# SUPPLEMENTARY MATERIALS: A Rigorous Theory of Conditional Mean Embeddings*

Ilja Klebanov†, Ingmar Schuster‡, and T. J. Sullivan§

**SM1. Technical Results.** This section contains several technical results used in the proofs of the theorems given in the article. The following well-known result due to [SM10, Theorem 1] (see also [SM11, Theorem 2.1]) is used several times:

Theorem SM1.1. *Let $\mathcal{H}$, $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces and let $A\colon \mathcal{H}_1 \to \mathcal{H}$ and $B\colon \mathcal{H}_2 \to \mathcal{H}$ be bounded linear operators with $\operatorname{ran} A \subseteq \operatorname{ran} B$. Then $Q := B^\dagger A\colon \mathcal{H}_1 \to \mathcal{H}_2$ is a well-defined bounded linear operator, where $B^\dagger$ denotes the Moore–Penrose pseudo-inverse of $B$. It is the unique operator that satisfies the conditions*

$$A = BQ, \qquad \ker Q = \ker A, \qquad \operatorname{ran} Q \subseteq \overline{\operatorname{ran} B^*}. \tag{SM1.1}$$

*Remark* SM1.2. In the original work of [SM10] only the existence of a bounded operator $Q$ such that $A = BQ$ was shown. However, the construction of $Q$ in the proof is identical to that of $B^\dagger$ (multiplied by $A$). This connection has been observed before by [SM1, Corollary 2.2 and Remark 2.3], where it was proven in the case of closed range operators, leaving the proof of the general case to the reader.

The following result partially generalises [SM9, Proposition 4.1]:

Lemma SM1.3. *Let $\mathcal{H}$ be a separable Hilbert space, let $\mathcal{G}$ be an RKHS over $\mathcal{Y}$ with canonical feature map $\psi$, and suppose that $\mathcal{G}$ is a subset of $\mathcal{L}^2(\nu)$, where $\nu$ is a $\sigma$-finite measure on $\mathcal{Y}$. Then any bounded linear operator $A\colon \mathcal{H} \to \mathcal{G}$ is Hilbert–Schmidt as an operator $A\colon \mathcal{H} \to L^2(\nu)$.*

*Proof.* Let $h \in \mathcal{H}$ and $y \in \mathcal{Y}$. Then $(Ah)(y) = \langle \psi(y), Ah\rangle_{\mathcal{G}} = \langle A^*\psi(y), h\rangle_{\mathcal{H}}$. Thus $A$ is a Carleman operator and the claim follows from [SM20, Theorem 6.15]. ∎

The following results are used in the proofs of Sections 3, 4, and 7. Note that Lemma SM1.4 is essentially one direction of Proposition 5 in [SM13], but does not require $k$ to be bounded, which makes a separate proof necessary.

Lemma SM1.4. *Under Assumption 2.1, if $k$ is a characteristic kernel, then $\mathcal{H}_C$ is dense in $L^2_C(\mathbb{P}_X)$.*

*Submitted to the editors December 6, 2019.

**Funding:** The research presented here was supported in part by the German Research Foundation (Deutsche Forschungsgemeinschaft) through project TrU-2 "Demand modelling and control for e-commerce using RKHS transfer operator approaches" of the Excellence Cluster "MATH+ The Berlin Mathematics Research Centre" (EXC-2046/1, project ID: 390685689). The authors also wish to thank S. Klus, H. C. Lie, M. Mollenhauer, and B. Sprungk for helpful and collegial discussions.

†Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany (klebanov@zib.de).

‡Zalando SE, 11501 Berlin, Germany (ingmar.schuster@zalando.de).

§Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany (t.j.sullivan@fu-berlin.de); Mathematics Institute and School of Engineering, The University of Warwick, Coventry, CV4 7AL, United Kingdom (t.j.sullivan@warwick.ac.uk); and Zuse Institute Berlin, Takustraße 7, 14195 Berlin, Germany (sullivan@zib.de).

*Proof.* Suppose that $\mathcal{H}_\mathcal{C}$ is not dense in $L^2_\mathcal{C}(\mathbb{P}_X)$. Then there exists $f \in L^2(\mathbb{P}_X)$ that is not $\mathbb{P}_X$-a.e. constant such that $[f] \perp_{L^2_\mathcal{C}(\mathbb{P}_X)} \mathcal{H}_\mathcal{C}$. Choose $\tilde{f} \coloneqq f - \mathbb{E}[f(X)]$ and set

$$Q_1(E) \coloneqq \int_E |\tilde{f}| \,\mathrm{d}\mathbb{P}_X, \qquad Q_2(E) \coloneqq \int_E \big(|\tilde{f}| - \tilde{f}\big) \,\mathrm{d}\mathbb{P}_X$$

for every Borel-measurable subset $E \subseteq \mathcal{X}$. Since $\|\tilde{f}\|_{L^1(\mathbb{P}_X)} \neq 0$, we may assume without loss of generality that $\|\tilde{f}\|_{L^1(\mathbb{P}_X)} = 1$, making $Q_1$ and $Q_2$ two *distinct* probability distributions. Since, for every $h \in \mathcal{H}$,

