Sequential Kernelized Independence Testing

Aleksandr Podkopaev [1] **Patrick Blöbaum** [2] **Shiva Prasad Kasiviswanathan** [2] **Aaditya Ramdas** [1 2]

Abstract

Independence testing is a classical statistical problem that has been extensively studied in the batch setting when one fixes the sample size before collecting data. However, practitioners often prefer procedures that adapt to the complexity of a problem at hand instead of setting sample size in advance. Ideally, such procedures should (a) stop earlier on easy tasks (and later on harder tasks), hence making better use of available resources, and (b) continuously monitor the data and efficiently incorporate statistical evidence after collecting new data, while controlling the false alarm rate. Classical batch tests are not tailored for streaming data: valid inference after data peeking requires correcting for multiple testing which results in low power. Following the principle of testing by betting, we design sequential kernelized independence tests that overcome such shortcomings. We exemplify our broad framework using bets inspired by kernelized dependence measures, e.g., the Hilbert-Schmidt independence criterion. Our test is also valid under non-i.i.d. time-varying settings. We demonstrate the power of our approaches on both simulated and real data.

1. Introduction

Independence testing is a fundamental statistical problem that has also been studied within information theory and machine learning. Given paired observations (X, Y) sampled from some (unknown) joint distribution P_{XY}, the goal is to test the null hypothesis that X and Y are independent. The literature on independence testing is vast as there is no unique way to measure dependence, and different measures give rise to different tests. Traditional measures of dependence, such as Pearson's r, Spearman's ρ, and Kendall's τ, are limited to the case of univariate random variables. Kernel tests (Jordan & Bach, 2001; Gretton et al., 2005c;a) are amongst the most prominent modern tools for nonparametric independence testing that work for general $\mathcal{X}, \mathcal{Y}$ spaces.

In the literature, heavy emphasis has been placed on *batch* testing when one has access to a sample whose size is specified in advance. However, even if random variables are dependent, the sample size that suffices to detect dependence is never known a priori. If the results of a test are promising yet non-conclusive (e.g., a p-value is slightly larger than a chosen significance level), one may want to collect more data and re-conduct the study. This is not allowed by traditional batch tests. We focus on sequential tests that allow peeking at observed data to decide whether to stop and reject the null or to continue collecting data.

Problem Setup. Suppose that one observes a stream of data: $(Z_t)_{t\geq 1}$, where $Z_t = (X_t, Y_t) \overset{\text{iid}}{\sim} P_{XY}$. We design sequential tests for the following pair of hypotheses:

$$H_0 : Z_t \overset{\text{iid}}{\sim} P_{XY},\ t \geq 1 \text{ and } P_{XY} = P_X \times P_Y, \quad \text{(1a)}$$

$$H_1 : Z_t \overset{\text{iid}}{\sim} P_{XY},\ t \geq 1 \text{ and } P_{XY} \neq P_X \times P_Y. \quad \text{(1b)}$$

Following the framework of "tests of power one" (Darling & Robbins, 1968), we define a level-α sequential test as a mapping $\Phi : \cup_{t=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^t \to \{0, 1\}$ that satisfies

$$\mathbb{P}_{H_0}\left(\exists t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1\right) \leq \alpha.$$

As is standard, 0 stands for "do not reject the null yet" and 1 stands for "reject the null and stop". Defining the stopping time $\tau := \inf\{t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1\}$ as the first time that the test outputs 1, a sequential test must satisfy

$$\mathbb{P}_{H_0}\left(\tau < \infty\right) \leq \alpha.$$

We work in a very general composite nonparametric setting: H_0 and H_1 consist of huge classes of distributions (discrete/continuous) for which there may not be a common reference measure, making it impossible to define densities and thus ruling out likelihood-ratio based methods.

Our Contributions. Following the principle of testing by betting, we design consistent sequential nonparametric independence tests. Our bets are inspired by popular kernelized dependence measures: Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005a), constrained covariance

[1]Statistics & Data Science and Machine Learning Departments, Carnegie Mellon University [2]Amazon Research. Correspondence to: Aleksandr Podkopaev <podkopaev@cmu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

criterion (COCO) (Gretton et al., 2005c), and kernelized canonical correlation (KCC) (Jordan & Bach, 2001). We provide theoretical guarantees on *time-uniform* type I error control for these tests — the type I error is controlled even if the test is continuously monitored and adaptively stopped — and further establish consistency and asymptotic rates for our sequential HSIC under the i.i.d. setting. Our tests also remain valid even if the underlying distribution changes over time. Additionally, while the initial construction of our tests requires bounded kernels, we also develop variants based on symmetry-based betting that overcome this requirement. This strategy can be readily used with a linear kernel to construct a sequential linear correlation test. We justify the practicality of our tests through a detailed empirical study.

We start by highlighting two major shortcomings of existing tests that our new tests overcome.

(i) Limitations of Corrected Batch tests and Reduction to Two-sample Testing. Batch tests (without corrections for multiple testing) have an inflated false alarm rate under continuous monitoring (see Appendix A.1). Naïve Bonferroni corrections restore type I error control, but generally result in tests with low power. This motivates a direct design of sequential tests (not by correcting batch tests). It is tempting to reduce sequential independence testing to sequential two-sample testing, for which a powerful solution has been recently designed (Shekhar & Ramdas, 2021). This can be achieved by splitting a single data stream into two and permuting the X data in one of the streams (see Appendix A.2). Still, the splitting results in inefficient use of data and thus low power, compared to our new direct approach (Figure 1).

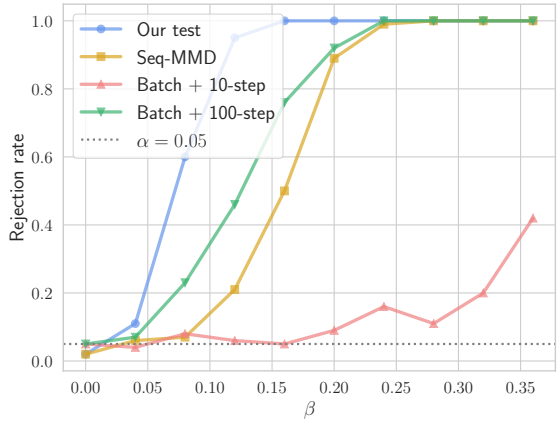


Figure 1. Valid sequential independence tests for: $Y_t = X_t\beta + \varepsilon_t$, $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$. Batch + n-step is batch HSIC with Bonferroni correction applied every n steps (allowing monitoring only at those steps). Seq-MMD refers to the reduction to two-sample testing (Appendix A.2). Our test outperforms other tests.

(ii) Time-varying Independence Testing: Beyond the i.i.d. Setting. A common practice of using a permutation p-value for batch independence testing requires (X, Y)-pairs to be i.i.d. (more generally, exchangeable). If data distribution drifts, the resulting test is no longer valid, and even under mild changes, an inflated false alarm rate is observed empirically. Our tests handle more general non-stationary settings. For a stream of independent data: $(Z_t)_{t\geq 1}$, where $Z_t \sim P_{XY}^{(t)}$, consider the following pair of hypotheses:

$$H_0 : P_{XY}^{(t)} = P_X^{(t)} \times P_Y^{(t)}, \ \forall t, \tag{2a}$$

$$H_1 : \exists t' : P_{XY}^{(t')} \neq P_X^{(t')} \times P_Y^{(t')}. \tag{2b}$$

Suppose that under H_0 in (2a), it holds that either $P_X^{(t-1)} = P_X^{(t)}$ or $P_Y^{(t-1)} = P_Y^{(t)}$ for each $t \geq 1$, meaning that either the distribution of X may change or that of Y may change, but not both simultaneously. In this case, our tests control the type I error, whereas batch independence tests fail to.

Example 1. *Let $((W_t, V_t))_{t\geq 1}$ be a sequence of i.i.d. jointly Gaussian random variables with zero mean and covariance matrix with ones on the diagonal and ρ off the diagonal. For $t = 1, 2, \ldots$ and $i \in \{0, 1\}$, consider the following stream:*

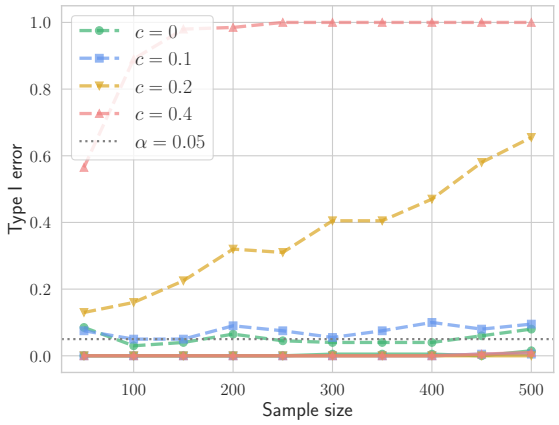
$$\begin{cases} X_{2t-i} = 2c\sin(t) + W_{2t-1}, \\ Y_{2t-i} = 3c\sin(t) + V_{2t-1}, \end{cases} \tag{3}$$

Setting $\rho = 0$ falls into the null case (2a), whereas any $\rho \neq 0$ implies dependence as per (2b). Visually, it is hard to distinguish between H_0 and H_1: the drift makes data seem dependent (see Appendix E.1). In Figure 2(a), we show that our test controls type I error, whereas batch test fails[1].

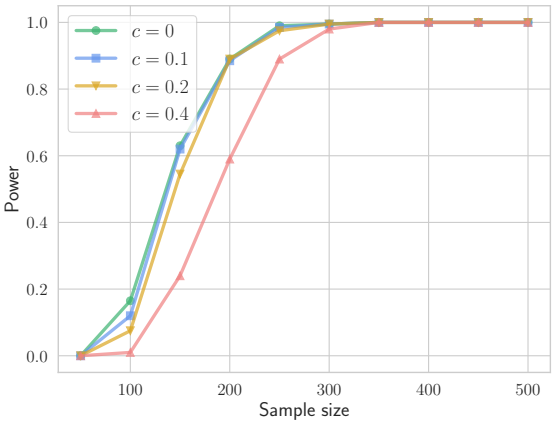
Related Work. In addition to the aforementioned papers on batch independence testing, our work is also related to methods for "safe, anytime-valid inference", e.g., confidence sequences (Waudby-Smith & Ramdas, 2023, and references therein) and e-processes (Grünwald et al., 2020; Ramdas et al., 2022). Sequential nonparametric two-sample tests of Balsubramani & Ramdas (2016), based on linear-time test statistics and empirical Bernstein inequality for random walks, are amongst the first results in this area. While such tests are valid in the same sense as ours, betting-based tests are much better empirically (Shekhar & Ramdas, 2021).

