(y|x)\mu(\mathrm{d}y) \leq \sup_{y\in\mathcal{Y}} |g_n(y) - g(y)|,$$

it follows, by (S11), that $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} |\tilde{f}_n(x) - \mathbb{E}_{Y\sim P_{Y|X=x}}[g(Y)]| = 0$, and the proof of the theorem is complete. □

4. PROOF OF COROLLARY ??

Corollary **??** is a direct consequence of Lemma **??**, Theorem **??** and of the following lemma:

LEMMA S4. *Assume that Assumptions* **??**-**??** *hold and that there exists a σ-finite measure $\mu(\mathrm{d}y)$ on $(\mathcal{Y}, \mathfrak{S}_\mathcal{Y})$ such that $P_{Y|x} = p(y|x)\mu(\mathrm{d}y)$ for all $x \in \mathcal{X}$, where $p(\cdot|\cdot)$ satisfies Assumptions* **??**-**??** *of Theorem* **??**. *Moreover, assume that there exists a bounded conditional mean embedding operator $\mathcal{C}_{Y|X}$ for $(P_{Y|x})_{x\in\mathcal{X}}$. Then, $\|\mathcal{C}_{Y|X}\|_\mathrm{o} \leq \int_\mathcal{Y} \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mu(\mathrm{d}y)$.*

Proof. Let $g \in \mathcal{H}_\mathcal{Y}$ and remark that

$$(\mathcal{C}^*_{Y|X} g)(x) =< \mathcal{C}^*_{Y|X} g, k_\mathcal{X}(x,\cdot) >_{\mathcal{H}_\mathcal{X}} =< g, \mathcal{C}_{Y|X} k_\mathcal{X}(x,\cdot) >_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_{Y\sim P_{Y|X=x}}[g(Y)], \quad \forall x \in \mathcal{X}$$

where the first equality uses the reproducing property of $k_\mathcal{X}$ and the third (S19).

Consequently,

$$\begin{aligned}\|\mathcal{C}^*_{Y|X} g\|_{\mathcal{H}_\mathcal{X}} &= \Big\| \int_\mathcal{Y} g(y) p(y|\cdot)\mathrm{d}y \Big\|_{\mathcal{H}_\mathcal{X}} \\ &\leq \int_\mathcal{Y} \|g(y)p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mathrm{d}y \\ &\leq \sup_{y\in\mathcal{Y}} |g(y)| \int_\mathcal{Y} \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mathrm{d}y \\ &\leq \|g\|_{\mathcal{H}_\mathcal{Y}} \int_\mathcal{Y} \|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}} \mathrm{d}y\end{aligned}$$

where, under the assumptions of the lemma, the first inequality holds by Cohn (2013, Proposition E.5) and where the last inequality uses the fact that $|k_{\mathcal{Y}}| \leq 1$.

Therefore,

$$\|\mathcal{C}_{Y|X}\|_{\mathrm{o}} = \|\mathcal{C}^*_{Y|X}\|_{\mathrm{o}} \leq \int_{\mathcal{Y}} \|p(y|\cdot)\|_{\mathcal{H}_{\mathcal{X}}} \mathrm{d}y$$

and the proof of the lemma is complete. □

5. A USEFUL COROLLARY OF THEOREM **??**

In order to state the next result we let $\Lambda_d(\mathrm{d}x)$ denote the Lebesgue measure on $\mathbb{R}^d$, $A_s = \big\{\tilde{a} \in \mathbb{N}_0^d : \sum_{i=1}^d \tilde{a}_i \leq s\big\}$ for all $s \in \mathbb{N}_0$ and $|a| = \sum_{i=1}^d a_i$ for all $a \in \mathbb{R}^d$.

COROLLARY S1. *Assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is bounded with Lipschitz boundary and, for some constants $m \in \mathbb{N}$ and $\gamma > 0$, let $k_{\mathcal{X}}$ be the restriction of the Matérn kernel $K_{\frac{m}{2},\gamma}$ on $\mathcal{X} \times \mathcal{X}$. Let $s = (d+m)/2$ if $(d+m)$ is even and $s = (d+m+1)/2$ if $(d+m)$ is odd, and assume that there exists a σ-finite measure $\mu(\mathrm{d}y)$ on $(\mathcal{Y}, \mathfrak{S}_{\mathcal{Y}})$ such that $P_{Y|x} = p(y|x)\mu(\mathrm{d}y)$ for all $x \in \mathcal{X}$, where $p(\cdot|\cdot)$ satisfies the following conditions:*

- *for all $y \in \mathcal{Y}$, the function $p(y|\cdot)$ is s times continuously differentiable on $\mathcal{X}$, with*

$$\max_{a \in A_s} \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \Big| \frac{\partial^{\sum_{i=1}^d a_i}}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} p(y|x) \Big| < \infty$$

  and with

$$\max_{a \in A_s} \int_{\mathcal{Y}} \bigg[\int_{\mathcal{X}} \Big\{ \frac{\partial^{\sum_{i=1}^d a_i}}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} p(y|x) \Big\}^2 \Lambda_d(\mathrm{d}x) \bigg]^{\frac{1}{2}} \mu(\mathrm{d}y) < \infty, \quad \forall a \in A_s,$$

- *the function $y \mapsto \frac{\partial^{\sum_{i=1}^d a_i}}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} p(y|x)$ is continuous on $\mathcal{Y}$, for all $x \in \mathcal{X}$ and $a \in A_s$.*

Assume also that the set $\mathcal{Y}$ is separable and that Assumptions **??**-**??** *hold. Then, conditions* **??**-**??** *of Theorem* **??** *hold and thus* (**??**) *is satisfied.*

Proof. Remark first that to prove the result it is enough to consider the case where $(m+d)$ is even. Indeed, if $(m+d)$ is odd then in what follows we can replace

- the set $\mathcal{X}$ by $\tilde{\mathcal{X}} = \mathcal{X} \times \mathbb{R}^1$,
- for all $y \in \mathcal{Y}$, the function $p(y|\cdot) : \mathcal{X} \to \mathbb{R}$ by the function $\tilde{p}(y|\cdot) : \tilde{\mathcal{X}} \to \mathbb{R}$ defined by $\tilde{p}(y|(x,v')) = p(y|x)$ for all $(x,u) \in \tilde{\mathcal{X}}$,
- d by $\tilde{d} = d+1$.

Recall that, since $(m+d)$ is even, the RKHS $\mathcal{H}_{\mathcal{X}}$ is norm-equivalent to the Sobolev space $W_2^s(\mathcal{X})$ (see e.g. Kanagawa et al., 2018, Example 2.6). In addition, recall that the norm $\|\cdot\|_{W_2^s(\mathcal{X})}$ is defined by

$$\|f\|_{W_2^s(\mathcal{X})} = \sum_{a \in A_s} \bigg(\int_{\mathcal{X}} \Big| \frac{\partial^{\sum_{i=1}^d a_i}}{\partial u_1^{a_1} \dots \partial u_d^{a_d}} f(x) \Big|^2 \Lambda_d(\mathrm{d}x) \bigg)^{\frac{1}{2}}, \quad f \in W_2^s(X)$$

and let

$$D_a p(y|x) = \frac{\partial^{\sum_{i=1}^d a_i}}{\partial x_1^{a_1}\dots\partial x_d^{a_d}} p(y|x), \quad \forall (a,x,y)\in A_s\times\mathcal{X}\times\mathcal{Y}.$$

To proof the corollary remark first that, under its assumptions, for all $y\in\mathcal{Y}$ we have $\|p(\cdot|y)\|_{W_2^s(\mathcal{X})}<\infty$. Thus, for all $y\in\mathcal{Y}$, the function $p(y|\cdot)$ belongs to the Sobolev space $W_2^s(\mathcal{X})$, and thus to the RKHS $\mathcal{H}_\mathcal{X}$. This shows that $p(\cdot|\cdot)$ verifies Condition **??** of Theorem **??**.