$$\langle \tilde{f}, h\rangle_{L^2(\mathbb{P}_X)} = \langle f - \mathbb{E}[f(X)], h\rangle_{L^2(\mathbb{P}_X)} \overset{[f]\perp\mathcal{H}_\mathcal{C}}{=} \langle f - \mathbb{E}[f(X)], \mathbb{E}[h(X)]\rangle_{L^2(\mathbb{P}_X)} = 0,$$

it follows that $\tilde{f} \perp_{L^2(\mathbb{P}_X)} \mathcal{H}$. Let $Z_1 \sim Q_1$ and $Z_2 \sim Q_2$ and $x \in \mathcal{X}$. Since $\varphi(x) \in \mathcal{H}$,

$$\big(\mathbb{E}[\varphi(Z_1)] - \mathbb{E}[\varphi(Z_2)]\big)(x) = \langle \tilde{f}, \varphi(x)\rangle_{L^2(\mathbb{P}_X)} = 0,$$

which contradicts the assumption that $k$ is characteristic. Note that, by Assumption 2.1, $\mathbb{E}[\varphi(Z_1)]$ and $\mathbb{E}[\varphi(Z_2)]$ are well defined. In fact, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\big[\|\varphi(Z_1)\|_\mathcal{H}\big] = \int_\mathcal{X} \|\varphi(x)\|_\mathcal{H} \,|\tilde{f}(x)| \,\mathrm{d}\mathbb{P}_X(x) \leq \mathbb{E}\big[\|\varphi(X)\|_\mathcal{H}^2\big]^{1/2} \mathbb{E}\big[\tilde{f}(X)^2\big]^{1/2} < \infty$$

and similarly for $Z_2$. ∎

Lemma SM1.5. *Under Assumption 2.1,* $\ker C_X = \{h \in \mathcal{H} \mid h \text{ is } \mathbb{P}_X\text{-a.e. constant in } \mathcal{X}\}$ *and* $\ker {}^uC_X = \{h \in \mathcal{H} \mid h = 0 \ \mathbb{P}_X\text{-a.e. in } \mathcal{X}\}$.

*Proof.* This is a direct consequence of the facts that $\langle h, C_X h\rangle_\mathcal{H} = \mathbb{V}[h(X)]$ and that $\langle h, {}^uC_X h\rangle_\mathcal{H} = \|h\|_{L^2(\mathbb{P}_X)}$. ∎

Lemma SM1.6. *Under Assumption 2.1, for all* $h \in \mathcal{H}$ *and* $g \in \mathcal{G}$,

$$\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \langle h, C_{XY} g\rangle_\mathcal{H}, \qquad {}^u\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \langle h, {}^uC_{XY} g\rangle_\mathcal{H}.$$

*Proof.* Let $h \in \mathcal{H}$ and $g \in \mathcal{G}$ be arbitrary. Then

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[h(X), f_g(X)] &= \mathbb{E}\big[h(X)\mathbb{E}[g(Y)|X]\big] - \mathbb{E}[h(X)]\mathbb{E}\big[\mathbb{E}[g(Y)|X]\big] \\
&= \mathbb{E}[h(X)g(Y)] - \mathbb{E}[h(X)]\mathbb{E}[g(Y)] \\
&= \mathbb{C}\mathrm{ov}[h(X), g(Y)] \\
&= \langle h, C_{XY} g\rangle_\mathcal{H},
\end{aligned}$$

as required. The second statement is proved analogously using uncentred covariance operators and without subtracting the (products of) expected values. ∎

Lemma SM1.7. *Under Assumption 2.1, let* $A\colon \mathcal{G} \to \mathcal{H}$ *be a bounded linear operator. Then, for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$,

$$(A\psi(y))(x) = (A^*\varphi(x))(y), \qquad \mathbb{E}[(A\psi(y))(X)] = (A^*\mu_X)(y).$$

*Proof.* By the reproducing properties of $\psi$, $\varphi$, and $\mu_X$,

$$(A\psi(y))(x) = \langle A\psi(y), \varphi(x)\rangle_{\mathcal{H}} = \langle \psi(y), A^*\varphi(x)\rangle_{\mathcal{G}} = (A^*\varphi(x))(y),$$
$$\mathbb{E}[(A\psi(y))(X)] = \langle A\psi(y), \mu_X\rangle_{\mathcal{H}} = \langle \psi(y), A^*\mu_X\rangle_{\mathcal{G}} = (A^*\mu_X)(y),$$

as claimed. ∎

Lemma SM1.8. *Let $V$ be a Hilbert space, let $U_1 \subseteq U_2 \subseteq \cdots$ be an increasing sequence of closed subspaces $U_n \subseteq V$, $n \in \mathbb{N}$, and let $U \coloneqq \bigcup_{n\in\mathbb{N}} U_n$. Further, let $P_{U_n} \colon V \to U_n$ denote the orthogonal projection onto $U_n$. Then, for all $v \in \overline{U}$,*