The roots of the principle of testing by betting can be traced back to Ville's 1939 doctoral thesis (Ville, 1939) and was recently popularized by Shafer (2021). The latter work considered it mainly in the context of parametric and simple hypotheses, far from our setting. The most closely related works to the current paper are (Shekhar & Ramdas, 2021; Shaer et al., 2023; Grünwald et al., 2023) which also handle composite and nonparametric hypotheses. Shekhar & Ramdas (2021) use testing by betting to design sequential nonparametric two-sample tests, including a state-of-the-art

[1]This is also related to Yule's nonsense correlation (Yule, 1926; Ernst et al., 2017), which would not pose a problem for our method.



(a) The null is true, meaning that $W \perp\!\!\!\perp V$ in Example 1. (Batch) HSIC: dashed lines, SKIT: solid lines.



(b) The alternative is true with $\rho = 1/2$.

Figure 2. Under distribution drift (3), SKIT controls type I error under H_0 and has high power under H_1. Batch HSIC fails to control type I error under H_0 (hence we do not plot its power).

sequential kernel maximum mean discrepancy test. Two recent works by Shaer et al. (2023); Grünwald et al. (2023), developed in parallel to the current paper, extend these ideas to the setting of sequential conditional independence tests ($H_0 : X \perp\!\!\!\perp Y \mid Z$) under the model-X assumption, i.e., when the distribution $X \mid Z$ is assumed to be known. Our methods are very different from the aforementioned papers because when $Z = \emptyset$, the model-X assumption reduces to assuming P_X is known, which we of course avoid. The current paper can be seen as extending the ideas from (Shekhar & Ramdas, 2021) to nonparametric independence testing.

2. Sequential Kernel Independence Test

We begin with a brief summary of the principle of testing by betting (Shafer, 2021; Shafer & Vovk, 2019). Suppose that one observes a sequence of random variables $(Z_t)_{t\geq 1}$, where $Z_t \in \mathcal{Z}$. A player begins with initial capital $\mathcal{K}_0 = 1$. At round t of the game, she selects a payoff function $f_t : \mathcal{Z} \to [-1, \infty)$ that satisfies $\mathbb{E}_{Z\sim P_Z}[f_t(Z) \mid \mathcal{F}_{t-1}] = 0$ for all $P_Z \in H_0$, where $\mathcal{F}_{t-1} = \sigma(Z_1, \ldots, Z_{t-1})$, and bets a fraction of her wealth $\lambda_t \mathcal{K}_{t-1}$ for an $\mathcal{F}_{t-1}$-measurable $\lambda_t \in [0, 1]$. Once Z_t is revealed, her wealth is updated as

$$\mathcal{K}_t = \mathcal{K}_{t-1}(1 + \lambda_t f_t(Z_t)). \tag{4}$$

A level-α sequential test is obtained using the following stopping rule: $\Phi(Z_1, \ldots, Z_t) = \mathbb{1}\{\mathcal{K}_t \geq 1/\alpha\}$, i.e., the null is rejected once the player's capital exceeds $1/\alpha$. If the null is true, imposed constraints on sequences of payoffs $(f_t)_{t\geq 1}$ and betting fractions $(\lambda_t)_{t\geq 1}$ prevent the player from making money. Formally, the wealth process $(\mathcal{K}_t)_{t\geq 0}$ is a nonnegative martingale. The validity of the resulting test then follows from Ville's inequality (Ville, 1939).

To ensure that the resulting test has power under the alternative, payoffs and betting fractions have to be chosen carefully. Inspired by sequential two-sample tests of Shekhar & Ramdas (2021), our construction relies on dependence measures: $m(P_{XY}; \mathcal{C})$, which admit a variational representation:

$$\sup_{c\in\mathcal{C}} \left[\mathbb{E}_{P_{XY}} c(X, Y) - \mathbb{E}_{P_X \times P_Y} c(X, Y)\right], \tag{5}$$

for some class $\mathcal{C}$ of bounded functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The supremum above is often achieved at some $c^* \in \mathcal{C}$, and in this case, c^* is called the "witness function". In what follows, we use sufficiently rich functional classes $\mathcal{C}$ for which the following characteristic condition holds:

$$\begin{cases} m(P_{XY}; \mathcal{C}) = 0, & \text{under } H_0, \\ m(P_{XY}; \mathcal{C}) > 0, & \text{under } H_1, \end{cases} \tag{6}$$

for H_0 and H_1 defined in (1). To proceed, we bet on *pairs* of points from P_{XY}. Swapping Y-components in a pair of points from P_{XY}: Z_{2t-1} and Z_{2t}, gives two points from $P_X \times P_Y$: $\tilde{Z}_{2t-1} = (X_{2t-1}, Y_{2t})$ and $\tilde{Z}_{2t} = (X_{2t}, Y_{2t-1})$. We consider payoffs $f(Z_{2t-1}, Z_{2t})$ of the form:

$$s \cdot \Big((c(Z_{2t-1}) + c(Z_{2t})) - (c(\tilde{Z}_{2t-1}) - c(\tilde{Z}_{2t})) \Big), \tag{7}$$

where the scaling factor $s > 0$ ensures that $f(z, z') \in [-1, 1]$ for any $z, z' \in \mathcal{X} \times \mathcal{Y}$. When the witness function c^* is used in the above, we denote the resulting function as the "oracle payoff" f^*. Let the oracle wealth process $(\mathcal{K}^*_t)_{t\geq 0}$ be defined by using f^* along with the betting fraction

$$\lambda^\star = \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[f^\star(Z_1, Z_2)\right] + \mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]}. \tag{8}$$

We have the following result regarding the above quantities, whose proof is presented in Appendix B.2.2.

Theorem 2.1. *Let $\mathcal{C}$ denote a class of functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for measuring dependence as per (5).*

1. *Under H_0 in (1a) and (2a), any payoff f of the form (7) satisfies $\mathbb{E}_{H_0}[f(Z_1, Z_2)] = 0$.*

2. *Suppose that $\mathcal{C}$ satisfies (6). Under H_1 in (1b), the oracle payoff f^* based on the witness function c^* satisfies $\mathbb{E}_{H_1}[f^\star(Z_1, Z_2)] > 0$. Further, for $\lambda^\star$ defined in (8), it holds that $\mathbb{E}_{H_1}[\log(1 + \lambda^\star f^\star(Z_1, Z_2)] > 0$. Hence, $\mathcal{K}_t^\star \xrightarrow{\text{a.s.}} +\infty$, which implies that the oracle test is consistent: $\mathbb{P}_{H_1}(\tau^\star < \infty) = 1$, where $\tau^\star = \inf\{t \geq 1 : \mathcal{K}_t^\star \geq 1/\alpha\}$.*

Remark 2.2. While the betting fraction (8) suffices to guarantee the consistency of the corresponding test, the fastest growth rate of the wealth process is ensured by considering

$$\lambda_{\mathrm{K}}^\star \in \operatorname*{arg\,max}_{\lambda \in (0,1)} \mathbb{E}\left[\log(1 + \lambda f^\star(Z_1, Z_2)\right].$$

Overshooting with the betting fraction may, however, result in the wealth tending to zero almost surely.

Example 2. *Consider a sequence $(W_t)_{t\geq 1}$, where*

$$W_t = \begin{cases} 1, & \text{with probability } 3/5, \\ -1, & \text{with probability } 2/5. \end{cases}$$

In this case, we have $\lambda_{\mathrm{K}}^\star = 1/5$ and $\mathbb{E}[\log(1 + \lambda^\star W_t)] > 0$, implying that $\mathcal{K}_t \xrightarrow{\text{a.s.}} +\infty$. On the other hand, it is easy to check that for $\tilde{\lambda} = 2\lambda_{\mathrm{K}}^\star$ we have: $\mathbb{E}[\log(1 + \tilde{\lambda} W_t)] < 0$. As a consequence, for the wealth process $\mathcal{K}_t$ corresponding to the pair $(f^*, \tilde{\lambda})$ it holds that $\mathcal{K}_t \xrightarrow{\text{a.s.}} 0$.

To construct a practical test, we select an appropriate class $\mathcal{C}$ for which the condition (6) holds and replace the oracle $f^\star$ and $\lambda^\star$ with predictable estimates $(f_t)_{t\geq 1}$ and $(\lambda_t)_{t\geq 1}$, meaning that those are computed using data observed prior to a given round of the game. We begin with a particular dependence measure, namely HSIC (Gretton et al., 2005a), and defer extensions to other measures to Section 3.

HSIC-based Sequential Kernel Independence Test (SKIT). Let $\mathcal{G}$ be a separable RKHS[2] with positive-definite kernel $k(\cdot, \cdot)$ and feature map $\varphi(\cdot)$ on $\mathcal{X}$. Let $\mathcal{H}$ be a separable RKHS with positive-definite kernel $l(\cdot, \cdot)$ and feature map $\psi(\cdot)$ on $\mathcal{Y}$.

Assumption 2.3. Suppose that:

(A1) Kernels k and l are nonnegative and bounded by one: $\sup_{x\in\mathcal{X}} k(x, x) \leq 1$ and $\sup_{y\in\mathcal{Y}} l(y, y) \leq 1$.

(A2) The product kernel $k \otimes l : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$, defined as $(k \otimes l)((x, y), (x', y')) := k(x, x')l(y, y')$, is a characteristic kernel on the joint domain.