In addition, under the assumptions of the corollary we have

$$\int_\mathcal{Y}\|p(y|\cdot)\|_{W_2^s(\mathcal{X})}\mu(\mathrm{d}y) = \sum_{a\in A_s}\int_\mathcal{Y}\left\{\int_\mathcal{X} D_a p(y|x)^2\Lambda_d(\mathrm{d}x)\right\}^{\frac{1}{2}}\mu(\mathrm{d}y)<\infty \tag{S12}$$

and thus, since the norm $\|\cdot\|_{\mathcal{H}_\mathcal{X}}$ is equivalent to the norm $\|\cdot\|_{W_2^s(\mathcal{X})}$, it follows that

$$\int_\mathcal{Y}\|p(y|\cdot)\|_{\mathcal{H}_\mathcal{X}}\mu(\mathrm{d}y)<\infty$$

showing that $p(\cdot|\cdot)$ verifies Condition **??** of Theorem **??**.

To proceed further recall that the image of a separable space by a continuous function is separable. Hence, since $\mathcal{Y}$ is assumed to be separable, to show that Condition **??** of Theorem **??** holds it suffices to show that, for every $y'\in\mathcal{Y}$, the function

$$\mathcal{Y}\ni y\mapsto k_\mathcal{Y}(y',y)p(y|\cdot)\in\mathcal{H}_\mathcal{X} \tag{S13}$$

is continuous. To this aim, let $y'\in\mathcal{Y}$ and $(y_i')_{i\geq 1}$ be a sequence in $\mathcal{Y}$ such that $\lim_{i\to\infty} y_i' = y'$. Then, since $k_\mathcal{Y}$ is continuous by assumption, to show that the function defined in (S13) is continuous it is enough to show that

$$\limsup_{i\to\infty}\|p(y_i'|\cdot)-p(y'|\cdot)\|_{\mathcal{H}_\mathcal{X}}=0. \tag{S14}$$

The norm $\|\cdot\|_{\mathcal{H}_\mathcal{X}}$ being norm-equivalent to the norm $\|\cdot\|_{W_2^s(\mathcal{X})}$, there exists a constant $C<\infty$ such that $\|f\|_{\mathcal{H}_\mathcal{X}}\leq C\|f\|_{W_2^s(\mathcal{X})}$ for all $f\in\mathcal{H}_\mathcal{X}$ and thus, for all $i\geq 1$, we have

$$\begin{aligned}\|p(y_i'|\cdot)-p(y'|\cdot)\|^2_{\mathcal{H}_\mathcal{X}} &\leq C^2\|p(y_i'|\cdot)-p(y'|\cdot)\|^2_{W_2^s(\mathcal{X})}\\ &\leq C^2\sum_{a\in A_s}\int_\mathcal{X}\left|D_a p(y_i'|x)-D_a p(y'|x)\right|^2\Lambda_d(\mathrm{d}x).\end{aligned} \tag{S15}$$

By assumption, for all $x\in\mathcal{X}$ and all $a\in A_s$, the function $y\mapsto D_a p(y|x)$ is continuous on $\mathcal{Y}$ while, for all $(a,x)\in A_s\times\mathcal{X}$ we have

$$\sup_{i\geq 1}\left|D_a p(y_i'|x)-D_a p(y'|x)\right|^2\leq 2\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|D^a p(y|x)|<\infty.$$

Consequently, since $\mathcal{X}$ is bounded, (S14) follows from (S15) and the dominated convergence theorem, and thus $p(\cdot|\cdot)$ satisfies Condition **??** of Theorem **??**.

Finally, since as shown above the mapping $\mathcal{Y}\ni y\mapsto p(y|\cdot)$ is continuous, it follows that this mapping is Borel measurable and thus $p(\cdot|\cdot)$ satisfies Condition **??** of Theorem **??**. Hence, all the conditions of Theorem **??** and the proof is complete. □

6. Proof of Proposition ??

6.1. *Preliminary result*

LEMMA S5. *Assume that $\mathcal{X}\subseteq\mathbb{R}^d$ for some integer d and that $\mathcal{X}$ is path-wise connected and such that $\Lambda_d(\mathcal{X})>0$. Assume also that $k_{\mathcal{X}}$ is continuous on $\mathcal{X}^2$. Then, there exists a distribution $P_X\in\mathcal{P}(\mathcal{X})$ such that*

$$\big\{f\in\mathcal{H}_{\mathcal{X}}:\,\mathbb{E}_{X\sim P_X}\big[f(X)h(X)\big]=0,\,\forall h\in\mathcal{X}\big\}=\{0\}. \tag{S16}$$

Proof. Remark first that since $k_{\mathcal{X}}$ is continuous on $\mathcal{X}^2$ any function $f\in\mathcal{H}_{\mathcal{X}}$ is continuous on $\mathcal{X}$ (Paulsen and Raghupathi, 2016, Theorem 2.17). Let P_X denote the $\mathcal{N}_d(0,I_d)$ distribution, truncated on $\mathcal{X}$ if $\mathcal{X}\neq\mathbb{R}^d$. Assume that there exists a non-zero function $f\in\mathcal{H}_{\mathcal{X}}$ such that

$$\mathbb{E}_{X\sim P_X}\big[f(X)h(X)\big]=0,\quad\forall h\in\mathcal{H}_{\mathcal{X}}.$$

Then, $\mathbb{E}_{X\sim P_X}[f(X)^2]=0$ and, since P_X admits a strictly positive density p_X on $\mathcal{X}$ w.r.t. Λ_d, we have $f(x)=0$ for Λ_d-almost every $x\in\mathcal{X}$. However, as f is assumed to be continuous, and $\mathcal{X}$ is path-wise connected, the function f is zero everywhere. □

6.2. *Proof of the proposition*

Proof:

The fact that k is characteristic follows from Szabó and Sriperumbudur (2018) and the properties of the Matérn kernel.

Next, remark that, by Lemma S5, under the assumption of the proposition there exists a distribution $P_X\in\mathcal{P}(\mathcal{X})$ such that the only function $f\in\mathcal{H}_{\mathcal{X}}$ for which we have $\mathbb{E}_{X\sim P_X}[f(X)^2]=0$ is the zero function. In addition, since $\mathcal{X}$ is bounded with Lipschitz boundary we can use Corollary S1 to check that there exists a bounded linear conditional mean operator for $(P_{Y|x})_{x\in\mathcal{X}}$.

To this aim, for all $x\in\mathcal{X}$ we let $p(y|x)$ be the density of $P_{Y|x}(\mathrm{d}y)$ w.r.t. $\mu(\mathrm{d}y)$. The σ-finite measure $\mu(\mathrm{d}y)$ on $\mathcal{Y}$ will be specified below for each example but, for all the considered examples, for all $y\in\Theta\times\mathcal{Y}$ the mapping $x\mapsto p(y|\cdot)$ is infinitely many times differentiable. Consequently, letting s and A_s be as defined in Corollary S1, we can define

$$D^a p(y|x)=\frac{\partial^{\sum_{i=1}^d a_i}}{\partial x_1^{a_1}\dots\partial x_d^{a_d}}p(y|x),\quad\forall(a,x,y)\in\Theta\times A_s\times\mathcal{X}\times\mathcal{Y}.$$

Then, by Corollary S1, a bounded linear conditional mean operator for $(P_{Y|x})_{x\in\mathcal{X}}$ exists if

1. the mapping $y\mapsto D^a p(y|x)$ is continuous for all $(a,x)\in A_s\times\mathcal{X}$,
2. the following two conditions hold:

$$\max_{a\in A_s}\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|D^a p(y|x)|<\infty \tag{S17}$$

$$\max_{a\in A_s}\int_{\mathcal{Y}}\bigg[\int_{\mathcal{X}}\{D^a p(y|x)\}^2\Lambda_d(\mathrm{d}x)\bigg]^{\frac{1}{2}}\mu(\mathrm{d}y)<\infty. \tag{S18}$$

For all the examples considered in the proposition it is trivial to see that the mapping $y\mapsto D^a p(y|x)$ is continuous for all $(a,x)\in A_s\times\mathcal{X}$. Under the assumptions made on $\mathcal{X}$, Conditions (S17) and (S18) are easily checked from the definition of $p(y|x)$ given below for each examples

