# A Wild Bootstrap for Degenerate Kernel Tests

**Kacper Chwialkowski**
Department of Computer Science
University College London
London, Gower Street, WC1E 6BT
`kacper.chwialkowski@gmail.com`

**Dino Sejdinovic**
Gatsby Computational Neuroscience Unit, UCL
17 Queen Square, London WC1N 3AR
`dino.sejdinovic@gmail.com`

**Arthur Gretton**
Gatsby Computational Neuroscience Unit, UCL
17 Queen Square, London WC1N 3AR
`arthur.gretton@gmail.com`

## Abstract

A wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes, for which the naive permutation-based bootstrap fails. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler. The code is available at `https://github.com/kacperChwialkowski/wildBootstrap`.

## 1 Introduction

Statistical tests based on distribution embeddings into reproducing kernel Hilbert spaces have been applied in many contexts, including two sample testing [19, 15, 31], tests of independence [17, 32, 4], tests of conditional independence [14, 32], and tests for higher order (Lancaster) interactions [24]. For these tests, consistency is guaranteed if and only if the observations are independent and identically distributed. Much real-world data fails to satisfy the i.i.d. assumption: audio signals, EEG recordings, text documents, financial time series, and samples obtained when running Markov Chain Monte Carlo, all show significant temporal dependence patterns.

The asymptotic behaviour of kernel test statistics becomes quite different when temporal dependencies exist within the samples. In recent work on independence testing using the Hilbert-Schmidt Independence Criterion (HSIC) [8], the asymptotic distribution of the statistic under the null hypothesis is obtained for a pair of independent time series, which satisfy an absolute regularity or a $\phi$-mixing assumption. In this case, the null distribution is shown to be an infinite weighted sum of *dependent* $\chi^2$-variables, as opposed to the sum of *independent* $\chi^2$-variables obtained in the i.i.d. setting [17]. The difference in the asymptotic null distributions has important implications in practice: under the i.i.d. assumption, an empirical estimate of the null distribution can be obtained by repeatedly permuting the time indices of one of the signals. This breaks the temporal dependence within the permuted signal, which causes the test to return an elevated number of false positives, when used for testing time series. To address this problem, an alternative estimate of the null distribution is proposed in [8], where the null distribution is simulated by repeatedly *shifting* one signal relative to the other. This preserves the temporal structure within each signal, while breaking the cross-signal dependence.

A serious limitation of the shift procedure in [8] is that it is specific to the problem of independence testing: there is no obvious way to generalise it to other testing contexts. For instance, we might have two time series, with the goal of comparing their marginal distributions - this is a generalization of the two-sample setting to which the shift approach does not apply.

We note, however, that many kernel tests have a test statistic with a particular structure: the Maximum Mean Discrepancy (MMD), HSIC, and the Lancaster interaction statistic, each have empirical estimates which can be cast as normalized $V$-statistics, $\frac{1}{n^{m-1}} \sum_{1 \leq i_1, \ldots, i_m \leq n} h(Z_{i_1}, \ldots, Z_{i_m})$, where $Z_{i_1}, \ldots, Z_{i_m}$ are samples from a random process at the time points $\{i_1, \ldots, i_m\}$. We show that a method of external randomization known as the *wild bootstrap* may be applied [22, 28] to simulate from the null distribution. In brief, the arguments of the above sum are repeatedly multiplied by random, user-defined time series. For a test of level $\alpha$, the $1 - \alpha$ quantile of the empirical distribution obtained using these perturbed statistics serves as the test threshold. This approach has the important advantage over [8] that it may be applied to *all* kernel-based tests for which $V$-statistics are employed, and not just for independence tests.

The main result of this paper is to show that the wild bootstrap procedure yields consistent tests for time series, i.e., tests based on the wild bootstrap have a Type I error rate (of wrongly rejecting the null hypothesis) approaching the design parameter $\alpha$, and a Type II error (of wrongly accepting the null) approaching zero, as the number of samples increases. We use this result to construct a two-sample test using MMD, and an independence test using HSIC. The latter procedure is applied both to testing for instantaneous independence, and to testing for independence across multiple time lags, for which the earlier shift procedure of [8] cannot be applied.

We begin our presentation in Section 2, with a review of the $\tau$-mixing assumption required of the time series, as well as of $V$-statistics (of which MMD and HSIC are instances). We also introduce the form taken by the wild bootstrap. In Section 3, we establish a general consistency result for the wild bootstrap procedure on $V$-statistics, which we apply to MMD and to HSIC in Section 4. Finally, in Section 5, we present a number of empirical comparisons: in the two sample case, we test for differences in audio signals with the same underlying pitch, and present a performance diagnostic for the output of a Gibbs sampler (the MCMC M.D.); in the independence case, we test for independence of two time series sharing a common variance (a characteristic of econometric models), and compare against the test of [4] in the case where dependence may occur at multiple, potentially unknown lags. Our tests outperform both the naive approach which neglects the dependence structure within the samples, and the approach of [4], when testing across multiple lags. Detailed proofs are found in the appendices of an accompanying technical report [9], which we reference from the present document as needed.

## 2 Background

The main results of the paper are based around two concepts: $\tau$-mixing [10], which describes the dependence within the time series, and $V$-statistics [27], which constitute our test statistics. In this section, we review these topics, and introduce the concept of wild bootstrapped $V$-statistics, which will be the key ingredient in our test construction.

**$\tau$-mixing.** The notion of $\tau$-mixing is used to characterise weak dependence. It is a less restrictive alternative to classical mixing coefficients, and is covered in depth in [10]. Let $\{Z_t, \mathcal{F}_t\}_{t \in \mathbb{N}}$ be a stationary sequence of integrable random variables, defined on a probability space $\Omega$ with a probability measure $P$ and a natural filtration $\mathcal{F}_t$. The process is called $\tau$-dependent if

$$\tau(r) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{r \leq i_1 \leq \ldots \leq i_l} \tau(\mathcal{F}_0, (Z_{i_1}, \ldots, Z_{i_l})) \overset{r \to \infty}{\longrightarrow} 0, \text{ where}$$

$$\tau(\mathcal{M}, X) = E \left( \sup_{g \in \Lambda} \left| \int g(t) P_{X|\mathcal{M}}(dt) - \int g(t) P_X(dt) \right| \right)$$

and $\Lambda$ is the set of all one-Lipschitz continuous real-valued functions on the domain of $X$. $\tau(\mathcal{M}, X)$ can be interpreted as the minimal $L_1$ distance between $X$ and $X^*$ such that $X \overset{d}{=} X^*$ and $X^*$ is independent of $\mathcal{M} \subset \mathcal{F}$. Furthermore, if $\mathcal{F}$ is rich enough, this $X^*$ can be constructed (see Proposition 4 in the Appendix). More information is provided in the Appendix B.

*V*-**statistics.** The test statistics considered in this paper are always $V$-statistics. Given the observations $Z = \{Z_t\}_{t=1}^n$, a $V$-statistic of a symmetric function $h$ taking $m$ arguments is given by

$$V(h, Z) = \frac{1}{n^m} \sum_{i \in N^m} h(Z_{i_1}, ..., Z_{i_m}), \tag{1}$$

where $N^m$ is a Cartesian power of a set $N = \{1, ..., n\}$. For simplicity, we will often drop the second argument and write simply $V(h)$.

We will refer to the function $h$ as to the *core* of the $V$-statistic $V(h)$. While such functions are usually called kernels in the literature, in this paper we reserve the term kernel for positive-definite functions taking two arguments. A core $h$ is said to be $j$-degenerate if for each $z_1, \ldots, z_j$ $Eh(z_1, \ldots, z_j, Z_{j+1}^*, \ldots, Z_m^*) = 0$, where $Z_{j+1}^*, \ldots, Z_m^*$ are independent copies of $Z_1$. If $h$ is $j$-degenerate for all $j \leq m - 1$, we will say that it is *canonical*. For a one-degenerate core $h$, we define an auxiliary function $h_2$, called the second component of the core, and given by $h_2(z_1, z_2) = Eh(z_1, z_2, Z_3^*, \ldots, Z_m^*)$. Finally we say that $nV(h)$ is a normalized $V$-statistic, and that a $V$-statistic with a one-degenerate core is a degenerate $V$-statistic. This degeneracy is common to many kernel statistics when the null hypothesis holds [15, 17, 24].

Our main results will rely on the fact that $h_2$ governs the asymptotic behaviour of normalized degenerate $V$-statistics. Unfortunately, the limiting distribution of such $V$-statistics is quite complicated - it is an infinite sum of *dependent* $\chi^2$-distributed random variables, with a dependence determined by the temporal dependence structure within the process $\{Z_t\}$ and by the eigenfunctions of a certain integral operator associated with $h_2$ [5, 8]. Therefore, we propose a bootstrapped version of the $V$-statistics which will allow a consistent approximation of this difficult limiting distribution.

**Bootstrapped *V*-statistic.** We will study two versions of the bootstrapped $V$-statistics

$$V_{b1}(h, Z) = \frac{1}{n^m} \sum_{i \in N^m} W_{i_1,n} W_{i_2,n} h(Z_{i_1}, ..., Z_{i_m}), \tag{2}$$

$$V_{b2}(h, Z) = \frac{1}{n^m} \sum_{i \in N^m} \tilde{W}_{i_1,n} \tilde{W}_{i_2,n} h(Z_{i_1}, ..., Z_{i_m}), \tag{3}$$

where $\{W_{t,n}\}_{1 \leq t \leq n}$ is an auxiliary wild bootstrap process and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n} \sum_{j=1}^n W_{j,n}$. This auxiliary process, proposed by [28, 22], satisfies the following assumption:

*Bootstrap assumption:* $\{W_{t,n}\}_{1 \leq t \leq n}$ is a row-wise strictly stationary triangular array independent of all $Z_t$ such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^{2+\sigma}| < \infty$ for some $\sigma > 0$. The autocovariance of the process is given by $EW_{s,n}W_{t,n} = \rho(|s - t|/l_n)$ for some function $\rho$, such that $\lim_{u \to 0} \rho(u) = 1$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = O(l_n)$. The sequence $\{l_n\}$ is taken such that $l_n = o(n)$ but $\lim_{n \to \infty} l_n = \infty$. The variables $W_{t,n}$ are $\tau$-weakly dependent with coefficients $\tau(r) \leq C\zeta^{\frac{r}{l_n}}$ for $r = 1, ..., n$, $\zeta \in (0, 1)$ and $C \in \mathbb{R}$.

As noted in in [22, Remark 2], a simple realization of a process that satisfies this assumption is $W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \epsilon_t$ where $W_{0,n}$ and $\epsilon_1, \ldots, \epsilon_n$ are independent standard normal random variables. For simplicity, we will drop the index $n$ and write $W_t$ instead of $W_{t,n}$. A process that fulfils the *bootstrap assumption* will be called *bootstrap process*. Further discussion of the wild bootstrap is provided in the Appendix A. The versions of the bootstrapped $V$-statistics in (2) and (3) were previously studied in [22] for the case of canonical cores of degree $m = 2$. We extend their results to higher degree cores (common within the kernel testing framework), which are not necessarily one-degenerate. When stating a fact that applies to both $V_{b1}$ and $V_{b2}$, we will simply write $V_b$, and the argument $Z$ will be dropped when there is no ambiguity.

## 3 Asymptotics of wild bootstrapped *V*-statistics

In this section, we present main Theorems that describe asymptotic behaviour of $V$-statistics. In the next section, these results will be used to construct kernel-based statistical tests applicable to dependent observations. Tests are constructed so that the $V$-statistic is degenerate under the null hypothesis and non-degenerate under the alternative. Theorem 1 guarantees that the bootstrapped $V$-statistic will converge to the same limiting null distribution as the simple $V$-statistic. Following [22], we will establish the convergence of the bootstrapped distribution to the desired asymptotic

distribution in the Prokhorov metric $\varphi$ [13, Section 11.3]), and ensure that this distance approaches zero *in probability* as $n \to \infty$. This two-part convergence statement is needed due to the additional randomness introduced by the $W_{j,n}$.