Assumption (A1) is used to justify that the mean embeddings introduced later are well-defined elements of RKHSs, and the particular bounds are used to simplify the constants. Assumption (A2) is a sufficient condition for the characteristic condition (6) to hold (Fukumizu et al., 2007b), and we use it to argue about the consistency of our test. Under mild assumptions, it can be further relaxed to characteristic property of the kernels on the respective domains (Gretton, 2015). We note that the most common kernels on $\mathbb{R}^d$: Gaussian (RBF) and Laplace, satisfy both (A1) and (A2). Define mean embeddings of the joint and marginal distributions:

$$\begin{aligned} \mu_{XY} &:= \mathbb{E}_{P_{XY}}[\varphi(X) \otimes \psi(Y)], \\ \mu_X := \mathbb{E}_{P_X}[\varphi(X)], \quad \mu_Y &:= \mathbb{E}_{P_Y}[\psi(Y)]. \end{aligned} \tag{9}$$

The cross-covariance operator $C_{XY} : \mathcal{H} \to \mathcal{G}$ associated with the joint measure P_{XY} is defined as

$$C_{XY} := \mu_{XY} - \mu_X \otimes \mu_Y,$$

where $\otimes$ is the outer product operation. This operator generalizes the covariance matrix. *Hilbert-Schmidt independence criterion* (HSIC) is a criterion defined as Hilbert-Schmidt norm, a generalization of Frobenius norm for matrices, of the cross-covariance operator (Gretton et al., 2005a):

$$\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) := \|C_{XY}\|_{\mathrm{HS}}^2. \tag{10}$$

HSIC is simply the squared kernel maximum mean discrepancy (MMD) between mean embeddings of P_{XY} and $P_X \times P_Y$ in the *product RKHS* $\mathcal{G} \otimes \mathcal{H}$ on $\mathcal{X} \times \mathcal{Y}$, defined by a product kernel $k \otimes l$. We can rewrite (10) as

$$\left(\sup_{\substack{g\in\mathcal{G}\otimes\mathcal{H} \\ \|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1}} \mathbb{E}_{P_{XY}}[g(X, Y)] - \mathbb{E}_{P_X\times P_Y}[g(X', Y')] \right)^2,$$

which matches the form (5). The witness function for HSIC admits a closed form (see Appendix D):

$$g^\star = \frac{\mu_{XY} - \mu_X \otimes \mu_Y}{\|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}}}, \tag{11}$$

where μ_{XY}, μ_X and μ_Y are defined in (9). The oracle payoff based on HSIC: $f^\star(Z_{2t-1}, Z_{2t})$, is given by

$$\frac{1}{2}\left(g^\star(Z_{2t-1}) + g^\star(Z_{2t}) - g^\star(\tilde{Z}_{2t-1}) - g^\star(\tilde{Z}_{2t})\right), \tag{12}$$

which has the form (7) with $s = 1/2$. To construct the test, we use estimators $(f_t)_{t\geq 1}$ of the oracle payoff $f^\star$ obtained by replacing $g^\star$ in (12) with the plug-in estimator:

$$\hat{g}_t = \frac{\hat{\mu}_{XY} - \hat{\mu}_X \otimes \hat{\mu}_Y}{\|\hat{\mu}_{XY} - \hat{\mu}_X \otimes \hat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}}}, \tag{13}$$

[2]Recall that an RKHS is a Hilbert space $\mathcal{G}$ of real-valued functions over $\mathcal{X}$, for which the evaluation functional $\delta_x : \mathcal{G} \to \mathbb{R}$, which maps $g \in \mathcal{G}$ to $g(x)$, is a continuous map, and this fact must hold for every $x \in \mathcal{X}$. Each RKHS is associated with a unique positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which can be viewed as a generalized inner product on $\mathcal{X}$. We refer the reader to (Muandet et al., 2017) for an extensive recent survey of kernel methods.

where $\hat{\mu}_{XY}, \hat{\mu}_X, \hat{\mu}_Y$ denote the empirical mean embeddings (plug-in estimators of (9)) computed at round t as[3]

$$\hat{\mu}_{XY} = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \otimes \psi(Y_i),$$
$$\hat{\mu}_X = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i), \quad \hat{\mu}_Y = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i). \tag{14}$$

Note that in (13) the witness function is defined as an operator. We clarify this point in Appendix D. To select betting fractions, we follow Cutkosky & Orabona (2018) who state the problem of choosing the optimal betting fraction for coin betting as an online optimization problem with exp-concave losses and propose a strategy based on online Newton step (ONS) (Hazan et al., 2007) as a solution. ONS betting fractions are inexpensive to compute while being supported by strong theoretical guarantees. We also consider other strategies for selecting betting fractions and defer a detailed discussion to Appendix C. We conclude with formal guarantees on time-uniform type I error control and consistency of HSIC-based SKIT. In fact, we establish a stronger result: we show that the wealth process grows exponentially and characterize the rate of the growth of wealth in terms of the true HSIC score. The proof is deferred to Appendix B.2.2.

Algorithm 1 Online Newton step (ONS) strategy for selecting betting fractions

Input: sequence of payoffs $(f_t(Z_{2t-1}, Z_{2t}))_{t\geq 1}$, $\lambda_1^{\mathrm{ONS}} = 0$, $a_0 = 1$.
for $t = 1, 2, \ldots$ **do**
  Observe $f_t(Z_{2t-1}, Z_{2t})$;
  Set $z_t = f_t(Z_{2t-1}, Z_{2t})/(1 - \lambda_t^{\mathrm{ONS}} f_t(Z_{2t-1}, Z_{2t}))$;
  Set $a_t = a_{t-1} + z_t^2$;
  Set $\lambda_{t+1}^{\mathrm{ONS}} := \frac{1}{2} \wedge \left(0 \vee \left(\lambda_t^{\mathrm{ONS}} - \frac{2}{2-\log 3} \cdot \frac{z_t}{a_t}\right)\right)$;
end for

Theorem 2.4. *Suppose that Assumption 2.3 is satisfied. The following claims hold for HSIC-based SKIT (Algorithm 2):*

1. *Suppose that H_0 in (1a) or (2a) is true. Then SKIT ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in (1b) is true. Then it holds that $\mathcal{K}_t \overset{\text{a.s.}}{\longrightarrow} +\infty$ and thus SKIT is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the wealth grows exponentially, and the corresponding growth rate satisfies*

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \overset{\text{a.s.}}{\geq} \frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{4} \cdot \left(\frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{\mathbb{E}[(f^\star(Z_1,Z_2))^2]} \wedge 1\right), \tag{15}$$

where $f^\star$ is the oracle payoff defined in (12).

[3] At round t, evaluating HSIC-based payoff requires a number of operations that is linear in t (see Appendix F.2). Thus after T steps, we have expended a total of $O(T^2)$ computation, the same as batch HSIC. However, our test threshold is $1/\alpha$, but batch HSIC requires permutations to determine the right threshold, requiring recomputing HSIC hundreds of times. Thus, our test is actually more computationally feasible than batch HSIC.

Algorithm 2 HSIC-based SKIT

Input: significance level $\alpha \in (0,1)$, data stream $(Z_i)_{i\geq 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Use $Z_1, \ldots, Z_{2(t-1)}$ to compute $\hat{g}_t$ as in (13);
  Compute HSIC payoff $f_t(Z_{2t-1}, Z_{2t})$;
  Update the wealth process $\mathcal{K}_t$ as in (4);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 1);
  end if
end for

Since $\mathbb{E}[f^\star(Z_1,Z_2)] = \sqrt{\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})}$ and $\mathbb{E}\left[(f^\star(Z_1,Z_2))^2\right] \leq 1$, Theorem 2.4 implies that:

$$\liminf_{t\to\infty} \left(\frac{1}{t}\log \mathcal{K}_t\right) \overset{\text{a.s.}}{\geq} \frac{1}{4} \cdot \mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}).$$

However, the lower bound (15) is never worse. In particular, if the variance of the oracle payoffs: $\sigma^2 = \mathbb{V}[f^\star(Z_1,Z_2)]$, is small, meaning that $\sigma^2 \leq \mathbb{E}[f^\star(Z_1,Z_2)](1 - \mathbb{E}[f^\star(Z_1,Z_2)])$, we get a faster rate: $\sqrt{\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})}/4$, reminiscent of an empirical Bernstein type adaptation. Up to some small constants, we show that this is the best possible exponent that adapts automatically between the low- and high-variance regimes. We do this by considering the oracle test, i.e., assuming that the oracle HSIC payoff $f^\star$ is known. Amongst the betting fractions that are constrained to lie in $[-0.5, 0.5]$, like ONS bets, the optimal growth rate is ensured by taking

$$\lambda^\star = \operatorname*{arg\,max}_{\lambda\in[-0.5,0.5]} \mathbb{E}\left[\log(1 + \lambda f^\star(Z_1,Z_2))\right]. \tag{16}$$

We have the following result about the growth rate of the oracle test, whose proof is deferred to Appendix B.2.2.

Proposition 2.5. *The optimal log-wealth $S^\star := \mathbb{E}\left[\log(1 + \lambda^\star f^\star(Z_1,Z_2))\right]$ — that can be achieved by an oracle betting scheme (16) which knows $f^\star$ from (12) and the underlying distribution — satisfies:*

$$S^\star \leq \frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{2}\left(\frac{8\mathbb{E}[f^\star(Z_1,Z_2)]}{3\mathbb{E}[(f^\star(Z_1,Z_2))^2]} \wedge 1\right). \tag{17}$$

Remark 2.6 (Minibatching). While our test processes the data stream in pairs, it is possible to use larger batches of

points from P_{XY}. For a batch size is $b \geq 2$, at round t, the bet is placed on $\{(X_{b(t-1)+1}, Y_{b(t-1)+1}), \ldots, (X_{bt}, Y_{bt})\}$. In this case, the empirical mean embeddings are computed analogous to (14) but using $\{(X_i, Y_i)\}_{i \leq b(t-1)}$. We defer the details to Appendix D. Such payoff function satisfies the necessary conditions for the wealth process to be a nonnegative martingale, and hence, the resulting sequential test has time-uniform type I error control. The same argument as in the proof of Theorem 2.4 can be used to show that the resulting test is consistent. The main downside of minibatching is that monitoring of the test (and hence, optional stopping) is allowed only after processing every b points from P_{XY}.

Distribution Drift. As discussed in Section 1, batch independence tests have an inflated false alarm rate even under mild changes in distribution. In contrast, SKIT remains valid even when the data distribution drifts over time. For a stream of independent points, we claimed that our test controls the type I error as long as only one of the marginal distributions changes at each round. In Appendix D, we provide an example that shows that this assumption is necessary for the validity of our tests. Our tests can also be used to test instantaneous independence between two streams. Formally, define $\mathcal{D}_t := \{(X_i, Y_i)\}_{i \leq 2t}$ and consider:

$$H_0 : \forall t,\ X_{2t-1} \perp\!\!\!\perp Y_{2t-1} \mid \mathcal{D}_{t-1} \text{ and } X_{2t} \perp\!\!\!\perp Y_{2t} \mid \mathcal{D}_{t-1}, \tag{18a}$$

$$H_1 : \exists t' : X_{2t'-1} \not\perp\!\!\!\perp Y_{2t'-1} \mid \mathcal{D}_{t-1} \text{ or } X_{2t'} \not\perp\!\!\!\perp Y_{2t'} \mid \mathcal{D}_{t-1}. \tag{18b}$$

Assumption 2.7. Suppose that under H_0 in (18a), it also holds that:

(a) The cross-links between X and Y streams are not allowed, meaning that for all $t \geq 1$,

$$\begin{aligned} Y_t &\perp\!\!\!\perp X_{t-1} \mid Y_{t-1}, \{(X_i, Y_i)\}_{i \leq t-2}, \\ X_t &\perp\!\!\!\perp Y_{t-1} \mid X_{t-1}, \{(X_i, Y_i)\}_{i \leq t-2}. \end{aligned} \tag{19}$$

(b) For all $t \geq 1$, either (X_t, X_{t-1}) or (Y_t, Y_{t-1}) are exchangeable conditional on $\{(X_i, Y_i)\}_{i \leq t-2}$.