$$P_{U_n} v \xrightarrow[n\to\infty]{} v.$$

*Proof.* Let $v \in \overline{U}$ and $\varepsilon > 0$. Then there exists $u \in U$ such that $\|u - v\| < \varepsilon$. Since the sequence $(U_n)_{n\in\mathbb{N}}$ is increasing and $U$ is its union, there exists an $n_0 \in \mathbb{N}$ such that $u \in U_n$ and thereby $P_{U_n} u = u$ for all $n \geq n_0$. We therefore obtain, for $n \geq n_0$,

$$\|P_{U_n} v - v\| \leq \|P_{U_n} v - P_{U_n} u\| + \|P_{U_n} u - u\| + \|u - v\| \leq \|P_{U_n}\|\|v - u\| + \|u - v\| < 2\varepsilon,$$

by the triangle inequality and non-expansivity of orthogonal projection. ∎

Lemma SM1.9. *Under Assumption 2.1, with $\mathcal{H}^{(n)}$, $C^{(n)}$, $h_g^{(n)}$ as in Theorem 4.4 and ${}^u C^{(n)}$, ${}^u h_g^{(n)}$ as in Theorem 5.4,*

*(a)* $\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \lim_{n\to\infty} \mathbb{C}\mathrm{ov}[h(X), h_g^{(n)}(X)]$ *for all $h \in \mathcal{H}$;*

*(b)* $[h_g^{(n)}]$ *is the $L^2_{\mathcal{C}}$-orthogonal projection of $[f_g]$ onto $\mathcal{H}_{\mathcal{C}}^{(n)}$ for all $g \in \mathcal{G}$;*

*(c)* ${}^u\mathbb{C}\mathrm{ov}[h(X), f_g(X)] = \lim_{n\to\infty} {}^u\mathbb{C}\mathrm{ov}[h(X), {}^u h_g^{(n)}(X)]$ *for all $h \in \mathcal{H}$;*

*(d)* ${}^u h_g^{(n)}$ *is the $L^2$-orthogonal projection of $f_g$ onto $\mathcal{H}^{(n)}$ for all $g \in \mathcal{G}$.*

*Proof.* We only give the proofs of (a) and (b); (c) and (d) can be proven similarly. It is clear that that $C^{(n)} \to C$ (in the strong and thereby in the weak sense) as $n \to \infty$ and that $C_X$ and $C_X^{(n)}$ agree on $\mathcal{H}^{(n)} \ni h_g^{(n)}$. Using Lemma SM1.6 we obtain, for all $h \in \mathcal{H}$,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[h(X), f_g(X)] &= \langle h, C_{XY} g\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \langle h, C_{XY}^{(n)} g\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \langle h, C_X^{(n)} h_g^{(n)}\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \langle h, C_X h_g^{(n)}\rangle_{\mathcal{H}} \\
&= \lim_{n\to\infty} \mathbb{C}\mathrm{ov}[h(X), h_g^{(n)}(X)],
\end{aligned}$$

which yields (a). Also, for arbitrary $h^{(n)} \in \mathcal{H}^{(n)}$, Lemma SM1.6 yields

$$\begin{aligned}
\langle [h^{(n)}], [f_g] \rangle_{L^2_{\mathcal{C}}} &= \mathbb{C}\mathrm{ov}\big[h^{(n)}(X), f_g(X)\big] \\
&= \langle h^{(n)}, C_{XY} g \rangle_{\mathcal{H}} \\
&= \langle C_{YX} h^{(n)}, g \rangle_{\mathcal{G}} \\
&= \langle C^{(n)}_{YX} h^{(n)}, g \rangle_{\mathcal{G}} \\
&= \langle h^{(n)}, C^{(n)}_{XY} g \rangle_{\mathcal{H}} \\
&= \langle h^{(n)}, C^{(n)}_{X} h^{(n)}_g \rangle_{\mathcal{H}} \\
&= \langle h^{(n)}, C_{X} h^{(n)}_g \rangle_{\mathcal{H}} \\
&= \mathbb{C}\mathrm{ov}\big[h^{(n)}(X), h^{(n)}_g(X)\big] \\
&= \langle [h^{(n)}], [h^{(n)}_g] \rangle_{L^2_{\mathcal{C}}},
\end{aligned}$$

which yields (b). ∎

Lemma SM1.10. *Let Assumption 2.1 hold and $\mathcal{H}^{(1)} \subseteq \mathcal{H}^{(2)} \subseteq \cdots$ be an increasing sequence of closed subspaces $\mathcal{H}^{(n)}$ of $L^2(\mathbb{P}_X)$, $n \in \mathbb{N}$. Further, let $\mathfrak{m}, \mathfrak{m}^{(n)} \in L^2(\mathbb{P}_X; \mathcal{G}) \simeq L^2(\mathbb{P}_X) \otimes \mathcal{G}$ and denote $\overline{f} := f - \mathbb{E}[f(X)]$ for $f \in L^2(\mathbb{P}_X)$ and $\overline{\mathfrak{f}} := \mathfrak{f} - \mathbb{E}[\mathfrak{f}(X)]$ for $\mathfrak{f} \in L^2(\mathbb{P}_X; \mathcal{G})$.*