**Theorem 1.** *Assume that the stationary process $\{Z_t\}$ is $\tau$-dependent with $\tau(r) = O(r^{-6-\epsilon})$ for some $\epsilon > 0$. If the core $h$ is a Lipschitz continuous, one-degenerate, and bounded function of $m$ arguments and its $h_2$-component is a positive definite kernel, then $\varphi(n\binom{m}{2}V_b(h, Z), nV(h, Z)) \to 0$ in probability as $n \to \infty$, where $\varphi$ is Prokhorov metric.*

*Proof.* By Lemma 3 and Lemma 2 respectively, $\varphi(nV_b(h), nV_b(h_2))$ and $\varphi(nV(h), n\binom{m}{2}V(h_2))$ converge to zero. By [22, Theorem 3.1], $nV_b(h_2)$ and $nV(h_2, Z)$ have the same limiting distribution, i.e., $\varphi(nV_b(h_2), nV(h_2, Z)) \to 0$ in probability under certain assumptions. Thus, it suffices to check these assumptions hold: *Assumption A2.* (i) $h_2$ is one-degenerate and symmetric - this follows from Lemma 1; (ii) $h_2$ is a kernel - is one of the assumptions of this Theorem; (iii) $Eh_2(Z_1, Z_1) \leq \infty$ - by Lemma 7, $h_2$ is bounded and therefore has a finite expected value; (iv) $h_2$ is Lipschitz continuous - follows from Lemma 7. *Assumption B1.* $\sum_{r=1}^{n} r^2 \sqrt{\tau(r)} < \infty$. Since $\tau(r) = O(r^{-6-\epsilon})$ then $\sum_{r=1}^{n} r^2 \sqrt{\tau(r)} \leq C \sum_{r=1}^{n} r^{-1-\epsilon/2} \leq \infty$. *Assumption B2.* This assumption about the auxiliary process $\{W_t\}$ is the same as our *Bootstrap assumption*. $\square$

On the other hand, if the $V$-statistic is not degenerate, which is usually true under the alternative, it converges to some non-zero constant. In this setting, Theorem 2 guarantees that the bootstrapped $V$-statistic will converge to zero in probability. This property is necessary in testing, as it implies that the test thresholds computed using the bootstrapped $V$-statistics will also converge to zero, and so will the corresponding Type II error. The following theorem is due to Lemmas 4 and 5.

**Theorem 2.** *Assume that the process $\{Z_t\}$ is $\tau$-dependent with a coefficient $\tau(r) = O(r^{-6-\epsilon})$. If the core $h$ is a Lipschitz continuous, symmetric and bounded function of $m$ arguments, then $nV_{b2}(h)$ converges in distribution to some non-zero random variable with finite variance, and $V_{b1}(h)$ converges to zero in probability.*

Although both $V_{b2}$ and $V_{b1}$ converge to zero, the rate and the type of convergence are not the same: $nV_{b2}$ converges in law to some random variable while the behaviour of $nV_{b1}$ is unspecified. As a consequence, tests that utilize $V_{b2}$ usually give lower Type II error then the ones that use $V_{b1}$. On the other hand, $V_{b1}$ seems to better approximate $V$-statistic distribution under the null hypothesis. This agrees with our experiments in Section 5 as well as with those in [22, Section 5]).

## 4  Applications to Kernel Tests

In this section, we describe how the wild bootstrap for $V$-statistics can be used to construct kernel tests for independence and the two-sample problem, which are applicable to weakly dependent observations. We start by reviewing the main concepts underpinning the kernel testing framework.

For every symmetric, positive definite function, i.e., *kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, there is an associated reproducing kernel Hilbert space $\mathcal{H}_k$ [3, p. 19]. The kernel embedding of a probability measure $P$ on $\mathcal{X}$ is an element $\mu_k(P) \in \mathcal{H}_k$, given by $\mu_k(P) = \int k(\cdot, x) \, dP(x)$ [3, 29]. If a measurable kernel $k$ is bounded, the mean embedding $\mu_k(P)$ exists for all probability measures on $\mathcal{X}$, and for many interesting bounded kernels $k$, including the Gaussian, Laplacian and inverse multi-quadrics, the kernel embedding $P \mapsto \mu_k(P)$ is injective. Such kernels are said to be *characteristic* [30]. The RKHS-distance $\|\mu_k(P_x) - \mu_k(P_y)\|^2_{\mathcal{H}_k}$ between embeddings of two probability measures $P_x$ and $P_y$ is termed the Maximum Mean Discrepancy (MMD), and its empirical version serves as a popular statistic for non-parametric two-sample testing [15]. Similarly, given a sample of paired observations $\{(X_i, Y_i)\}_{i=1}^{n} \sim P_{xy}$, and kernels $k$ and $l$ respectively on $X$ and $Y$ domains, the RKHS-distance $\|\mu_\kappa(P_{xy}) - \mu_\kappa(P_x P_y)\|^2_{\mathcal{H}_\kappa}$ between embeddings of the joint distribution and of the product of the marginals, measures dependence between $X$ and $Y$. Here, $\kappa((x, y), (x', y')) = k(x, x')l(y, y')$ is the kernel on the product space of $X$ and $Y$ domains. This quantity is called Hilbert-Schmidt Independence Criterion (HSIC) [16, 17]. When characteristic RKHSs are used, the HSIC is zero iff $X \perp\!\!\!\perp Y$: this follows from [18]. The empirical statistic is written $\widehat{\mathrm{HSIC}}_\kappa = \frac{1}{n^2}\mathrm{Tr}(KHLH)$ for kernel matrices $K$ and $L$ and the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

4

## 4.1 Wild Bootstrap For MMD

Denote the observations by $\{X_i\}_{i=1}^{n_x} \sim P_x$, and $\{Y_j\}_{j=1}^{n_y} \sim P_y$. Our goal is to test the null hypothesis $\mathbf{H}_0 : P_x = P_y$ vs. the alternative $\mathbf{H}_1 : P_x \neq P_y$. In the case where samples have equal sizes, i.e., $n_x = n_y$, application of the wild bootstrap to MMD-based tests on dependent samples is straightforward: the empirical MMD can be written as a $V$-statistic with the core of degree two on pairs $z_i = (x_i, y_i)$ given by $h(z_1, z_2) = k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2)$. It is clear that whenever $k$ is Lipschitz continuous and bounded, so is $h$. Moreover, $h$ is a valid positive definite kernel, since it can be represented as an RKHS inner product $\langle k(\cdot, x_1) - k(\cdot, y_1), k(\cdot, x_2) - k(\cdot, y_2) \rangle_{\mathcal{H}_k}$. Under the null hypothesis, $h$ is also one-degenerate, i.e., $Eh((x_1, y_1), (X_2, Y_2)) = 0$. Therefore, we can use the bootstrapped statistics in (2) and (3) to approximate the null distribution and attain a desired test level.

When $n_x \neq n_y$, however, it is no longer possible to write the empirical MMD as a one-sample $V$-statistic. We will therefore require the following bootstrapped version of MMD

$$\widehat{\mathrm{MMD}}_{k,b} = \frac{1}{n_x^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} \tilde{W}_i^{(x)} \tilde{W}_j^{(x)} k(x_i, x_j) - \frac{1}{n_x^2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} \tilde{W}_i^{(y)} \tilde{W}_j^{(y)} k(y_i, y_j)$$
$$- \frac{2}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \tilde{W}_i^{(x)} \tilde{W}_j^{(y)} k(x_i, y_j), \tag{4}$$

where $\tilde{W}_t^{(x)} = W_t^{(x)} - \frac{1}{n_x} \sum_{i=1}^{n_x} W_i^{(x)}$, $\tilde{W}_t^{(y)} = W_t^{(y)} - \frac{1}{n_y} \sum_{j=1}^{n_y} W_j^{(y)}$; $\{W_t^{(x)}\}$ and $\{W_t^{(y)}\}$ are two auxiliary wild bootstrap processes that are independent of $\{X_t\}$ and $\{Y_t\}$ and also independent of each other, both satisfying the bootstrap assumption in Section 2. The following Proposition shows that the bootstrapped statistic has the same asymptotic null distribution as the empirical MMD. The proof follows that of [22, Theorem 3.1], and is given in the Appendix.

**Proposition 1.** *Let $k$ be bounded and Lipschitz continuous, and let $\{X_t\}$ and $\{Y_t\}$ both be $\tau$-dependent with coefficients $\tau(r) = O(r^{-6-\epsilon})$, but independent of each other. Further, let $n_x = \rho_x n$ and $n_y = \rho_y n$ where $n = n_x + n_y$. Then, under the null hypothesis $P_x = P_y$, $\varphi\left(\rho_x \rho_y n \widehat{MMD}_k, \rho_x \rho_y n \widehat{MMD}_{k,b}\right) \to 0$ in probability as $n \to \infty$, where $\varphi$ is the Prokhorov metric and $\widehat{MMD}_k$ is the MMD between empirical measures.*

## 4.2 Wild Bootstrap For HSIC

Using HSIC in the context of random processes is not new in the machine learning literature. For a 1-approximating functional of an absolutely regular process [6], convergence in probability of the empirical HSIC to its population value was shown in [33]. No asymptotic distributions were obtained, however, nor was a statistical test constructed. The asymptotics of a normalized $V$-statistic were obtained in [8] for absolutely regular and $\phi$-mixing processes [12]. Due to the intractability of the null distribution for the test statistic, the authors propose a procedure to approximate its null distribution using circular shifts of the observations leading to tests of instantaneous independence, i.e., of $X_t \perp\!\!\!\perp Y_t$, $\forall t$. This was shown to be consistent under the null (i.e., leading to the correct Type I error), however consistency of the shift procedure under the alternative is a challenging open question (see [8, Section A.2] for further discussion). In contrast, as shown below in Propositions 2 and 3 (which are direct consequences of the Theorems 1 and 2), the wild bootstrap guarantees test consistency under both hypotheses: null and alternative, which is a major advantage. In addition, the wild bootstrap can be used in constructing a test for the harder problem of determining independence across multiple lags simultaneously, similar to the one in [4].

Following symmetrisation, it is shown in [17, 8] that the empirical HSIC can be written as a degree four $V$-statistic with core given by

$$h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})[l(y_{\pi(1)}, y_{\pi(2)}) + l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})],$$

where we denote by $S_n$ the group of permutations over $n$ elements. Thus, we can directly apply the theory developed for higher-order $V$-statistics in Section 3. We consider two types of tests: instantaneous independence and independence at multiple time lags.

Table 1: Rejection rates for two-sample experiments. **MCMC**: sample size=500; a Gaussian kernel with bandwidth $\sigma = 1.7$ is used; every second Gibbs sample is kept (i.e., after a pass through both dimensions). **Audio**: sample sizes are $(n_x, n_y) = \{(300, 200), (600, 400), (900, 600)\}$; a Gaussian kernel with bandwidth $\sigma = 14$ is used. **Both**: wild bootstrap uses blocksize of $l_n = 20$; averaged over at least 200 trials. The Type II error for all tests was zero

| | experiment \ method | permutation | $\widehat{\text{MMD}}_{k,b}$ | $V_{b1}$ | $V_{b2}$ |
|---|---|---|---|---|---|
| **MCMC** | i.i.d. vs i.i.d. ($\mathbf{H_0}$) | .040 | .025 | .012 | .070 |
| | i.i.d. vs Gibbs ($\mathbf{H_0}$) | .528 | .100 | .052 | .105 |
| | Gibbs vs Gibbs ($\mathbf{H_0}$) | .680 | .110 | .060 | .100 |
| **Audio** | $\mathbf{H_0}$ | $\{.970,.965,.995\}$ | $\{.145,.120,.114\}$ | | |
| | $\mathbf{H_1}$ | $\{1,1,1\}$ | $\{.600,.898,.995\}$ | | |

**Test of instantaneous independence**  Here, the null hypothesis $\mathbf{H_0}$ is that $X_t$ and $Y_t$ are independent at all times $t$, and the alternative hypothesis $\mathbf{H_1}$ is that they are dependent.