Example 1: For this example $\mathcal{Y}=\mathbb{R}$ and we let $\mu(\mathrm{d}y)$ be the Lebesgue measure on $\mathbb{R}$ so that

$$p(y|x)=\sum_{m=1}^M w_m\frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\Big\{-\frac{(y-\beta_m^\top x)^2}{2\sigma_m^2}\Big\},\quad\forall(x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Example 2: For this example, $\mathcal{Y}=\mathbb{N}_0$ and we let $\mu(\mathrm{d}y)$ be the counting measure on $\mathbb{N}_0$ so that

$$p(y|x)=\frac{\exp\big\{y\,\beta^\top x-\exp(\beta^\top x)\big\}}{y!},\quad \forall (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Example 3: For this example, $\mathcal{Y}=\{0,1\}$ and we let $\mu(\mathrm{d}y)$ be the counting measure on $\{0,1\}$ so that

$$p(y|x)=\Big(\frac{1}{1+\exp(-\beta^\top x)}\Big)^{y}\Big(\frac{1}{1+\exp(\beta^\top x)}\Big)^{1-y},\quad \forall (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Example 4: For this example, $\mathcal{Y}=(0,\infty)$ and we let $\mu(\mathrm{d}y)$ be the Lebesgue measure on $\mathbb{R}$ so that

$$p(y|x)=\frac{1}{\Gamma(\nu)}y^{\nu-1}\exp(-\nu\beta^\top x)\exp\big\{-\nu y\exp(-\beta^\top x)\big\},\quad \forall (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Example 5: For this example, $\mathcal{Y}=\mathbb{R}\times\{0,1\}$ and we let

$$\mu(\mathrm{d}y)=\big(\Lambda_1(\mathrm{d}y_1)+\delta_{\{0\}}(\mathrm{d}y_1))\otimes\delta_{\{0\}}(\mathrm{d}y_2)$$

so that

$$P_\lambda(\mathrm{d}y)=\check{p}_\lambda(y)\mu(\mathrm{d}y)$$

where, denoting by $\phi(\cdot;\mu,\sigma^2)$ the probability density function of the $\mathcal{N}_1(\mu,\sigma^2)$ distribution w.r.t. Λ_1, for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ we have

$$\begin{aligned}p(y|x)&=\phi\big(y_1;\beta^\top x,\sigma^2\big)\Phi\Big(\{\gamma^\top x+(\rho/\sigma)\beta^\top x\}/\sqrt{1-\rho^2}\Big)\mathbb{1}_{\mathbb{R}\setminus\{0\}}(y_1)\big(1-\mathbb{1}_{\{0\}}(y_2)\big)\\&\quad+\Phi(-\gamma^\top x)\mathbb{1}_{\{0\}}(y_1)\mathbb{1}_{\{0\}}(y_2).\end{aligned}$$

□

7. PROOF OF LEMMA ??

Proof. Let $\mathcal{C}_{Y|X}:\mathcal{H}_\mathcal{X}\to\mathcal{H}_\mathcal{Y}$ be a bounded linear operator such that

$$\mu(P_{Y|x})=\mathcal{C}_{Y|X}k_\mathcal{X}(x,\cdot),\quad \forall x\in\mathcal{X}. \tag{S19}$$

and let $\tilde{\mathcal{C}}_{P_{Y|X}}:\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}\to\mathcal{H}$ be the (unique) linear operator on $\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{X}$ such that

$$\tilde{\mathcal{C}}_{P_{Y|X}}(f_1\otimes f_2)=f_1\otimes\mathcal{C}_{Y|X}f_2,\quad f_1\in\mathcal{H}_\mathcal{X},f_2\in\mathcal{H}_\mathcal{X}.$$

For all $f_1\in\mathcal{H}_\mathcal{X}$ and $f_2\in\mathcal{H}_{\star,\mathcal{X}}$ we have

$$\begin{aligned}\|\tilde{\mathcal{C}}_{P_{Y|X}}(f_1\otimes f_2)\|_\mathcal{H}&=\|f_1\otimes\mathcal{C}_{Y|X}f_2\|_\mathcal{H}\\&=\|f_1\|_{\mathcal{H}_\mathcal{X}}\|\mathcal{C}_{Y|X}f_2\|_{\mathcal{H}_\mathcal{Y}}\\&\leq\|f_1\|_{\mathcal{H}_\mathcal{X}}\|f_2\|_{\mathcal{H}_\mathcal{X}}\|\mathcal{C}_{Y|X}\|_\mathrm{o}\\&=\|f_1\otimes f_2\|_\mathcal{H}\|\mathcal{C}_{Y|X}\|_\mathrm{o}\end{aligned}$$

showing that

$$\|\tilde{\mathcal{C}}_{P_{Y|X}}\|_\mathrm{o}\leq\|\mathcal{C}_{Y|X}\|_\mathrm{o}<\infty. \tag{S20}$$

where the last inequality holds by assumption.

Next, remark that for every $f \in \mathcal{H}_\mathcal{X}$ the linear operator $(f \otimes \cdot) : \mathcal{H}_\mathcal{Y} \to \mathcal{H}$ is such that

$$\|f \otimes \cdot\|_o \leq \|f\|_{\mathcal{H}_\mathcal{X}} < \infty. \tag{S21}$$

since

$$\|f \otimes g\|_\mathcal{H} = \|f\|_{\mathcal{H}_\mathcal{X}} \|g\|_{\mathcal{H}_\mathcal{Y}}, \quad \forall f \in \mathcal{H}_\mathcal{X}, \forall g \in \mathcal{H}_\mathcal{Y}.$$

Let $\tilde{\mu}(P_X) = \mathbb{E}_{X \sim P_X}\left[k_\mathcal{X}(X, \cdot) \otimes k_\mathcal{X}(X, \cdot)\right]$ be the embedding of $P_X \in \mathcal{P}(\mathcal{X})$ in $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}$ Then, for $P'_X \in \mathcal{P}(\mathcal{X})$ and using the shorthand $P' = P'_X P_{Y|\cdot}$, we have

$$\begin{aligned}
\mu(P') :&= \mathbb{E}_{(X,Y) \sim P'}\left[k_\mathcal{X}(X, \cdot) \otimes k_\mathcal{Y}(Y, \cdot)\right] \\
&= \mathbb{E}_{X \sim P'_X}\Big[\mathbb{E}_{Y \sim P_{Y|X}}\left[k_\mathcal{X}(X, \cdot) \otimes k_\mathcal{Y}(Y, \cdot)\right]\Big] \\
&= \mathbb{E}_{X \sim P'_X}\Big[k_\mathcal{X}(X, \cdot) \otimes \mathbb{E}_{Y \sim P_{Y|X}}\left[k_\mathcal{Y}(Y, \cdot)\right]\Big] \\
&= \mathbb{E}_{X \sim P'_X}\Big[k_\mathcal{X}(X, \cdot) \otimes \mu(P_{Y|X})\Big] \\
&= \mathbb{E}_{X' \sim P'_X}\Big[k_\mathcal{X}(X, \cdot) \otimes \mathcal{C}_{Y|X} k_\mathcal{X}(X, \cdot)\Big] \\
&= \mathbb{E}_{X' \sim P'_X}\Big[\tilde{\mathcal{C}}_{P_{Y|X}}\Big(k_\mathcal{X}(X', \cdot) \otimes k_\mathcal{X}(X', \cdot)\Big)\Big] \\
&= \tilde{\mathcal{C}}_{P_{Y|X}} \mathbb{E}_{X \sim P'_X}\left[k_\mathcal{X}(X, \cdot) \otimes k_\mathcal{X}(X, \cdot)\right] \\
&= \tilde{\mathcal{C}}_{P_{Y|X}} \tilde{\mu}(P'_X)
\end{aligned}$$

where the interchange between expectation and tensor product between the second and the third equality is justified by Lemma S1 and by (S21), where the interchanges between expectation and tensor product between the fifth and the sixth equality is justified by Lemma S1 and by (S20), while the fifth equality holds by (S19).

