A Generalised Jensen’s Inequality

In Section[4] we require a version of Jensen’s inequality generalised to (possibly) infinite-dimensional vector spaces, because our random variable takes values in $\mathcal{H}_X$, and our convex function is $\|\cdot\|_{\mathcal{H}_X} : \mathcal{H}_X \to \mathbb{R}$. Note that this square norm function is indeed convex, since, for any $t \in [0, 1]$ and any pair $f, g \in \mathcal{H}_X$.

$$\|tf + (1-t)g\|_{\mathcal{H}_X}^2 \leq t \|f\|_{\mathcal{H}_X}^2 + (1-t) \|g\|_{\mathcal{H}_X}^2$$

by the triangle inequality

$$\leq t \|f\|_{\mathcal{H}_X}^2 + (1-t) \|g\|_{\mathcal{H}_X}^2,$$

by the convexity of $x \mapsto x^2$.

The following theorem generalises Jensen’s inequality to infinite-dimensional vector spaces.

**Theorem A.1** (Generalised Jensen’s Inequality, [38], Theorem 3.10). Suppose $\mathcal{T}$ is a real Hausdorff locally convex (possibly infinite-dimensional) linear topological space, and let $C$ be a closed convex subset of $\mathcal{T}$. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $V : \Omega \to \mathcal{T}$ a Pettis-integrable random variable such that $V(\Omega) \subseteq C$. Let $f : C \to [-\infty, \infty)$ be a convex, lower semi-continuous extended-real-valued function such that $\mathbb{E}_V[f(V)]$ exists. Then

$$f(\mathbb{E}_V[V]) \leq \mathbb{E}_V[f(V)].$$

We will actually apply generalised Jensen’s inequality with conditional expectations, so we need the following theorem.

**Theorem A.2** (Generalised Conditional Jensen’s Inequality). Suppose $\mathcal{T}$ is a real Hausdorff locally convex (possibly infinite-dimensional) linear topological space, and let $C$ be a closed convex subset of $\mathcal{T}$. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and $V : \Omega \to \mathcal{T}$ a Pettis-integrable random variable such that $V(\Omega) \subseteq C$. Let $f : C \to [-\infty, \infty)$ be a convex, lower semi-continuous extended-real-valued function such that $\mathbb{E}_V[f(V)]$ exists. Suppose $\mathcal{E}$ is a sub-$\sigma$-algebra of $\mathcal{F}$. Then

$$f(\mathbb{E}_V[V \mid \mathcal{E}]) \leq \mathbb{E}_V[f(V) \mid \mathcal{E}].$$

**Proof.** Let $\mathcal{T}^*$ be the dual space of all real-valued continuous linear functionals on $\mathcal{T}$. The first part of the proof of [38] Theorem 3.6] tells us that, for all $v \in \mathcal{T}$, we can write

$$f(v) = \sup\{m(v) \mid m \text{ affine, } m \leq f \text{ on } C\},$$

where an affine function $m$ on $\mathcal{T}$ is of the form $m(v) = v^*(v) + \alpha$ for some $v^* \in \mathcal{T}^*$ and $\alpha \in \mathbb{R}$. If we define the subset $Q$ of $\mathcal{T}^* \times \mathbb{R}$ as

$$Q := \{(v^*, \alpha) : v^* \in \mathcal{T}^*, \alpha \in \mathbb{R}, v^*(v) + \alpha \leq f(v) \text{ for all } v \in \mathcal{T}\},$$

then we can rewrite $f$ as

$$f(v) = \sup_{(v^*, \alpha) \in Q} \{v^*(v) + \alpha\}, \quad \text{for all } v \in \mathcal{T}. \quad (5)$$

See that, for any $(v^*, \alpha) \in Q$, we have

$$\mathbb{E} \left[f(V) \mid \mathcal{E} \right] \geq \mathbb{E} \left[v^*(V) + \alpha \mid \mathcal{E} \right] \quad \text{almost surely, by assumption (*)}$$

$$= \mathbb{E} \left[v^*(V) \mid \mathcal{E} \right] + \alpha \quad \text{almost surely, by linearity (**).}$$

Here, (*) and (**) use the properties of conditional expectation of vector-valued random variables given in [12] pp.45-46, Properties 43 and 40 respectively.

We want to show that $\mathbb{E} \left[v^*(V) \mid \mathcal{E} \right] = v^* \left(\mathbb{E} \left[V \mid \mathcal{E} \right] \right)$ almost surely, and in order to so, we show that the right-hand side is a version of the left-hand side. The right-hand side is clearly $\mathcal{E}$-measurable, since we have a linear operator on an $\mathcal{E}$-measurable random variable. Moreover, for any $A \in \mathcal{E}$,

$$\int_A v^* \left(\mathbb{E} \left[V \mid \mathcal{E} \right] \right) \ dP = v^* \left(\int_A \mathbb{E} \left[V \mid \mathcal{E} \right] \ dP \right) \quad \text{by [10] p.403, Proposition E.11}$$

$$= v^* \left(\int_A V \ dP \right) \quad \text{by the definition of conditional expectation}$$

$$= \int_A v^*(V) \ dP \quad \text{by [10] p.403, Proposition E.11}$$
(here, all the equalities are almost-sure equalities). Hence, by the definition of the conditional expectation, we have that $E \left[ \nu^*(V) \mid \mathcal{E} \right] = \nu^* \left( E \left[ V \mid \mathcal{E} \right] \right)$ almost surely. Going back to our above work, this means that

$$E \left[ f(V) \mid \mathcal{E} \right] \geq \nu^* \left( E \left[ V \mid \mathcal{E} \right] \right) + \alpha.$$  

Now take the supremum of the right-hand side over $Q$. Then (5) tells us that

$$E \left[ f(V) \mid \mathcal{E} \right] \geq f \left( E \left[ V \mid \mathcal{E} \right] \right),$$

as required.

In the context of Section 4, $\mathcal{H}_X$ is real and Hausdorff, and locally convex (because it is a normed space). We take the closed convex subset to be the whole space $\mathcal{H}_X$ itself. The function $\| \cdot \|_{\mathcal{H}_X}^2 : \mathcal{H}_X \to \mathbb{R}$ is convex (as shown above) and continuous, and finally, since Bochner-integrability implies Pettis integrability, all the conditions of Theorem A.2 are satisfied.

### B Generalisation Error Bounds

Caponnetto and De Vito [5] give an optimal rate of convergence of vector-valued RKHS regression estimators, and its results are quoted by Grünewälder et al. [22] as the state of the art convergence rates, $O(\frac{\log n}{n})$. In particular, this implies that the learning algorithm is consistent. However, the lower rate uses an assumption that the output space is a finite-dimensional Hilbert space [5, Theorem 2]; and in our case, this will mean that $\mathcal{H}_X$ is finite-dimensional. This is not true if, for example, we take $k_X$ to be the Gaussian kernel; indeed, this is noted as a limitation by Grünewälder et al. [22], stating that “It is likely that this (finite-dimension) assumption can be weakened, but this requires a deeper analysis”. In this paper, we do not want to restrict our attention to finite-dimensional $\mathcal{H}_X$.