In the above, (a) relaxes the independence assumption within each pair, and (b) generalizes the assumption about allowed changes in the marginal distributions of X and Y. Under the above setting, we deduce that our test retains type-1 error control, and the proof is deferred to Appendix B.2.2.

Theorem 2.8. *Suppose that H_0 in (18a) is true. Further, assume that Assumption 2.7 holds. Then HSIC-based SKIT (Algorithm 2) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Chwialkowski & Gretton (2014) considered a related (at a high level) problem of testing instantaneous independence between a pair of time series. Similar to distribution drift, HSIC fails to test independence between innovations in time series since naively permuting one series destroys the underlying structure. Chwialkowski & Gretton (2014) used a subset of permutations — rotations by circular shifting (allowed by their assumption of strict stationarity) of one series for preserving the structure — to design a p-value and used the assumption of mixing (decreasing memory of a process) to justify the asymptotic validity. The setting we consider is very different, and we make no assumptions of mixing or stationarity anywhere. Related works on independence testing for time series also include (Chwialkowski et al., 2014; Besserve et al., 2013). In the next section, we extend the methodology to other dependence measures.

3. Alternative Dependence Measures

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ denote some classes of bounded functions $c_1 : \mathcal{X} \to \mathbb{R}$ and $c_2 : \mathcal{Y} \to \mathbb{R}$ respectively. For a class $\mathcal{C}$ of functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that factorize into the product: $c(x, y) = c_1(x)c_2(y)$ for some $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$, the general form of dependence measures (5) reduces to

$$m(P_{XY}; \mathcal{C}_1, \mathcal{C}_2) = \sup_{c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2} \operatorname{Cov}\left(c_1(X), c_2(Y)\right).$$

Next, we develop SKITs based on two dependence measures of this form: COCO and KCC. While the corresponding witness functions do not admit a closed form, efficient algorithms for computing the plug-in estimates are available.

Witness Functions for COCO. *Constrained covariance* (COCO) is a criterion for measuring dependence based on covariance between smooth functions of random variables:

$$\sup_{\substack{g,h: \\ \|g\|_{\mathcal{G}} \leq 1, \\ \|h\|_{\mathcal{H}} \leq 1}} \operatorname{Cov}\left(g(X), h(Y)\right) = \sup_{\substack{g,h: \\ \|g\|_{\mathcal{G}} \leq 1, \\ \|h\|_{\mathcal{H}} \leq 1}} \langle g, C_{XY} h \rangle_{\mathcal{G}}, \tag{20}$$

where the supremum is taken over the unit balls in the respective RKHSs (Gretton et al., 2005c;b). At round t, we are interested in empirical witness functions computed from $\{(X_i, Y_i)\}_{i \leq 2(t-1)}$. The key observation is that maximizing the objective function in (20) with the plug-in estimator of the cross-covariance operator requires considering only functions in $\mathcal{G}$ and $\mathcal{H}$ that lie in the span of the data:

$$\begin{aligned} \hat{g}_t &= \sum_{i=1}^{2(t-1)} \alpha_i \left(\varphi(X_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \varphi(X_j) \right), \\ \hat{h}_t &= \sum_{i=1}^{2(t-1)} \beta_i \left(\psi(Y_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \psi(Y_j) \right). \end{aligned} \tag{21}$$

Coefficients α and β that solve the maximization problem (20) define the leading eigenvector of the following

generalized eigenvalue problem (see Appendix D):

$$\begin{pmatrix} 0 & \frac{1}{2(t-1)}\tilde{K}\tilde{L} \\ \frac{1}{2(t-1)}\tilde{L}\tilde{K} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \gamma \begin{pmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{22}$$

where $\tilde{K} = HKH$, $\tilde{L} = HLH$, and $H = \mathbf{I}_{2(t-1)} - (1/(2(t-1))\mathbf{1}\mathbf{1}^\top$ is centering projection matrix. Computing the leading eigenvector for (22) is computationally demanding for moderately large t. A common practice is to use low-rank approximations of K and L with fast-decaying spectrum (Jordan & Bach, 2001). We present an approach based on incomplete Cholesky decomposition in Appendix F.1.

Witness Functions for KCC. *Kernelized canonical correlation* (KCC) relies on the regularized correlation between smooth functions of random variables:

$$\sup_{\substack{g\in\mathcal{G},\\ h\in\mathcal{H}}} \frac{\mathrm{Cov}\left(g(X), h(Y)\right)}{\sqrt{\mathbb{V}\left[g(X)\right] + \kappa_1 \left\|g\right\|_{\mathcal{G}}^2} \cdot \sqrt{\mathbb{V}\left[h(Y)\right] + \kappa_2 \left\|h\right\|_{\mathcal{H}}^2}}, \tag{23}$$

where regularization is necessary for obtaining meaningful estimates of correlation (Jordan & Bach, 2001; Fukumizu et al., 2007a). Witness functions for KCC have the same form as for COCO (21), but α and β define the leading eigenvector of a modified problem (Appendix D).

SKIT based on COCO or KCC. Given a pair of the witness functions $g^\star$ and $h^\star$ for COCO (or KCC) criterion, the corresponding oracle payoff: $f^\star(Z_{2t-1}, Z_{2t})$, is given by

$$\frac{1}{2}\left(g^\star(X_{2t}) - g^\star(X_{2t-1})\right)\left(h^\star(Y_{2t}) - h^\star(Y_{2t-1})\right), \tag{24}$$

which has the form (7) with $s = 1/2$. To construct the test, we rely on estimates $(f_t)_{t\geq 1}$ of the oracle payoff $f^\star$ obtained by using $\hat{g}_t$ and $\hat{h}_t$, defined in (21), in (24). We assume that α and β in (22) are normalized: $\alpha^\top \tilde{K}\alpha = 1$ and $\beta^\top \tilde{L}\beta = 1$. We conclude with a guarantee on time-uniform false alarm rate control of SKITs based on COCO (Algorithm 3), whose proof is deferred to Appendix B.3.

Theorem 3.1. *Suppose that (A1) in Assumption 2.3 is satisfied. Then, under H_0 in (1a) and (18a), COCO/KCC-based SKIT (Algorithm 3) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Remark 3.2. The above result does not contain a claim regarding the consistency of the corresponding tests. If (A2) in Assumption 2.3 holds, the same argument as in the proof of Theorem 2.4 can be used to deduce that SKITs based on the *oracle* payoffs (with oracle witness functions $g^\star$ and $h^\star$) are consistent. In contrast to HSIC, for which the oracle witness function is closed-form and the respective plug-in estimator is amenable for the analysis, to argue about the consistency of SKITs based on COCO/KCC, it is necessary to place additional assumptions, especially since low-rank approximations of kernel matrices are involved. We note that a sufficient condition for consistency is that the payoffs are positive on average: $\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^t f_i(Z_{2i-1}, Z_{2i}) \overset{\text{a.s.}}{>} 0$.

Algorithm 3 SKIT based on COCO (or KCC)

Input: significance level $\alpha \in (0,1)$, data stream $(Z_i)_{i\geq 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Use $Z_1, \ldots, Z_{2(t-1)}$ to compute $\hat{g}_t$ and $\hat{h}_t$ as in (21);
  Compute COCO payoff $f_t(Z_{2t-1}, Z_{2t})$;
  Update the wealth process $\mathcal{K}_t$ as in (4);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 1);
  end if
end for

Synthetic Experiments. To compare SKITs based on HSIC, COCO, and KCC payoffs, we use RBF kernel with hyperparameters taken to be inversely proportional to the second moment of the underlying variables; we observed no substantial difference when such selection is data-driven (median heuristic). We consider settings where the complexity of a task is controlled through a single univariate parameter:

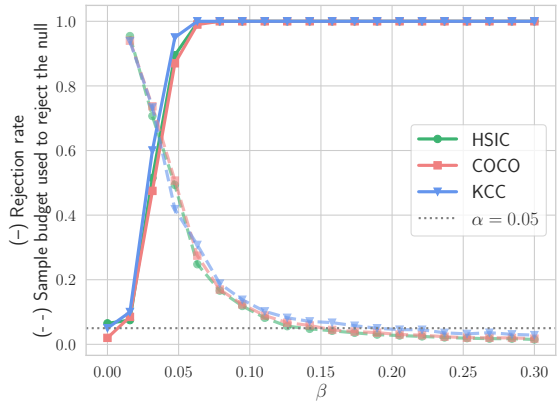
(a) *Gaussian model.* For $t \geq 1$, we consider $Y_t = X_t\beta + \varepsilon_t$, where $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$. We have that $\beta \neq 0$ implies nonzero linear correlation (hence dependence). We consider 20 values for β, spaced uniformly in [0,0.3], and use $\lambda_X = 1/4$ and $\lambda_Y = 1/(4(1+\beta^2))$ as kernel hyperparameters.

(b) *Spherical model.* We generate a sequence of dependent but linearly uncorrelated random variables by taking $(X_t, Y_t) = ((U_t)_{(1)}, (U_t)_{(2)})$, where $U_t \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathbb{S}^d)$, for $t \geq 1$. $\mathbb{S}^d$ denotes a unit sphere in $\mathbb{R}^d$ and $u_{(i)}$ is the i-th coordinate of u. We consider $d \in \{3, \ldots, 15\}$, and use $\lambda_X = \lambda_Y = d/4$ as kernel hyperparameters.

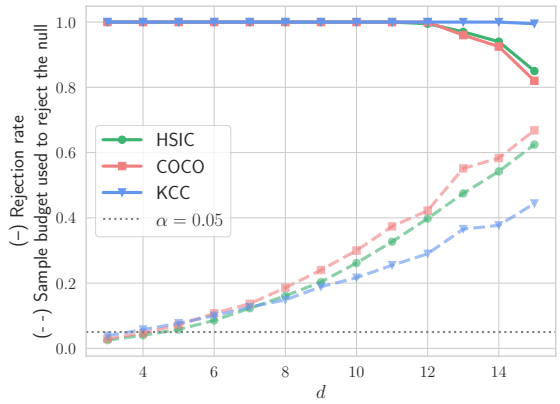
We stop monitoring after observing 20000 points from P_{XY} (if SKIT does not stop by that time, we retain the null) and aggregate the results over 200 runs for each value of β and d. In Figure 3, we confirm that SKITs control the type I error and adapt to the complexity of a task. In settings with a very low signal-to-noise ratio (small β or large d), SKIT's power drops, but in such cases, both sequential and batch independence tests inevitably require a lot of data to reject the null. We defer additional experiments to Appendix E.4.

4. Symmetry-based Betting Strategies

In this section, we develop a betting strategy that relies on symmetry properties, whose advantage is that it overcomes the kernel boundedness assumption that underlined



(a) Gaussian model.



(b) Spherical model.