Similarly, for $P''_X \in \mathcal{P}(\mathcal{X})$ and with $P'' = P'_X P_{Y|\cdot}$, we have

$$\mu(P'') := \mathbb{E}_{(X,Y) \sim P''}\left[k_\mathcal{X}(X, \cdot) \otimes k_\mathcal{Y}(Y, \cdot)\right] \tilde{\mathcal{C}}_{P_{Y|X}} \tilde{\mu}(P''_X)$$

and thus,

$$\begin{aligned}
\mathbb{D}_k(P', P'') &= \|\mu(P') - \mu(P'')\|_\mathcal{H} \\
&= \left\|\tilde{\mathcal{C}}_{P_{Y|X}}\Big(\tilde{\mu}(P'_X) - \tilde{\mu}(P''_X)\Big)\right\|_\mathcal{H} \\
&\leq \|\tilde{\mathcal{C}}_{P_{Y|X}}\|_o \left\|\tilde{\mu}(P'_X) - \tilde{\mu}(P''_X)\right\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}} \\
&\leq \|\mathcal{C}_{Y|X}\|_o \, \mathbb{D}_{k^2_\mathcal{X}}(P'_X, P''_X)
\end{aligned} \tag{S22}$$

where the last inequality holds by (S20). The proof is complete.

8. Proof of Lemma ??

8.1. *Preliminary results*

The following lemma is adapted from Lemma 5 in Chérief-Abdellatif and Alquier (2020). While the proof is quite similar, the statement is more general.

LEMMA S6. *Let $\mathcal{S}$ be a set (equipped with a σ-algebra). Let K be any symmetric function $\mathcal{S}^2 \to [-1, 1]$ that can be written $K(s, s') = \langle \varphi(s), \varphi(s') \rangle_\mathcal{H}$ for some Hilbert space $\mathcal{H}$ and some function φ (note that we do not assume that K is a characteristic kernel). Let*

$S_1, \dots, S_n$ be independent random variables on $\mathcal{S}$ with respective distributions $Q_1, \dots, Q_n$. Define $\bar{Q} = (1/n)\sum_{i=1}^n Q_i$ and $\hat{Q} = (1/n)\sum_{i=1}^n \delta_{S_i}$. We define, for any Q and Q' probability distributions on $\mathcal{S}$,

$$\mathbb{D}_K^2(Q,Q') = \mathbb{E}_{S\sim Q, S'\sim Q}[K(Z,Z')] - 2\mathbb{E}_{S\sim Q,S'\sim Q'}[K(Z,Z')] + \mathbb{E}_{S\sim Q', S'\sim Q'}[K(Z,Z')]$$

(which is indeed a metric if K is a characteristic kernel). We have:

$$\mathbb{E}\left[\mathbb{D}_K(\bar{Q},\hat{Q})\right] \leq \frac{1}{\sqrt{n}} \text{ and } \mathbb{E}\left[\mathbb{D}_K^2(\bar{Q},\hat{Q})\right] \leq \frac{1}{n}.$$

Proof. Jensen's inequality gives $\mathbb{E}[\mathbb{D}_K(\bar{Q},\hat{Q})] \leq \sqrt{\mathbb{E}[\mathbb{D}_K^2(\bar{Q},\hat{Q})]}$. Put $m_i = \mathbb{E}_{S\sim Q_i}[\varphi(S)]$, then

$$\begin{aligned}
\mathbb{E}\left[\mathbb{D}_K^2(\bar{Q},\hat{Q})\right] &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n [\varphi(S_i) - m_i]\right\|_{\mathcal{H}}^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|\varphi(S_i) - m_i\|_{\mathcal{H}}^2\right] + \frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{E}\left[\langle \varphi(S_i) - m_i, \varphi(S_j) - m_j\rangle_{\mathcal{H}}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \Big(\mathbb{E}\left[\|\varphi(S_i)\|_{\mathcal{H}}^2\right] - \|m_i\|_{\mathcal{H}}^2\Big) + 0 \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\|\varphi(S_i)\|_{\mathcal{H}}^2\right] = \frac{1}{n^2}\sum_{i=1}^n K(S_i,S_i) \leq \frac{1}{n}.
\end{aligned}$$

Our proof strategy to study $\hat{\theta}_n(D_n)$ actually relies on the fact that despite contamination, the performance of $\hat{\theta}_n(D_n)$ remains close to the one of $\hat{\theta}_n(D_n)$. The following lemma will help to formalize this claim.

LEMMA S7. *Let $\hat{P}^{n,0} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i^0, Y_i^0)}$ be the non-contaminated empirical distribution and $\hat{P}_\theta^{n,0} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^0} P_{g(\theta, X_i^0)}$ be the uncontaminated counterpart of $\hat{P}_\theta^n$. Then, for any probability distribution Q on $\mathcal{X}\times\mathcal{Y}$, we have*

$$\left|\mathbb{D}_k\left(\hat{P}^{n,0}, Q\right) - \mathbb{D}_k\left(\hat{P}^n, Q\right)\right| < 2\epsilon \tag{S23}$$

and

$$\left|\mathbb{D}_k\left(\hat{P}_\theta^{n,0}, Q\right) - \mathbb{D}_k\left(\hat{P}_\theta^n, Q\right)\right| < 2\epsilon. \tag{S24}$$

Proof. For the first inequality (S23),

$$\begin{aligned}\left|\mathbb{D}_k\left(\hat{P}^{n,0},Q\right)-\mathbb{D}_k\left(\hat{P}^{n},Q\right)\right| &\leq \mathbb{D}_k\left(\hat{P}^{n,0},\hat{P}^{n}\right)\\ &= \left\|\frac{1}{n}\sum_{i=1}^{n}\Big[k((X_i^0,Y_i^0),\cdot)-k((X_i,Y_i),\cdot)\Big]\right\|_{\mathcal{H}}\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\Big[k((X_i^0,Y_i^0),\cdot)-k((X_i,Y_i),\cdot)\Big]\right\|_{\mathcal{H}}\\ &= \frac{1}{n}\sum_{i\in I}\left\|\Big[k((X_i^0,Y_i^0),\cdot)-k((X_i,Y_i),\cdot)\Big]\right\|_{\mathcal{H}}\\ &\leq \frac{1}{n}\sum_{i\in I}2=\frac{2|I|}{n}<2\epsilon.\end{aligned}$$

The proof of (S7) is exactly the same. □

8.2. *Proof of the lemma*

Proof. Thanks to (S23) of Lemma S7 we have, for any fixed $\theta\in\Theta$,

$$\begin{aligned}\mathbb{D}_k(\hat{P}^n_{\hat{\theta}_n},\bar{P}^0_n) &\leq \mathbb{D}_k(\hat{P}^n_{\hat{\theta}_n},\hat{P}^n)+\mathbb{D}_k(\hat{P}^n,\bar{P}^0_n) \text{ (triangle inequality)}\\ &\leq \mathbb{D}_k(\hat{P}^n_{\hat{\theta}_n},\hat{P}^n)+\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)+2\epsilon \text{ where we used (S23) with } Q=\bar{P}^0_n\\ &\leq \mathbb{D}_k(\hat{P}^n_{\theta},\hat{P}^n)+\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)+2\epsilon \text{ (by definition of } \hat{\theta}_n)\\ &\leq \mathbb{D}_k(\hat{P}^n_{\theta},\hat{P}^{n,0})+\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)+4\epsilon \text{ by (S23) with } Q=\hat{P}^n_\theta\\ &\leq \mathbb{D}_k(\hat{P}^n_{\theta},\bar{P}^0_n)+2\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)+4\epsilon \text{ (triangle inequality).}\end{aligned} \tag{S25}$$

Taking the expectation in (S25) gives:

$$\mathbb{E}\left[\mathbb{D}_k(\hat{P}^n_{\hat{\theta}_n},\bar{P}^0_n)\right]\leq 4\epsilon+\mathbb{D}_k(\hat{P}^n_\theta,\bar{P}^0_n)+2\mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)\right]. \tag{S26}$$