The upper bound would have been sufficient to guarantee consistency, but an assumption used in the upper bound requires the operator $l_{XZ,z} : \mathcal{H}_X \to \mathcal{G}_{XZ}$ defined by

$$l_{XZ,z}(f)(z') = l_{XZ}(z,z')(f)$$

to be Hilbert-Schmidt for all $z \in Z$. However, for each $z \in Z$, taking any orthonormal basis $\{\varphi_i\}_{i=1}^\infty$ of $\mathcal{H}_X$, we see that

$$\sum_{i=1}^\infty (l_{XZ,z}(\varphi_i), l_{XZ,z}(\varphi_i))_{\mathcal{G}_{XZ}} = \sum_{i=1}^\infty (k_Z(z,z'), \varphi_i, k_Z(z', \cdot) \varphi_i)_{\mathcal{H}_X}$$

$$= \sum_{i=1}^\infty (k_Z(z,z) \varphi_i, \varphi_i)_{\mathcal{H}_X}$$

$$= k_Z(z,z) \sum_{i=1}^\infty 1$$

$$= \infty,$$

meaning this assumption is not fulfilled with our choice of kernel either. Hence, results in [5], used by [22], are not applicable to guarantee consistency in our context.

Kadri et al. [26] address the problem of generalisability of function-valued learning algorithms, using the concept of uniform algorithmic stability [4]. Let us write

$$\mathcal{D} := \{(x_1, z_1), \ldots, (x_n, z_n)\}$$

for our training set of size $n$ drawn i.i.d. from the distribution $P_{XZ}$, and we denote by $\mathcal{D}^i = \mathcal{D} \setminus (x_i, z_i)$ the set $\mathcal{D}$ from which the data point $(x_i, z_i)$ is removed. Further, we denote by $\hat{F}_{P_{X|Z}, \mathcal{D}}(\cdot) = \sum_{i=1}^n k_X(x_i, \cdot) - F(z_i)\|\|_{\mathcal{H}_X} + \lambda \|F\|_{\mathcal{G}_{XZ}}^2$ the estimate produced by our learning algorithm from the dataset $\mathcal{D}$ by minimising the loss $\mathcal{L}_{X|Z,n,\lambda}(F) = \sum_{i=1}^n k_X(x_i, \cdot) - F(z_i)\|\|_{\mathcal{H}_X} + \lambda \|F\|_{\mathcal{G}_{XZ}}^2$.

The assumptions used in this paper, with notations translated to our context, are
1. There exists $\kappa_1 > 0$ such that for all $z \in Z$, 
\[ \|l_{XZ}(z, z)\|_{\text{op}} = \sup_{f \in \mathcal{H}_X} \frac{\|l_{XZ}(z, z)(f)\|_{\mathcal{H}_X}}{\|f\|_{\mathcal{H}_X}} \leq \kappa_1^2. \]

2. The real function $Z \times Z \to \mathbb{R}$ defined by 
\[(z_1, z_2) \mapsto (l_{XZ}(z_1, z_2)f_1, f_2)_{\mathcal{H}_X} \]
is measurable for all $f_1, f_2 \in \mathcal{H}_X$.

3. The map $(f, F, z) \mapsto \|f - F(z)\|_{\mathcal{H}_X}^2$ is $\tau$-admissible, i.e. convex with respect to $F$ and Lipschitz continuous with respect to $F(z)$, with $\tau$ as its Lipschitz constant.

4. There exists $\kappa_2 > 0$ such that for all $(z, f) \in Z \times \mathcal{H}_X$ and any training set $D$,
\[ \|f - \hat{F}_{P_X\mid Z, D}(z)\|_{\mathcal{H}_X}^2 \leq \kappa_2. \]

The concept of uniform stability, with notations translated to our context, is defined as follows.

**Definition B.1** (Uniform algorithmic stability, [26, Definition 6]). For each $F \in \mathcal{G}_{XZ}$, define the function 
\[ \mathcal{R}(F) : Z \times \mathcal{H}_X \to \mathbb{R} \]
\[ (z, x) \mapsto \|k_X(z, \cdot) - F(z)\|_{\mathcal{H}_X}^2. \]

A learning algorithm that calculates the estimate $\hat{F}_{P_X\mid Z, D}$ from a training set has uniform stability $\beta$ with respect to the squared loss if the following holds: for all $n \geq 1$, all $i \in \{1, \ldots, n\}$ and any training set $D$ of size $n$,
\[ \|\mathcal{R}(\hat{F}_{P_X\mid Z, D}) - \mathcal{R}(\hat{F}_{P_X\mid Z, D})\|_{\infty} \leq \beta. \]

The next two theorems are quoted from [26].

**Theorem B.2** ([26, Theorem 7]). Under assumptions 1, 2 and 3, a learning algorithm that maps a training set $D$ to the function $\hat{F}_{P_X\mid Z, D} = \hat{F}_{P_X\mid Z, n, \lambda}$ is $\beta$-stable with
\[ \beta = \tau^2 \kappa_1^2 \]
\[ 2\lambda n. \]

**Theorem B.3** ([26, Theorem 8]). Let $D \mapsto \hat{F}_{P_X\mid Z, D} = \hat{F}_{P_X\mid Z, n, \lambda}$ be a learning algorithm with uniform stability $\beta$, and assume Assumption 4 is satisfied. Then, for all $n \geq 1$ and any $0 < \delta < 1$, the following bound holds with probability at least $1 - \delta$ over the random draw of training samples:
\[ \hat{\mathcal{E}}_{X\mid Z}(\hat{F}_{P_X\mid Z, n, \lambda}) \leq \frac{1}{n} \hat{\mathcal{E}}_{X\mid Z,n}(\hat{F}_{P_X\mid Z, n, \lambda}) + 2\beta + (4n\beta + \kappa_2)\sqrt{\frac{\ln \frac{1}{\delta}}{2n}}. \]

Theorems B.2 and B.3 give us results about the generalisability of our learning algorithm. It remains to check whether the assumptions are satisfied.

Assumption 2 is satisfied thanks to our assumption that point embeddings are measurable functions, and Assumption 1 is satisfied if we assume that $k_Z$ is a bounded kernel (i.e. there exists $B_Z > 0$ such that $k_Z(z_1, z_2) \leq B_Z$ for all $z_1, z_2 \in Z$), because
\[ \|l_{XZ}(z, z)\|_{\text{op}} = \sup_{f \in \mathcal{H}_X, \|f\|_{\mathcal{H}_X} = 1} \|k_Z(z, z)(f)\|_{\mathcal{H}_X} \leq B_Z. \]

In [26], a general loss function is used rather than the squared loss, and it is noted that Assumption 3 is in general not satisfied with the squared loss, which is what we use in our context. However, this issue can be addressed if we restrict the output space to a bounded subset. In fact, the only elements in $\mathcal{H}_X$ that appear as the output samples in our case are $k_X(x, \cdot)$ for $x \in X$, so if we place the assumption that $k_X$ is a bounded kernel (i.e. there exists $B_X > 0$ such that $k_X(x_1, x_2) \leq B_X$ for all $x_1, x_2 \in X$), then by the reproducing property,
\[ \|k_X(x, \cdot)\|_{\mathcal{H}_X} = \sqrt{k_X(x, x)} \leq \sqrt{B_X}. \]
So it is no problem, in our case, to place this boundedness assumption. [26, Appendix D] tells us that Assumption 1 with this boundedness assumption implies Assumption 4 with
\[ \kappa_2 = B_x \left( 1 + \frac{\kappa_1}{\sqrt{\lambda}} \right)^2, \]
while [26, Lemma 2] provides us with a condition which can replace Assumption 3 in Theorem B.2, giving us the uniform stability of our algorithm with
\[ \beta = \frac{2\kappa_2^2 B_x \left( 1 + \frac{\kappa_1}{\sqrt{\lambda}} \right)^2}{\lambda \kappa}. \]
Then the result of Theorem B.3 holds with this new \( \beta \).