*(a) If $([\mathfrak{m}^{(n)}](\cdot))(y)$ is the orthogonal projection in $L^2_{\mathcal{C}}(\mathbb{P}_X)$ of $([\mathfrak{m}](\cdot))(y)$ onto $\mathcal{H}^{(n)}_{\mathcal{C}}$ for each $y \in \mathcal{Y}$, then $[\mathfrak{m}^{(n)}]$ is the orthogonal projection in $L^2_{\mathcal{C}}(\mathbb{P}_X; \mathcal{G})$ of $[\mathfrak{m}]$ onto $(\mathcal{H}^{(n)} \otimes \mathcal{G})_{\mathcal{C}}$.*

*(b) If, in addition to the assumption in (a), $([\mathfrak{m}^{(n)}](\cdot))(y) \to ([\mathfrak{m}](\cdot))(y)$ in $L^2_{\mathcal{C}}(\mathbb{P}_X)$ as $n \to \infty$ for each $y \in \mathcal{Y}$, then $[\mathfrak{m}^{(n)}] \to [\mathfrak{m}]$ in $L^2_{\mathcal{C}}(\mathbb{P}_X; \mathcal{G})$, or, in other words, $\overline{\mathfrak{m}}^{(n)}(X) \to \overline{\mathfrak{m}}(X)$ in $L^2(\mathbb{P}; \mathcal{G})$.*

*(c) If $(\mathfrak{m}^{(n)}(\cdot))(y)$ is the orthogonal projection in $L^2(\mathbb{P}_X; \mathbb{R})$ of $(\mathfrak{m}(\cdot))(y)$ onto $\mathcal{H}^{(n)}$ for each $y \in \mathcal{Y}$, then $\mathfrak{m}^{(n)}$ is the orthogonal projection in $L^2(\mathbb{P}_X; \mathcal{G})$ of $\mathfrak{m}$ onto $\mathcal{H}^{(n)} \otimes \mathcal{G}$.*

*(d) If, in addition to the assumption in (c), $(\mathfrak{m}^{(n)}(\cdot))(y) \to (\mathfrak{m}(\cdot))(y)$ in $L^2(\mathbb{P}_X; \mathbb{R})$ as $n \to \infty$ for each $y \in \mathcal{Y}$, then $\mathfrak{m}^{(n)} \to \mathfrak{m}$ in $L^2(\mathbb{P}_X; \mathcal{G})$, or, in other words, $\mathfrak{m}^{(n)}(X) \to \mathfrak{m}(X)$ in $L^2(\mathbb{P}; \mathcal{G})$.*

*Proof.* We only give the proofs of (a) and (b); (c) and (d) can be proven similarly with fewer technicalities. Let $h \in \mathcal{H}^{(n)}$ and $y \in \mathcal{Y}$. Then

$$\begin{aligned}
\langle [\mathfrak{m}^{(n)}] - [\mathfrak{m}], [h \otimes \psi(y)] \rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X; \mathcal{G})} &= \mathbb{E}\big[\big\langle \overline{\mathfrak{m}}^{(n)}(X) - \overline{\mathfrak{m}}(X), \psi(y) \big\rangle_{\mathcal{G}} \, \overline{h}(X)\big] \\
&= \mathbb{E}\left[\Big(\big(\overline{\mathfrak{m}}^{(n)}(X)\big)(y) - \big(\overline{\mathfrak{m}}(X)\big)(y)\Big) \, \overline{h}(X)\right] \\
&= \Big\langle \big([\mathfrak{m}^{(n)}](\cdot)\big)(y) - \big([\mathfrak{m}](\cdot)\big)(y), [h] \Big\rangle_{L^2_{\mathcal{C}}(\mathbb{P}_X; \mathbb{R})} \\
&= 0,
\end{aligned}$$

which proves (a). Hence, by Lemma SM1.8, $[\mathfrak{m}^{(n)}]$ converges in $L^2_{\mathcal{C}}(\mathbb{P}_X; \mathcal{G})$ to some limit $[\mathfrak{m}']$.

This implies pointwise convergence for each $y \in \mathcal{Y}$ in the following sense:

$$\begin{aligned}\big\|\big([\mathfrak{m}^{(n)}](\cdot)\big)(y) - \big([\mathfrak{m}'](\cdot)\big)(y)\big\|^2_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathbb{R})} &= \mathbb{E}\big[|\langle \psi(y), \overline{\mathfrak{m}}^{(n)}(X) - \overline{\mathfrak{m}}'(X)\rangle_{\mathcal{G}}|^2\big] \\ &\leq \|\psi(y)\|^2_{\mathcal{G}}\, \mathbb{E}\big[\|\overline{\mathfrak{m}}^{(n)}(X) - \overline{\mathfrak{m}}'(X)\|^2_{\mathcal{G}}\big] \\ &= \|\psi(y)\|^2_{\mathcal{G}}\, \|\mathfrak{m}^{(n)} - \mathfrak{m}'\|^2_{L^2_{\mathcal{C}}(\mathbb{P}_X;\mathcal{G})} \\ &\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Therefore, by assumption, $[\mathfrak{m}']$ agrees with $[\mathfrak{m}]$ $\mathbb{P}_X$-a.e., proving (b). ∎