We can control the expectation in the right-hand side by an application of Lemma S6, where 330 $S_i=(X_i^0,Y_i^0)\sim Q_i:=\delta_{X_i^0}P^0_{Y|X_i^0}$ that are indeed independent, and where $K=k$. The lemma gives:

$$\mathbb{E}\left[\mathbb{D}_k(\hat{P}^n,\bar{P}^0_n)\right]\leq\frac{1}{\sqrt{n}}. \tag{S27}$$

We take the infinimum with respect to θ to obtain:

$$\mathbb{E}\left[\mathbb{D}_k(\hat{P}^n_{\hat{\theta}_n},\bar{P}^0_n)\right]\leq 4\epsilon+\inf_{\theta\in\Theta}\mathbb{D}_k(\hat{P}^n_\theta,\bar{P}^0_n)+\frac{2}{\sqrt{n}}. \tag{S28}$$

In order to prove (**??**), take any $z_i'\in\mathcal{Z}$ and define

$$\hat{P}^{n,0}_{(i)}=\frac{1}{n}\Big(\sum_{j\neq i}\delta_{(X_j^0,Y_j^0)}+\delta_{z_i'}\Big).$$

We note that:

$$\left|\mathbb{D}_k(\hat{P}^{n,0},\bar{P}^0_n)-\mathbb{D}_k(\hat{P}^{n,0}_{(i)},\bar{P}^0_n)\right|\leq\mathbb{D}_k(\hat{P}^{n,0},\hat{P}^{n,0}_{(i)})\leq\frac{2}{n}.$$

This allows to use the McDiarmind's bounded difference inequality McDiarmid (1989), which gives:

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^n, \bar{P}_n^0) - \mathbb{E}\left[\mathbb{D}_k(\hat{P}^n, \bar{P}_n^0)\right] \geq t\right\} \leq \exp\left(-\frac{nt^2}{2}\right), \quad \forall t > 0. \tag{S29}$$

Put $\eta = \exp(-nt^2/2)$ to get

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0}, \bar{P}_n^0) - \mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0}, \bar{P}_n^0)\right] \geq \sqrt{\frac{2\log(1/\eta)}{n}}\right\} \leq \eta,$$

which, together with (S26)-(S27), gives the statement of the theorem. □

9. PROOF OF THEOREM ??

9.1. *Preliminary result*

LEMMA S8. *Let $\|\cdot\|$ be a semi-norm on Θ. Let $M : \Theta \to [0,2]$ be such that there exists a unique $\theta_\star \in \Theta$ verifying $\inf_{\theta\in\Theta} M(\theta) = M(\theta_\star)$ and such that there exists a neighborhood U of $\theta_\star$ and a constant $\mu > 0$ for which*

$$M(\theta) - M(\theta_\star) \geq \mu\|\theta - \theta_\star\|, \quad \forall \theta \in U.$$

Let $(\check{\theta}_n)_{n\geq 1}$ be a sequence of random variables taking values in Θ and such that there exist a strictly increasing function $h_1 : (0,\infty) \to (0,\infty)$ with $\lim_{x\to\infty} h_1(x) = \infty$, a continuous and strictly decreasing function $h_2 : (0,1) \to (0,\infty)$, and a constant $x \geq 0$ such that

$$\mathbb{P}\Big\{M(\check{\theta}_n) < M(\theta_\star) + x + \frac{h_2(\eta)}{h_1(n)}\Big\} \geq 1 - \eta, \quad \forall \eta \in (0,1), \quad \forall n \geq 1. \tag{S30}$$

Then for any $t > 0$,

$$\mathbb{P}\Big\{\|\check{\theta}_n - \theta_\star\| \geq x/\mu + t\Big\} \leq 2h_2^{-1}\left[\left((\mu t) \wedge (\alpha - x)_+\right) h_1(n)\right],$$

and

$$\mathbb{P}\Big\{\|\check{\theta}_n - \theta_\star\| < \frac{x}{\mu} + \frac{h_2\left(\frac{\eta}{2}\right)}{\mu h_1(n)}\Big\} \geq 1 - \eta, \quad \forall n \geq 1, \quad \forall \eta \in \big[2h_2^{-1}((\alpha - x)_+ h_1(n)), 1\big)$$

where $\alpha = \inf_{\theta\in U^c} M(\theta) - M(\theta_\star) \in (0,2]$.

Remark S2. It would also be possible to get a result on $\mathbb{E}[\|\check{\theta}_n - \theta_\star\|]$, but at the price of the additional assumption that the parameter space Θ is bounded: $\sup_{(\theta,\theta')\in\Theta^2} \|\theta - \theta'\|_\Theta < \infty$.

Proof. Note that (S30) is equivalent to

$$\mathbb{P}\Big\{M(\check{\theta}_n) - M(\theta_\star) - x > t\Big\} \leq h_2^{-1}(t h_1(n)), \quad \forall t > 0, \quad \forall n \geq 1. \tag{S31}$$

Remind that $\alpha = \inf_{\theta\in U^c} M(\theta) - M(\theta_\star)$. It is immediate to see that $\alpha \leq 2$. Moreover, $\alpha > 0$, otherwise, U^c being a closed set, there would be a $\theta' \in U^c$ such that $M(\theta') - M(\theta_\star) = 0$.

Now, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| \geq t + x/\mu\right\} \\
&\quad = \mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| \geq t + x/\mu, \check{\theta}_n \in U\right\} + \mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| \geq t + x/\mu, \check{\theta}_n \notin U\right\} \\
&\quad \leq \mathbb{P}\left\{M(\check{\theta}) - M(\theta_\star) \geq \mu t + x, \check{\theta}_n \in U\right\} + \mathbb{P}\left\{\check{\theta}_n \notin U\right\} \\
&\quad \leq \mathbb{P}\left\{M(\check{\theta}) - M(\theta_\star) - x \geq \mu t\right\} + \mathbb{P}\left\{M(\check{\theta}) - M(\theta_\star) \geq \alpha\right\} \\
&\quad \leq h_2^{-1}\left(\mu t h_1(n)\right) + h_2^{-1}\left((\alpha - x)_+ h_1(n)\right)
\end{aligned}$$

where we used (S31) for the last inequality. As h_2^{-1} is strictly decreasing, we obtain:

$$\mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| \geq t + x/\mu\right\} \leq 2h_2^{-1}\left[\left((\mu t) \wedge (\alpha - x)_+\right) h_1(n)\right]. \tag{S32}$$

Fix $\eta \in \left[2h_2^{-1}((\alpha - x)_+ h_1(n)), 1\right)$ as in the statement of the lemma, and note that

$$2h_2^{-1}\left[\left((\mu t) \wedge (\alpha - x)_+\right) h_1(n)\right] = \eta \Leftrightarrow t = \frac{h_2\left(\frac{\eta}{2}\right)}{\mu h_1(n)}.$$

Plugging these values in (S32), we obtain:

$$\mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| < \frac{x}{\mu} + \frac{h_2\left(\frac{\eta}{2}\right)}{\mu h_1(n)}\right\} \geq 1 - \eta.$$

9.2. *Proof of the theorem*

Proof. From Lemma (**??**), (S30) in Lemma S8 holds with $\theta_\star = \theta_0$, $x = 4\epsilon$, $h_1(n) = \sqrt{n}$, $h_2(\eta) = 2 + \sqrt{2\log(1/\eta)}$ and $\check{\theta}_n = \hat{\theta}_n$. Apply Lemma S8 to get:

$$\sum_{n\geq 1} \mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| \geq + \frac{4\epsilon}{\mu} + t\right\} \leq 2\sum_{n\geq 1} \exp\left[-\frac{\left[\left((\mu t) \wedge (\alpha - x)_+\right)\sqrt{n} - 2\right]^2}{2}\right] < \infty, \quad \forall t > 0$$

showing that $\mathbb{P}\big(\limsup_{n\to\infty} \|\check{\theta}_n - \theta_\star\| \leq 4\epsilon/\mu\big) = 1$. Lemma S8 also states

$$\mathbb{P}\left\{\|\check{\theta}_n - \theta_\star\| < \frac{h_2\left(\frac{\eta}{2}\right)}{\mu h_1(n)}\right\} \geq 1 - \eta, \quad \forall n \geq 1, \quad \forall \eta \in \left[2h_2^{-1}((\alpha - x)_+ h_1(n)), 1\right).$$