C Proofs

**Lemma 2.1.** For each \( f \in \mathcal{H}_x \), \( \int_X f(x) dP(x) = \langle f, \mu_{P_x} \rangle_{\mathcal{H}_x} \).

**Proof.** Let \( L_P \) be a functional on \( \mathcal{H} \) defined by \( L_P(f) := \int_X f(x) dP(x) \). Then \( L_P \) is clearly linear, and moreover,
\[ |L_P(f)| = \left| \int_X f(x) dP(x) \right| \]
\[ = \left| \int_X \langle f, k(x, \cdot) \rangle_{\mathcal{H}} dP(x) \right| \quad \text{by the reproducing property} \]
\[ \leq \int_X |\langle f, k(x, \cdot) \rangle_{\mathcal{H}}| dP(x) \quad \text{by Jensen's inequality} \]
\[ \leq \|f\|_{\mathcal{H}} \int_X \|k(x, \cdot)\|_{\mathcal{H}} dP(x) \quad \text{by Cauchy-Schwarz inequality.} \]
Since the map \( x \mapsto k(x, \cdot) \) is Bochner \( P \)-integrable, \( L_P \) is bounded, i.e. \( L_P \in \mathcal{H}^* \). So by the Riesz Representation Theorem, there exists a unique \( h \in \mathcal{H} \) such that \( L_P(f) = \langle f, h \rangle_{\mathcal{H}} \) for all \( f \in \mathcal{H} \).
Choose \( f(\cdot) = k(x, \cdot) \) for some \( x \in X \). Then
\[ h(x) = \langle k(x, \cdot), h \rangle_{\mathcal{H}} \]
\[ = L_P(k(x, \cdot)) \]
\[ = \int_X k(x', x) dP(x'), \]
which means \( h(\cdot) = \int_X k(x, \cdot) dP(x) = \mu_{P_x}(\cdot) \) (implicitly applying [12, Corollary 37]). \( \square \)

**Lemma 2.3.** For \( f \in \mathcal{H}_X, g \in \mathcal{H}_Y, \langle f \otimes g, \mu_{P_{XY}} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} = E_{XY} [f(X) g(Y)] \).

**Proof.** For Bochner integrability, we see that
\[ E_{XY} \left[ \|k_X(X, \cdot) \otimes k_Y(Y, \cdot)\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} \right] = E_{XY} \left[ \sqrt{k_X(X, X)} \sqrt{k_Y(Y, Y)} \right] \]
\[ \leq \sqrt{E_X [k_X(X, X)]} \sqrt{E_Y [k_Y(Y, Y)]}, \]
by Cauchy-Schwarz inequality. [2] now implies that \( k_X(X, \cdot) \otimes k_Y(Y, \cdot) \) is Bochner \( P_{XY} \)-integrable.

Let \( L_{P_{XY}} \) be a functional on \( \mathcal{H}_X \otimes \mathcal{H}_Y \) defined by \( L_{P_{XY}} \left( \sum_i f_i \otimes g_i \right) := E_{XY} \left[ \sum_i f_i(X) g_i(Y) \right] \). Then \( L_{P_{XY}} \) is clearly linear, and moreover,
\[ |L_{P_{XY}} \left( \sum_i f_i \otimes g_i \right) | = |E_{XY} \left( \sum_i f_i(X) g_i(Y) \right) | \]
\[ \leq E_{XY} \left[ \sum_i f_i(X) g_i(Y) \right] \quad \text{by Jensen's inequality} \]

17
Take an element $T \in \mathcal{H}_X \otimes \mathcal{H}_Y$. Then the map $T_{\varphi} = \{ \sum_{i=1}^{\infty} c_{i,j} (\varphi_i \otimes \psi_j) \}$ is an isometric isomorphism.

Hence, by Bochner integrability shown above, $L_{P_{XY}} \in (\mathcal{H}_X \otimes \mathcal{H}_Y)^*$. So by the Riesz Representation Theorem, there exists $h \in \mathcal{H}_X \otimes \mathcal{H}_Y$ such that $L_{P_{XY}} (\sum_i f_i \otimes g_i) = \langle \sum_i f_i \otimes g_i, h \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}$ for all $\sum_i f_i \otimes g_i \in \mathcal{H}_X \otimes \mathcal{H}_Y$.

Choose $k_X(x, \cdot) \otimes k_Y(y, \cdot) \in \mathcal{H}_X \otimes \mathcal{H}_Y$ for some $x \in X$ and $y \in Y$. Then
\[
h(x, y) = \langle k_X(x, \cdot) \otimes k_Y(y, \cdot), h \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}
= L_{P_{XY}} (k_X(x, \cdot) \otimes k_Y(y, \cdot))
= \mathbb{E}_{XY} [k_X(x, X) \otimes k_Y(y, Y)]
= \mu_{P_{XY}} (x, y),
\]
as required.

\[\square\]

**Lemma C.1.** Let $\{ \varphi_i \}_{i=1}^{\infty}$ and $\{ \psi_j \}_{j=1}^{\infty}$ be orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively (note that they are countable, since the RKHSs are separable). Then the map
\[
\Phi : \mathcal{H}_X \otimes \mathcal{H}_Y \to HS(\mathcal{H}_X, \mathcal{H}_Y)
\]
is an isometric isomorphism.

**Proof.** $\Phi$ is clearly linear. We first show isometry:
\[
\left\| \Phi \left( \sum_{i=1,j=1}^{\infty} c_{i,j} (\varphi_i \otimes \psi_j) \right) \right\|_{HS}^2
= \sum_{i=1,j=1}^{\infty} \left\| c_{i,j} (\varphi_i \otimes \psi_j) \right\|_{\mathcal{H}_X \otimes \mathcal{H}_Y}^2
= \sum_{i=1,j=1}^{\infty} c_{i,j}^2
\]
by orthonormality,
\[
= \left\| \sum_{i=1,j=1}^{\infty} c_{i,j} (\varphi_i \otimes \psi_j) \right\|_{\mathcal{H}_X \otimes \mathcal{H}_Y}^2
\]
by orthonormality,
as required. It remains to show surjectivity.