**Proposition 2.** *Under the null hypothesis, if the stationary process $Z_t = (X_t, Y_t)$ is $\tau$-dependent with a coefficient $\tau(r) = O\left(r^{-6-\epsilon}\right)$ for some $\epsilon > 0$, then $\varphi(6nV_b(h), nV(h)) \to 0$ in probability, where $\varphi$ is the Prokhorov metric.*

*Proof.* Since $k$ and $l$ are bounded and Lipschitz continuous, the core $h$ is bounded and Lipschitz continuous. One-degeneracy under the null hypothesis was stated in [17, Theorem 2], and that $h_2$ is a kernel is shown in [17, section A.2, following eq. (11)]. The result follows from Theorem 1. $\square$

The following proposition holds by the Theorem 2, since the core $h$ is Lipschitz continuous, symmetric and bounded.

**Proposition 3.** *If the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(r) = O\left(r^{-6-\epsilon}\right)$ for some $\epsilon > 0$, then under the alternative hypothesis $nV_{b2}(h)$ converges in distribution to some random variable with a finite variance and $V_{b1}$ converges to zero in probability.*

**Lag-HSIC**  Propositions 2 and 3 also allow us to construct a test of time series independence that is similar to one designed by [4]. Here, we will be testing against a broader null hypothesis: $X_t$ and $Y_{t'}$ are independent for $|t - t'| < M$ for an arbitrary large but fixed $M$. In the Appendix, we show how to construct a test when $M \to \infty$, although this requires an additional assumption about the uniform convergence of cumulative distribution functions.

Since the time series $Z_t = (X_t, Y_t)$ is stationary, it suffices to check whether there exists a dependency between $X_t$ and $Y_{t+m}$ for $-M \leq m \leq M$. Since each lag corresponds to an individual hypothesis, we will require a Bonferroni correction to attain a desired test level $\alpha$. We therefore define $q = 1 - \frac{\alpha}{2M+1}$. The shifted time series will be denoted $Z_t^m = (X_t, Y_{t+m})$. Let $S_{m,n} = nV(h, Z^m)$ denote the value of the normalized HSIC statistic calculated on the shifted process $Z_t^m$. Let $F_{b,n}$ denote the empirical cumulative distribution function obtained by the bootstrap procedure using $nV_b(h, Z)$. The test will then reject the null hypothesis if the event $\mathcal{A}_n = \left\{\max_{-M \leq m \leq M} S_{m,n} > F_{b,n}^{-1}(q)\right\}$ occurs. By a simple application of the union bound, it is clear that the asymptotic probability of the Type I error will be $\lim_{n \to \infty} P_{\mathbf{H_0}}(\mathcal{A}_n) \leq \alpha$. On the other hand, if the alternative holds, there exists some $m$ with $|m| \leq M$ for which $V(h, Z^m) = n^{-1}S_{m,n}$ converges to a non-zero constant. In this case

$$P_{\mathbf{H_1}}(\mathcal{A}_n) \geq P_{\mathbf{H_1}}(S_{m,n} > F_{b,n}^{-1}(q)) = P_{\mathbf{H_1}}(n^{-1}S_{m,n} > n^{-1}F_{b,n}^{-1}(q)) \to 1 \tag{5}$$

as long as $n^{-1}F_{b,n}^{-1}(q) \to 0$, which follows from the convergence of $V_b$ to zero in probability shown in Proposition 3. Therefore, the Type II error of the multiple lag test is guaranteed to converge to zero as the sample size increases. Our experiments in the next Section demonstrate that while this procedure is defined over a finite range of lags, it results in tests more powerful than the procedure for an infinite number of lags proposed in [4]. We note that a procedure that works for an infinite number of lags, although possible to construct, does not add much practical value under the present assumptions. Indeed, since the $\tau$-mixing assumption applies to the joint sequence $Z_t = (X_t, Y_t)$,
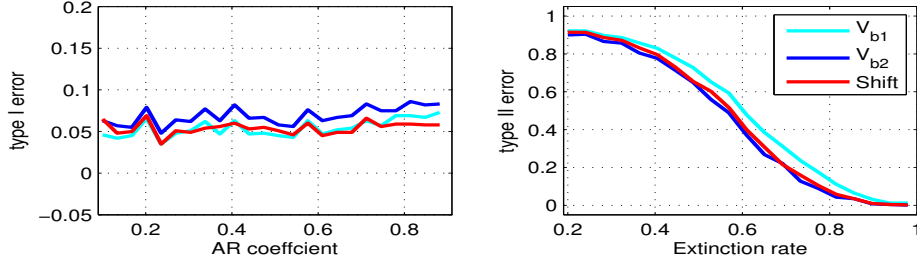
Figure 1: Comparison of Shift-HSIC and tests based on $V_{b1}$ and $V_{b2}$. The left panel shows the performance under the null hypothesis, where a larger AR coefficient implies a stronger temporal dependence. The right panel show the performance under the alternative hypothesis, where a larger extinction rate implies a greater dependence between processes.
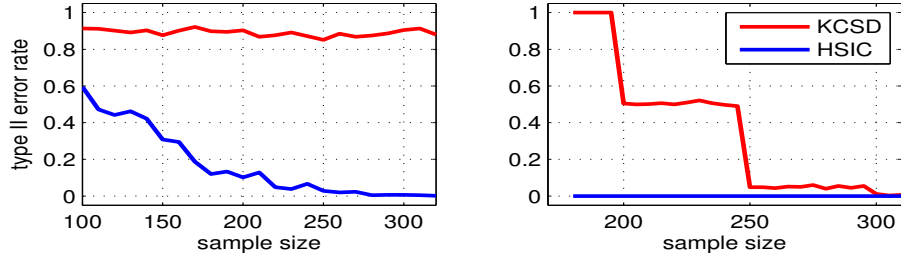


Figure 2: In both panel Type II error is plotted. The left panel presents the error of the lag-HSIC and KCSD algorithms for a process following dynamics given by the equation (6). The errors for a process with dynamics given by equations (7) and (8) are shown in the right panel. The X axis is indexed by the time series length, i.e., sample size. The Type I error was around 5%.

dependence between $X_t$ and $Y_{t+m}$ is bound to disappear at a rate of $o(m^{-6})$, i.e., the variables both within and across the two series are assumed to become gradually independent at large lags.

## 5 Experiments

**The MCMC M.D.** We employ MMD in order to diagnose how far an MCMC chain is from its stationary distribution [26, Section 5], by comparing the MCMC sample to a benchmark sample. A hypothesis test of whether the sampler has converged based on the standard permutation-based bootstrap leads to too many rejections of the null hypothesis, due to dependence within the chain. Thus, one would require heavily thinned chains, which is wasteful of samples and computationally burdensome. Our experiments indicate that the wild bootstrap approach allows consistent tests directly on the chains, as it attains a desired number of false positives.

To assess performance of the wild bootstrap in determining MCMC convergence, we consider the situation where samples $\{X_i\}$ and $\{Y_i\}$ are bivariate, and both have the identical marginal distribution given by an elongated normal $P = \mathcal{N}\left(\left[\begin{array}{cc} 0 & 0 \end{array}\right], \left[\begin{array}{cc} 15.5 & 14.5 \\ 14.5 & 15.5 \end{array}\right]\right)$. However, they could have arisen either as independent samples, or as outputs of the Gibbs sampler with stationary distribution $P$. Table 1 shows the *rejection rates* under the significance level $\alpha = 0.05$. It is clear that in the case where at least one of the samples is a Gibbs chain, the permutation-based test has a Type I error much larger than $\alpha$. The wild bootstrap using $V_{b1}$ (without artificial degeneration) yields the correct Type I error control in these cases. Consistent with findings in [22, Section 5], $V_{b1}$ mimics the null distribution better than $V_{b2}$. The bootstrapped statistic $\widehat{\mathrm{MMD}}_{k,b}$ in (4) which also relies on the artificially degenerated bootstrap processes, behaves similarly to $V_{b2}$. In the alternative scenario where $\{Y_i\}$ was taken from a distribution with the same covariance structure but with the mean set to $\mu = \left[\begin{array}{cc} 2.5 & 0 \end{array}\right]$, the Type II error for all tests was zero.

**Pitch-evoking sounds** Our second experiment is a two sample test on sounds studied in the field of pitch perception [20]. We synthesise the sounds with the fundamental frequency parameter of treble C, subsampled at 10.46kHz. Each $i$-th period of length $\Omega$ contains $d = 20$ audio samples

7

at times $0 = t_1 < \ldots < t_d < \Omega$ – we treat this whole vector as a single observation $X_i$ or $Y_i$, i.e., we are comparing distributions on $\mathbb{R}^{20}$. Sounds are generated based on the AR process $a_i = \lambda a_{i-1} + \sqrt{1 - \lambda^2} \epsilon_i$, where $a_0, \epsilon_i \sim \mathcal{N}(0, I_d)$, with $X_{i,r} = \sum_j \sum_{s=1}^{d} a_{j,s} \exp\left( -\frac{(t_r - t_s - (j-i)\Omega)^2}{2\sigma^2} \right)$. Thus, a given pattern – a smoothed version of $a_0$ – slowly varies, and hence the sound deviates from periodicity, but still evokes a pitch. We take $X$ with $\sigma = 0.1\Omega$ and $\lambda = 0.8$, and $Y$ is either an independent copy of $X$ (null scenario), or has $\sigma = 0.05\Omega$ (alternative scenario) (Variation in the smoothness parameter changes the width of the spectral envelope, i.e., the brightness of the sound). $n_x$ is taken to be different from $n_y$. Results in Table 1 demonstrate that the approach using the wild bootstrapped statistic in (4) allows control of the Type I error and reduction of the Type II error with increasing sample size, while the permutation test virtually always rejects the null hypothesis. As in [22] and the MCMC example, the artificial degeneration of the wild bootstrap process causes the Type I error to remain above the design parameter of $0.05$, although it can be observed to drop with increasing sample size.

**Instantaneous independence**   To examine instantaneous independence test performance, we compare it with the Shift-HSIC procedure [8] on the 'Extinct Gaussian' autoregressive process proposed in the [8, Section 4.1]. Using exactly the same setting we compute type I error as a function of the temporal dependence and type II error as a function of extinction rate. Figure 1 shows that all three tests (Shift-HSIC and tests based on $V_{b1}$ and $V_{b2}$) perform similarly.

**Lag-HSIC**   The KCSD [4] is, to our knowledge, the only test procedure to reject the null hypothesis if there exist $t, t'$ such that $Z_t$ and $Z_{t'}$ are dependent. In the experiments, we compare lag-HSIC with KCSD on two kinds of processes: one inspired by econometrics and one from [4].

In lag-HSIC, the number of lags under examination was equal to $\max\{10, \log n\}$, where $n$ is the sample size. We used Gaussian kernels with widths estimated by the median heuristic. The cumulative distribution of the $V$-statistics was approximated by samples from $nV_{b2}$. To model the tail of this distribution, we have fitted the generalized Pareto distribution to the bootstrapped samples ([23] shows that for a large class of underlying distribution functions such an approximation is valid). The first process is a pair of two time series which share a common variance,

$$X_t = \epsilon_{1,t}\sigma_t^2, \quad Y_t = \epsilon_{2,t}\sigma_t^2, \sigma_t^2 = 1 + 0.45(X_{t-1}^2 + Y_{t-1}^2), \quad \epsilon_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad i \in \{1, 2\}. \quad (6)$$

The above set of equations is an instance of the VEC dynamics [2] used in econometrics to model market volatility. The left panel of the Figure 2 presents the Type II error rate: for KCSD it remains at 90% while for lag-HSIC it gradually drops to zero. The Type I error, which we calculated by sampling two independent copies $(X_t^{(1)}, Y_t^{(1)})$ and $(X_t^{(2)}, Y_t^{(2)})$ of the process and performing the tests on the pair $(X_t^{(1)}, Y_t^{(2)})$, was around 5% for both of the tests.