Note that

$$\frac{h_2\left(\frac{\eta}{2}\right)}{\mu h_1(n)} = \frac{1}{\mu\sqrt{n}}\left(2 + \sqrt{2\log(2/\eta)}\right)$$

and $2h_2^{-1}((\alpha - x)_+ h_1(n)) = 2\exp(-((\alpha - x)_+\sqrt{n} - 2)^2/2)$. For the sake of simplicity, we only consider $n \geq 16/(\alpha - x)_+^2$, in this case, we have $(\alpha - x)_+\sqrt{n} - 2 \geq (\alpha - x)_+\sqrt{n}/2$ and thus the result holds in particular for any $\eta \in [2\exp(-n(\alpha - x)_+^2/8), 1)$. Finally, remind that $x = 4\epsilon < \alpha/8$ so it holds in particular for $n \geq 64/\alpha^2$ and $\eta \in [2\exp(-n\alpha^2/32), 1)$. □

10. Proof of Lemma **??**

10.1. *Preliminary result*

We start with a result that will be an essential tool in the proof of Lemma **??**. Essentially, it quantifies how well $\hat{P}^{n,0}_{\hat{\theta}_n} = (1/n)\sum_{i=1}^n \delta_{X_i^0} P_{g(\hat{\theta}_n, X_i^0)}$ approximates P^0. Usually, in regression literature, we focus mostly on the estimation of the distribution of $Y|X$ rather than on the estimation of the distribution of the pair (X,Y). Still, we believe that this result has in interpretation on its own, so we state is as a theorem.

THEOREM S1. *Under Assumption* **??** *we have*

$$\mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, P^0)\right] \leq 8\epsilon + \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0) + \frac{3}{\sqrt{n}}$$

and, for any $\eta \in (0,1)$*,*

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, P^0) \leq 8\epsilon + \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0) + \frac{3}{\sqrt{n}}\left(1+\sqrt{2\log(2/\eta)}\right)\right\} \geq 1-\eta.$$

Proof. The proof is quite similar to the proof of Lemma **??**, but requires some adaptations, in particular in the application of Lemma S6.

First,

$$\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, P^0) \leq \mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, \hat{P}^{n,0}) + \mathbb{D}_k(\hat{P}^{n,0}, P^0). \tag{S33}$$

Let us deal with the first term of this upper bound in a first time. Here, we will use both (S23) and (S24) of Lemma S7. We have:

$$\begin{aligned}
\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, \hat{P}^{n,0}) \leq \mathbb{D}_k\left(\hat{P}^{n,0}_{\hat{\theta}_n}, \hat{P}^{n}\right) + 2\epsilon \leq \mathbb{D}_k\left(\hat{P}^{n}_{\hat{\theta}_n}, \hat{P}^{n}\right) + 4\epsilon &\leq \mathbb{D}_k\left(\hat{P}^{n}_{\hat{\theta}_n(D_n^0)}, \hat{P}^{n}\right) + 4\epsilon \\
&\leq \mathbb{D}_k\left(\hat{P}^{n,0}_{\hat{\theta}_n(D_n^0)}, \hat{P}^{n}\right) + 6\epsilon \\
&\leq \mathbb{D}_k\left(\hat{P}^{n,0}_{\hat{\theta}_n(D_n^0)}, \hat{P}^{n,0}\right) + 8\epsilon \\
&= \inf_{\theta\in\Theta} \mathbb{D}_k\left(\hat{P}^{n,0}_{\theta}, \hat{P}^{n,0}\right) + 8\epsilon.
\end{aligned}$$

where the first inequality uses (S23), the second (S24), the third the definition of $\hat{\theta}_n$, the fourth (S24), the fifth (S23) and the sixth the definition of $\hat{\theta}_n$.

Together with (S33), this shows that

$$\begin{aligned}
\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, P^0) &\leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(\hat{P}^{n,0}_{\theta}, \hat{P}^{n,0}\right) + \mathbb{D}_k(\hat{P}^{n,0}, P^0) + 8\epsilon \\
&\leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(\hat{P}^{n,0}_{\theta}, P^0\right) + 2\mathbb{D}_k(\hat{P}^{n,0}, P^0) + 8\epsilon \\
&\leq \inf_{\theta\in\Theta} \left[\mathbb{D}_k\left(\hat{P}^{n,0}_{\theta}, P_\theta\right) + \mathbb{D}_k(P_\theta, P^0)\right] + 2\mathbb{D}_k(\hat{P}^{n,0}, P^0) + 8\epsilon \qquad \text{(S34)}
\end{aligned}$$

and so, taking expectations on both sides,

$$\mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n}, P^0)\right] \leq \inf_{\theta\in\Theta} \left\{\mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0}_{\theta}, P_\theta)\right] + \mathbb{D}_k(P_\theta, P^0)\right\} 2\mathbb{E}\left[\mathbb{D}_k(\hat{P}^{n,0}, P^0)\right] + 8\epsilon. \tag{S35}$$

We tackle the term $\mathbb{E}\left[\mathbb{D}_k(\hat{P}_\theta^{n,0}, P_\theta)\right]$. Letting Φ denote the function such that $k((x,y),(x',y')) = \langle \Phi(x,y), \Phi(x',y')\rangle_{\mathcal{H}}$, we have

$$\begin{aligned}
\mathbb{D}_k(\hat{P}_\theta^{n,0}, P_\theta) &= \sqrt{\mathbb{D}_k^2(\hat{P}_\theta^{n,0}, P_\theta)} \\
&= \Bigg(\mathbb{E}_{(X,Y)\sim\hat{P}_\theta^{n,0},(X',Y')\sim\hat{P}_\theta^{n,0}} \left\langle \Phi(X,Y), \Phi(X',Y')\right\rangle_{\mathcal{H}} \\
&\quad - 2\mathbb{E}_{(X,Y)\sim\hat{P}_\theta^{n,0},(X',Y')\sim P_\theta} \left\langle \Phi(X,Y), \Phi(X',Y')\right\rangle_{\mathcal{H}} \\
&\quad + \mathbb{E}_{(X,Y)\sim P_\theta,(X',Y')\sim P_\theta} \left\langle \Phi(X,Y), \Phi(X',Y')\right\rangle_{\mathcal{H}} \Bigg)^{\frac{1}{2}} \\
&= \Bigg(\mathbb{E}_{X\sim\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}, X'\sim\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}} \left\langle \mathbb{E}_{Y\sim P_{g(\theta,X)}}[\Phi(X,Y)], \mathbb{E}_{Y'\sim P_{g(\theta,X')}}[\Phi(X',Y')]\right\rangle_{\mathcal{H}} \\
&\quad - 2\mathbb{E}_{X\sim\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}, X'\sim P_X^0} \left\langle \mathbb{E}_{Y\sim P_{g(\theta,X)}}[\Phi(X,Y)], \mathbb{E}_{Y'\sim P_{g(\theta,X')}}[\Phi(X',Y')]\right\rangle_{\mathcal{H}} \\
&\quad + \mathbb{E}_{X\sim P_X^0, X'\sim P_X^0} \left\langle \mathbb{E}_{Y\sim P_{g(\theta,X)}}[\Phi(X,Y)], \mathbb{E}_{Y'\sim P_{g(\theta,X')}}[\Phi(X',Y')]\right\rangle_{\mathcal{H}} \Bigg)^{\frac{1}{2}} \\
&= \sqrt{\mathbb{D}_{\bar{k}}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}, P_X^0\right)} = \mathbb{D}_{\bar{k}}\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}, P_X^0\right)
\end{aligned}$$

where the function $\bar{k}$ is given by:

$$\bar{k}(x,x') = \left\langle \mathbb{E}_{Y\sim P_{g(\theta,x)}}[\Phi(x,Y)], \mathbb{E}_{Y'\sim P_{g(\theta,x')}}[\Phi(x',Y')]\right\rangle_{\mathcal{H}}.$$