Lemma SM1.11. *Under the assumptions and notation of Theorem 4.4, $(\mu^{(n)}(X,\cdot))_{n\in\mathbb{N}}$ is a martingale in $L^2(\Omega,\Sigma,\mathbb{P};\mathcal{G})$ with respect to the filtration $(\sigma(V^{(n)}))_{n\in\mathbb{N}}$ of $\Sigma$, where $V^{(n)} := P_{\mathcal{H}^{(n)}}(\varphi(X))$ and $P_{\mathcal{H}^{(n)}}\colon \mathcal{H} \to \mathcal{H}^{(n)}$ denotes the orthogonal projection in $\mathcal{H}$ onto $\mathcal{H}^{(n)}$.*

*Proof.* Consider the Karhunen–Loève expansion of $\varphi(X)$,

$$\varphi(X) = \mu_X + \sum_{i\in\mathbb{N}} Z_i\, h_i,$$

where $Z_i\colon (\Omega,\Sigma,\mathbb{P}) \to \mathbb{R}$ are uncorrelated real-valued random variables with $\mathbb{E}[Z_i] = 0$ and $\mathbb{V}[Z_i] = \sigma_i$ for all $i \in \mathbb{N}$, $\sigma_i \geq 0$ denoting the eigenvalue of $C_X$ corresponding to the eigenvector $h_i$. We observe that $\sigma(V^{(n)}) = \sigma(Z^{(n)})$, where $Z^{(n)} := (Z_1,\ldots,Z_n)$. Now let $A^{(n)} = (C_X^{(n)\dagger} C_{XY}^{(n)})^*$, $n \in \mathbb{N}$, and observe that $A^{(n)}v = A^{(n)}P_{\mathcal{H}^{(n)}}v$ and that $A^{(n+1)}$ and $A^{(n)}$ agree on $\mathcal{H}^{(n)}$. Hence, for $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[\mu^{(n+1)}(X,\cdot)\,|\,V^{(n)}] &= \mu_Y + A^{(n+1)}\,\mathbb{E}[V - \mu_X\,|\,Z^{(n)}] \\ &= \mu_Y + A^{(n+1)}\,\mathbb{E}\Big[\sum_{i\in\mathbb{N}} Z_i\, h_i \,\Big|\, Z^{(n)}\Big] \\ &= \mu_Y + A^{(n)} \sum_{i=1}^{n} Z_i\, h_i \\ &= \mu_Y + A^{(n)}\,(\varphi(X) - \mu_X) \\ &= \mu^{(n)}(X,\cdot),\end{aligned}$$

proving the martingale property. ∎

Lemma SM1.12. *Let Assumptions 2.1 and B* hold. Then $\mathbb{E}[C_{Y|X}] = \int_{\mathcal{X}} C_{Y|X=x}\,\mathrm{d}\mathbb{P}_X(x)$ is well defined as a strong (Bochner) integral, i.e. $\int_{\mathcal{X}} \|C_{Y|X=x}\|\,\mathrm{d}\mathbb{P}_X(x) < \infty$.*

*Proof.* The Cauchy–Schwarz inequality and (2.2) imply that, for $x \in \mathcal{X}_Y$,

$$\begin{aligned}\|C_{Y|X=x}\| &= \sup_{\|g\|_{\mathcal{G}}\leq 1, \|\tilde{g}\|_{\mathcal{G}}\leq 1} \langle g, C_{Y|X=x}\tilde{g}\rangle_{\mathcal{G}} \\ &= \sup_{\|g\|_{\mathcal{G}}\leq 1, \|\tilde{g}\|_{\mathcal{G}}\leq 1} \langle g, \tilde{g}\rangle_{\mathcal{L}^2_{\mathcal{C}}(\mathbb{P}_{Y|X=x})} \\ &\leq \sup_{\|g\|_{\mathcal{G}}\leq 1, \|\tilde{g}\|_{\mathcal{G}}\leq 1} \|g\|_{\mathcal{L}^2(\mathbb{P}_{Y|X=x})}\|\tilde{g}\|_{\mathcal{L}^2(\mathbb{P}_{Y|X=x})} \\ &\leq \mathbb{E}\big[\|\psi(Y)\|^2_{\mathcal{G}}\big|X = x\big],\end{aligned}$$

which, by the law of total expectation and (2.1), yields that

$$\mathbb{E}[\|C_{Y|X}\|] \leq \mathbb{E}\big[\mathbb{E}[\|\psi(Y)\|_{\mathcal{G}}^2 \mid X]\big] = \mathbb{E}[\|\psi(Y)\|_{\mathcal{G}}^2] < \infty,$$

as claimed. ∎

**SM2. Empirical Estimates for CMEs.** In practice, the kernel mean embeddings and kernel (cross-)covariance operators will often be estimated empirically from observed data, and so empirical versions of the CME, along with convergence guarantees, are of great importance. As mentioned in Remark 1.2, this topic is beyond the scope of this paper. However, we wish to point out why this is a complex problem and briefly address the main difficulties.