Take an element $T \in HS(\mathcal{H}_X, \mathcal{H}_Y)$. Then $T$ is completely determined by $\{ T \varphi_i \}_{i=1}^{\infty}$. For each $i$, suppose $T \varphi_i = \sum_{j=1}^{\infty} d_{ij} \psi_j$, with $d_{ij} \in \mathbb{R}$ for all $i$ and $j$. Then
\[
\Phi\left( \sum_{i'=1,j=1}^{\infty} d_{i'j} (\varphi_{i'} \otimes \psi_j) \right)
= \left[ \varphi_i \mapsto \sum_{i'=1,j=1}^{\infty} \langle d_{i'j}, \varphi_{i'}, \varphi_i \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \psi_j \right]
= \left[ \varphi_i \mapsto \sum_{j=1}^{\infty} d_{ij} \psi_j \right]
= T.
\]
So $\Phi$ is surjective, and hence an isometric isomorphism. \[\square\]
Before we prove Theorem 2.9, we state the following definition and theorems related to measurable functions for Banach-space valued functions.

**Definition C.2 ([12] p.4, Definition 5).** A function \( H : \Omega \to H \) is called an \( \mathcal{F} \)-simple function if it has the form \( H = \sum_{i=1}^{n} h_i 1_{B_i} \) for some \( h_i \in H \) and \( B_i \in \mathcal{F} \).

A function \( H : \Omega \to H \) is said to be \( \mathcal{F} \)-measurable if there is a sequence \((H_n)\) of \( H \)-valued, \( \mathcal{F} \)-simple functions such that \( H_n \to H \) pointwise.

**Theorem C.3 ([12] p.4, Theorem 6).** If \( H : \Omega \to H \) is \( \mathcal{F} \)-measurable, then there is a sequence \((H_n)\) of \( H \)-valued, \( \mathcal{F} \)-simple functions such that \( H_n \to H \) pointwise and \( |H_n| \leq |H| \) for every \( n \).

**Theorem C.4 ([12] p.19, Theorem 48), Lebesgue Convergence Theorem.** Let \((H_n)\) be a sequence in \( L_0^q(P) \), \( H : \Omega \to H \) a \( \mathcal{F} \)-measurable function, and \( g \in L_1^q(P) \) such that \( H_n \to H \) \( P \)-almost everywhere and \( |H_n| \leq g \), \( P \)-almost everywhere, for each \( n \). Then \( H \in L_0^q(P) \) and \( H_n \to H \) in \( L_0^q(P) \), i.e. \( \int_{\Omega} H_n dP \to \int_{\Omega} H dP \).

**Theorem 2.9.** Suppose that \( P(\cdot \mid \mathcal{E}) \) admits a regular version \( Q \). Then \( QH : \Omega \to H \) with \( \omega \mapsto Q(\omega,H) = \int_{\Omega} H(\omega')Q(\omega)(d\omega') \) is a version of \( \mathbb{E}[H | \mathcal{E}] \) for every Bochner \( P \)-integrable \( H \).

**Proof.** Suppose \( H \) is Bochner \( P \)-integrable. Since \( Q \) is a regular version of \( P(\cdot \mid \mathcal{E}) \), it is a probability transition kernel from \((\Omega, \mathcal{E})\) to \((\Omega, \mathcal{F})\).

We first show that \( QH \) is measurable with respect to \( \mathcal{E} \). The map \( Q : \Omega \to H \) is well-defined, since, for each \( \omega \in \Omega \), \( Q(\omega,H) \) is the Bochner-integral of \( H \) with respect to the measure \( B \to Q(\omega)(B) \). Since \( H \) is \( \mathcal{F} \)-measurable, by Theorem C.3 there is a sequence \((H_n)\) of \( H \)-valued, \( \mathcal{F} \)-simple functions such that \( H_n \to H \) pointwise. Then for each \( \omega \in \Omega \), \( Q(\omega,H) = \lim_{n \to \infty} Q(\omega,H_n) \) by Theorem C.4. But for each \( n \), we can write \( H_n = \sum_{j=1}^{m} h_j 1_{B_j} \) for some \( h_j \in H \) and \( B_j \in \mathcal{F} \), and so \( Q(\omega,H_n) = \sum_{j=1}^{m} h_j Q(\omega)(B_j) \). For each \( B_j \) the map \( \omega \mapsto Q(\omega)(B_j) \) is \( \mathcal{E} \)-measurable (by the definition of transition probability kernel, Definition 2.7), and so as a linear combination of \( \mathcal{E} \)-measurable functions, \( Q(\omega,H_n) \) is \( \mathcal{E} \)-measurable. Hence, as a pointwise limit of \( \mathcal{E} \)-measurable functions, \( QH \) is also \( \mathcal{E} \)-measurable, by [12] p.6, Theorem 10).

Next, we show that, for all \( A \in \mathcal{E} \), \( \int_{A} QH dP = \int_{A} QH_n dP \). Fix \( A \in \mathcal{E} \). By Theorem C.3 there is a sequence \((H_n)\) of \( H \)-valued, \( \mathcal{F} \)-simple functions such that \( H_n \to H \) pointwise. For each \( n \), we can write \( H_n = \sum_{j=1}^{m} h_j 1_{B_j} \) for some \( h_j \in H \) and \( B_j \in \mathcal{F} \), and

\[
\int_{A} QH_n dP = \int_{A} \sum_{j=1}^{m} h_j Q(B_j) dP = \int_{A} \sum_{j=1}^{m} h_j P(B_j \mid \mathcal{E}) dP \quad \text{since } Q \text{ is a version of } P(\cdot \mid \mathcal{E})
\]

\[
= \sum_{j=1}^{m} h_j \int_{A} \mathbb{E}[1_{B_j} \mid \mathcal{E}] dP \quad \text{by the definition of conditional probability measures}
\]

\[
= \int_{A} \sum_{j=1}^{m} h_j 1_{B_j} dP \quad \text{by the definition of conditional expectations, since } A \in \mathcal{E}
\]

\[
= \int_{A} H_n dP.
\]

We have \( H_n \to H \) pointwise by assertion, and as before, \( QH_n \to QH \) pointwise. Hence,

\[
\int_{A} QH dP = \lim_{n \to \infty} \int_{A} QH_n dP \quad \text{by Theorem C.4}
\]

\[
= \lim_{n \to \infty} \int_{A} H_n dP \quad \text{by above}
\]

\[
= \int_{A} H dP \quad \text{by Theorem C.4}
\]

Hence, by the definition of the conditional expectation, \( QH \) is a version of \( \mathbb{E}[H | \mathcal{E}] \).
Lemma 3.2. For any \( f \in \mathcal{H}_X, \mathbb{E}_{X|Z}[f(X) \mid Z] = \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_X} \) almost surely.