Figure 3. Rejection rate and scaled sample size used to reject the null hypothesis for synthetic data. Inspecting the rejection rate for $\beta = 0$ (independence holds) confirms that the type I error is controlled. Further, we confirm that SKITs are adaptive to the complexity (smaller β and larger d correspond to harder settings).

the SKIT construction. For example, using this betting strategy with a linear kernel: $k(x, y) = l(x, y) = \langle x, y \rangle$ readily implies a valid sequential linear correlation test. Consider

$$W_t = \hat{g}_t(Z_{2t-1}) + \hat{g}_t(Z_{2t}) - \hat{g}_t(\tilde{Z}_{2t-1}) - \hat{g}_t(\tilde{Z}_{2t}), \quad (25)$$

where $\hat{g}_t = \hat{\mu}_{XY} - \hat{\mu}_X \otimes \hat{\mu}_Y$ is the *unnormalized* plug-in witness function computed from $\{Z_i\}_{i \le 2(t-1)}$. Symmetry-based betting strategies rely on the following key fact.

Proposition 4.1. *Under any distribution in H_0, W_t is symmetric around zero, conditional on $\mathcal{F}_{t-1}$.*

By construction, we expect the sign and magnitude of W_t to be positively correlated under the alternative. We consider three payoff functions that aim to exploit this fact.

1. *Composition with an odd function.* This approach is based on the idea from sequential symmetry testing (Ramdas et al., 2020) that composition with an odd function of a symmetric around zero random variable is mean-zero. Absent knowledge regarding the scale of considered random variables, it is natural to standardize $\{W_i\}_{i \ge 1}$ in a predictable way. We consider

$$f_t^{\text{odd}}(W_t) = \tanh\left(W_t / N_{t-1}\right), \quad (26)$$

where $N_t = \mathbf{Q}_{0.9}(\{|W_i|\}_{i \le t}) - \mathbf{Q}_{0.1}(\{|W_i|\}_{i \le t})$, and $\mathbf{Q}_\alpha(\{|W_i|\}_{i \le t})$ is the α-quantile of the empirical distribution of the absolute values of $\{W_i\}_{i \le t}$. (The choices of 0.1 and 0.9 are heuristic, and can be replaced by other constants without violating the validity of the test.) The composition approach has demonstrated promising empirical performance for the betting-based two-sample tests of Shekhar & Ramdas (2021) and conditional independence tests of Shaer et al. (2023).

2. *Rank-based approach.* Inspired by sequential signed-rank test of symmetry around zero of Reynolds Jr. (1975), we consider the following payoff function:

$$f_t^{\text{rank}}(W_t) = \text{sign}(W_t) \cdot \frac{\text{rk}(|W_t|)}{t}, \quad (27)$$

where $\text{rk}(|W_t|) = \sum_{i=1}^{t} \mathbb{1}\{|W_i| \le |W_t|\}$.

3. *Predictive approach.* At round t, we fit a probabilistic predictor $p_t : \mathbb{R}_+ \to [0, 1]$, e.g., logistic regression, using $\{|W_i|, \text{sign}(W_i)\}_{i \le t-1}$ as feature-label pairs. We consider the following payoff function:

$$f_t^{\text{pred}}(W_t) = (2p_t(|W_t|) - 1)_+ \cdot (1 - 2\ell_t(W_t)), \quad (28)$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ and $\ell_t(|W_t|, \text{sign}(W_t))$ is the misclassification loss of the predictor p_t.

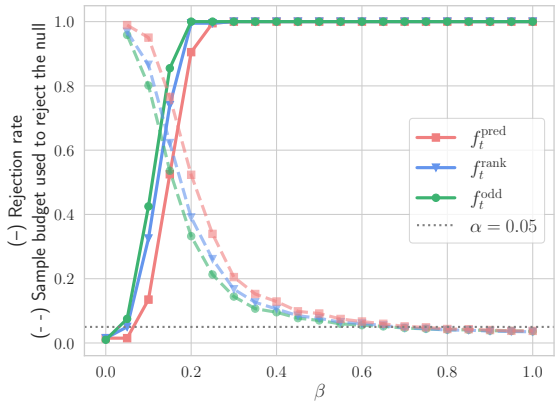
In the next result, whose proof is deferred to Appendix B.4, we show that symmetry-based SKITs are valid.

Algorithm 4 SKIT with symmetry-based betting

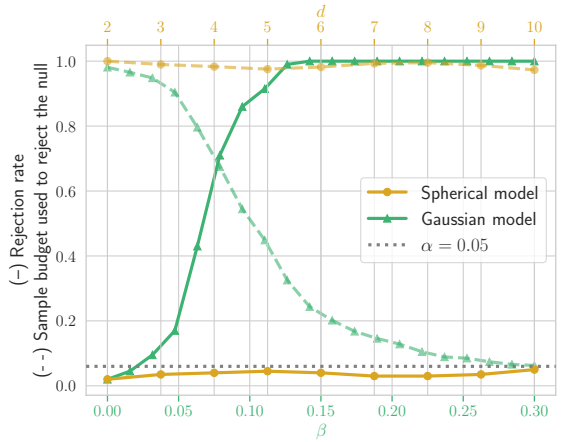
Input: significance level $\alpha \in (0, 1)$, data stream $(Z_i)_{i \ge 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\text{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Observe Z_{2t-1}, Z_{2t} and compute W_t as in (25);
  Compute payoff $f_t^{\text{odd}}(W_t)$ as in (26);
  Update the wealth process $\mathcal{K}_t$ as in (4);
  if $\mathcal{K}_t \ge 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\text{ONS}}$ (Algorithm 1);
  end if
end for

Theorem 4.2. *Under H_0 in (1a) and (18a), the symmetry-based SKIT (Algorithm 4) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \le \alpha$.*

Synthetic Experiments. To compare the symmetry-based payoffs, we consider the Gaussian model along with aGRAPA betting fractions. For visualization purposes, we complete monitoring after observing 2000 points from the joint distribution. In Figure 4(a), we observe that the resulting SKITs demonstrate similar performance. In Figure 4(b), we demonstrate that SKIT with a linear kernel has high power under the Gaussian model, whereas its false alarm rate does not exceed α under the spherical model. Additional synthetic experiments can be found in Appendix E.3.



(a)



(b)

Figure 4. (a) SKITs with symmetry-based payoffs have high power under the Gaussian model. (b) SKIT with linear kernel has high power under the Gaussian model (X and Y are linearly correlated for $\beta \neq 0$), and its false alarm rate is controlled under the spherical model (X and Y are linearly uncorrelated but dependent).

Real Data Experiments. We analyze average daily temperatures[4] in four European cities: London, Amsterdam, Zurich, and Nice, from January 1, 2017, to May 31, 2022. The processes underlying temperature formation are complex and act both on macro (e.g., solar phase) and micro (e.g., local winds) levels. While average daily temperatures in selected cities share similar cyclic patterns, one may still expect the temperature fluctuations occurring in nearby locations to be dependent. We use SKIT for testing instantaneous independence (as per (18)) between fluctuations (assuming that the conditions that underlie our test hold).

We run SKIT with the rank-based payoff and ONS betting fractions for each pair of cities using $6/\alpha$ as a rejection threshold (accounting for multiple testing). We select the kernel hyperparameters via the median heuristic using recordings for the first 20 days. In Figures 5, we illustrate that SKIT supports our conjecture that temperature fluctuations are dependent in nearby locations. We also run this experiment for four cities in South Africa (see Appendix E.5).

In addition, we analyze the performance of SKIT on MNIST data; the details are deferred to Appendix E.6.

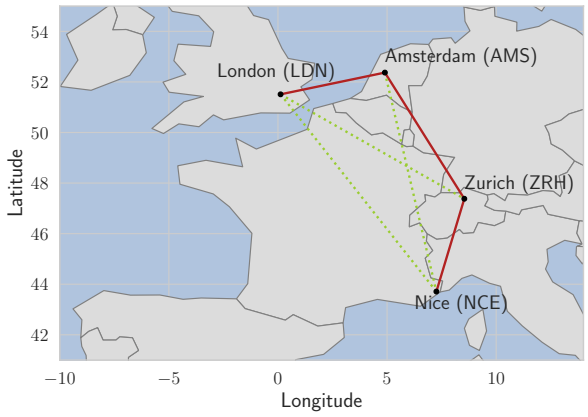


Figure 5. Solid lines connect cities for which the null is rejected. SKIT supports the conjecture regarding dependent temperature fluctuations in nearby locations.

5. Conclusion

A key advantage of sequential tests is that they can be continuously monitored, allowing an analyst to adaptively decide whether to stop and reject the null hypothesis or to continue collecting data, without inflating the false positive rate. In this paper, we design consistent sequential kernel independence tests (SKITs) following the principle of testing by betting. SKITs are also valid beyond the i.i.d. setting, allowing the data distribution to drift over time. Experiments on synthetic and real data confirm the power of SKITs.

Acknowledgements. The authors thank Ian Waudby-Smith, Tudor Manole and the anonymous reviewers for constructive feedback. The authors also acknowledge Lucas Janson and Will Hartog for thoughtful questions and comments at the International Seminar for Selective Inference.

[4] data source: https://www.wunderground.com

References

Balsubramani, A. and Ramdas, A. Sequential nonparametric testing with the law of the iterated logarithm. In *Conference on Uncertainty in Artificial Intelligence*, 2016.

Besserve, M., Logothetis, N. K., and Schölkopf, B. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems*, 2013.

Breiman, L. Optimal gambling systems for favorable games. *Berkeley Symposium on Mathematical Statistics and Probability*, 1962.

Chwialkowski, K. and Gretton, A. A kernel independence test for random processes. In *International Conference on Machine Learning*, 2014.

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, 2014.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in banach spaces. In *Conference on Learning Theory*, 2018.

Darling, D. A. and Robbins, H. E. Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 1968.

Ernst, P. A., Shepp, L. A., and Wyner, A. J. Yule's "nonsense correlation" solved! *The Annals of Statistics*, 2017.

Fan, X., Grama, I., and Liu, Q. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 2015.

Fukumizu, K., Bach, F. R., and Gretton, A. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 2007a.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, 2007b.

Gretton, A. A simpler condition for consistency of a kernel independence test. *arXiv preprint: 1501.06103*, 2015.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, 2005a.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 2005b.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. Kernel constrained covariance for dependence measurement. In *Tenth International Workshop on Artificial Intelligence and Statistics*, 2005c.

Grünwald, P., Henzi, A., and Lardy, T. Anytime-valid tests of conditional independence under model-X. *Journal of the American Statistical Association*, 2023.

Grünwald, P., de Heide, R., and Koolen, W. M. Safe testing. In *Information Theory and Applications Workshop*, 2020.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007.

Hoeffding, W. The strong law of large numbers for U–statistics. Technical report, University of North Carolina, 1961.