Note that $-1 \leq \bar{k} \leq 1$ so we can apply Lemma S6 to $S_i = X_i^0 \sim Q_i = P_X^0$ and $K = \bar{k}$ to get:

$$\mathbb{E}\left[\mathbb{D}_{\bar{k}}\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i^0}, P_X^0\right)\right] \leq \frac{1}{\sqrt{n}}.$$

Combining this last result with (S35), and applying Lemma S6 with $S_i = (X_i^0, Y_i^0) \sim Q_i = P^0$ and $K = k$ that gives $\mathbb{E}[\mathbb{D}_k(\hat{P}^{n,0}, P^0)] \leq 1/\sqrt{n}$, we finally obtain:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{D}_k(\hat{P}_{\hat{\theta}_n}^{n,0}, P^0)\right] &\leq \inf_{\theta\in\Theta}\left\{\frac{1}{\sqrt{n}} + \mathbb{D}_k(P_\theta, P^0)\right\} + \frac{2}{\sqrt{n}} + 8\epsilon \\
&= \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0) + \frac{3}{\sqrt{n}} + 8\epsilon,
\end{aligned}$$

that is the first inequality of the theorem.

In order to prove the second inequality let $\theta_0 \in \operatorname{argmin}_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0)$. Then (S34) implies

$$\begin{aligned}
&\mathbb{D}_k(\hat{P}_{\hat{\theta}_n}^{n,0}, P^0) \\
&\quad \leq \mathbb{D}_k(\hat{P}_{\theta_0}^{n,0}, P_{\theta_0}) + \mathbb{D}_k(P_{\theta_0}, P^0) + 2\mathbb{D}_k(\hat{P}^{n,0}, P^0) + 8\epsilon \\
&\quad = \mathbb{D}_k(\hat{P}_{\theta_0}^{n,0}, P_{\theta_0}) + \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, P^0) + 2\mathbb{D}_k(\hat{P}^{n,0}, P^0) + 8\epsilon.
\end{aligned}$$

McDiarmid's bounded difference inequality leads to

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0},P^0)-\mathbb{E}\big[\mathbb{D}_k(\hat{P}^{n,0},P^0)\big]\geq t\right\}\leq \exp\left(-\frac{nt^2}{2}\right)$$

and to

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0},P^0)-\mathbb{E}\left(\mathbb{D}_k(\hat{P}^{n,0},P^0)\right)\geq t\right\}\leq \exp\left(-\frac{nt^2}{2}\right).$$

By a union bound, the probability that one of the two events hold is smaller or equal to $2\exp(-nt^2/2)$, which leads to

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n},P^0)\leq \inf_{\theta\in\Theta}\mathbb{D}_k(P_\theta,P^0)+\frac{3}{\sqrt{n}}\left(1+\sqrt{2\log(2/\eta)}\right)+8\epsilon\right\}\geq 1-\eta.$$

This ends the proof.

10.2. *Proof of the lemma*

Proof. By Lemma **??** applied to $P'_X=\hat{P}^{n,0}_X$ and $P''_X=P^0_X$, we have

$$\mathbb{D}_k(\hat{P}^{n}_{\hat{\theta}_n},P_{\hat{\theta}_n})\leq \mathfrak{C}\,\mathbb{D}_{k^2_X}(\hat{P}^{n,0}_X,P^0_X) \tag{S36}$$

and thus

$$\mathbb{E}\Big[\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n},P_{\hat{\theta}_n})\Big]\leq \mathfrak{C}\mathbb{E}\Big[\mathbb{D}_{k^2_X}(\hat{P}^{n,0}_X,P^0_X)\Big].$$

Applying Lemma S6 with $Z_i=X_i\sim Q_i=P^0_X$ and $K=k^2_{\mathcal{X}}$, we obtain

$$\mathbb{E}\Big[\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n},P_{\hat{\theta}_n})\Big]\leq \frac{\mathfrak{C}}{\sqrt{n}}. \tag{S37}$$

Now:

$$\begin{aligned}\mathbb{E}\left[\mathbb{D}_k(P_{\hat{\theta}_n},P^0)\right]&\leq \mathbb{E}\left[\mathbb{D}_k(P_{\hat{\theta}_n},\hat{P}_{\hat{\theta}_n})\right]+\mathbb{E}\left[\mathbb{D}_k(\hat{P}_{\hat{\theta}_n},P^0)\right]\\&\leq \frac{\mathfrak{C}}{\sqrt{n}}+\left(\inf_{\theta\in\Theta}\mathbb{D}_k(P_\theta,P^0)+8\epsilon+\frac{3}{\sqrt{n}}\right)\end{aligned}$$

where we used (S37) to upper bound the first term, and Theorem S1 for the second term. This ends the proof of the bound in expectation.

Let us now prove the inequality in probability. Let $\eta\in(0,1)$ and use the bounded difference inequality to get

$$\mathbb{P}\left\{\mathbb{D}_{k^2_{\mathcal{X}}}(\hat{P}^{n,0}_X,P^0_X)-\mathbb{E}\Big[\mathbb{D}_{k^2_X}(\hat{P}^{n,0}_X,P^0_X)\Big]\leq\sqrt{\frac{2\log(2/\eta)}{n}}\right\}\geq 1-\frac{\eta}{2}$$

while, by Theorem S1,

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}^{n,0}_{\hat{\theta}_n},P^0)\leq 8\epsilon+\inf_{\theta\in\Theta}\mathbb{D}_k(P_\theta,P^0)+\frac{3}{\sqrt{n}}\left(1+\sqrt{2\log(4/\eta)}\right)\right\}\leq 1-\frac{\eta}{2}.$$

Together with (S36), and using a union bound, we obtain

$$\mathbb{P}\left\{\mathbb{D}_k(\hat{P}_{\hat{\theta}_n},P^0)\leq\inf_{\theta\in\Theta}\mathbb{D}_k(P_\theta,P^0)+\frac{3\big(1+\sqrt{2\log(4/\eta)}\big)+\mathfrak{C}\big(1+\sqrt{2\log(2/\eta)}\big)}{\sqrt{n}}\right\}\geq 1-\eta.$$

11. Proof of Theorem ??

Proof. From Lemma **??**, (S30) in Lemma S8 holds with $h_1(n) = \sqrt{n}$, $h_2(\eta) = (\mathfrak{C}+3)(1+\sqrt{2\log(4/\eta)})$ and $\breve{\theta}_n = \hat{\theta}_n$. Then, the result is proved following the computations done in the proof of Theorem **??**. □

12. Proof of Proposition ??

Proof. Let $f : \Theta \to [0,4]$ be defined by

$$f(\theta) = \big(\mathbb{D}_k(P_\theta, \tilde{P}^0) - \mathbb{D}_k(P_{\theta_0}, \tilde{P}^0)\big)^2, \quad \theta \in \Theta$$

and let U be an open set containing θ_0 such that f is twice continuously differentiable on U. Let H_θ be the Hessian matrix of f evaluated at $\theta \in U$.