Proof. The left-hand side is the conditional expectation of the real-valued random variable \( f(X) \) given \( Z \). We need to check that the right-hand side is also that. Note that \( \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_X} \) is clearly \( Z \)-measurable, and \( P \)-integrable (by the Cauchy-Schwarz inequality and the integrability condition \([1]\)). Take any \( A \in \sigma(Z) \). Then

\[
\int_A \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_X} dP = \int_A \left\langle \frac{f, \mathbb{E}_{X|Z}[k_X(\cdot, X) \mid Z]}{\mathcal{H}_X} dP \right\rangle_{\mathcal{H}_X} dP \text{ by definition} \\
= \left\langle \frac{f, \mathbb{E}_{X|Z}[k_X(\cdot, X) \mid Z]}{\mathcal{H}_X} \right\rangle_{\mathcal{H}_X} dP \text{ (+)} \\
= \left\langle \frac{f, k_X(\cdot, X) dP}{\mathcal{H}_X} \right\rangle_{\mathcal{H}_X} \text{ see Definition 2.5} \\
= \int_A \langle f, k_X(\cdot, X) \rangle_{\mathcal{H}_X} dP \text{ (+)} \\
= \int_A f(X) dP \text{ by the reproducing property.}
\]

Here, in (+), we used the fact that the order of a continuous linear operator and Bochner integration can be interchanged \([12]\) p.30, Theorem 36]. Hence \( \langle f, \mu_{P_{X|Z}} \rangle_{\mathcal{H}_X} \) is a version of the conditional expectation \( \mathbb{E}_{X|Z}[f(X) \mid Z] \).

Lemma 3.3. For any pair \( f \in \mathcal{H}_X \) and \( g \in \mathcal{H}_Y, \mathbb{E}_{XY|Z}[f(X)g(Y) \mid Z] = \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \) almost surely.

Proof. The left-hand side is the conditional expectation of the real-valued random variable \( f(X)g(Y) \) given \( Z \). We need to check that the right-hand side is also that. Note that \( \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \) is clearly \( Z \)-measurable, and \( P \)-integrable (by the Cauchy-Schwarz inequality and the integrability condition \([2]\)). Take any \( A \in \sigma(Z) \). Then

\[
\int_A \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} dP = \int_A \left\langle \frac{f \otimes g, \mathbb{E}_{XY|Z}[k_X(\cdot, X) \otimes k_Y(\cdot, Y) \mid Z]}{\mathcal{H}_X \otimes \mathcal{H}_Y} dP \right\rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} dP \\
= \left\langle \frac{f \otimes g, \mathbb{E}_{XY|Z}[k_X(\cdot, X) \otimes k_Y(\cdot, Y) \mid Z]dP}{\mathcal{H}_X \otimes \mathcal{H}_Y} \right\rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} dP \\
= \left\langle \frac{f \otimes g, k_X(\cdot, X) \otimes k_Y(\cdot, Y) dP}{\mathcal{H}_X \otimes \mathcal{H}_Y} \right\rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} dP \\
= \int_A \langle f \otimes g, k_X(\cdot, X) \otimes k_Y(\cdot, Y) \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} dP \\
= \int_A f(X) g(Y) dP.
\]

So \( \langle f \otimes g, \mu_{P_{XY|Z}} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \) is a version of the conditional expectation \( \mathbb{E}_{XY|Z}[f(X)g(Y) \mid Z] \).

Theorem 4.1. Assume that \( \mathcal{H}_X \) is separable, and denote its Borel \( \sigma \)-algebra by \( \mathcal{B}(\mathcal{H}_X) \). Then we can write

\[\mu_{P_{X|Z}} = F_{P_{X|Z}} \circ Z,\]

where \( F_{P_{X|Z}} : Z \to \mathcal{H}_X \) is some deterministic function, measurable with respect to \( \mathfrak{F} \) and \( \mathcal{B}(\mathcal{H}_X) \).

Proof. Let \( \text{Im}(Z) \subseteq Z \) be the image of \( Z : \Omega \to Z \), and let \( \mathfrak{F} \) denote the \( \sigma \)-algebra on \( \text{Im}(Z) \) defined by \( \mathfrak{F} = \{ A \cap \text{Im}(Z) : A \in \mathfrak{F} \} \) (see \([9]\) page 5, 1.15]). We will first construct a function \( F : \text{Im}(Z) \to \mathcal{H}_X \), measurable with respect to \( \mathfrak{F} \) and \( \mathcal{B}(\mathcal{H}_X) \), such that \( \mu_{P_{X|Z}} = F \circ Z \).

For a given \( z \in \text{Im}(Z) \subseteq Z \), we have \( Z^{-1}(z) \subseteq \Omega \). Suppose for contradiction that there are two distinct elements \( \omega_1, \omega_2 \in Z^{-1}(z) \) such that \( \mu_{P_{X|Z}}(\omega_1) = \mu_{P_{X|Z}}(\omega_2) \). Since \( \mathcal{H}_X \) is Hausdorff,
there are disjoint open neighbourhoods $N_1$ and $N_2$ of $\mu_{P_{X|Z}}(\omega_1)$ and $\mu_{P_{X|Z}}(\omega_2)$ respectively. By definition of a Borel $\sigma$-algebra, we have $N_1, N_2 \in \mathcal{B}(\mathcal{H}_X)$, and since $\mu_{P_{X|Z}}$ is $\sigma(Z)$-measurable,

$$\mu_{P_{X|Z}}^{-1}(N_1), \mu_{P_{X|Z}}^{-1}(N_2) \in \sigma(Z). \quad (6)$$

Furthermore, $\mu_{P_{X|Z}}^{-1}(N_1)$ and $\mu_{P_{X|Z}}^{-1}(N_2)$ are neighbourhoods of $\omega_1$ and $\omega_2$ respectively, and are disjoint.

(i) For any $B \in \tilde{\mathcal{F}}$ with $z \in B$, since $Z(\omega_1) = z = Z(\omega_2)$, we have $\omega_1, \omega_2 \in Z^{-1}(B)$. So $Z^{-1}(B) \neq \mu_{P_{X|Z}}^{-1}(N_1)$ and $Z^{-1}(B) \neq \mu_{P_{X|Z}}^{-1}(N_2)$, as $\omega_2 \notin \mu_{P_{X|Z}}^{-1}(N_1)$ and $\omega_1 \notin \mu_{P_{X|Z}}^{-1}(N_2)$.

(ii) For any $B \in \tilde{\mathcal{F}}$ with $z \notin B$, we have $\omega_1 \notin Z^{-1}(B)$ and $\omega_2 \notin Z^{-1}(B)$. So $Z^{-1}(B) \neq \mu_{P_{X|Z}}^{-1}(N_1)$ and $Z^{-1}(B) \neq \mu_{P_{X|Z}}^{-1}(N_2)$.

Since $\sigma(Z) = \{Z^{-1}(B) \mid B \in \tilde{\mathcal{F}}\}$ (see [9], page 11, Exercise 2.20), we can’t have $\mu_{P_{X|Z}}^{-1}(N_1) \in \sigma(Z)$ nor $\mu_{P_{X|Z}}^{-1}(N_2) \in \sigma(Z)$. This is a contradiction to (6). We therefore conclude that, for any $z \in \mathcal{Z}$, if $Z(\omega_1) = z = Z(\omega_2)$ for distinct $\omega_1, \omega_2 \notin \Omega$, then $\mu_{P_{X|Z}}(\omega_1) = \mu_{P_{X|Z}}(\omega_2)$.

We define $\tilde{F}(z)$ to be the unique value of $\mu_{P_{X|Z}}(\omega)$ for all $\omega \in Z^{-1}(z)$. Then for any $\omega \in \Omega$, $\mu_{P_{X|Z}}(\omega) = \tilde{F}(Z(\omega))$ by construction. It remains to check that $\tilde{F}$ is measurable with respect to $\tilde{\mathcal{F}}$ and $\mathcal{B}(\mathcal{H}_X)$.