Jordan, M. I. and Bach, F. R. Kernel independent component analysis. *Journal of Machine Learning Research*, 2001.

Kelly, J. L. A new interpretation of information rate. *The Bell System Technical Journal*, 1956.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 2017.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint: 2009.03167*, 2020.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2022.

Reynolds Jr., M. R. A sequential signed-rank test for symmetry. *The Annals of Statistics*, 1975.

Shaer, S., Maman, G., and Romano, Y. Model-free sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Shafer, G. Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2021.

Shafer, G. and Vovk, V. *Game-Theoretic Foundations for Probability and Finance*. Wiley Series in Probability and Statistics. Wiley, 2019.

Shekhar, S. and Ramdas, A. Nonparametric two-sample testing by betting. *arXiv preprint: 2112.09162*, 2021.

Ville, J. *Étude critique de la notion de collectif*. Thèses de l'entre-deux-guerres. Gauthier-Villars, 1939.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2023.

Yule, G. U. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*, 1926.

Appendix

A. Independence Testing for Streaming Data

In Section A.1, we describe a permutation-based approach for conducting batch HSIC and show that continuous monitoring of batch HSIC (without corrections for multiple testing) leads to an inflated false alarm rate. In Section A.2, we introduce the sequential two-sample testing (2ST) problem and describe a reduction of sequential independence testing to sequential 2ST. In Section A.3, we compare our test to HSIC in the batch setting.

A.1. Failure of Batch HSIC under Continuous Monitoring

To conduct independence testing using batch HSIC, we use permutation p-value (with $M = 1000$ random permutations): $P = \frac{1}{M+1}(1 + \sum_{m=1}^{M} \mathbb{1}\{T_m \geq T\})$, where T_m is the value of HS-norm computed from the m-th permutation and T is HS-norm value on the original data. In other words, suppose that we are given a sample $Z_1, \dots, Z_t$, where $Z_i = (X_i, Y_i)$. Let $\mathcal{S}_t$ denote the set of all permutations of t indices and let $\sigma \sim \mathrm{Unif}(\mathcal{S}_t)$ be a random permutation of indices. Then:

$$(X_1, Y_1), \dots, (X_t, Y_t) \Longrightarrow T = \widehat{\mathrm{HSIC}}_b\left((X_1, Y_1), \dots, (X_t, Y_t)\right)$$
$$(X_1, Y_{\sigma_m(1)}), \dots, (X_t, Y_{\sigma_m(t)}), \Longrightarrow T_m = \widehat{\mathrm{HSIC}}_b\left((X_1, Y_{\sigma_m(1)}), \dots, (X_t, Y_{\sigma_m(t)})\right), \quad m \in \{1, \dots, M\},$$

where we use a biased estimator of HSIC:

$$\widehat{\mathrm{HSIC}}_b = \frac{1}{t^2}\sum_{i,j} K_{ij}L_{ij} + \frac{1}{t^4}\sum_{i,j,q,r} K_{ij}L_{qr} - \frac{2}{t^3}\sum_{i,j,q} K_{ij}L_{iq} = \frac{1}{t^2}\mathrm{tr}\left(KHLH\right).$$

For brevity, we use $K_{ij} = k(X_i, X_j)$, $L_{ij} = l(Y_i, Y_j)$ for $i, j \in \{1, \dots, t\}$. Next, we study batch HSIC under (a) *fixed-time* and (b) *continuous* monitoring. We consider a simple case when X and Y are independent standard Gaussian random variables. We consider (re)conducting a test at 12 different sample sizes: $t \in \{50, 100, \dots, 600\}$:

(a) Under fixed-time monitoring, for each value of t, we sample a sequence $Z_1, \dots, Z_t$ (100 times for each t) and conduct batch-HSIC test. The goal is to confirm that batch-HSIC controls type I error by tracking the standard miscoverage rate.

(b) Under continuous monitoring, we sample new datapoints and re-conduct the test. We illustrate inflated type I error by tracking the *cumulative miscoverage rate*, that is, the fraction of times, the test falsely rejects the independence null.

The results are presented in Figure 6. For Bonferroni correction, we decompose the error budget as: $\alpha = \sum_{i=1}^{\infty} \frac{\alpha}{i(i+1)}$, that is, for t-th test we use threshold $\alpha_t = \alpha/(t(t+1))$ for testing.

A.2. Sequential Independence Testing via Sequential Two-Sample Testing

First, we introduce the sequential two-sample testing problem. Suppose that we observe a stream of data: $(\tilde{X}_1, \tilde{Y}_1), (\tilde{X}_2, \tilde{Y}_2), \dots$, where $(\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q$. Two-sample testing refers to testing:

$$H_0: (\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q \text{ and } P = Q, \quad \text{vs.} \quad H_1: (\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q \text{ and } P \neq Q.$$

In Figure 1, we compared our test against the approach based on the reduction of independence testing to two-sample testing. We used the sequential two-sample kernel MMD test of Shekhar & Ramdas (2021) with the product kernel $\tilde{K}$ (that is, a product of Gaussian kernels) and the same set of hyperparameters as for our test for a fair comparison. To reduce sequential independence testing to any off-the-shelf sequential two-sample testing procedure, we convert the original sequence of points from P_{XY} to a sequence of i.i.d. $(\tilde{X}_t, \tilde{Y}_t)$-pairs, where $\tilde{X}_t \sim P_{XY}$ and $\tilde{Y}_t \sim P_X \times P_Y$ respectively; see Figure 7(a). At t-th round, we randomly choose one point as $\tilde{X}_t$, e.g., (X_1, Y_1) for the first triple. Next, we obtain $\tilde{Y}_t$ by randomly matching X

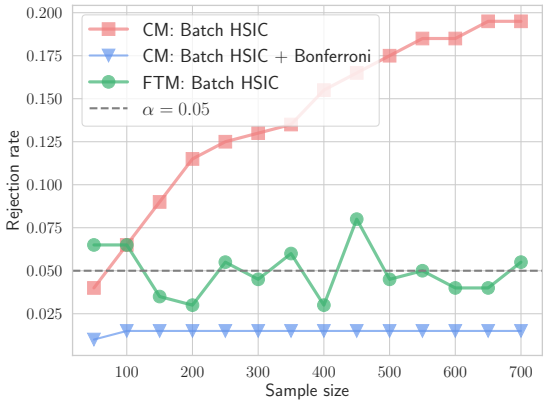


Figure 6. Inflated false alarm rate of batch HSIC under continuous monitoring (CM, red line with squares) for the case when X and Y are independent standard Gaussian random variables. Bonferroni correction (CM, blue line with triangles) restores type I error control. As expected, type I error is controlled at a specified level under fixed-time monitoring (FTM, green line with circles).

and Y from two other pairs, e.g., (X_2, Y_3) or (X_3, Y_2) for the first triple. In fact, the betting-based sequential two-sample test of (Shekhar & Ramdas, 2021) allows removing the effect of randomization (i.e., throwing away one observation in each triple), by averaging payoffs evaluated on $(\tilde{X}_t, \tilde{Y}_t^{(1)})$ and $(\tilde{X}_t, \tilde{Y}_t^{(2)})$. Other approaches — which do not require throwing data away — are also available (Figures 7(b)) but those only yield an i.i.d. sequence only under the null.

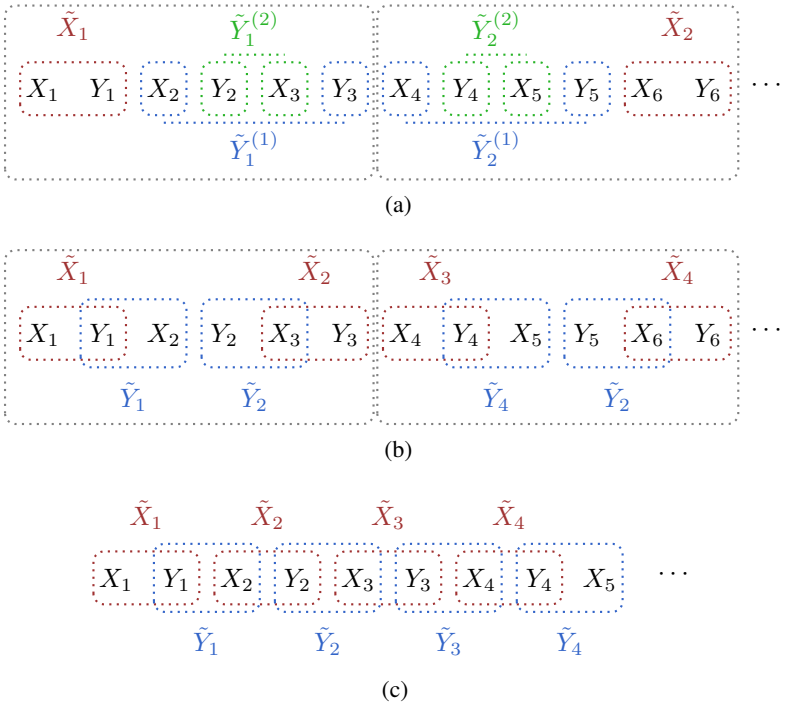


Figure 7. Reducing sequential independence testing to sequential two-sample testing. Processing as per (a) results in a sequence of i.i.d. observations both under the null and under the alternative (making the results about power valid). Processing data as per (b) gives an i.i.d. sequence only under the null. Reduction (b) is very similar to reduction (c). However, the latter makes $\tilde{X}_i$, $i \geq 2$, dependent on the past, and thus can not be used directly for considered sequential two-sample tests.

Additional Details of the Simulation Presented in Figure 1. We consider the Gaussian model: $Y_t = X_t\beta + \varepsilon_t$, where $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$, $t \geq 1$. We consider 10 values of β: $\beta \in \{0, 0.04, \ldots, 0.36\}$, and for each β we repeat the simulation 100

times. In this simulation, we compare three approaches for testing independence (valid under continuous monitoring):

1. **HSIC-based SKIT** proposed in this work (Algorithm 2);
2. **Batch HSIC** adapted to continuous monitoring via Bonferroni correction. We allow monitoring after processing every n, $n \in \{10, 100\}$, new points from P_{XY}, that is, the permutation p-value (computed over 2500 randomly sampled permutations) is compared against rejection thresholds: $\alpha_n = \alpha/(n(n+1))$, $n = 1, 2, \ldots$
3. **Sequential independence testing via reduction sequential 2ST** as described above.

We use RBF kernel with the same set of kernel hyperparameters for all testing procedures: $\lambda_X = 1/4$, $\lambda_Y = 1/(4(1+\beta^2))$.