Our next experiment is a process sampled according to the dynamics proposed by [4],

$$X_t = \cos(\phi_{t,1}), \qquad \phi_{t,1} = \phi_{t-1,1} + 0.1\epsilon_{1,t} + 2\pi f_1 T_s, \quad \epsilon_{1,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad (7)$$

$$Y_t = [2 + C\sin(\phi_{t,1})]\cos(\phi_{t,2}), \quad \phi_{t,2} = \phi_{t-1,2} + 0.1\epsilon_{2,t} + 2\pi f_2 T_s, \quad \epsilon_{2,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad (8)$$

with parameters $C = .4$, $f_1 = 4Hz$, $f_2 = 20Hz$, and frequency $\frac{1}{T_s} = 100Hz$. We compared performance of the KCSD algorithm, with parameters set to vales recommended in [4], and the lag-HSIC algorithm. The Type II error of lag-HSIC, presented in the right panel of the Figure 2, is substantially lower than that of KCSD. The Type I error ($C = 0$) is equal or lower than 5% for both procedures. Most oddly, KCSD error seems to converge to zero in steps. This may be due to the method relying on a spectral decomposition of the signals across a fixed set of bands. As the number of samples increases, the quality of the spectrogram will improve, and dependence will become apparent in bands where it was undetectable at shorter signal lengths.

## References

[1] M.A. Arcones. The law of large numbers for U-statistics under absolute regularity. *Electron. Comm. Probab*, 3:13–19, 1998.

[2] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH models: a survey. *J. Appl. Econ.*, 21(1):79–109, January 2006.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[4] M. Besserve, N.K. Logothetis, and B. Schlkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*, pages 2535–2543. 2013.

[5] I.S. Borisov and N.V. Volodko. Orthogonal series and limit theorems for canonical U- and V-statistics of stationary connected observations. *Siberian Adv. Math.*, 18(4):242–257, 2008.

[6] S. Borovkova, R. Burton, and H. Dehling. Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Trans. Amer. Math. Soc.*, 353(11):4261–4318, 2001.

[7] R. Bradley et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.

[8] K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In *ICML*, 2014.

[9] Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. tech. report. *arXiv preprint arXiv:1408.5404*, 2014.

[10] J. Dedecker, P. Doukhan, G. Lang, S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190. Springer, 2007.

[11] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

[12] P. Doukhan. *Mixing*. Springer, 1994.

[13] R.M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

[14] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[15] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[16] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[17] A. Gretton, K. Fukumizu, C Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *NIPS*, volume 20, pages 585–592, 2007.

[18] Arthur Gretton. A simpler condition for consistency of a kernel independence test. arXiv:1501.06103, 2015.

[19] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*. 2008.

[20] P. Hehrmann. *Pitch Perception as Probabilistic Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2011.

[21] A. Leucht. Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli*, 18(2):552–585, 2012.

[22] A. Leucht and M.H. Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

[23] J. Pickands III. Statistical inference using extreme order statistics. *Ann. Statist.*, pages 119–131, 1975.

[24] D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *NIPS*, pages 1124–1132, 2013.

[25] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2702, 2013.

[26] D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.

[27] R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[28] X. Shao. The dependent wild bootstrap. *J. Amer. Statist. Assoc.*, 105(489):218–235, 2010.

[29] A. J Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume LNAI4754, pages 13–31, Berlin/Heidelberg, 2007. Springer-Verlag.

[30] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

[31] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[32] K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, pages 804–813, 2011.

[33] X. Zhang, L. Song, A. Gretton, and A. Smola. Kernel measures of independence for non-iid data. In *NIPS*, volume 22, 2008.

# A  An Introduction to the Wild Bootstrap

Bootstrap methods aim to evaluate the accuracy of the sample estimates - they are particularly useful when dealing with complicated distributions, or when the assumptions of a parametric procedure are in doubt. Bootstrap methods randomize the dataset used for the sample estimate calculation, so that a new dataset with a similar statistical properties is obtained, e.g. one popular method is resampling. In the wild bootstrap method the observations in the dataset are multiplied by appropriate random numbers. To present the idea behind the wild bootstrap we will discuss a toy example similar to that in [28], and then relate it to the wild bootstrap method used in this article.

Consider a stationary autoregressive-moving-average random process $\{X_i\}_{i \in \mathbf{Z}}$ with zero mean. The normalized sample mean of the process $X_t$ has normal distribution

$$\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_\infty^2), \tag{9}$$

where $\sigma_\infty^2 = \sum_{j=-\infty}^{j=\infty} cov(X_0, X_j)$. The variance $\sigma_\infty^2$ is not easy to estimate (the naive approach of approximating different covariances separately and summing them has several drawbacks, e.g. how many empirical covariances should be calculated?). Using the wild bootstrap method we will construct processes that mimic behaviour of the $X_t$ process and then use them to approximate the distribution of the normalized sample mean, $\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}}$. The bootstrap process used to to randomize the sample meets the following criteria:

- $\{W_{t,n}\}_{1 \leq t \leq n}$ is a row-wise strictly stationary triangular array independent of all $X_t$, such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^{2+\sigma}| < \infty$ for some $\sigma > 0$.
- The autocovariance of the process is given by $EW_{s,n}W_{t,n} = \rho(|s-t|/l_n)$ for some function $\rho$, such that $\lim_{u \to 0} \rho(u) = 1$.
- The sequence $\{l_n\}$ is taken such that $\lim_{n \to \infty} l_n = \infty$.

A process that fulfils those criteria, given also in the main article, is

$$W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \epsilon_t \tag{10}$$

We need to show that the distribution of the normalized sample mean of the process $Y_t^n = W_t^n X_t$, where $|t| \leq n$, mimics the distribution $N(0, \sigma_\infty^2)$. It suffices to calculate the expected value and correlations:

$$EY_t^n = EW_t^n X_t = 0, \tag{11}$$
$$cov(Y_0^n, Y_t^n) = cov(X_0, X_t)cov(Y_0^n, Y_t^n) = cov(X_0, X_t)\rho(|t|/l_n) \to cov(X_0, X_t) \tag{12}$$

The asymptotic auto-covariance structure of the process $Y_t$ is the same as the auto-covariance structure of the process $X_t$. Therefore

$$\frac{\sum_{i=1}^{N} Y_i}{\sqrt{n}} \xrightarrow{d} N(0, \sigma_\infty). \tag{13}$$

This mechanism is used in [22]. Recall that, under some assumptions, a normalized V-statistic can be written as

$$\sum_{k=0}^{\infty} \lambda_k \left( \frac{\sum_{i=1}^{n} \phi_k(X_i)}{\sqrt{n}} \right)^2 \overset{P}{=} \frac{1}{n} \sum_{1 \leq i,j \leq n} h(X_i, X_j)$$

where $\lambda_k$ are eigenvalues and $\phi_k$ are eigenfunction of the kernel $h$, respectively. Since $E\phi_k(X_i) = 0$ (degeneracy condition) one may replace

$$\frac{\sum_{i=1}^{n} \phi_k(X_i)}{\sqrt{n}}$$

with a bootstrapped version

$$\frac{\sum_{i=1}^{n} W_t^n \phi_k(X_i)}{\sqrt{n}},$$

and conclude, as in the toy example, that the limiting distribution of the single component of the sum $\sum_k \lambda_k$... remains the same. One of the main contributions of [22] is in showing that the distribution of the whole sum $\sum_k \lambda_k \left( \frac{\sum_{i=1}^n W_t^n \phi_k(X_i)}{\sqrt{n}} \right)^2$ with the components bootstrapped converges in a particular sense (in probability in Prokhorov metric) to the distribution of the normalized V-statistic, $\frac{1}{n} \sum_{1 \le i,j \le n} h(X_i, X_j)$.

## B   Relation between $\beta$,$\phi$ and $\tau$ mixing

**Strong mixing coefficients.**   A process is called absolutely regular ($\beta$-mixing) if $\beta(m) \to 0$, where

$$\beta(m) = \frac{1}{2} \sup_n \sup \sum_{i=1}^I \sum_{j=1}^J |P(A_i \cap B_j) - P(A_i)P(B_j)|.$$

The second supremum in the $\beta(m)$ definition is taken over all pairs of finite partitions $\{A_1, \cdots, A_I\}$ and $\{B_1, \cdots, B_J\}$ of the sample space such that $A_i \in \mathcal{A}_1^n$ and $B_j \in \mathcal{A}_{n+m}^\infty$, and $\mathcal{A}_b^c$ is a sigma field spanned by a subsequence, $\mathcal{A}_b^c = \sigma(Z_b, Z_{b+1}, ..., Z_c)$. A process is called uniform mixing ($\phi$-mixing) if $\phi(m) \to 0$, where

$$\phi(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^\infty} |P(B|A) - P(B)|.$$

The process is called strongly mixing ($\alpha$-mixing) if $\alpha(m) \to 0$, where

$$\alpha(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^\infty} |P(B \cap A) - P(B)P(A)|.$$

By [7] we have $\alpha(m) \le \beta(m) \le \phi(m)$ .

**Weak mixing The expected value and variance ofcoefficients.**   The process is called $\tilde{\alpha}$-mixing if $\tilde{\alpha}(m) \to 0$, where

$$\tilde{\alpha}(m) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{m \le i_1 \le ... \le i_l} \tilde{\alpha}(\mathcal{F}_0, (Z_{i_1}, ..., Z_{i_l})) \overset{r \to \infty}{\longrightarrow} 0, \text{ where}$$

$$\tilde{\alpha}(\mathcal{M}, X) = \sup_{g \in \Lambda} \| E(g(X)|\mathcal{M}) - Eg(X) \|_1$$

and $\Lambda$ is the set of all one-Lipschitz continuous real-valued functions on the domain of $X$. The other weak mixing coefficient, already introduced, is $\tau$-mixing. [10, Remark 2.4] show that $\tilde{\alpha}(m) \le 2\alpha(m)$. [11, Proposition 2] relates $\tau$-mixing and $\tilde{\alpha}$ mixing, as follows: if $Q_x$ is the generalized inverse of the tail function

$$Q_x(u) = \inf_{t \in R} \{P(|X| > t) \le u\},$$

then

$$\tau(\mathcal{M}, X) \le 2 \int_0^{\tilde{\alpha}(\mathcal{M},X)} Q_x(u)du.$$

While this definition can be hard to interpret, it can be simplified in the case $E|X|^p = M$ for some $p > 1$, since via Markov's inequality $P(|X| > t) \le \frac{M}{t^p}$, and thus $\frac{M}{t^p} \le u$ implies $P(|X| > t) \le u$. Therefore $Q'(u) = \frac{M}{\sqrt[p]{u}} \ge Q_x(u)$. As a result, under the assumption that the real valued random variable is $p$-integrable for some $p > 1$, we have the following inequality

$$\frac{\sqrt[p]{\tilde{\alpha}(\mathcal{M}, X)}}{M} \ge C\tau(\mathcal{M}, X)$$

## C   Proofs

The proofs are organized as follows

- Subsection C.1 introduces notation and states basic results concerning $V$-statistics.

- Subsection C.2 provides proofs of Lemmas 2 and 3 that constitute for most of the poof of the Theorem 1.
- Subsection C.2 provides proofs of Lemmas 1 and 5 from which Theorem 2 follows.
- Subsection C.4 provides proof of the Proposition 1.
- Finally, section C.5 lists all auxiliary lemmas.

## C.1 Notation and basic facts.

The following Lemma introduces notation and basic facts about $V$-statistics. These can be found can be found in the [27, Section 5.1.5] (page 178).