Take any $N \in \mathcal{B}(\mathcal{H}_X)$. Since $\mu_{P_{X|Z}}$ is $\sigma(Z)$-measurable, $\mu_{P_{X|Z}}^{-1}(N) = Z^{-1}(\tilde{F}^{-1}(N)) \in \sigma(Z)$.

Since $\sigma(Z) = \{Z^{-1}(B) \mid B \in \tilde{\mathcal{F}}\}$, we have $Z^{-1}(\tilde{F}^{-1}(N)) = Z^{-1}(C)$ for some $C \in \tilde{\mathcal{F}}$. Since the mapping $Z : \Omega \to \text{Im}(Z)$ is surjective, $\tilde{F}^{-1}(N) = C$. Hence $\tilde{F}^{-1}(N) \in \tilde{\mathcal{F}}$, and so $\tilde{F}$ is measurable with respect to $\tilde{\mathcal{F}}$ and $\mathcal{B}(\mathcal{H}_X)$.

Finally, we can extend $\tilde{F} : \text{Im}(Z) \to \mathcal{H}_X$ to $F : Z \to \mathcal{H}_X$ by [13] page 128, Corollary 4.2.7] (note that $\mathcal{H}_X$ is a complete metric space, and assumed to be separable in this theorem). \qed

**Theorem 4.2.** $F_{P_{X|Z}} \in L^2(Z, P_Z; \mathcal{H}_X)$ minimises both $\tilde{\mathcal{E}}_{X|Z}$ and $\mathcal{E}_{X|Z}$, i.e.

$$F_{P_{X|Z}} = \arg \min_{F \in L^2(Z, P_Z; \mathcal{H}_X)} \mathcal{E}_{X|Z}(F) = \arg \min_{F \in L^2(Z, P_Z; \mathcal{H}_X)} \tilde{\mathcal{E}}_{X|Z}(F).$$

Moreover, it is almost surely unique, i.e. it is almost surely equal to any other minimiser of the objective functionals.

**Proof.** Recall that we have

$$\mathcal{E}_{X|Z}(F) := \mathbb{E}_Z \left[ \|F_{P_{X|Z}}(Z) - F(Z)\|_{\mathcal{H}_X}^2 \right].$$

So clearly, $\mathcal{E}_{X|Z}(F_{P_{X|Z}}) = 0$, meaning $F_{P_{X|Z}}$ minimises $\mathcal{E}_{X|Z}$ in $L^2(Z, P_Z; \mathcal{H}_X)$. So it only remains to show that $\tilde{\mathcal{E}}_{X|Z}$ is minimised in $L^2(Z, P_Z; \mathcal{H}_X)$ by $F_{P_{X|Z}}$.

Let $F$ be any element in $L^2(Z, P_Z; \mathcal{H}_X)$. Then we have

$$\tilde{\mathcal{E}}_{X|Z}(F) - \tilde{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) = \mathbb{E}_{X,Z}[\|k_X(X, \cdot) - F(Z)\|_{\mathcal{H}_X}^2] - \mathbb{E}_{X,Z}[\|k_X(X, \cdot) - F_{P_{X|Z}}(Z)\|_{\mathcal{H}_X}^2]$$

$$= \mathbb{E}_{Z}[\|F(Z)\|_{\mathcal{H}_X}^2] - 2\mathbb{E}_{X,Z}[\langle k_X(X, \cdot), F(Z) \rangle_{\mathcal{H}_X}] + 2\mathbb{E}_{X,Z}[\langle k_X(X, \cdot), F_{P_{X|Z}}(Z) \rangle_{\mathcal{H}_X}] - \mathbb{E}_{Z}[\|F_{P_{X|Z}}(Z)\|_{\mathcal{H}_X}^2].$$

Here,

$$\mathbb{E}_{X,Z}[\langle k_X(X, \cdot), F(Z) \rangle_{\mathcal{H}_X}] = \mathbb{E}_{Z}\left[\mathbb{E}_{X|Z}[F(Z)(X) \mid Z]\right]$$

by the reproducing property.
which immediately implies that

$$\| \mathcal{F}(Z) \|_{\mathcal{H}} = \mathbb{E}_Z \left[ \langle F(Z), \mu_{P_{X|Z}} \rangle_{\mathcal{H}_X} \right]$$

by Lemma 3.2

and similarly,

$$\mathbb{E}_{X,Z}[\langle k_X(X,\cdot), F_{P_{X|Z}}(Z) \rangle_{\mathcal{H}}] = \mathbb{E}_Z[\mathbb{E}_{X}[F_{P_{X|Z}}(Z)(X) \mid Z]]$$

by the reproducing property

$$= \mathbb{E}_Z \left[ \langle F_{P_{X|Z}}(Z), F_{P_{X|Z}}(Z) \rangle_{\mathcal{H}_X} \right]$$

by Lemma 3.2

$$= \mathbb{E}_Z \left[ \| F_{P_{X|Z}}(Z) \|_{\mathcal{H}_X}^2 \right].$$

Substituting these expressions back into (7), we have

$$\hat{\mathcal{E}}_{X|Z}(F) - \mathcal{E}_{X|Z}(F_{P_{X|Z}})$$

$$= \mathbb{E}_Z[\| F(Z) \|_{\mathcal{H}_X}^2] - 2\mathbb{E}_Z[\| F(Z), F_{P_{X|Z}}(Z) \rangle_{\mathcal{H}_X}] + \mathbb{E}_Z[\| F_{P_{X|Z}}(Z) \|_{\mathcal{H}_X}^2]$$

$$= \mathbb{E}_Z[\| F(Z) - F_{P_{X|Z}}(Z) \|_{\mathcal{H}_X}^2]$$

$$\geq 0.$$

Hence, $F_{P_{X|Z}}$ minimises $\hat{\mathcal{E}}_{X|Z}$ in $L^2(\mathcal{Z}, P_Z; \mathcal{H}_X)$. The minimiser is further more $P_Z$-almost surely unique; indeed, if $F' \in L^2(\mathcal{Z}, P_Z; \mathcal{H}_X)$ is another minimiser of $\hat{\mathcal{E}}_{X|Z}$, then the calculation in (7) shows that

$$\mathbb{E}_Z \left[ \| F_{P_{X|Z}}(Z) - F'(Z) \|_{\mathcal{H}_X}^2 \right] = 0,$$

which immediately implies that $\| F_{P_{X|Z}}(Z) - F'(Z) \|_{\mathcal{H}_X} = 0$ $P_Z$-almost surely, which in turn implies that $P_{X|Z} = F' P_Z$-almost surely.

**Theorem 4.4.** Suppose that $k_X$ and $k_Z$ are bounded kernels, i.e. there exist $B_Z, B_X > 0$ such that $\sup_{z \in Z} k_Z(z, z') \leq B_Z$ and $\sup_{x \in X} k_X(x, x') \leq B_X$, and that the operator-valued kernel $l_{X,Z}$ is $C_0$-universal. Let the regularisation parameter $\lambda_n$ decay to 0 at a slower rate than $O(n^{-1/2})$. Then our learning algorithm that produces $\hat{F}_{P_{X|Z},n,\lambda_n}$ is universally consistent (in the surrogate loss $\hat{\mathcal{E}}_{X|Z}$), i.e. for any joint distribution $P_{X,Z}$ and constants $\epsilon > 0$ and $\delta > 0$,

$$P_{X,Z}(\hat{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \hat{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) > \epsilon) < \delta$$

for large enough $n$.

