A SUPPLEMENTARY MATERIAL

This supplementary section contains proofs omitted from the main paper and includes a proof that the HSIC statistic asymptotically satisfies the hypothesis of the Wild Bootstrap.

A.1 HILBERT SPACE RANDOM VARIABLE CLT

In this paper we exploit a Central Limit Theorem for Hilbert space valued random variables that are functions of random processes [Dehling et al., 2015]. One of the conditions required to apply this theorem concerns appropriate β-mixing of the underlying processes. This theorem is used as a black-box, and it is hoped by the authors that as further theorems concerning CLT-properties of Hilbert space random variables are developed, the conditions required of the processes may be weakened.

Proof. (Lemma 1) We exploit Theorem 1.1 from Dehling et al. [2015]. Using the language of this paper, $\bar{\phi}(X_i)$ is a 1-approximating functional of $(X_i)_i$, following straightforwardly from the definition of 1-approximating functionals given.

Since our kernels are bounded, $\exists C: \quad \|\bar{\phi}(X_i)\| < C$ and so

$$\mathbb{E}\|\bar{\phi}(X_1)\|^{2+\delta} < C^{2+\delta} < \infty \ \ \forall \delta > 0$$

Thus condition (1) is satisfied.

We can take $f_m = \bar{\phi}(X_0) \ \ \forall m$ and so achieve $a_m = 0 \ \ \forall m$, thus condition (2) is satisfied.

By assumption on the time series, condition (3) is satisfied.

Thus, by Theorem 1.1 in Dehling et al. [2015]

$$\sqrt{n}(\tilde{\mu}_X - \mu_X) \overset{n}{\rightsquigarrow} N$$

where N is a Hilbert space valued Gaussian random variable and convergence is in distribution. Thus

$$\|\tilde{\mu}_X - \mu_X\| = O_P(\frac{1}{\sqrt{n}})$$

□

A.2 SUB-PROCESSES OF β-MIXING PROCESSES ARE β-MIXING

Lemma 2. *Suppose that the process $(X_t, Y_t, Z_t)_t$ is β-mixing. Then any 'sub-process' is also β-mixing (for example $(X_t, Y_t)_t$ or $(X_t)_t$)*

Proof. (Lemma 2)

Let us consider $(X_t, Y_t)_t$. Let us call $\beta_{XYZ}(m)$ the coefficients for the process $(X_t, Y_t, Z_t)_t$, and $\beta_{XY}(m)$ the coefficients for the process $(X_t, Y_t)_t$.

Observe that for $A \in \sigma((X_b, Y_b), \ldots, (X_c, Y_c))$, it is the case that $A \times \mathcal{Z} \in \sigma((X_b, Y_b, Z_b), \ldots, (X_c, Y_c, Z_c))$ and $\mathbb{P}_{XY}(A) = \mathbb{P}_{XYZ}(A \times \mathcal{Z})$.

Thus

$$
\begin{aligned}
\beta_{XY}(m) &= \frac{1}{2}\sup_n \sup_{\{A_i^{XY}\},\{B_j^{XY}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XY}(A_i^{XY} \cap B_j^{XY}) - \mathbb{P}_{XYZ}(A_i^{XY})\mathbb{P}_{XYZ}(B_j^{XY})| \\
&= \frac{1}{2}\sup_n \sup_{\{A_i^{XY}\},\{B_j^{XY}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XYZ}((A_i^{XY} \times \mathcal{Z}) \cap (B_j^{XY} \times \mathcal{Z})) \\
&\qquad\qquad - \mathbb{P}_{XYZ}(A_i^{XY} \times \mathcal{Z})\mathbb{P}_{XYZ}(B_j^{XY} \times \mathcal{Z})| \\
&\leq \frac{1}{2}\sup_n \sup_{\{A_i^{XYZ}\},\{B_j^{XYZ}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XYZ}(A_i^{XYZ} \cap B_j^{XYZ}) - \mathbb{P}_{XYZ}(A_i^{XYZ})\mathbb{P}_{XYZ}(B_j^{XYZ})| \\
&= \beta_{XYZ}(m)
\end{aligned}
$$