**Lemma 1.** *[27, Section 5.1.5] Any core $h$ can be written as a sum of canonical cores $h_1, ..., h_m$ and a constant $h_0$*

$$h(z_1, ..., z_m) = h_m(z_1, ..., z_m) + \sum_{1 \leq i_1 < ... < i_{m-1} \leq m} h_{m-1}(z_{i_1}, ..., z_{i_{m-1}})$$
$$+ ... + \sum_{1 \leq i_1 < i_2 \leq m} h_2(z_{i_1}, z_{i_2}) + \sum_{1 \leq i \leq m} h_1(z_i) + h_0$$

*We call $h_1, ..., h_m$ components of a core $h$. We do not call $h_0$ a component, its simply a constant. The components are defined in terms of auxiliary functions $g_c$*

$$g_c(z_1, ... z_c) = Eh(z_1, ..., z_c, Z^*_{c+1}, ..., Z^*_m)$$

*for each $c = 0, ..., m - 1$ and we put $g_m = h$. We define components as follows*

$$h_0 = g_0, \tag{14}$$
$$h_1(z_1) = g_1(z_1) - h_0, \tag{15}$$
$$h_2(z_1, z_2) = g_2(z_1, z_2) - h_1(z_1) - h_1(z_2) - h_0, \tag{16}$$
$$h_3(z_1, z_2, z_3) = g_3(z_1, z_2, z_3) - \sum_{1 \leq i < j \leq 3} h_2(z_i, z_j) - \sum_{1 \leq i \leq 3} h_1(z_i) - h_0, \tag{17}$$
$$\cdots, \tag{18}$$
$$h_m(z_1, ..., z_m) = g_m(z_1, ..., z_m) - \sum_{1 \leq i_1 < ... < i_{m-1} \leq m} h_{m-1}(z_{i_1}, ..., z_{i_{m-1}}) \tag{19}$$
$$- ... - \sum_{1 \leq i_1 < i_2 \leq m} h_2(z_{i_1}, z_{i_2}) - \sum_{1 \leq i \leq m} h_1(z_i) - h_0. \tag{20}$$
$$\tag{21}$$

*[27, Section 5.1.5] shows that components $h_c$ are symmetric (and therefore cores) and canonical. Finally a V-statistic of a core function $h$ can be written as a sum of V-statistics with canonical cores*

$$V(h) = V(h_m) + \binom{m}{1} V(h_{m-1}) + ... + \binom{m}{m-2} V(h_2) + \binom{m}{m-1} V(h_1) + h_0. \tag{22}$$

## C.2 Proof of Theorem 1

**Lemma 2.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. If core $h$ is Lipschitz continuous, one-degenerate, bounded, and its $h_2$ component is a kernel then its normalized $V$ statistic limiting distribution is proportional to its second component normalized $V$-statistic distribution. Shortly*

$$\lim_{n \to \infty} \varphi(nV(h), \binom{m}{2} nV(h_2)) = 0 \tag{23}$$

*where $\varphi$ denotes Prokhorov metric.*

*Proof.* Lemma 1 shows how to write core $h$ as a sum of its components $h_i$,

$$nV(h) = nV(h_m) + \binom{m}{1}nV(h_{m-1}) + ... + \binom{m}{m-2}nV(h_2) + \binom{m}{m-1}nV(h_1) + h_0. \tag{24}$$

By Lemma 7 all components of $h$ are bounded and Lipschitz continuous. Since $h$ is one-degenerate, $h_0 = 0$ and component $h_1(z)$ is equal to zero everywhere

$$h_1(z) = Eh(z, Z_2^*, ..., Z_m^*) = 0. \tag{25}$$

By Lemma 14, for $c \geq 3$, $nV(h_c)$ converges to zero in probability. Therefore the behaviour of $V(h)$ is determined by $\binom{m}{m-2}V(h_2) = \binom{m}{2}V(h_2)$. Convergence of $V(h_2)$ follows from the [22, Theorem 2.1]. $\qquad\square$

**Lemma 3.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. If core $h$ is Lipschitz continuous, one-degenerate, bounded, and its $h_2$ component is a kernel then its normalized and bootstrapped $V$ statistic limiting distribution is same its second component normalized and bootstrapped $V$-statistic distribution. Shortly*

$$\lim_{n \to \infty} \varphi(nV_b(h), nV_b(h_2)) = 0, \tag{26}$$

*in probability, where $\varphi$ denotes Prokhorov metric.*

*Proof.* We show that the proposition holds for $V_{b1}$ and then we prove that $\varphi(nV_{b2}(h), nV_{b1}(h)) = 0$ converges to zero - the concept is to [22].

$nV_{b1}$ **convergence.** We write core $h$ as a sum of components $h_i$ ( $h_0, h_1$ are equal to zero and therefore omitted). By Lemma 1

$$nV_{b1}(h) = \frac{1}{n^{m-1}} \sum_{i \in N^m} \Big[ W_{i_1} W_{i_2} h_m(Z_{i_1}, ..., Z_{i_m}) + \tag{27}$$

$$\sum_{1 \leq j_1 < ... < j_{m-1} \leq m} W_{i_1} W_{i_2} h_{m-1}(Z_{i_{j_1}}, ..., Z_{i_{j_{m-1}}}) + ... + \sum_{1 \leq j_1 < j_2 \leq m} W_{i_1} W_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \Big]. \tag{28}$$

Consider a sum associated with $h_2$

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} \sum_{1 \leq j_1 < j_2 \leq m} W_{i_1} W_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}). \tag{29}$$

Fix $j_1, j_2$. If $j_1 \neq 1, j_2 \neq 2$ then the sum

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} W_{i_1} W_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \overset{L.8}{=\!=\!=} \frac{1}{n^3} \sum_{i \in N^4} W_{i_1} W_{i_2} h_2(Z_{i_3}, Z_{i_4}) = \tag{30}$$

$$\left( \frac{1}{n} \sum_{i \in N^2} h_2(Z_{i_1}, Z_{i_2}) \right) \left( \frac{1}{n} \sum_{i=1}^n W_i \right)^2 \overset{L.16}{\longrightarrow} 0 \text{ in probability.} \tag{31}$$

If $j_1 = 1$ and $j_2 \neq 2$, then the sum

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} W_{i_1} W_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \overset{L.8}{=\!=\!=} \frac{1}{n^2} \sum_{i \in N^3} W_{i_1} W_{i_3} h_2(Z_{i_1}, Z_{i_3}) = \tag{32}$$

$$\left( \frac{1}{n} \sum_{i \in N^2} W_{i_1} h_2(Z_{i_1}, Z_{i_2}) \right) \left( \frac{1}{n} \sum_{i=1}^n W_i \right) \overset{L.15,18}{\longrightarrow} 0 \text{ in probability.} \tag{33}$$

The similar reasoning holds for $j_i = 2$ and $j_2 > 2$. The sum associated with $h_c$ for $c > 2$

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} \sum_{1 \leq j_1 < ... < j_c \leq m} W_{i_1} W_{i_2} h_c(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \overset{L.13}{\longrightarrow} 0 \text{ in probability.} \tag{34}$$

Therefore

$$\lim_{n\to\infty}\left(nV_b(h)-\sum_{i\in N^2}W_{i_1}W_{i_2}h_2(Z_{i_1},Z_{i_2})\right)\overset{P}{=}0. \tag{35}$$

what proofs the proposition for $V_{b1}$.

$nV_{b1}$ **convergence.** To prove that $nV_{b2}$ converges to the same distribution as $nV_{b1}$ we investigate the difference

$$V_{b1}-V_{b2}=\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}W_{i_2}h(Z_{i_1},...,Z_{i_m})-\frac{1}{n^{m-1}}\sum_{i\in N^m}\tilde{W}_{i_1}\tilde{W}_{i_2}h(Z_{i_1},...,Z_{i_m})= \tag{36}$$

$$\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}W_{i_2}h(\cdot)-\frac{1}{n^{m-1}}\sum_{i\in N^m}(W_{i_1}-\frac{1}{n}\sum_{j=1}^n W_j)(W_{i_2}-\frac{1}{n}\sum_{j=1}^n W_j)h(\cdot)= \tag{37}$$

$$-\left(\frac{2}{n}\sum_{j=1}^n W_j\right)\left(\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}h(\cdot)\right)+\left(\frac{1}{n^{m-1}}\sum_{i\in N^m}h(\cdot)\right)\left(\frac{1}{n}\sum_{j=1}^n W_j\right)^2. \tag{38}$$

The second term

$$\left(\frac{1}{n^{m-1}}\sum_{i\in N^m}h(Z_{i_1},...,Z_{i_m})\right)\left(\frac{1}{n}\sum_{j=1}^n W_j\right)^2\overset{L.16}{\longrightarrow}0\text{ in probability.} \tag{39}$$

Therefore we only need to show that the first term converges to zero

$$\left(\frac{2}{n}\sum_{j=1}^n W_j\right)\left(\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}h(Z_{i_1},...,Z_{i_m})\right). \tag{40}$$

Since $\frac{2}{n}\sum_{j=1}^n W_j$ converges in probability to zero (by Lemma 15) we only nee to show that $\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}h(Z_{i_1},...,Z_{i_m})$ converges. Using decomposition from Lemma 1 we write

$$\frac{1}{n^{m-1}}\sum_{i\in N^m}W_{i_1}h(Z_{i_1},...,Z_{i_m})=\frac{1}{n^{m-1}}\sum_{i\in N^m}\Big[W_{i_1}h_m(Z_{i_1},...,Z_{i_m})+ \tag{41}$$

$$\sum_{1\le j_1<...<j_{m-1}\le m}W_{i_1}h_{m-1}(Z_{i_{j_1}},...,Z_{i_{j_{m-1}}})+...+\sum_{1\le j_1<j_2\le m}W_{i_1}h_2(Z_{i_{j_1}},Z_{i_{j_2}})\Big]. \tag{42}$$

Term associated with $h_2$ can be written as

$$\frac{1}{n^{m-1}}\sum_{i\in N^m}\sum_{1\le j_1<j_2\le m}W_{i_1}h_2(Z_{i_{j_1}},Z_{i_{j_2}})= \tag{43}$$

$$=\begin{cases}n^{-1}\sum_{i\in N^2}W_{i_1}h_2(Z_{i_1},Z_{i_2}):j_1=1\text{ or }j_1=2\\\left(n^{-1}\sum_{i\in N^2}h_2(Z_{i_1},Z_{i_2})\right)\left(\frac{1}{n}\sum_{j=1}^n W_j\right):\text{ otherwise }.\end{cases} \tag{44}$$

In the first case ($j_1=1$ or $j_1=2$) Lemma 18 assures convergence. In the second case we use Lemma 16 to show convergence to zero. Other terms with $h_c$ for $c>2$

$$\frac{1}{n^{m-1}}\sum_{i\in N^m}\sum_{1\le j_1<...<j_c\le m}W_{i_1}h_{m-1}(Z_{i_{j_1}},...,Z_{i_{j_c}})\overset{L.13}{\longrightarrow}0\text{ in probability.} \tag{45}$$

$\square$

## C.3   Proof of Proposition 3

**Lemma 4.** $nV_{b2}(h)$ *converges to some non-zero random variable with finite variance.*

*Proof.* Using decomposition from the Lemma 1 we write core $h$ as a sum of components $h_c$ and $h_0$

$$nV_{b2}(h) = \frac{1}{n^{m-1}} \sum_{i \in N^m} \left[ h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} + \sum_{1 \leq j \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}) \right. \tag{46}$$

$$\left. \sum_{1 \leq j_1 < j_2 \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) + ... + \tilde{W}_{i_1} \tilde{W}_{i_2} h_m(Z_{i_1}, ..., Z_{i_m}) \right]. \tag{47}$$

We examine terms of the above sum starting form the one with $h_0$ - it is equal to zero

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} \overset{L.8}{=\!=\!=} \frac{1}{n} h_0 \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} = \frac{1}{n} h_0 \left( \sum_{i=1} \tilde{W}_i \right)^2 \overset{L.19}{=\!=\!=} 0. \tag{48}$$

Term with $h_1$ is zero as well, to see that fix $j$ and consider

$$T_j = \frac{1}{n^{m-1}} \sum_{i \in N^m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}). \tag{49}$$

If $j = 1$ then

$$T_1 \overset{L.8}{=\!=\!=} \frac{1}{n} \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_1}) = \frac{1}{n} \left( \sum_{i=1}^n \tilde{W}_i h_1(Z_i) \right) \left( \sum_{i=1} \tilde{W}_i \right) \overset{L.19}{=\!=\!=} 0. \tag{50}$$