**Proof.** Follows immediately from [7], Theorem 2.3.

**Theorem 4.5.** In addition to the setting in Theorem 4.4, assume that $F_{P_{X|Z}} \in \mathcal{G}_{X,Z}$. Let the regularisation parameter $\lambda_n$ decay to 0 with rate $O(n^{-1/4})$. Then $\hat{\mathcal{E}}_{X|Z}(\hat{F}_{P_{X|Z},n,\lambda_n}) - \hat{\mathcal{E}}_{X|Z}(F_{P_{X|Z}}) = O_P(n^{-1/4})$.

**Proof.** Follows immediately from [7], Theorem 2.4.

**Theorem 5.2.** Suppose that $k_X$ is a characteristic kernel, that $P_Z$ and $P_{Z'}$ are absolutely continuous with respect to each other, and that $P(\cdot \mid Z)$ and $P(\cdot \mid Z')$ admit regular versions. Then $MCMD_{P_{X|Z}Z'} \neq 0$ if and only if, for $P_Z$ (or $P_{Z'}$)-almost all $z \in Z$, $P_{X|Z}(X \mid z) = P_{X'|Z}(X' \mid z)$ for all $B \in \mathcal{X}$.

**Proof.** Write $Q$ and $Q'$ for some regular versions of $P(\cdot \mid Z)$ and $P(\cdot \mid Z')$ respectively, and assume without loss of generality that the conditional distributions $P_{X|Z}$ and $P_{X'|Z'}$ are given by $P_{X|Z}(\omega)(B) = Q_{\omega}(X \in B)$ and $P_{X'|Z}(\omega)(B) = Q'_{\omega}(X' \in B)$ for $B \in \mathcal{X}$. By the definition of regular versions, for each $B \in \mathcal{X}$, the real-valued random variables $\omega \mapsto P_{X|Z}(\omega)(B)$ and $\omega \mapsto P_{X'|Z}(\omega)(B)$ are measurable with respect to $Z$ and $Z'$ respectively, and so there are functions $R_B : \mathcal{Z} \to \mathbb{R}$ and $R'_B : \mathcal{Z} \to \mathbb{R}$ such that $P_{X|Z}(\omega)(B) = R_B(\omega)$ and $P_{X'|Z}(\omega)(B) = R'_B(\omega)$.
$R'_B(Z'(\omega))$. Moreover, for each fixed $z \in \mathbb{Z}$, the mappings $B \mapsto P_{X|Z}(Z^{-1}(z))(B) = R_B(z)$ and $B \mapsto P_{X'|Z'}(Z'^{-1}(z))(B) = R'_B(z)$ are measures. We write $R_B(z) = P_{X|Z = z}(B)$ and $R'_B(z) = P_{X'|Z' = z}(B)$.

By Theorem 2.9 there exists an event $A_1 \in \mathcal{F}$ with $P(A_1) = 1$ such that for all $\omega \in A_1$,

$$
\mu_{P_{X|Z}}(\omega) := \mathbb{E}_{X|Z}[k_X(X, \cdot) | Z](\omega) = \int_{\mathcal{X}} k_X(X(\omega'), \cdot) q_{\omega'}(d\omega') = \int_{\mathcal{X}} k_X(x, \cdot) P_{X|Z}(\omega)(dx),
$$

and an event $A_2 \in \mathcal{F}$ with $P(A_2) = 1$ such that for all $\omega \in A_2$,

$$
\mu_{P_{X'|Z'}}(\omega) := \mathbb{E}_{X'|Z'}[k_X'(X', \cdot) | Z'](\omega) = \int_{\mathcal{X}} k_X'(x', \cdot) q_{\omega'}(d\omega') = \int_{\mathcal{X}} k_X(x', \cdot) P_{X'|Z'}(\omega)(dx').
$$

Suppose for contradiction that there exists some $D \in \mathcal{F}$ with $P_Z(D) > 0$ such that for all $z \in D$, $F_{P_{X|Z}}(z) \neq \int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z)$. Then $P(Z^{-1}(D)) = P_Z(D) > 0$, and hence $P(Z^{-1}(D) \cap A_1) > 0$. For all $\omega \in Z^{-1}(D) \cap A_1$, we have $Z(\omega) \in D$, and hence

$$
\mu_{P_{X|Z}}(\omega) = F_{P_{X|Z}}(Z(\omega)) \neq \int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(Z(\omega)) = \int_{\mathcal{X}} k_X(x, \cdot) P_{X|Z}(\omega)(dx).
$$

This contradicts our assertion that $\mu_{P_{X|Z}}(\omega) = \int_{\mathcal{X}} k_X(x, \cdot) P_{X|Z}(\omega)(dx)$ for all $\omega \in A_1$, hence there does not exist $D \in \mathcal{F}$ with $P_Z(D) > 0$ such that for all $z \in D$, $F_{P_{X|Z}}(z) \neq \int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z)$. Therefore, there must exist some $C_1 \in \mathcal{F}$ with $P_Z(C_1) = 1$ such that for all $z \in C_1$, $F_{P_{X|Z}}(z) = \int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z)$. Similarly, there must exist some $C_2 \in \mathcal{F}$ with $P_Z(C_2) = 1$ such that for all $z \in C_2$, $F_{P_{X'|Z'}}(z) = \int_{\mathcal{X}} k_X(x, \cdot) R'_{dz}(z)$. Since $P_Z$ and $P_{X'}$ are absolutely continuous with respect to each other, we also have $P_Z(C_2) = 1 = P_{X'}(C_1)$.

($\implies$) Suppose first that $\text{MCMD}_{P_{X|Z}, P_{X'|Z'}} = \|F_{P_{X|Z}} - F_{P_{X'|Z'}}\|_{\mathcal{H}_X} = 0$ $P_Z$-almost everywhere, i.e. there exists $C \in \mathcal{F}$ with $P_Z(C) = 1$ such that for all $z \in C$, $\|F_{P_{X|Z}}(z) - F_{P_{X'|Z'}}(z)\|_{\mathcal{H}_X} = 0$. Then for each $z \in C \cap C_1 \cap C_2$,

$$
\int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z) = F_{P_{X|Z}}(z) = F_{P_{X'|Z'}}(z) = \int_{\mathcal{X}} k_X(x, \cdot) R'_{dz}(z)
$$

since $z \in C_1$.

Since the kernel $k_X$ is characteristic, this means that $B \mapsto R_B(z)$ and $B \mapsto R'_B(z)$ are the same probability measure on $(\mathcal{X}, \mathcal{X})$. By countable intersection, we have $P_Z(C \cap C_1 \cap C_2) = 1$, so $P_Z$-almost everywhere,

$$
P_{X|Z = z}(B) = P_{X'|Z' = z}(B)
$$

for all $B \in \mathcal{X}$.