Thus we have shown that $\beta_{XYZ}(m) \longrightarrow 0 \implies \beta_{XY}(m) \longrightarrow 0$. That is, if $(X_t, Y_t, Z_t)_t$ is β-mixing then so is $(X_t, Y_t)_t$

A similar argument holds for any other sub-process. □

A.3 CONTROL OF TYPE I ERROR

Theorem 3 shows that the quantiles of the bootstrapped statistic nV_b (which we can estimate by drawing a large number of samples) converge to those of the test statistic $\|\hat{\mu}_L\|^2$ under the null hypothesis. Therefore, we can estimate rejection thresholds to appropriately control Type I error.

Proof. (Theorem 3)

We use Theorem 3.1 from Leucht and Neumann [2013]. By assumption, condition (B2) is satisfied by the random matrix W. (A2) is satisfied due to Theorem 2. (B1) is satisfied due to the suitable mixing assumptions.

Therefore, Theorem 3.1 implies that nV_b converges in probability to the null distribution of $n\|\hat{\mu}_{L,2}^{(Z)}\|^2$. Since $n\|\mu_L\|^2$ also converges in probability to $n\|\hat{\mu}_{L,2}^{(Z)}\|^2$, it follows that nV_b converges to $n\|\mu_L\|^2$ in probability, and thus also in distribution. Convergence in distribution implies that the quantiles converge. □

A.4 SEMI-CONSISTENCY

Theorem 4 provides a consistency result: if $\Delta_L P \neq 0$, then we correctly reject $\mathcal{H}_0$ with probability 1 in the limit $n \longrightarrow \infty$.

Proof. By Theorem 2 from Chwialkowski et al. [2014], nV_b converges to some random variable with finite variance, while $n\|\hat{\mu}_L\|^2 \longrightarrow \infty$. Thus if Q_α is the α-quantile of nV_b, then $P(n\|\hat{\mu}_L\|^2 > Q_\alpha) \longrightarrow 1$ for any α. □

A.5 PROOF THAT BOUNDEDNESS AND LIPSCHITZ CONTINUITY IS PRESERVED

Recall that a kernel k defined on $\mathcal{X}$ is Lipschitz continuous iff $\exists C_k : \forall w \;\; |k(x,w) - k(x',w)| \leq C_k d_{\mathcal{X}}(x,x')$ where $d_{\mathcal{X}}$ is the metric on $\mathcal{X}$ with respect to which k is Lipschitz continuous.

Claim 1. *k bounded and Lipschitz continuous $\implies$ $\bar{k}$ is bounded and Lipschitz continuous*

Proof. k bounded implies there exists B_k such that $|k(x,w)| \leq B_k \; \forall x, w \in \mathcal{X}$. It follows that

$$
\begin{aligned}
|\bar{k}(x,w)| &= |k(x,w) - \mathbb{E}_X[k(X,w)] - \mathbb{E}_W[k(x,W)] + \mathbb{E}_{XW}[k(X,W)]| \\
&\leq |k(x,w)| + \mathbb{E}_X|k(X,w)| + \mathbb{E}_W|k(x,W)| + \mathbb{E}_{XW}|k(X,W)| \\
&\leq 4B_k
\end{aligned}
$$

And thus $\bar{k}$ is bounded. For Lipschitz continuity, observe that for any $w \in \mathcal{X}$

$$
\begin{aligned}
|\bar{k}(x,w) - \bar{k}(x',w)| &= |k(x,w) - \mathbb{E}_X[k(X,w)] - \mathbb{E}_W[k(x,W)] + \mathbb{E}_{XW}[k(X,W)] \\
&\quad - k(x',w) + \mathbb{E}_X[k(X,w)] + \mathbb{E}_W[k(x',W)] - \mathbb{E}_{XW}[k(X,W)]| \\
&= |k(x,w) - k(x',w) + \mathbb{E}_W[k(x',W)] - \mathbb{E}_W[k(x,W)]| \\
&\leq |k(x,w) - k(x',w)| + |\mathbb{E}_W[k(x',W)] - \mathbb{E}_W[k(x,W)]| \\
&\leq |k(x,w) - k(x',w)| + \mathbb{E}_W|k(x',W) - k(x,W)| \\
&\leq 2C_k d_{\mathcal{X}}(x,x')
\end{aligned}
$$

and thus $\bar{k}$ is Lipschitz continuous.