Then, using Taylor's theorem, for every $\theta \in U$ we have, for some $\tau \in [0,1]$

$$\begin{aligned}
f(\theta) &= f(\theta_0) + (\theta-\theta_0)^\top \nabla f(\theta_0) + \frac{1}{2}(\theta-\theta_0)^\top H_{\theta_0+\tau(\theta-\theta_0)}(\theta-\theta_0)\\
&= (\theta-\theta_0)^\top H_{\theta_0+\tau(\theta-\theta_0)}(\theta-\theta_0)\\
&\geq \|\theta-\theta_0\|^2 \frac{\lambda_{\min}\big(H_{\theta_0+\tau(\theta-\theta_0)}\big)}{2}\\
&\geq \|\theta-\theta_0\|^2 \frac{\inf_{\theta\in U,\tau\in[0,1]} \lambda_{\min}\big(H_{\theta_0+\tau(\theta-\theta_0)}\big)}{2}
\end{aligned}$$

where for every $\theta \in U$ we denote by $\lambda_{\min}(H_\theta)$ the minimum eigenvalue of H_θ. Under the assumptions of the proposition, we can take U sufficiency small so that $c := \inf_{\theta\in U,\tau\in[0,1]} \lambda_{\min}\big(H_{\theta_0+\tau(\theta-\theta_0)}\big) > 0$. Then,

$$\mathbb{D}_k(P_\theta, \tilde{P}^0) - \mathbb{D}_k(P_{\theta_0}, \tilde{P}^0) = \sqrt{f(\theta)} \geq \sqrt{c/2}\,\|\theta-\theta_0\|$$

showing that (**??**) holds for $\mu = \sqrt{c/2}$. □

13. Proof of Proposition ??

Proof. For all $(\theta, x, y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$, let

$$m_\theta(x,y) = \mathbb{E}_{Y,Y'\overset{\text{iid}}{\sim}P_{g(\theta,X)}}\big[k_{\mathcal{Y}}(Y,Y') - 2k_{\mathcal{Y}}(Y,y)\big] + \mathbb{E}_{X\sim P^0_X}\Big[\mathbb{E}_{Y,Y'\overset{\text{iid}}{\sim}P^0_{Y|X}}\big[k_{\mathcal{Y}}(Y,Y')\big]\Big]$$

and remark that

$$\mathbb{E}_{(X,Y)\sim P^0}[m_\theta(X,Y)] = \mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_{\mathcal{Y}}}(P_{g(\theta,X)}, P^0_{Y|X})^2\big], \quad \forall \theta \in \Theta.$$

Under the assumptions of the theorem, the mapping $\theta \mapsto m_\theta(x,y)$ is continuous on the compact set Θ and is such that $|m_\theta(x,y)| \leq 4$ for all $(\theta,x,y) \in \Theta \times \mathcal{X} \times \mathcal{Y}$. Then (see e.g Van der Vaart, 2000, page 46)

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n m_\theta(X_i,Y_i) - \mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_{\mathcal{Y}}}(P_{g(\theta,X)}, P^0_{Y|X})^2\big]\Big| \to 0, \quad \text{in } \mathbb{P}\text{-probability}$$

and therefore, noting that $\tilde{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n m_\theta(X_i,Y_i)$, the result follows by Van der Vaart (2000, Theorem 5.7). □

14. PROOF OF THEOREM ??

Proof. Let $\epsilon\in[0,1)$ and, for all $x\in\mathcal{X}$, let $\tilde P^0_{Y|x}=(1-\epsilon)P^0_{Y|x}+\epsilon Q_{Y|x}$ and $\tilde P^0_X=(1-\epsilon)P^0_X+\epsilon Q_X$ where Q_X denotes the distribution of X under Q.

Then, for all $\theta\in\Theta$ we have

$$\begin{aligned}
&\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]\\
&\qquad\le \mathbb{E}_{X\sim P^0_X}\Big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})+\mathbb{D}_{k_\mathcal{Y}}(P^0_{Y|X},\tilde P^0_{Y|X})\big)^2\Big]\\
&\qquad\le \mathbb{E}_{X\sim P^0_X}\big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})+2\epsilon\big)^2\big]\\
&\qquad\le \mathbb{E}_{X\sim P^0_X}\big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})^2\big]+8\epsilon+4\epsilon^2\\
&\qquad\le \mathbb{E}_{X\sim P^0_X}\big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})^2\big]+12\epsilon\\
&\qquad\le \frac{1}{1-\epsilon}\mathbb{E}_{X\sim \tilde P^0_X}\big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})^2\big]+12\epsilon
\end{aligned}\tag{S38}$$

where the third inequality the fact uses the that, since $|k_\mathcal{Y}|\le 1$, $\mathbb{P}(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})\le 2)=1$, the penultimate inequality holds since $\epsilon<1$ and the last inequality holds since $\mathbb{E}_{X\sim Q_X}\big[\big(\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})^2\big]\ge 0$ for all $\theta\in\Theta$.

Then, applying (S38) with $\theta=\tilde\theta_{Q,\epsilon}$ yields

$$\begin{aligned}
&\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_{Q,\epsilon},X)},P^0_{Y|X})^2\big]\\
&\qquad\le \frac{1}{1-\epsilon}\inf_{\theta\in\Theta}\mathbb{E}_{X\sim\tilde P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0_{Y|X})^2\big]+12\epsilon\\
&\qquad\le \frac{1}{1-\epsilon}\inf_{\theta\in\Theta}\mathbb{E}_{X\sim\tilde P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]+\frac{12\epsilon}{1-\epsilon}+12\epsilon\\
&\qquad\le \inf_{\theta\in\Theta}\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]+\frac{24\epsilon}{1-\epsilon}+\frac{12\epsilon}{1-\epsilon}+16\epsilon\\
&\qquad\le \mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_0,X)},P^0_{Y|X})^2\big]+\frac{52\epsilon}{1-\epsilon}
\end{aligned}\tag{S39}$$

where the second inequality follows by swapping $\tilde P^0_{Y|X}$ and $P^0_{Y|X}$ in (S38) and the third one uses the fact that, since $|k_\mathcal{Y}|\le 1$,

$$\mathbb{E}_{X\sim Q_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]\le 4,\quad\forall\theta\in\Theta.$$

By assumption, $\tilde\theta_0$ is the unique minimizer of the function $\theta\mapsto\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]$ and therefore (see the proof of Lemma S8)

$$\alpha=\inf_{\theta\in U^c}\Big(\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},P^0_{Y|X})^2\big]-\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_0,X)},P^0_{Y|X})^2\big]\Big)>0.$$

Together with (S39), this shows that if

$$\frac{52\epsilon}{1-\epsilon}<\alpha\Leftrightarrow\epsilon<\frac{\alpha}{52+\alpha}$$

then

$$\begin{aligned}
&\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_{Q,\epsilon},X)},P^0_{Y|X})^2\big]-\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_0,X)},P^0_{Y|X})^2\big]\\
&\qquad<\inf_{\theta\in U^c}\Big(\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\theta,X)},\tilde P^0)^2\big]-\mathbb{E}_{X\sim P^0_X}\big[\mathbb{D}_{k_\mathcal{Y}}(P_{g(\tilde\theta_0,X)},\tilde P^0)^2\big]\Big)
\end{aligned}$$

implying that $\tilde{\theta}_{Q,\epsilon} \in U$. Consequently, using again (S39),

$$\frac{52\epsilon}{1-\epsilon} \geq \mathbb{E}_{X\sim P_X^0}\left[\mathbb{D}_{k_{\mathcal{Y}}}(P_{g(\tilde{\theta}_{Q,\epsilon},X)}, P^0_{Y|X})^2\right] - \mathbb{E}_{X\sim P_X^0}\left[\mathbb{D}_{k_{\mathcal{Y}}}(P_{g(\tilde{\theta}_0,X)}, P^0_{Y|X})^2\right] \geq \mu\|\tilde{\theta}_{Q,\epsilon} - \theta_0\|$$

and the result follows. □

References

Arias, M. L. and Gonzalez, M. C. (2009). Reduced solutions of douglas equations and angles between subspaces. *Journal of mathematical analysis and applications*, 355(1):426–433.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR.

Chérief-Abdellatif, B.-E. and Alquier, P. (2022). Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213.

Cohn, D. L. (2013). *Measure theory*. Springer.

Da Prato, G. and Zabczyk, J. (2014). *Stochastic equations in infinite dimensions*. Cambridge university press.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99.

Gupta, P. and Bhattacharjee, G. (1984). An efficient algorithm for random sampling without replacement. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 435–442. Springer.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Klebanov, I., Schuster, I., and Sullivan, T. J. (2020). A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Paulsen, V. I. and Raghupathi, M. (2016). *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press.

Szabó, Z. and Sriperumbudur, B. (2018). Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:233.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[*Received on. Editorial decision on*]