In the simplest setting, given $J \in \mathbb{N}$ independent samples $(X_1, Y_1), \ldots, (X_J, Y_J) \sim \mathbb{P}_{XY}$, we have the empirical estimators

$$\mu_X \approx \widehat{\mu}_X \coloneqq \frac{1}{J}\sum_{j=1}^{J} \varphi(X_j), \qquad C_{XY} \approx \widehat{C}_{XY} \coloneqq \frac{1}{J}\sum_{j=1}^{J} (\varphi(X_j) - \widehat{\mu}_X) \otimes (\psi(Y_j) - \widehat{\mu}_Y),$$

and so on. (To simplify the notation, we suppress the obvious $J$-dependence of these estimators.) Laws of large numbers for these empirical estimators have already been established — see e.g. [SM19, Theorem 2] and [SM17, Lemma 5.8] — but the impact of this approximation error upon conditioning is, to the best of our knowledge, not yet fully quantified. One natural approach to approximate the CME $\mu_{Y|X=x}$ is the regularisation of $\widehat{C}_X$ or $\widehat{{}^{u}C}_X$,

$$\mu_{Y|X=x} \approx \Big(\big(\widehat{{}^{u}C}_X + \varepsilon \operatorname{Id}_{\mathcal{H}}\big)^{\dagger}\, \widehat{{}^{u}C}_{XY}\Big)^{*} \varphi(x) = \widehat{{}^{u}C}_{YX}\big(\widehat{{}^{u}C}_X + \varepsilon \operatorname{Id}_{\mathcal{H}}\big)^{-1}\varphi(x),$$

where $\varepsilon > 0$ is a regularisation parameter which may depend on $J$. Note that such a regularisation can be viewed as an approximation both to the new CME formula derived in Theorem 5.3, $\mu_{Y|X=x} = ({}^{u}C_X^{\dagger}\, {}^{u}C_{XY})^{*}\varphi(x)$, as well as to the original (uncentred) one, $\mu_{Y|X=x} = {}^{u}C_{YX}\, {}^{u}C_X^{-1}\varphi(x)$. Therefore, this approach is rather well studied and convergence rates for this strategy have been established under certain conditions [SM12, SM15, SM18].

However, the new formulae (4.3), (4.6), and (5.3) relying on the Moore–Penrose pseudo-inverse suggest another type of approximation, where we will focus on the centred case from now on. The naïve estimate would be

$$\mu_{Y|X=x} \approx \widehat{\mu}_Y + \big(\widehat{C}_X^{\dagger}\widehat{C}_{XY}\big)^{*}\big(\varphi(x) - \widehat{\mu}_X\big). \tag{SM2.1}$$

Note that $\operatorname{ran}\widehat{C}_{XY} \subseteq \operatorname{ran}\widehat{C}_X$ and so (SM2.1) is well defined. However, the convergence of $\widehat{C}_X$ to $C_X$ (e.g. in the Hilbert–Schmidt norm, as $J \to \infty$) translates badly to the convergence of $\widehat{C}_X^{\dagger}$ to the pseudo-inverse $C_X^{\dagger}$. One problem is that small eigenvalues of $C_X$ might be approximated by eigenvalues of $\widehat{C}_X$ that are orders of magnitude smaller, causing $\widehat{C}_X^{\dagger}$ to "blow up". So, in addition to the convergence of $\widehat{C}_X$ in the classical norms (such as the Hilbert–Schmidt norm or operator norm), we need to control the the smallest eigenvalue of $\widehat{C}_X$.

A natural workaround, inspired by the finite-rank approximation in Theorem 4.4, is to truncate[1] the (cross-)covariance operators to a subspace $\mathcal{H}^{(n)} = \operatorname{span}\{h_1, \dots, h_n\}$ of $\mathcal{H}$ with $\dim \mathcal{H}^{(n)} = n = n(J) \ll J$. One might thus hope to approximate the dominant $n$ eigenvalues of $C_X$ well while artificially setting the others to zero and preventing the blow-up of $\widehat{C}_X^{\dagger}$. There are several results from random matrix theory that control the behaviour of the $n^{\text{th}}$ eigenvalue of (truncated) empirical covariance matrices for growing $J$ and $n = n(J)$ [SM2, SM3, SM4, SM16]. Most of these results are formulated for the case where the true mean is known to be zero and the true covariance matrix is the identity matrix and are typically of the following form, where $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the largest and smallest eigenvalues of a matrix $M$, respectively:

Theorem SM2.1 ([SM4, Theorem 2]). *Let $(\xi_{ij})_{i,j\in\mathbb{N}}$ be a double array of independent and identically distributed random variables with zero mean and unit variance. For $J \in \mathbb{N}$, let $n = n(J)$ be such that $n(J) \to \infty$ and $n(J)/J \to \gamma \in (0, 1)$ for $J \to \infty$ and let*

$$A_J = (\xi_{ij})_{i=1,\dots,n(J),\, j=1,\dots,J}, \qquad S_J = \tfrac{1}{J} A_J A_J^{\top}. \tag{SM2.2}$$

*Then, if $\mathbb{E}[\xi_{11}^4] < \infty$,*

$$\lambda_{\max}(S_J) \xrightarrow[J\to\infty]{a.e.} (1+\sqrt{\gamma})^2, \qquad \lambda_{\min}(S_J) \xrightarrow[J\to\infty]{a.e.} (1-\sqrt{\gamma})^2.$$