□

Claim 2. *k and l bounded and Lipschitz continuous with respect to the metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ respectively $\implies$ $k \otimes l$ is bounded and Lipschitz continuous with respect to any metric on $\mathcal{X} \times \mathcal{Y}$ equivalent to $d\left((x,y),(x',y')\right) = d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y')$*

Note that all norms on finite dimensional vector spaces are equivalent, and so if $\mathcal{X}$ and $\mathcal{Y}$ are finite dimensional vector spaces then $k \otimes l$ is Lipschitz continuous with respect to *any* norm on $\mathcal{X} \times \mathcal{Y}$

Proof. Let k and l be bounded by B_k and B_l respectively. Then

$$
\begin{aligned}
|k \otimes l\left((x,y),(w,z)\right)| &= |k(x,w)l(y,z)| \\
&= |k(x,w)||l(y,z)| \\
&\leq B_k B_l
\end{aligned}
$$

Let k and l have Lipschitz constants C_k and C_l respectively. Then, for any $(w,z) \in \mathcal{X} \times \mathcal{Y}$

$$
\begin{aligned}
&|k \otimes l\left((x,y),(w,z)\right) - k \otimes l\left((x',y'),(w,z)\right)| \\
&\qquad = |k(x,w)l(y,z) - k(x',w)l(y',z)| \\
&\qquad = |k(x,w)l(y,z) - k(x',w)l(y,z) + k(x',w)l(y,z) - k(x',w)l(y',z)| \\
&\qquad \leq |l(y,z)||k(x,w) - k(x',w)| + |k(x',w)||l(y,z) - l(y',z)| \\
&\qquad \leq B_l C_k d_{\mathcal{X}}(x,x') + B_k C_l d_{\mathcal{Y}}(y,y') \\
&\qquad \leq \max(B_l C_k, B_k C_l)\ \ d\left((x,y),(x',y')\right)
\end{aligned}
$$

□

A.6 PROOF THAT HSIC CAN BE WILD BOOTSTRAPPED

Given samples $\{(X_i, Y_i)\}_{i=1}^n$, and taking all notation involving kernels and base spaces as before, the HSIC statistic is defined to be the squared RKHS distance between the empirical embeddings of the distributions $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$:

$$
\begin{aligned}
HSIC_b &= \| \frac{1}{n}\sum_i \phi_X(X_i) \otimes \phi_Y(Y_i) - \left(\frac{1}{n}\sum_i \phi_X(X_i)\right) \otimes \left(\frac{1}{n}\sum_i \phi_Y(Y_i)\right) \|^2 \\
&= \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^4}K_{++}L_{++} \\
&= \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}
\end{aligned}
$$

where the last equality can be shown easily by expanding $\tilde{K}$ (and $\tilde{L}$ similarly) as

$$\begin{aligned}\tilde{K}_{ij} &= \langle \phi_X(X_i) - \frac{1}{n}\sum_k \phi_X(X_k), \phi_X(X_j) - \frac{1}{n}\sum_k \phi_X(X_k)\rangle \\ &= K_{ij} - \frac{1}{n}\sum_k K_{ik} - \frac{1}{n}\sum_k K_{jk} + \frac{1}{n^2}\sum_{kl} K_{kl}\end{aligned}$$

Theorem 5. *Suppose that $(X_i, Y_i)_{i=1}^n$ are drawn from a process that is β-mixing with coefficients $\beta(m)$ satisfying $\sum_{m=1}^{\infty} \beta(m)^{\frac{\delta}{2+\delta}} < \infty$ for some $\delta > 0$. Under $\mathcal{H}_0 = \{\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y\}$, $\lim_{n\to\infty}(nHSIC_b - \frac{1}{n}(\bar{K} \circ \bar{L})_{++}) = 0$ in probability.*