A.3. Comparison in the Batch Setting

Sequential tests are complementary to batch tests and are not intended to replace them, and hence comparing the two on equal footing is hard. To highlight this, consider two simple scenarios. If we have 2000 data points, and HSIC fails to reject, there is not much we can do to rescue the situation. But if SKIT fails to reject, an analyst can collect more data and continue testing, retaining type I error control. In contrast, with 2 million points, HSIC will take forever to run, especially due to permutations. But if the alternative is true and the signal is strong, then SKIT may reject within 200 samples and stop. In short, the ability of SKIT to continue collecting and analyzing data is helpful for hard problems, and the ability of SKIT to stop early is helpful for easy problems. There is no easy sense in which one can compare them apples to apples and there is no sense in which batch HSIC uniformly dominates SKIT or vice versa. In a real setting, if an analyst has a strong hunch that the null is false and has the ability to collect data and run HSIC, the question is how much data should be collected? The answer depends on the underlying data distribution, which is of course unknown. With SKIT, data can be collected and analyzed sequentially. Theorem 2.4 implies that SKIT will stop early on easy problems and later on harder problems, all without knowing anything about the problem in advance. If however, one has a fixed batch of data, no chance to collect more, and no computational constraints, then running HSIC makes more sense.

To illustrate that batch HSIC can be superior to SKIT, we compare tests on a dataset with a prespecified sample size (500 observations from the Gaussian model) and track the empirical rejection rates of two tests. In Figure 8, we show that HSIC actually has higher power than SKIT. However, for $\beta = 0.1$ (where all tests have low power), Figure 3(a) shows that collecting just a bit more data (which is allowed) is needed for SKIT to reach perfect power. We also added a third method (D-SKIT) which removes the effect of the ordering of random variables under the assumptions that $\{(X_i, Y_i)\}_{i=1}^n$ are independent draws from P_{XY}. Let $\{\sigma_b\}_{b=1}^B$ define B random permutations of n indices, and let $\mathcal{K}_n^b$ denote the wealth after betting on a sequence ordered according to σ_b. For each b, $\mathcal{K}_n^b$ has expectation at most one, and hence (by linearity of expectation and Markov's inequality) $\mathbb{1}\left\{\frac{1}{B}\sum_{i=1}^B \mathcal{K}_n^b \geq 1/\alpha\right\}$ is a valid level-α batch test. This test is a bit more stable: it improves SKIT's power on moderate-complexity setups at the cost of a slight power loss on more extreme ones.

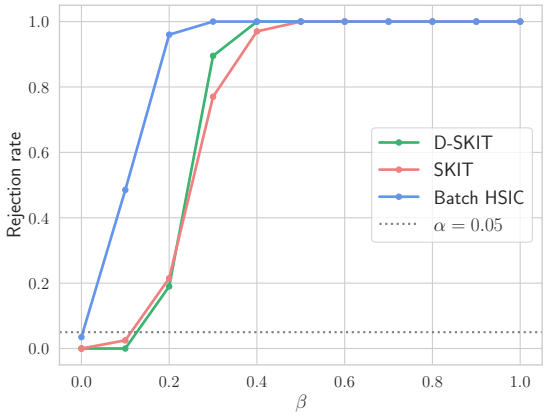


Figure 8. Comparison of SKIT and HSIC under Gaussian model in the batch setting. Non-surprisingly, batch HSIC performs best. D-SKIT improves over SKIT's power on moderate-complexity setups at the cost of a slight power loss on more extreme ones.

B. Proofs

Section B.1 contains auxiliary results needed to prove the results presented in this paper. In Section B.2, we prove the results from Section 2. In Secton B.3, we prove the results from Section 3.

B.1. Auxiliary Results

Theorem B.1 (Ville's inequality (Ville, 1939))**.** *Suppose that $(\mathcal{M}_t)_{t\geq 0}$ is a nonnegative supermartingale process adapted to a filtration $\{\mathcal{F}_t : t \geq 0\}$. Then, for any $a > 0$ it holds that:*

$$\mathbb{P}\left(\exists t \geq 1 : \mathcal{M}_t \geq a\right) \leq \frac{\mathbb{E}\left[\mathcal{M}_0\right]}{a}.$$

Theorem B.2 (Theorem 3 in (Gretton et al., 2005a))**.** *Assume that k and l are bounded almost everywhere by 1, and are nonnegative. Then for $n > 1$ and any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$ that:*

$$\left|\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) - \widehat{\mathrm{HSIC}}_b(P_{XY}; \mathcal{G}, \mathcal{H})\right| \leq \sqrt{\frac{\log(6/\delta)}{\alpha^2 n}} + \frac{C}{n},$$

where $\alpha^2 > 0.24$ and C are some absolute constants.

B.2. Proofs for Section 2

In Section B.2.1, we prove several intermediate results. The proofs of the main results are deferred to Section B.2.2.

B.2.1. Supporting Lemmas

Before we state the first result, recall the definition of the empirical mean embeddings computed from the first $2(t-1)$ datapoints:

$$\begin{aligned} \hat{\mu}_{XY}^{(t)} &= \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \otimes \psi(Y_i), \\ \hat{\mu}_X^{(t)} = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i), &\quad \hat{\mu}_Y^{(t)} = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i), \end{aligned} \tag{29}$$

where we highlight the dependence on the number of processed datapoints. We have the following result.

Lemma B.3. *For the empirical (29) and population (9) mean embeddings, it holds that:*

$$\left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \left\|\mu_{XY} - \mu_X \otimes \mu_Y\right\|_{\mathcal{G}\otimes\mathcal{H}}. \tag{30}$$

Proof. We have

$$\begin{aligned} \left\|\mu_{XY} - \mu_X \otimes \mu_Y\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 &= \mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}), \\ \left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 &= \widehat{\mathrm{HSIC}}_b^{(t)}(P_{XY}; \mathcal{G}, \mathcal{H}), \end{aligned}$$

where the latter is a biased estimator of HSIC, computed from $2(t-1)$ datapoints from P_{XY}. From Theorem B.2 and the Borel-Cantelli lemma, it follows that:

$$\left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 \xrightarrow{\text{a.s.}} \left\|\mu_{XY} - \mu_X \otimes \mu_Y\right\|_{\mathcal{G}\otimes\mathcal{H}}^2.$$

The result then follows from the continuous mapping theorem. □

Lemma B.4. *Suppose that H_1 in (1b) is true. Then for the oracle (11) and plug-in (13) witness functions, it holds that:*

$$\langle \hat{g}_t, g^\star \rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 1. \tag{31}$$

As a consequence, $\|\hat{g}_t - g^\star\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 0$.

Proof. Suppose that the alternative in (1b) happens to be true. Then since k and l are characteristic kernels, it follows that:

$$\|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}} > 0.$$

We aim to show that:

$$\left\langle \frac{\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}}{\left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}}}, \frac{\mu_{XY} - \mu_X \otimes \mu_Y}{\|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}}} \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 1.$$

From Lemma B.3, we know that: $\left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}}$. Hence it suffices to show that

$$\left\langle \hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}, \mu_{XY} - \mu_X \otimes \mu_Y \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}}^2 . \tag{32}$$

Recall that: $\mu_{XY} - \mu_X \otimes \mu_Y = \mathbb{E}\left[\varphi(\tilde{X}) \otimes \psi(\tilde{Y})\right] - \mathbb{E}\left[\varphi(\tilde{X})\right] \otimes \mathbb{E}\left[\psi(\tilde{Y})\right]$. We have:

$$\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)} = \left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \varphi(X_i)\otimes\psi(Y_i) - \frac{1}{4(t-1)^2 - 2(t-1)} \sum_{\substack{j,k=1: \\ j\neq k}}^{2(t-1)} \varphi(X_j)\otimes\psi(Y_k)\right).$$

Further, it holds that:

$$\begin{aligned}
&\left\langle \hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}, \mu_{XY} - \mu_X \otimes \mu_Y \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \\
= &\left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X_i)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y_i)\rangle_{\mathcal{H}}\right]\right) \\
- &\left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{4(t-1)^2 - 2(t-1)} \sum_{\substack{j,k=1: \\ j\neq k}}^{2(t-1)} \mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X_j)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y_k)\rangle_{\mathcal{H}}\right]\right),
\end{aligned}$$

For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have:

$$\begin{aligned}
\left|\mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right]\right| &\leq \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\left|\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right|\right] \\
&\leq \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\sqrt{k(\tilde{X},\tilde{X})k(x,x)l(\tilde{Y},\tilde{Y})k(y,y)}\right] \\
&\leq 1,
\end{aligned}$$

and similarly, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ it holds that:

$$\left|\mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right]\right| \leq 1.$$

Hence, by the SLLN, it follows that ($(X, Y), (\tilde{X}, \tilde{Y}) \overset{\text{iid}}{\sim} P_{XY}$):

$$\begin{aligned}
\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X_i)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y_i)\rangle_{\mathcal{H}}\right] &\xrightarrow{\text{a.s.}} \mathbb{E}_{X,Y,\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y)\rangle_{\mathcal{H}}\right] \\
&= \langle\mu_{XY},\mu_{XY}\rangle_{\mathcal{G}\otimes\mathcal{H}}.
\end{aligned}$$

Similarly, by the SLLN for U-statistics with bounded kernel (Hoeffding, 1961), it follows that:

$$\begin{aligned}
&\frac{1}{4(t-1)^2 - 2(t-1)} \sum_{\substack{j,k=1: \\ j\neq k}}^{2(t-1)} \mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X_j)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y_k)\rangle_{\mathcal{H}}\right] \\
\xrightarrow{\text{a.s.}} &\ \mathbb{E}_{X,\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X)\rangle_{\mathcal{G}}\right]\mathbb{E}_{Y,\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y)\rangle_{\mathcal{H}}\right] \\
= &\ \langle\mu_X\otimes\mu_Y,\mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}.
\end{aligned}$$

Hence, we deduce that:

$$\begin{aligned}\left\langle \hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}, \mu_{XY} - \mu_X \otimes \mu_Y \right\rangle_{\mathcal{G}\otimes\mathcal{H}} &\xrightarrow{\text{a.s.}} \langle \mu_{XY}, \mu_{XY} \rangle_{\mathcal{G}\otimes\mathcal{H}} - \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &= \langle \mu_{XY} - \mu_X \otimes \mu_Y, \mu_{XY} - \mu_X \otimes \mu_Y \rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &= \|\mu_{XY} - \mu_X \otimes \mu_Y\|^2_{\mathcal{G}\otimes\mathcal{H}}.\end{aligned}$$

Recalling (32), the proof of (31) is complete. To establish the consequence, simply note that:

$$\|\hat{g}_t - g^\star\|_{\mathcal{G}\otimes\mathcal{H}} = \sqrt{2\left(1 - \langle \hat{g}_t, g^\star \rangle_{\mathcal{G}\otimes\mathcal{H}}\right)},$$

and hence the result follows. □

Lemma B.5. *Suppose that $(x_t)_{t\geq 1}$ is a sequence of numbers such that $\lim_{t\to\infty} x_t = x$. Then the corresponding sequence of partial averages also converges to x, that is, $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n x_t = x$. This also implies that if $(X_t)_{t\geq 1}$ is a sequence of random variables such that $X_t \xrightarrow{\text{a.s.}} X$, then $(\sum_{t=1}^n X_t)/n \xrightarrow{\text{a.s.}} X$.*