If $j = 2$ the same reasoning holds and if $j > 2$

$$T_j \overset{L.8}{=\!=\!=} \frac{1}{n^2} \sum_{i \in N^3} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_3}) = \frac{1}{n} \left( \sum_{i=1}^n h_1(Z_i) \right) \left( \sum_{i=1} \tilde{W}_i \right)^2 \overset{L.19}{=\!=\!=} 0. \tag{51}$$

Term containing $h_2$

$$T_{j_1,j_2} = \frac{1}{n^{m-1}} \sum_{i \in N^m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \tag{52}$$

is not zero. In the Lemma 20 we show that for $j_1 = 1$ and $j_2 = 2$ it converges to some non-zero variable. For $j_1 = 1$ and $j_2 > 2$ we have

$$T_{1,j_2} \overset{L.8}{=\!=\!=} \frac{1}{n^2} \sum_{i \in N^3} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_1}, Z_{i_{j_2}}) = \frac{1}{n^2} \left( \sum_{i \in N^2} \tilde{W}_{i_1} h_2(Z_{i_1}, Z_{i_2}) \right) \left( \sum_{i=1} \tilde{W}_i \right) \overset{L.19}{=\!=\!=} 0. \tag{53}$$

Exactly the same argument works for $T_{j_2,1}$. If both $j_1 \neq 1$ and $j_2 \neq 2$ then

$$T_{j_1,j_2} \overset{L.8}{=\!=\!=} \frac{1}{n^3} \sum_{i \in N^4} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) = \frac{1}{n^3} \left( \sum_{i \in N^2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) \right) \left( \sum_{i=1} \tilde{W}_i \right)^2 \overset{L.19}{=\!=\!=} 0. \tag{54}$$

Terms containing $h_c$ for $c > 2$

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} \sum_{1 \leq j_1 < ... < j_c \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_{m-1}(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \overset{L.13}{\longrightarrow} 0 \tag{55}$$

converge to zero in probability. $\qquad \square$

**Lemma 5.** $V_{b1}$ *converges to zero in probability.*

*Proof.* The expected value and variance of $V_{b1}$ converge to 0, therefore $V_{b1}$ converges to zero in probability. Indeed for an expected value we have

$$EV_{b1} = \frac{1}{n^m} \sum_{i \in N^m} EW_{i_1} W_{i_2} Eh(Z_{i_1}, ..., Z_{i_m}) = \frac{1}{n^m} \sum_{i \in N^m} \rho(|i_2 - i_1|/l_n) Eh(\cdot) \leq \tag{56}$$

$$\frac{1}{n^m} \sum_{i \in N^m} \rho(|i_2 - i_1|/l_n) \parallel h \parallel_\infty = \parallel h \parallel_\infty \frac{1}{n^2} \sum_{i \in N^2} \rho(|i_2 - i_1|/l_n) \to 0. \tag{57}$$

Last convergence follows from the fact that $\sum_{i \in N^2} \rho(|i_2 - i_1|/l_n) \leq \sum_{r=1}^{n-1} n\rho(|r|/l_n) = O(nl_n)$. Similar reasoning shows convergence of $EV_{b1}^2$. $\qquad \square$

## C.4 Proof of Proposition 1

**Proposition.** *Let $k$ be bounded and Lipschitz continuous, and let $\{X_t\}$ and $\{Y_t\}$ both be $\tau$-dependent with coefficients $\tau(i) = O(i^{-6-\epsilon})$, but independent of each other. Further, let $n_x = \rho_x n$ and $n_y = \rho_y n$ where $n = n_x + n_y$. Then, under the null hypothesis $P_x = P_y$, $\varphi\left(\rho_x\rho_y n\widehat{MMD}_k, \rho_x\rho_y n\widehat{MMD}_{k,b}\right) \to 0$ in probability as $n \to \infty$, where $\varphi$ is the Prokhorov metric.*

*Proof.* Since $\widehat{MMD}_k$ is just the MMD between empirical measures using kernel $k$, it must be the same as the empirical MMD $\widehat{MMD}_{\tilde{k}}$ with centred kernel $\tilde{k}(x,x') = \langle k(\cdot,x) - Ek(\cdot,X), k(\cdot,x') - Ek(\cdot,X)\rangle_{\mathcal{H}_k}$ according to [25, Theorem 22]. Using the Mercer expansion, we can write

$$\rho_x\rho_y n\widehat{MMD}_k = \rho_x\rho_y n \sum_{r=1}^{\infty} \lambda_r \left(\frac{1}{n_x}\sum_{i=1}^{n_x}\Phi_r(x_i) - \frac{1}{n_y}\sum_{j=1}^{n_y}\Phi_r(y_j)\right)^2$$

$$= \sum_{r=1}^{\infty} \lambda_r \left(\sqrt{\frac{\rho_y}{n_x}}\sum_{i=1}^{n_x}\Phi_r(x_i) - \sqrt{\frac{\rho_x}{n_y}}\sum_{j=1}^{n_y}\Phi_r(y_j)\right)^2,$$

where $\{\lambda_r\}$ and $\{\Phi_r\}$ are the eigenvalues and the eigenfunctions of the integral operator $f \mapsto \int f(x)\tilde{k}(\cdot,x)dP_x(x)$ on $L_2(P_x)$. Similarly as in [22, Theorem 2.1], the above converges in distribution to $\sum_{r=1}^{\infty}\lambda_r Z_r^2$, where $\{Z_r\}$ are marginally standard normal, jointly normal and given by $Z_r = \sqrt{\rho_x}A_r - \sqrt{\rho_y}B_r$. $\{A_r\}$ and $\{B_r\}$ are in turn also marginally standard normal and jointly normal, with a dependence structure induced by that of $\{X_t\}$ and $\{Y_t\}$ respectively. This suggests individually bootstrapping each of the terms $\Phi_r(x_i)$ and $\Phi_r(y_j)$, giving rise to

$$\widehat{MMD}_{\tilde{k},b} = \sum_{r=1}^{\infty}\lambda_r\left(\frac{1}{n_x}\sum_{i=1}^{n_x}\Phi_r(x_i)\tilde{W}_i^{(x)} - \frac{1}{n_y}\sum_{j=1}^{n_y}\Phi_r(y_j)\tilde{W}_j^{(y)}\right)^2$$

$$= \frac{1}{n_x^2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_x}\tilde{W}_i^{(x)}\tilde{W}_j^{(x)}\tilde{k}(x_i,x_j) - \frac{1}{n_x^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\tilde{W}_i^{(y)}\tilde{W}_j^{(y)}\tilde{k}(y_i,y_j)$$

$$- \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\tilde{W}_i^{(x)}\tilde{W}_j^{(y)}\tilde{k}(x_i,y_j).$$

Now, since $\tilde{k}$ is degenerate under the null distribution, the first two terms (after appropriate normalization) converge in distribution to $\rho_x\sum_{r=1}^{\infty}\lambda_r A_r^2$ and $\rho_y\sum_{r=1}^{\infty}\lambda_r B_r^2$ by [22, Theorem 3.1] as required. The last term follows the same reasoning - it suffices to check part (b) of [22, Theorem 3.1] (which is trivial as processes $\{X_t\}$ and $\{Y_t\}$ are assumed to be independent of each other) and apply the continuous mapping theorem to obtain convergence to $-2\sqrt{\rho_x\rho_y}\sum_{r=1}^{\infty}\lambda_r A_r B_r$ implying that $\widehat{MMD}_{\tilde{k},b}$ has the same limiting distribution as $\widehat{MMD}_k$. While we cannot compute $\tilde{k}$ as it depends on the underlying probability measure $P_x$, it is readily checked that due to the empirical centering of processes $\{\tilde{W}_t^{(x)}\}$ and $\{\tilde{W}_t^{(y)}\}$, $\widehat{MMD}_{\tilde{k},b} = \widehat{MMD}_{k,b}$ holds and the claim follows. Note that the result fails to be valid for wild bootstrap processes that are not empirically centred. $\square$

## C.5 Auxiliary results

The following section lists all auxiliary lemmas.

**Proposition 4.** *[22, p.259, Equation 2.1] If process $\{Z_t, \mathcal{F}_t\}_{t\in\mathbb{N}}$ is $\tau$-dependent and $\mathcal{F}$ is rich enough (see [10, Lemma 5.3]), then there exists, for all $t < t_1 < ... < t_l$, $l \in \mathbb{N}$, a random vector $(Z_{t_1}^*, ..., Z_{t_l}^*)$ that is independent of $\mathcal{F}_t$, has the same distribution as $(Z_{t_1}, ..., Z_{t_l})$ and*

$$E\|(Z_{t_1}^*, ..., Z_{t_l}^*) - (Z_{t_1}, ..., Z_{t_l})\|_1 \leq l\tau(t_1 - t).$$

**Lemma 6.** *Let $\{Z_i, \mathcal{F}_i\}$ be a $\tau$-mixing sequence, $\{\delta_i\}$ a sequence of i.i.d random variables independent of filtration $\mathcal{F}$, $(i_1 \leq ... \leq i_m)$ a non-decreasing sequence of positive integers, $k$ a positive integer such that $1 < k < m$ and $(Z_{i_1}, ..., Z_{i_m})$ some random vector. Further let $A = (Z_{i_1}, ..., Z_{i_{k-1}})$, $B = Z_{i_k}$ and $C = (Z_{i_{k+1}}, ..., Z_{i_m})$, $\mathcal{F}_A = \mathcal{F}_{k-1}$. Then, there exist the independent random variables $B^*$ and $C^*$, both independent of $\mathcal{F}_A$, such that*

$$E|B - B^*| = \tau(i_k - i_{k+1}) \text{ and } \frac{1}{m-k} E \parallel C - C^* \parallel_1 \leq \tau(i_{k+1} - i_k) \tag{58}$$

*Proof.* Let $\mathcal{F}_B = \mathcal{F}_k$. We first use [22, Equation 2.1]] (also [10, Lemma 5.3]) to construct $C^*$ such that $\frac{1}{m-k} E \parallel C - C^* \parallel_1 \leq \tau(i_{k+1} - i_k)$. By construction $C^*$ is independent of $\mathcal{F}_B$. Since $\mathcal{F}_A \subset \mathcal{F}_B$ and $\sigma(B) \subset \mathcal{F}_B$ and $C^* \perp\!\!\!\perp \sigma(\delta_k)$, then $C^* \perp\!\!\!\perp (\mathcal{F}_A \vee \sigma(B) \vee \sigma(\delta_k))$. Next by [10, Lemma 5.2] we construct $B^*$ such that $E|B - B^*| = \tau(i_{k+1} - i_k)$, and $B^*$ independent of $\mathcal{F}_A$, but $\mathcal{F}_A \vee \sigma(B) \vee \sigma(\delta_k)$ measurable. Since $\sigma(C^*) \perp\!\!\!\perp (\mathcal{F}_A \vee \sigma(B) \vee \sigma(\delta))$ then $C^*$ and $B^*$ are independent. Finally both $C^*$ and $B^*$ are independent of $\mathcal{F}_A$ $\qquad\square$

**Lemma 7.** *If $h$ is bounded and Lipschitz continuous core then its components are also bounded and Lipschitz continuous.*

*Proof.* Note that

$$g_c(z_1, ... z_c) = Eh(z_1, ..., z_c, Z^*_{c+1}, ..., Z^*_m) \leq E \parallel h \parallel_\infty . \tag{59}$$

To prove boundedness we use induction - we assume that components with low index are bounded and use the fact that sum of bounded functions is bounded to obtain the required results.