Similar to the case with the Lancaster statistic, $\frac{1}{n}(\bar{K} \circ \bar{L})_{++}$ is much easier to study than $nHSIC_b$ under the non-*i.i.d.* assumption. It can be written as a normalised V-statistic as:

$$nV_n = \frac{1}{n}\sum_{1\le i,j\le n} \bar{k} \otimes \bar{l}(S_i, S_j)$$

where $S_i = (X_i, Y_i)$. Again, the crucial observation is that

$$h = \bar{k} \otimes \bar{l}$$

is well behaved in the following sense

Theorem 6. *Suppose that k and l are bounded symmetric Lipschitz contentious kernels. Then h is also bounded symmetric and Lipschitz continuous, which is moreover degenerate under $\mathcal{H}_0$.*

Together, Theorems 5 and 6 justify use of the Wild Bootstrap in estimating the quantiles of the null distribution of the test statistic $nHSIC_b$.

Proof. (Theorem 5) We can equivalently write $HSIC_b$ as the norm of the empirically centred covariance operator, which is invariant to population centering the feature maps:

$$\begin{aligned}HSIC_b &= \left\| \frac{1}{n}\sum_i \left(\phi_X(X_i) - \frac{1}{n}\sum_j \phi_X(X_j)\right) \otimes \frac{1}{n}\sum_i \left(\phi_Y(Y_i) - \frac{1}{n}\sum_j \phi_Y(Y_j)\right)\right\|^2 \\ &= \left\| \frac{1}{n}\sum_i \left(\bar{\phi}_X(X_i) - \frac{1}{n}\sum_j \bar{\phi}_X(X_j)\right) \otimes \frac{1}{n}\sum_i \left(\bar{\phi}_Y(Y_i) - \frac{1}{n}\sum_j \bar{\phi}_Y(Y_j)\right)\right\|^2\end{aligned}$$

Expanding this, we can rewrite the above in terms of inner products involving the population centred covariance operator and the population centred mean embeddings:

$$nHSIC_b = n\|\bar{C}_{XY}\|^2 - 2n\langle \bar{C}_{XY}, \bar{\mu}_X \otimes \bar{\mu}_Y\rangle + n\|\bar{\mu}_X \otimes \bar{\mu}_Y\|^2$$

The first term in this expression can be written as $n\|\bar{C}_{XY}\|^2 = \frac{1}{n}\sum_{ij} \bar{k}(X_i, X_j)\bar{l}(Y_i, Y_j) = \frac{1}{n}\sum_{ij} h(S_i, S_j)$. We show that the remaining two terms decay to zero in probability.

By assumption, $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ and thus the expectation operator factorises similarly. Therefore, for any $A \in HS(\mathcal{F}_Y, \mathcal{F}_X)$,

$$\begin{aligned}
\mathbb{E}_{XY}\langle A, \bar{C}_{XY}\rangle &= \frac{1}{n}\sum_i \mathbb{E}_X\mathbb{E}_Y\langle A, (\phi_X(X_i) - \mu_X) \otimes (\phi_Y(Y_i) - \mu_Y)\rangle_{HS} \\
&= \frac{1}{n}\sum_i \mathbb{E}_X\mathbb{E}_Y\langle \phi_X(X_i) - \mu_X, A\,(\phi_Y(Y_i) - \mu_Y)\rangle_{\mathcal{F}_X} \\
&= \frac{1}{n}\sum_i \mathbb{E}_Y\langle \mathbb{E}_X\,(\phi_X(X_i) - \mu_X)\,, A\,(\phi_Y(Y_i) - \mu_Y)\rangle_{\mathcal{F}_X} \\
&= 0
\end{aligned}$$

where the commutativity of $\mathbb{E}_X$ with the inner product in the penultimate line follows from the Bochner integrability of the quantity $\phi_X(X) - \mu_X$, which in turn follows from the conditions under which μ_X exists [Steinwart and Christmann, 2008]. It follows that $\mathbb{E}_{XY}\bar{C}_{XY} = 0$.

Thus by Lemma 1 as before, it follows that $\|\bar{C}_{XY}\|, \|\bar{\mu}_X\|, \|\bar{\mu}_Y\| = O_P(n^{-\frac{1}{2}})$.