In our case the covariance operator $C_X$ is not the identity; however, our case follows partially from Theorem SM2.1 using the eigendecomposition of $C_X$,

$$C_X = \sum_{i\in\mathbb{N}} \sigma_i \, h_i \otimes h_i, \qquad \sigma_1 \geq \sigma_2 \geq \dots \geq 0,$$

and the Karhunen–Loève expansion of $V = \varphi(X) - \mu_X$,

$$V = \sum_{i\in\mathbb{N}} \sqrt{\sigma_i}\, \xi_i \, h_i,$$

where $\xi_n \colon \Omega \to \mathbb{R}$ are uncorrelated random variables with $\mathbb{E}[\xi_n] = 0$ and $\mathbb{V}[\xi_i] = 1$ for each $i \in \mathbb{N}$ and $(h_i)_{i\in\mathbb{N}}$ is an orthonormal eigenbasis of $\mathcal{H}$. To this end, let $V_j \coloneqq \varphi(X_j) - \mu_X$ be i.i.d. copies of $V$ and let $V_j^{(n)}$ be their respective orthogonal projections onto $\mathcal{H}^{(n)} \coloneqq \operatorname{span}\{h_1, \dots, h_n\}$, i.e.

$$V_j = \sum_{i\in\mathbb{N}} \sqrt{\sigma_i}\, \xi_{ij} \, h_i \text{ with } \xi_{ij} \overset{\text{i.i.d.}}{\sim} \xi_i \text{ for all } i, j, \qquad V_j^{(n)} = \sum_{i=1}^{n} \sqrt{\sigma_i}\, \xi_{ij} \, h_i.$$

To simplify notation, let us work with $(n \times n)$-matrices instead of operators, expressed in the basis $(h_1, \dots, h_n)$ of $\mathcal{H}^{(n)}$. Then

$$\widehat{C}_X^{(n)} = \frac{1}{J} \sum_{j=1}^{J} V_j \otimes V_j = \left(C_X^{(n)}\right)^{1/2} S_J \left(C_X^{(n)}\right)^{1/2}, \tag{SM2.3}$$

[1] Naturally, truncation can be viewed as another form of regularisation. For further regularised estimates of large covariance and precision matrices by tapering, banding, sparsifying or similar see e.g. [SM5, SM6, SM7, SM21] and references therein.

where $S_J$ is defined by (SM2.2). So, we are *nearly* in the setup of Theorem SM2.1 and ready to conclude

$$\lim_{J\to\infty} \lambda_{\min}(\widehat{C}_X^{(n)}) \geq \lim_{J\to\infty} \sigma_n^{1/2}\lambda_{\min}(S_J)\sigma_n^{1/2} \geq \sigma_n(1-\sqrt{\gamma})^2,$$

with $n = n(J)$ and $\gamma$ as in Theorem SM2.1. (Here we make use of the fact that $\|Ax\| \geq \lambda_{\min}(A)\|x\|$ for all $x$ when $A$ is positive semi-definite.) However, some obstacles remain:

- Since the random variables $V_j$ are independent copies of $V$, the distributions of $\xi_{ij}$ and $\xi_{ij'}$ agree for all $j, j'$, but we cannot expect the distributions of $\xi_{ij}$ and $\xi_{i'j}$ to agree for all $i, i'$. Hence, $\xi_{ij}$ do not fulfil the requirement of being identically distributed. However, there are generalisations of Theorem SM2.1 to this case under technical assumptions; see e.g. [SM3, Theorem 2.8].
- It is unclear when the condition $\mathbb{E}[\xi_{11}^4] < \infty$ is fulfilled. There are, however, some results in the case of infinite fourth moments; see [SM16] and references therein.
- The random variables $\xi_i$ are uncorrelated but, in general, not independent. We are not aware of a result similar to Theorem SM2.1 for the uncorrelated case.[2]

Even if we manage to formulate a version of Theorem SM2.1 which suits our needs, note that the considerations so far, if executed rigorously, only guarantee that $C_X^\dagger$ does not blow up. This is still some way short of establishing the convergence of the corresponding CME estimator

$$\mu_{Y|X=x} \approx \widehat{\mu}_Y + \left(\widehat{C}_X^{(n)\dagger}\widehat{C}_{XY}^{(n)}\right)^* \left(\varphi(x) - \widehat{\mu}_X\right), \qquad n = n(J). \tag{SM2.4}$$

We conclude here by summarising some of the necessary steps:

- In the above considerations we assumed $\mu_X$ to be known. In practice, however, we may only employ its empirical estimate $\widehat{\mu}_X$. Since $\mu_X \otimes \mu_X$ is a rank-one estimator, this issue might be partially resolved by [SM14, Corollary 2.1], which implies that the eigenvalues of ${}^uC_X$ and $C_X = {}^uC_X - \mu_X \otimes \mu_X$ have a similar decay rate.
- In order to project onto $\mathcal{H}^{(n)} = \operatorname{span}\{h_1, \ldots, h_n\}$, we require the eigenvectors $h_i$ of $C_X$. While the $n$ dominant eigenvectors of $\widehat{C}_X$ can be used as estimates of $h_i$, it is unclear how the approximation error affects the theoretical results presented above.
- Controlling the eigenvalues alone is insufficient. For a reasonable approximation of $C_X^{(n)}$ in (SM2.3), we would need a result which tells us that $S_J$ becomes close to the identity matrix for large $J$, which requires us to understand the behaviour of its eigenvectors as well. The investigation of the eigenvectors of $S_J$ turns out to be extremely challenging; see [SM2, Chapter 10] for a survey of existing results.