We prove Lipschitz continuity similarly, first by showing that $g_c(z_1, ... z_c)$ are Lipschitz continuous with the same coefficient as the core $h$ and then we use the fact that sum of Lipschitz continuous functions is Lipschitz continuous. $\qquad\square$

**Lemma 8.** *Let $f$ be a symmetric function and let $j = \{j_1, \ldots, j_q\}$ be a subset of $\{1, \ldots, m\}$. Then*

$$\sum_{i \in N^m} f(Z_{i_{j_1}}, ..., Z_{i_{j_q}}) = n^{m-q} \sum_{i \in N^q} f(Z_{i_1}, ..., Z_{i_q}) \tag{60}$$

*Proof.* Denote $r = \{r_1, \ldots, r_{m-q}\} = \{1, \ldots, m\} \backslash j$. Then

$$\sum_{i \in N^m} f(Z_{i_{j_1}}, ..., Z_{i_{j_q}}) = \sum_{(i_{j_1}, ..., i_{j_q}) \in N^q} \sum_{(i_{r_1}, ..., i_{r_{m-q}}) \in N^{m-q}} f(Z_{i_{j_1}}, ..., Z_{i_{j_q}})$$

$$= \sum_{(i_{j_1}, ..., i_{j_q}) \in N^q} \left( f(Z_{i_{j_1}}, ..., Z_{i_{j_q}}) \sum_{(i_{r_1}, ..., i_{r_{m-q}}) \in N^{m-q}} 1 \right)$$

$$= n^{m-q} \sum_{(i_{j_1}, ..., i_{j_q}) \in N^q} f(Z_{i_{j_1}}, ..., Z_{i_{j_q}}) = n^{m-q} \sum_{i \in N^q} f(Z_{i_1}, ..., Z_{i_q}).$$

$\qquad\square$

**Lemma 9.** *Let $\{Z_t\}$ be a $\tau$-dependent stationary process with $\tau(r) = O(r^{-6-\epsilon})$. Let $h$ be a bounded Lipschitz continuous function of $m \geq 3$ arguments (not necessarily symmetric) such that $\forall j \in \{1, \ldots, m\}$,*

$$\mathbb{E}h(z_1, \ldots, z_{j-1}, Z_j, z_{j+1}, \ldots, z_m) = 0. \tag{61}$$

*Then,*

$$\sum_{i \in [n]^m} |\mathbb{E}h(Z_{i_1}, \ldots, Z_{i_m})| = O\left( n^{\lfloor \frac{m}{2} \rfloor} + n^{2\lfloor \frac{m}{2} \rfloor - 4 - \epsilon} \right).$$

*Proof.* The proof uses the same technique as [1, Lemma 3]. We will focus on ordered $m$-tuples $1 \leq i_1 \leq \ldots \leq i_m \leq n$, and by considering all possible permutations of their indices, we obtain an upper bound

$$\sum_{i \in [n]^m} |Eh\left(Z_{i_1}, \ldots, Z_{i_m}\right)| \quad \leq \quad \sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} \sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right|, \qquad (62)$$

where (strict) inequality stems from the fact that the $m$-tuples $i$ with some coinciding entries appear multiple times on the right. Now denote $s = \left\lfloor \frac{m}{2} \right\rfloor + 1$ and

$$j_1 = i_2 - i_1; \; j_l = \min\{i_{2l} - i_{2l-1}, i_{2l-1} - i_{2l-2}\}, \; l = 2, \ldots, s-1; \; j_s = i_m - i_{m-1}.$$

Let $w(i) = \max\{j_1, \ldots, j_s\}$, i.e., $w(i)$ corresponds to the largest minimum gap between an individual entry in the ordered $m$-tuple $i$ and its neighbours. For example, $w([1, 2, 5, 9, 9]) = 3$. Note that $w(i) = 0$ means that no entry in $i$ appears exactly once. Let us assume that the maximum $w(i) = w > 0$ is obtained at $j_r$ for some $r \in \{1, \ldots, s\}$. Let us partition the vector $(Z_{i_1}, \ldots, Z_{i_m})$ into three parts:

$$A \quad = \quad \left(Z_{i_1}, \ldots, Z_{i_{2r-2}}\right), \; B = Z_{i_{2r-1}}, \; C = \left(Z_{i_{2r}}, \ldots, Z_{i_m}\right).$$

Note that if $r = 1$, $A$ is empty and if $r = s$ and $m$ is odd, $C$ is empty but this does not change our arguments below. Using Lemma 6, we can construct $B^*$ and $C^*$ that are independent of $A$ and independent of each other and

$$E \|(A, B, C) - (A, B^*, C^*)\|_1 \leq m\tau(w). \qquad (63)$$

Because $B^*$ consists of a singleton and is independent of both $A$ and $C^*$, (61) implies

$$Eh(A, B^*, C^*) = 0. \qquad (64)$$

Thus, for $w(i) = w > 0$, we have that

$$
\begin{aligned}
|Eh\left(Z_{i_1}, \ldots, Z_{i_m}\right)| \quad &\leq \quad E\left|h\left(A, B, C\right) - h\left(A, B^*, C^*\right)\right| + |Eh(A, B^*, C^*)| \\
&\leq \quad \text{Lip}(h)E\|(A, B, C) - (A, B^*, C^*)\|_1 + 0 \\
&\leq \quad m\text{Lip}(h)\tau(w).
\end{aligned}
$$

Finally, if the entries within the ordered $m$-tuple $i$ are permuted, $L_1$-norm in (63) remains the same and (64) still holds, so also $\left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right| \leq m\text{Lip}(h)\tau(w) \; \forall \pi \in S_m$ and

$$\sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right| \leq m!m\text{Lip}(h)\tau(w).$$

Let us upper bound the number of ordered $m$-tuples $i$ with $w(i) = w$. $i_1$ can take $n$ different values, but since $i_2 \leq i_1 + w$, $i_2$ can take at most $w + 1$ different values. For $2 \leq l \leq s - 1$, since $\min\{i_{2l} - i_{2l-1}, i_{2l-1} - i_{2l-2}\} \leq w$, we can either let $i_{2l-1}$ take up to $n$ different values and let $i_{2l}$ take up to $w + 1$ different values (if $i_{2l} - i_{2l-1} \leq i_{2l-1} - i_{2l-2}$) or let $i_{2l-1}$ take up to $w + 1$ different values and let $i_{2l}$ take up to $n$ different values (if $i_{2l} - i_{2l-1} > i_{2l-1} - i_{2l-2}$), upper bounding the total number of choices for $[i_{2l-1}, i_{2l}]$ by $2n(w + 1)$. Finally, the last term $i_m$ can always have at most $w + 1$ different values. This brings the total number of $m$-tuples with $w(i) = w$ to at most $2^{s-2}n^{s-1}(w + 1)^s$. Thus, the number of $m$-tuples with $w(i) = 0$ is $O(n^{s-1})$ and since $h$ is bounded, we have

$$\sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} \sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right|$$

$$\leq O(n^{s-1}) + \sum_{w=1}^{n-1} \sum_{\substack{i \in [n]^m: \\ w(i)=w}} \sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right|$$

$$\leq O(n^{s-1}) + 2^{s-2}m!m\text{Lip}(h)n^{s-1} \sum_{w=1}^{n-1} (w + 1)^s \tau(w)$$

$$\leq O(n^{s-1}) + Cn^{s-1} \sum_{w=1}^{n-1} w^{s-6-\epsilon}$$

$$\leq O(n^{s-1}) + O(n^{2s-6-\epsilon}),$$

which proves the claim. We have used $\tau(w) = O(w^{-6-\epsilon})$ and collated the constants into $C$. $\quad \square$

**Lemma 10.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. If $h$ is a canonical and Lipschitz continuous core of three or more arguments, then*

$$\lim_{n\to\infty} \frac{1}{n^{m-1}} \sum_{i\in N^m} |Eh(Z_{i_1}, ..., Z_{i_m})| = 0. \tag{65}$$

*Proof.* We use Lemma 9 to obtain a bound

$$\sum_{i\in N^m} |Eh(Z_{i_1}, ..., Z_{i_m})| = O\left(n^{\lfloor \frac{m}{2}\rfloor} + n^{2\lfloor \frac{m}{2}\rfloor - 4 - \epsilon}\right). \tag{66}$$

The right hand side of the above equation divided by $n^{m-1}$ converges to zero. $\qquad\square$

**Lemma 11.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. If $h$ is a canonical and Lipschitz continuous core of three or more arguments, then*

$$\lim_{n\to\infty} \frac{1}{n^{2m-2}} \sum_{i\in N^{2m}} E|h(Z_{i_1}, ..., Z_{i_m})h(Z_{i_{m+1}}, ..., Z_{i_{2m}})| = 0. \tag{67}$$

*Proof.* Let $g(z_{i_1}, ..., z_{i_m}, z_{i_{m+1}}, ..., z_{i_{2m}}) = h(z_{i_1}, ..., z_{i_m})h(z_{i_{m+1}}, ..., z_{i_{2m}})$. Since $g$ meets assumptions of the Lemma 9 and for $m \geq 3$

$$\lim_{n\to\infty} \frac{1}{n^{2m-2}} \left(n^m + n^{2m-4-\epsilon}\right) = 0, \tag{68}$$

the Lemma follows from the lemma 9. $\qquad\square$

**Lemma 12.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. Let $h$ be a canonical and Lipschitz continuous core of $c$ arguments, $3 \leq c \leq m$, and $1 \leq j_1 < ... < j_c \leq m$ be a sequence of $c$ integers. If $Q_{i_1,...,i_m}$ is a random variable independent of $(Z_{i_1}, ..., Z_{i_m})$ such that $\sup_{i\in N^m} E|Q_i| \leq \infty$ and $\sup_{i\in N^m} \sup_{o\in N^m} E|Q_iQ_o| \leq \infty$ then*

$$\lim_{n\to\infty} \frac{1}{n^{m-1}} E \sum_{i\in N^m} Q_i h(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \overset{P}{=} 0. \tag{69}$$

$$\lim_{n\to\infty} \frac{1}{n^{2m-2}} E \sum_{i\in N^m} \sum_{o\in N^m} Q_iQ_o h(Z_{i_{j_1}}, ..., Z_{i_{j_c}})h(Z_{i_{o_1}}, ..., Z_{i_{o_c}}) \overset{P}{=} 0. \tag{70}$$

$$\tag{71}$$

*For the first limit notice that*

$$\frac{1}{n^{m-1}} E \sum_{i\in N^m} Q_i h(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \leq \frac{1}{n^{m-1}} \sum_{i\in N^m} |EQ_i||Eh(Z_{i_{j_1}}, ..., Z_{i_{j_c}})| \overset{8}{\leq} \tag{72}$$

$$\sup_{i\in N^m} |EQ_i| \frac{1}{n^{c-1}} \sum_{i\in N^c} |Eh(Z_{i_1}, ..., Z_{i_c})| \overset{10}{\longrightarrow} 0 \text{ in probability}. \tag{73}$$

*Similar reasoning, which uses Lemma 11 instead of 10, shows convergence of the second limit.*

**Lemma 13.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. Let $h$ be a canonical and Lipschitz continuous core of $c$ arguments, $3 \leq c \leq m$. If $Q_{i_1,...,i_m}$ is a random variable independent of $(Z_{i_1}, ..., Z_{i_m})$ such that $\sup_{i\in N^m} E|Q_i| \leq \infty$ and $\sup_{i\in N^m} \sup_{o\in N^m} E|Q_iQ_o| \leq \infty$ then*

$$\lim_{n\to\infty} \frac{1}{n^{m-1}} \sum_{i\in N^m} \sum_{1\leq j_1<...<j_c<m} EQ_i h_c(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \overset{P}{=} 0. \tag{74}$$

*Proof.* For each sequence such that $1 \leq j_1 < ... < j_c < m$ we apply Lemma 12 and conclude that the random sum

$$\frac{1}{n^{m-1}} \sum_{i\in N^m} EQ_{i_{j_1}, ..., i_{j_c}} h_c(Z_{i_{j_1}}, ..., Z_{i_{j_c}}) \tag{75}$$

converges to zero in a probability - from this the proposition follows. $\qquad\square$

**Lemma 14.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$. If $h$ if a one-degenerate and Lipschitz continuous core of three or more arguments*

$$\lim_{n \to \infty} nV(h_c) = 0, \tag{76}$$

*for $2 < c \leq m$.*

*Proof.* For each $c$ satisfying $2 < c \leq m$, $h_c$ is canonical and Lipschitz continuous. It suffices to put $Q = 1$ and use Lemma 13. $\qquad\square$