($\Longleftarrow$) Now assume there exists $C \in \mathcal{F}$ with $P_Z(C) = 1$ such that for each $z \in C$, $R_B(z) = R'_B(z)$ for all $B \in \mathcal{X}$. Then for all $z \in C \cap C_1 \cap C_2$,

$$
\|F_{P_{X|Z}}(z) - F_{P_{X'|Z'}}(z)\|_{\mathcal{H}_X} = \|\int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z) - \int_{\mathcal{X}} k_X(x, \cdot) R'_{dz}(z)\|_{\mathcal{H}_X} = 0, \quad \text{since } z \in C_1 \cap C_2,
$$

$$
\|\int_{\mathcal{X}} k_X(x, \cdot) R_{dz}(z) - \int_{\mathcal{X}} k_X(x, \cdot) R'_{dz}(z)\|_{\mathcal{H}_X} = 0, \quad \text{since } z \in C
$$

and since $P_Z(C \cap C_1 \cap C_2) = 1$, $\|F_{P_{X|Z}} - F_{P_{X'|Z'}}\|_{\mathcal{H}_X} = 0$ $P_Z$-almost everywhere.
Theorem 5.4. Suppose \( k_X \otimes k_Y \) is a characteristic kernel on \( \mathcal{X} \times \mathcal{Y} \), and that \( P(\cdot \mid Z) \) admits a regular version. Then HSCIC\((X, Y \mid Z) = 0 \) almost surely if and only if \( X \perp Y \mid Z \).

Proof. Write \( Q \) for a regular version of \( P(\cdot \mid Z) \), and assume without loss of generality that the conditional distributions \( P_X|Z \), \( P_Y|Z \) and \( P_{XY}|Z \) are given by \( P_X|Z(\omega)(B) = Q(\omega)(X \in B) \) for \( B \in \mathcal{X} \), \( P_Y|Z(\omega)(C) = Q(\omega)(Y \in C) \) for \( C \in \mathcal{Y} \) and \( P_{XY}|Z(\omega)(D) = Q(\omega)((X, Y) \in D) \) for \( D \in \mathcal{X} \times \mathcal{Y} \). By Theorem 2.9, there exists an event \( A_1 \in \mathcal{F} \) with \( P(A_1) = 1 \) such that for all \( \omega \in A_1 \),

\[
\mu_{P|Z}(\omega) := \mathbb{E}_{X|Z}[k_X(X, \cdot) \mid Z](\omega) = \int_{X} k_X(X(\omega'), \cdot)Q_{\omega}(d\omega') = \int_{X} k_X(x, \cdot)P_X|Z(\omega)(dx),
\]

an event \( A_2 \in \mathcal{F} \) with \( P(A_2) = 1 \) such that for all \( \omega \in A_2 \),

\[
\mu_{P|Y}(\omega) := \mathbb{E}_{Y|Z}[k_Y(Y, \cdot) \mid Z](\omega) = \int_{Y} k_Y(Y(\omega'), \cdot)Q_{\omega}(d\omega') = \int_{Y} k_Y(y, \cdot)P_Y|Z(\omega)(dy),
\]

and an event \( A_3 \in \mathcal{F} \) with \( P(A_3) = 1 \) such that for all \( \omega \in A_3 \),

\[
\mu_{P|Z}(\omega) = \int_{X \times Y} k_X(x, \cdot) \otimes k_Y(y, \cdot)P_{XY}|Z(\omega)(d(x, y)).
\]

This means that, for each \( \omega \in A_1 \), \( \mu_{P_X|Z}(\omega) \) is the mean embedding of \( P_X|Z(\omega) \), and for each \( \omega \in A_2 \), \( \mu_{P_Y|Z}(\omega) \) is the mean embedding of \( P_Y|Z(\omega) \).

(\( \implies \)) Suppose first that HSCIC\((X, Y \mid Z) = \|\mu_{P_{XY}|Z} - \mu_{P_X|Z} \otimes \mu_{P_Y|Z}\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} = 0 \) almost surely, i.e. there exists \( A \in \mathcal{F} \) with \( P(A) = 1 \) such that for all \( \omega \in A \), \( \|\mu_{P_{XY}|Z}(\omega) - \mu_{P_X|Z}(\omega) \otimes \mu_{P_Y|Z}(\omega)\|_{\mathcal{H}_X \otimes \mathcal{H}_Y} = 0 \). Then for each \( \omega \in A \cap A_1 \cap A_2 \cap A_3 \),

\[
\int_{X \times Y} k_X(x, \cdot) \otimes k_Y(y, \cdot)P_{XY}|Z(\omega)(d(x, y)) = \mu_{P_{XY}|Z}(\omega) \quad \text{since } \omega \in A_3
\]

\[
= \mu_{P_X|Z}(\omega) \otimes \mu_{P_Y|Z}(\omega) \quad \text{since } \omega \in A
\]

\[
= \int_{X} k_X(x, \cdot)P_X|Z(\omega)(dx) \otimes \int_{Y} k_Y(y, \cdot)P_Y|Z(\omega)(dy) \quad \text{since } \omega \in A_1 \cap A_2
\]

\[
= \int_{X \times Y} k_X(x, \cdot) \otimes k_Y(y, \cdot)P_{XY}|Z(\omega)(d(x, y)) \quad \text{by Fubini.}
\]

Since the kernel \( k_X \otimes k_Y \) is characteristic, the distributions \( P_{XY}|Z(\omega) \) and \( P_X|Z(\omega)P_Y|Z(\omega) \) on \( X \times Y \) are the same. By countable intersection, we have \( P(A \cap A_1 \cap A_2 \cap A_3) = 1 \), so \( P_{XY}|Z \) and \( P_X|ZP_Y|Z \) are the same almost surely, and we have \( X \perp Y \mid Z \).

(\( \iff \)) Now assume \( X \perp Y \mid Z \), i.e. there exists \( A \in \mathcal{F} \) with \( P(A) = 1 \) such that for each \( \omega \in A \), the distributions \( P_{XY}|Z(\omega) \) and \( P_X|Z(\omega)P_Y|Z(\omega) \) are the same. Then for all \( \omega \in A \cap A_1 \cap A_2 \cap A_3 \),

\[
\mu_{P_{XY}|Z}(\omega) = \int_{X \times Y} k_X(x, \cdot) \otimes k_Y(y, \cdot)P_{XY}|Z(\omega)(d(x, y)) \quad \text{since } \omega \in A_3
\]

\[
= \int_{X \times Y} k_X(x, \cdot) \otimes k_Y(y, \cdot)P_{XY}|Z(\omega)(dx)P_Y|Z(\omega)(dy) \quad \text{since } \omega \in A
\]

\[
= \int_{X} k_X(x, \cdot)P_X|Z(\omega)(dx) \otimes \int_{Y} k_Y(y, \cdot)P_Y|Z(\omega)(dy) \quad \text{by Fubini.}
\]

\[
= \mu_{P_{XY}|Z}(\omega) \otimes \mu_{P_Y|Z}(\omega) \quad \text{since } \omega \in A_1 \cap A_2
\]

and since \( P(A \cap A_1 \cap A_2 \cap A_3) = 1 \), HSCIC\((X, Y \mid Z) = 0 \) almost surely.