It thus follows that the two latter quantities in the above expression for $nHSIC_b$ decay to 0 in probability.

$$\begin{aligned}
n\langle \bar{C}_{XY}, \bar{\mu}_X \otimes \bar{\mu}_Y\rangle &\leq n\|\bar{C}_{XY}\|\|\bar{\mu}_X\|\|\bar{\mu}_Y\| \\
&= O_P(n^{-\frac{1}{2}})
\end{aligned}$$

$$\begin{aligned}
\|\bar{\mu}_X \otimes \bar{\mu}_Y\|^2 &= n\|\bar{\mu}_X\|^2\|\bar{\mu}_Y\|^2 \\
&= nO_P(n^{-2}) \\
&= O_P(n^{-1})
\end{aligned}$$

It follows that $nHSIC_b \xrightarrow{O(n^{-\frac{1}{2}})} = n\|\bar{C}_{XY}\|^2 = \frac{1}{n}(\bar{K} \circ \bar{L})_{++}$, as required. □

Proof. (Theorem 6)

To show degeneracy, fix any s_i and observe that

$$\begin{aligned}
\mathbb{E}_S h(s_i, S) &= \mathbb{E}_X\mathbb{E}_Y\langle\bar{\phi}(x_i), \bar{\phi}(X)\rangle\langle\bar{\phi}(y_i), \bar{\phi}(Y)\rangle \\
&= \langle\bar{\phi}(x_i), \mathbb{E}_X\bar{\phi}(X)\rangle\langle\bar{\phi}(y_i), \mathbb{E}_Y\bar{\phi}(Y)\rangle \\
&= \langle\bar{\phi}(x_i), 0\rangle\langle\bar{\phi}(y_i), 0\rangle = 0
\end{aligned}$$

Symmetry is inherited from symmetry of k and l. Boundedness and Lipschitz continuity are implied by application of the claims in Section A.5.

□

A.7 DISCUSSION OF MIXING ASSUMPTIONS

Throughout this paper, there are (related) assumptions that need to be made on the random processes we consider in order to satisfy the conditions of (1) the wild bootstrap; and (2) the Hilbert space CLT. For simplicity, we wrapped up the assumptions into the single "suitable mixing assumptions". We discuss here the precise assumptions that are needed, how they relate to the suitable mixing assumptions and the applicability of the suitable mixing assumptions.

(1) For their proof of the consistency of the wild bootstrap, Leucht and Neumann [2013] invoke the notion of τ-mixing. We require that the τ-mixing coefficients $\tau(n)$ satisfy the hypothesis of their theorem, namely that $\sum_{n=1}^{\infty} n^2\tau(n) < \infty$.

Properties of τ-mixing, for example its relationship to other types of more commonly understood mixing or models that satisfy τ-mixing, are discussed in Dedecker and Prieur [2005]. In particular, under the assumption that X_i has finite pth moment for any $p > 1$, τ-mixing implies beta-mixing. Examples of systems that are τ-mixing are: causal functions of stationary sequences, iterated random functions, Markov chains and expanding maps.

(2) In order to use the Hilbert space CLT, we require that our processes are β-mixing with coefficients $\beta(n)$ satisfying $\sum_{n=1}^{\infty} \beta(n)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$ and $\sum_{n=1}^{\infty} n\beta(n) < \infty$

In particular, both (1) and (2) are satisfied by a process that is β-mixing with coefficients $\beta(n) = o(n^{-6})$ as stated in the "Suitable Mixing" section.

Many commonly studied processes satisfy the “suitable mixing” condition. In particular Corollary 3.6 of Bradley et al. [2005] states that Harris recurrent and aperiodic markov chains satisfy absolute regularity and Theorem 3.7 of Bradley et al. [2005] states that geometric ergodicity implies geometric decay of beta coefficients. Interestingly Theorem 3.3 of Bradley et al. [2005] describes situations in which a non-stationary chain mixes exponentially.

Note, however, that our novel proof idea relies on the Hilbert space CLT and so requires only assumption (2) above to be used. Therefore our proof idea could be applied to the asymptotic study of other V-statistics in the case that the processes are beta-mixing with $\beta(n) = o(n^{-3+\epsilon})$ for any $\epsilon > 0$.