**Lemma 15.** *If $W_i$ is a bootstrap process then*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} W_i \overset{P}{=} 0. \tag{77}$$

*Proof.* By the definition of $W_i$, $E(\sum_{i=1}^{n} W_i)^2 = O(nl_n)$, $\lim_{n\to\infty} \frac{l_n}{n} = 0$ and $E \sum_{i=1}^{n} W_i = 0$. Therefore $\frac{1}{n} \sum_{i=1}^{n} W_i$ converges to zero in probability. $\qquad\square$

**Lemma 16.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$ and $W_i$ is a bootstrap process. Let $f$ be a one-degenerate, Lipschitz continuous, bounded core of at least $m$ arguments, $m \geq 2$. Further assume that $f_2$ is a kernel. Then for a positive integer $p$*

$$\lim_{n \to \infty} nV(f) \left( \frac{1}{n} \sum_{i=1}^{n} W_i \right)^p \overset{P}{=} 0 \tag{78}$$

*Proof.* By the Lemma 15 $\frac{1}{n} \sum_{i=1}^{n} W_i$ converges to zero in probability. By Theorem 1 $\frac{1}{n^{m-1}} \sum_{i \in N^m} f(Z_{i_m}, ..., Z_{i_m})$ converges to some random variable. $\qquad\square$

**Lemma 17.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$ and $W_i$ is a bootstrap process. Let $x = (w, z)$ and suppose $f(x_1, x_2) = g(w_1)g(w_2)h(z_1, z_2)$ where $g$ is Lipschitz continuous, $E|g(W_1)|^k \leq \infty$ for any finite $k$ and $h$ is symmetric, Lipschitz continuous, degenerate and bounded. Then*

$$nV(f) = \frac{1}{n} \sum_{i,j} f(X_i, X_j) = \frac{1}{n} \sum_{i,j} g(W_i)g(W_j)h(Z_i, Z_j) \tag{79}$$

*converges in law to some random variable.*

*Proof.* We use [21, Theorem 2.1] to show that $nV(f)$ converges. We check the assumptions *A1 - A3*

*Assumption A1.* Point (i) requires that the process $(W_n, Z_n)$ is a strictly stationary sequence of $\mathbb{R}^d$-values integrable random variables - this follows from the assumptions of this Lemma. For the point (ii) we put $\delta = \frac{1}{3}$ and check condition

$$\sum_{r=1}^{\infty} r\tau(r)^\delta \leq \sum_{r=1}^{\infty} rr^{-6\frac{1}{3}} = \sum_{r=1}^{\infty} r^{-2} < \infty. \tag{80}$$

*Assumption A2.* Point (i) requires that the function $f$ is symmetric, measurable and degenerate. Symmetry and measurability are obvious and so we check degeneracy condition

$$Eg(W_1)g(w)h(Z_1, z) = Eg(W_1)g(w)Eh(Z_1, z) = 0. \tag{81}$$

Point (ii) requires that for $\nu > (2 - \delta)/(1 - \delta) = 2.5$ (since we have chosen $\delta = \frac{1}{3}$).

$$\sup_{k \in \mathbf{N}} E|f(X_1, X_k)|^\nu < \infty \text{ and } \sup_{k \in \mathbf{N}} E|f(X_1, X_k^*)|^\nu < \infty \tag{82}$$

Both requirements are met since $h$ is bounded and the process $E|g(W_i)|^k \leq \infty$ for any finite $k$.

*Assumption A3.* Function $f$ is Lipschitz continuous - this is met since both $g$ and $h$ are Lipschitz continuous. $\qquad\square$

**Lemma 18.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$ and $W_i$ is a bootstrap process. If $h$ is a Lipschitz continuous, degenerate and bounded core of two arguments then*

$$\frac{1}{n} \sum_{i \in N^2} W_{i_1} h(Z_{i_1}, Z_{i_2}) \tag{83}$$

*converges in distribution to some random variable.*

*Proof.*

$$\frac{1}{n} \sum_{i \in N^2} W_{i_1} h(Z_{i_1}, Z_{i_2}) = \frac{1}{4}(V_+ - V_-) \text{ where,} \tag{84}$$

$$V_- = n^{-1} \sum_{i \in N^2} (W_{i_1} - 1) h(Z_{i_1}, Z_{i_2})(W_{i_2} - 1), \tag{85}$$

$$V_+ = n^{-1} \sum_{i \in N^2} (W_{i_1} + 1) h(Z_{i_1}, Z_{i_2})(W_{i_2} + 1), \tag{86}$$

are normalized V statistics that converge. To see that we use Lemma 17 with $g_+(x) = x + 1$ and $g_-(x) = x - 1$ respectively. The only non-trivial assumption is that $E|g_+(W_i)|^k < \infty$ and $E|g_-(W_i)|^k < \infty$ - this follows from $E|W_i|^k$. $\qquad\square$

**Lemma 19.** *If $\{W_i\}$ is a bootstrap process then*

$$\sum_{i=1}^{n} \tilde{W}_i = \sum_{i=1}^{n} \left( W_i - \frac{1}{n} \sum_{j=1}^{n} W_j \right) = 0. \tag{87}$$

**Lemma 20.** *Assume that the stationary process $Z_t$ is $\tau$-dependent with a coefficient $\tau(i) = i^{-6-\epsilon}$ for some $\epsilon > 0$ and $W_i$ is a bootstrap process. If $f$ is canonical, Lipschitz continuous, bounded core then a random variable*

$$\frac{1}{n} \sum_{1 \le i,j \le n} \tilde{W}_i \tilde{W}_j f(Z_i, Z_j) \tag{88}$$

*converges in law.*

*Proof.*

$$\frac{1}{n} \sum_{1 \le i,j \le n} \tilde{W}_i \tilde{W}_j f(Z_i, Z_j) = \frac{1}{n} \sum_{1 \le i,j \le n} \left( W_i - \sum_{a=1}^{n} W_a \right) \left( W_i - \sum_{b=1}^{n} W_b \right) f(Z_i, Z_j) = \tag{89}$$

$$\frac{1}{n} \sum_{1 \le i,j \le n} W_i W_j f(Z_i, Z_j) - \left( \frac{2}{n} \sum_{1 \le i,j \le n} f(Z_i, Z_j) \right) \left( \frac{1}{n} \sum_{b=1}^{n} W_b \right) + \tag{90}$$

$$\left( \frac{1}{n} \sum_{1 \le i,j \le n} f(Z_i, Z_j) \right) \left( \frac{1}{n} \sum_{b=1}^{n} W_b \right)^2. \tag{91}$$

Last two terms converge to zero since by Lemma 19 $\left( \frac{1}{n} \sum_{b=1}^{n} W_b \right)$ converges to zero and by the Lemma 17 (with $g = 1$ ) $\frac{1}{n} \sum_{1 \le i,j \le n} f(Z_i, Z_j)$ converges in law. The first term converges by the Lemma 17. $\qquad\square$

# D  Lag-HSIC with $M \to \infty$

We here consider a multiple lags test described in Section 4.2 where the number of lags $M = M_n$ being considered goes to infinity with the sample size $n$. Thus, we will be testing if there exists a dependency between $X_t$ and $Y_{t+m}$ for $-M_n \le m \le M_n$ where $\{M_n\}$ is an increasing sequence of

positive numbers such that $M_n = o(n^r)$ for some $0 < r \leq 1$, but $\lim_{n \to \infty} M_n = \infty$. We denote $q_n = 1 - \frac{\alpha}{2M_n+1}$. As before, the shifted time series will be denoted $Z_t^m = (X_t, Y_{t+m})$ and $S_{m,n} = nV(h, Z^m)$ and $F_{b,n}$ is the empirical cumulative distribution function obtained from $nV_b(h, Z)$. We also let $F_n$ and $F$ denote respectively the finite-sample and the limiting distribution under the null hypothesis of $S_{0,n} = nV(h, Z)$ (or, equivalently, of any $S_{m,n}$ since the null hypothesis holds).

Let us assume that we have computed the empirical $q_n$-quantile based on the bootstrapped samples, denoted by $t_{b,q_n} = F_{b,n}^{-1}(q_n)$. The null hypothesis is then be rejected if the event $\mathcal{A}_n = \{\max_{-M_n \leq k \leq M_n} S_{m,n} > t_{b,q_n}\}$ occurs. By definition, since $F$ is continuous, $F_n(x) \to F(x)$, $\forall x$. In addition, our Theorem 1 implies that $F_{b,n}(x) \to F(x)$ in probability. Thus, $|F_{b,n}(x) - F_n(x)| \to 0$ in probability as well. However, in order to guarantee that $|q_n - F_n(t_{b,q_n})| \to 0$, which we require for the Type I error control, we require a stronger assumption of uniform convergence, that $\|F_{b,n} - F_n\|_\infty \leq \frac{C}{n^r}$, for some $C < \infty$. Then, by continuity and sub-additivity of probability, the asymptotic Type I error is given by

$$\lim_{n \to \infty} P_{\mathbf{H_0}}(\mathcal{A}_n) \leq \lim_{n \to \infty} \sum_{-M_n \leq m \leq M_n} P_{\mathbf{H_0}}(S_{m,n} > t_{b,q_n}) =$$

$$\lim_{n \to \infty} (2M_n + 1)(1 - F_n(t_{b,q_n})) \leq \lim_{n \to \infty} (2M_n + 1)\left(1 - (1 - \frac{\alpha}{2M_n+1}) + \frac{C}{n^r}\right) = \alpha, \quad (92)$$

as long as $M_n = o(n^r)$. Intuitively, we require that the number of tests being performed increases at a slower rate than the rate of distributional convergence of the bootstrapped statistics.

On the other hand, under the alternative, there exists some $m$ for which $n^{-1}S_{m,n}$ converges to some positive constant. In this case however, we do not have a handle on the asymptotic distribution $F$ of $S_{m,n} = nV(h, Z^m)$: cumulative distribution function obtained from sampling $nV_{b2}(h)$ converges to $G$ (possibly different from $F$) with a finite variance, while the behaviour of $nV_{b1}(h)$ is unspecified. We can however show that for any such cumulative distribution function $G$, the Type II error still converges to zero since

$$P_{\mathbf{H_1}}(\mathcal{A}_n) \geq P_{\mathbf{H_1}}(S_{m,n} > G^{-1}(q_n)) = P_{\mathbf{H_1}}(n^{-1}S_{m,n} > n^{-1}G^{-1}(q_n)) \to 1,$$

which follows from Lemma 21 below that shows that $n^{-1}G^{-1}(q_n)$ converges to zero.

**Lemma 21.** *If $X \sim G$ is a random variable such that $EX^2 < \infty$, $q_n = 1 - \frac{\alpha}{2M_n+1}$ and $M_n = o(n)$ then $n^{-1}G^{-1}(q_n) \to 0$.*

*Proof.* First observe that by Markov inequality $P(X \geq t) \leq \frac{EX^2}{t}$ and therefore $G(t) > g(t) = 1 - \frac{EX^2}{t}$. Therefore, on the interval $(EX, 1)$, $G^{-1}(x) < g^{-1}(x) = \frac{EX^2}{1-x}$. As a result

$$n^{-1}G^{-1}(q_n) \leq n^{-1}g^{-1}(q_n) = n^{-1}\frac{EX^2}{1 - (1 - \frac{\alpha}{2M_n+1})} = \frac{(2M_n + 1)EX^2}{\alpha n} \xrightarrow{n \to \infty} 0. \quad (93)$$

$\square$

# E  Various comments

## E.1  Time complexity

The original HSIC and MMD tests for i.i.d. data, the computational cost of the wild bootstrap approach scales quadratically in the number of samples, and linearly in the number of bootstrap iterations (in the i.i.d. case, these were permutations of the data). The main alternative approaches are the lagged bootstrap of [8], which has the same scaling with data and number of bootstraps, and the spectrogram approach of [4] (note, however, that both these alternative approaches apply only to the independence testing case). The cost of [4] is comparable to our approach, however the statistical power of [4] was much weaker on the data we examined.