## REFERENCES

[1] M. L. Arias, G. Corach, and M. C. Gonzalez, *Generalized inverses and Douglas equations*, Proc. Amer. Math. Soc., 136 (2008), pp. 3177–3183, https://doi.org/10.1090/S0002-9939-08-09298-8.

[2] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer Series in Statistics, Springer, New York, second ed., 2010, https://doi.org/10.1007/978-1-4419-0661-8.

[2] [SM8] discuss the estimation of covariance and precision matrices for time series data where the random variables $V_j$ are not assumed to be independent. However, we need to drop independence in the rows of $A_J$, not in its columns.

[3] Z. D. Bai, *Methodologies in spectral analysis of large-dimensional random matrices, a review*, Statist. Sinica, 9 (1999), pp. 611–677, https://doi.org/10.1142/9789812793096_0015. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.
[4] Z. D. Bai and Y. Q. Yin, *Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix*, Ann. Probab., 21 (1993), pp. 1275–1294, https://doi.org/10.1214/aop/1176989118.
[5] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, Ann. Statist., 36 (2008), pp. 2577–2604, https://doi.org/10.1214/08-AOS600.
[6] P. J. Bickel and E. Levina, *Regularized estimation of large covariance matrices*, Ann. Statist., 36 (2008), pp. 199–227, https://doi.org/10.1214/009053607000000758.
[7] T. T. Cai, C.-H. Zhang, and H. H. Zhou, *Optimal rates of convergence for covariance matrix estimation*, Ann. Statist., 38 (2010), pp. 2118–2144, https://doi.org/10.1214/09-AOS752.
[8] X. Chen, M. Xu, and W. B. Wu, *Covariance and precision matrix estimation for high-dimensional time series*, Ann. Statist., 41 (2013), pp. 2994–3021, https://doi.org/10.1214/13-AOS1182.
[9] E. De Vito, L. Rosasco, and A. Caponnetto, *Discretization error analysis for Tikhonov regularization*, Analysis and Applications, 04 (2006), pp. 81–99, https://doi.org/10.1142/S0219530506000711.
[10] R. G. Douglas, *On majorization, factorization, and range inclusion of operators on Hilbert space*, Proc. Amer. Math. Soc., 17 (1966), pp. 413–415, https://doi.org/10.2307/2035178.
[11] P. A. Fillmore and J. P. Williams, *On operator ranges*, Adv. Math., 7 (1971), pp. 254–281, https://doi.org/10.1016/S0001-8708(71)80006-3.
[12] K. Fukumizu, *Nonparametric Bayesian inference with kernel mean embedding*, in Modern Methodology and Applications in Spatial-Temporal Modeling, Springer Japan, Tokyo, 2015, pp. 1–24, https://doi.org/10.1007/978-4-431-55339-7_1.
[13] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Kernel dimension reduction in regression*, Ann. Statist., 37 (2009), pp. 1871–1905, https://doi.org/10.1214/08-AOS637.
[14] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, vol. 18 of Translations of Mathematical Monographs, American Mathematical Society, Providence, R.I., 1969. Transl. A. Feinstein.
[15] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil, *Conditional mean embeddings as regressors*, in Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 1823–1830, https://icml.cc/2012/papers/898.pdf.
[16] J. Heiny and T. Mikosch, *Almost sure convergence of the largest and smallest eigenvalues of high-dimensional sample correlation matrices*, Stochastic Process. Appl., 128 (2018), pp. 2779–2815, https://doi.org/10.1016/j.spa.2017.10.002.
[17] M. Mollenhauer, *Singular value decomposition of operators on reproducing kernel Hilbert spaces*, master's thesis, Freie Universität Berlin, 2018.
[18] J. Park and K. Muandet, *A measure-theoretic approach to kernel conditional mean embeddings*, 2020, https://arxiv.org/abs/2002.03689.
[19] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, *A Hilbert space embedding for distributions*, in Proceedings of the 18th International Conference on Algorithmic Learning Theory, Berlin, Heidelberg, 2007, Springer, pp. 13–31, https://doi.org/10.1007/978-3-540-75225-7_5.
[20] J. Weidmann, *Linear Operators in Hilbert Spaces*, vol. 68 of Graduate Texts in Mathematics, Springer-Verlag, New York-Berlin, 1980, https://doi.org/10.1007/978-1-4612-6027-1. Transl. J. Szücs.
[21] M. Yuan, *High dimensional inverse covariance matrix estimation via linear programming*, J. Mach. Learn. Res., 11 (2010), pp. 2261–2286, http://www.jmlr.org/papers/volume11/yuan10b/yuan10b.pdf.