Proof. Fix any $\varepsilon > 0$. Since $(x_t)_{t\geq 1}$ is converging, then $\exists M > 0$:

$$|x_t - x| \leq M, \quad \forall t \geq 1.$$

Now, let n_0 be such that $|x_t - x| \leq \varepsilon/2$ for all $n > n_0$. Further, choose any $n_1 > n_0$: $Mn_0/n_1 \leq \varepsilon/2$. Hence, for any $\tilde{n} > n_1$, it holds that:

$$\begin{aligned}\left|\frac{1}{\tilde{n}}\sum_{t=1}^{\tilde{n}} x_t - x\right| &\leq \left|\frac{1}{\tilde{n}}\sum_{t=1}^{n_0} x_t - x\right| + \left|\frac{1}{\tilde{n}}\sum_{t=n_0+1}^{\tilde{n}} x_t - x\right| \\ &\leq \frac{1}{\tilde{n}}\sum_{t=1}^{n_0}|x_t - x| + \frac{1}{\tilde{n}}\sum_{t=n_0+1}^{\tilde{n}}|x_t - x| \\ &\leq \frac{n_0}{\tilde{n}}M + \frac{\tilde{n} - n_0}{\tilde{n}}\frac{\varepsilon}{2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,\end{aligned}$$

which implies the result. □

Before we state the next result, recall that HSIC-based payoffs are based on the predictable estimates $\{\hat{g}_i\}_{i\geq 1}$ of the oracle witness function $g^\star$ and have the following form:

$$\begin{aligned}f_i(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(\hat{g}_i(Z_{2i-1}) + \hat{g}_i(Z_{2i})\right) - \frac{1}{2}\left(\hat{g}_i(\tilde{Z}_{2i-1}) + \hat{g}_i(\tilde{Z}_{2i})\right), \quad i \geq 1. \\ f^\star(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(g^\star(Z_{2i-1}) + g^\star(Z_{2i})\right) - \frac{1}{2}\left(g^\star(\tilde{Z}_{2i-1}) + g^\star(\tilde{Z}_{2i})\right).\end{aligned} \tag{33}$$

Lemma B.6. *Suppose that H_1 in (1b) is true. Then it holds that:*

$$\frac{1}{t}\sum_{i=1}^t f_i(Z_{2i-1}, Z_{2i}) \xrightarrow{\text{a.s.}} \mathbb{E}\left[f^\star(Z_1, Z_2)\right], \tag{34}$$

$$\frac{1}{t}\sum_{i=1}^t \left(f_i(Z_{2i-1}, Z_{2i})\right)^2 \xrightarrow{\text{a.s.}} \mathbb{E}\left[\left(f^\star(Z_1, Z_2)\right)^2\right]. \tag{35}$$

Proof. We start by proving (34). Note that:

$$\begin{aligned}f_i(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(\hat{g}_i(Z_{2i-1}) + \hat{g}_t(Z_{2i})\right) - \frac{1}{2}\left(\hat{g}_i(\tilde{Z}_{2i-1}) + \hat{g}_i(\tilde{Z}_{2i})\right) \\ &= \frac{1}{2}\left\langle \hat{g}_i, (\varphi(X_{2i}) - \varphi(X_{2i-1})) \otimes (\psi(Y_{2i}) - \psi(Y_{2i-1})) \right\rangle_{\mathcal{G}\otimes\mathcal{H}}.\end{aligned}$$

Next, observe that:

$$\left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right| \leq \left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i})\right| + \underbrace{\left|\frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i}) - \mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right|}_{\xrightarrow{\text{a.s.}} 0},$$

where the second term converges almost surely to 0 by the SLLN. For the first term, we have that:

$$\left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i})\right| \leq \frac{1}{t}\sum_{i=1}^{t} \left|f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})\right|.$$

Finally, note that:

$$\begin{aligned} \left|f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})\right| &= \frac{1}{2}\left|\langle \hat{g}_i - g^\star, (\varphi(X_{2i}) - \varphi(X_{2i-1})) \otimes (\psi(Y_{2i}) - \psi(Y_{2i-1}))\rangle_{\mathcal{G}\otimes\mathcal{H}}\right| \\ &\leq \|\hat{g}_i - g^\star\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 0, \end{aligned} \tag{36}$$

where the convergence result is due to Lemma B.4. The result (34) then follows after invoking Lemma B.5. Next, we prove (35). Note that:

$$\begin{aligned} \frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1}, Z_{2i}))^2 &= \frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i}) + f^\star(Z_{2i-1}, Z_{2i}))^2 \\ &= \underbrace{\frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i}))^2}_{\xrightarrow{\text{a.s.}} 0} \\ &\quad + \frac{2}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1}, Z_{2i}))(f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})) \\ &\quad + \underbrace{\frac{1}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1}, Z_{2i}))^2}_{\xrightarrow{\text{a.s.}} \mathbb{E}[(f^\star(Z_1,Z_2))^2]}, \end{aligned}$$

where the first convergence result is due to (36) and Lemma B.5 and the second convergence result is due to the SLLN. Using (36) and Lemma B.5, we deduce that:

$$\left|\frac{2}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1}, Z_{2i}))(f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i}))\right| \leq 2 \cdot \frac{1}{t}\sum_{i=1}^{t}\left|f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})\right| \xrightarrow{\text{a.s.}} 0,$$

and hence we conclude that the convergence (35) holds.

□

B.2.2. Main Results

Theorem 2.1. *Let $\mathcal{C}$ denote a class of functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for measuring dependence as per (5).*

1. *Under H_0 in (1a) and (2a), any payoff f of the form (7) satisfies $\mathbb{E}_{H_0}\left[f(Z_1, Z_2)\right] = 0$.*
2. *Suppose that $\mathcal{C}$ satisfies (6). Under H_1 in (1b), the oracle payoff f^* based on the witness function c^* satisfies $\mathbb{E}_{H_1}\left[f^\star(Z_1, Z_2)\right] > 0$. Further, for $\lambda^\star$ defined in (8), it holds that $\mathbb{E}_{H_1}\left[\log(1 + \lambda^\star f^\star(Z_1, Z_2)\right] > 0$. Hence, $\mathcal{K}_t^\star \xrightarrow{\text{a.s.}} +\infty$, which implies that the oracle test is consistent: $\mathbb{P}_{H_1}(\tau^\star < \infty) = 1$, where $\tau^\star = \inf\{t \geq 1 : \mathcal{K}_t^\star \geq 1/\alpha\}$.*

Proof. 1. Under H_0 in (1a), we have that:

$$(X_{2t-1}, Y_{2t-1}) \stackrel{d}{=} (X_{2t}, Y_{2t}) \stackrel{d}{=} (X_{2t-1}, Y_{2t}) \stackrel{d}{=} (X_{2t}, Y_{2t-1}),$$

and hence, the first part of the Proposition trivially follows from the linearity of expectation. Under distribution drift, we use that at least one of the marginal distributions does not change at each round. For example, suppose that at round t, it holds that: $P_X^{2t-1} = P_X^{2t}$. For the stream of independent observations, we have: $X_{2t} \perp\!\!\!\perp Y_{2t-1}$ and $X_{2t-1} \perp\!\!\!\perp Y_{2t}$. Further, under the H_0 in (2a), it holds that: $X_{2t-1} \perp\!\!\!\perp Y_{2t-1}$ and $X_{2t} \perp\!\!\!\perp Y_{2t}$. Hence, we have:

$$(X_{2t-1}, Y_{2t-1}) \stackrel{d}{=} (X_{2t}, Y_{2t-1}) \quad \text{and} \quad (X_{2t-1}, Y_{2t}) \stackrel{d}{=} (X_{2t}, Y_{2t}),$$

and hence, we get the result using linearity of expectation.

2. Under the i.i.d. setting, we have

$$\mathbb{E}\left[f^\star(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[f^\star(Z_1, Z_2)\right] = s \cdot m(P_{XY}; \mathcal{C}),$$

and hence the result follows from the fact that the functional class $\mathcal{C}$ satisfies the characteristic condition (6).

3. Let $W := f^\star(Z_1, Z_2)$, and consider $\mathbb{E}_{H_1}[\log(1 + \lambda W)]$. We know that $\mathbb{E}_{H_1}[W] > 0$. We use the following inequality (Fan et al., 2015, Equation (4.12)): for any $y \geq -1$ and $\lambda \in [0, 1)$, it holds:

$$\log(1 + \lambda y) \geq \lambda y + y^2 \left(\log(1 - \lambda) + \lambda\right)$$

Hence

$$\mathbb{E}\left[\log(1 + \lambda W)\right] \geq \lambda \mathbb{E}\left[W\right] + \mathbb{E}\left[W^2\right] \left(\log(1 - \lambda) + \lambda\right).$$

Finally, using that $\log(1 - x) + x \geq -x^2/(2(1 - x))$ for $x \in [0, 1)$, we get:

$$\mathbb{E}_{H_1}\left[\log(1 + \lambda^\star W)\right] \geq \frac{(\mathbb{E}_{H_1}[W])^2/2}{\mathbb{E}_{H_1}[W] + \mathbb{E}_{H_1}[W^2]} > 0,$$

where recall that:

$$\lambda^\star = \frac{\mathbb{E}[W]}{\mathbb{E}[W] + \mathbb{E}[W^2]} \in (0, 1).$$

The wealth process corresponding to the oracle test satisfies:

$$\mathcal{K}_t = \prod_{i=1}^{t} (1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i})) = \exp\left(t \cdot \frac{1}{t} \sum_{i=1}^{t} \log(1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i}))\right).$$

By the Strong Law of Large Numbers (SLLN), we have:

$$\frac{1}{t} \sum_{i=1}^{t} \log(1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i})) \xrightarrow{\text{a.s.}} \mathbb{E}\left[\log(1 + \lambda^\star W)\right] > 0.$$

Hence, we get that $\mathcal{K}_t \xrightarrow{\text{a.s.}} +\infty$, and hence, the oracle test is consistent.

□

Theorem 2.4. *Suppose that Assumption 2.3 is satisfied. The following claims hold for HSIC-based SKIT (Algorithm 2):*

1. *Suppose that H_0 in (1a) or (2a) is true. Then SKIT ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in (1b) is true. Then it holds that $\mathcal{K}_t \overset{\text{a.s.}}{\longrightarrow} +\infty$ and thus SKIT is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the wealth grows exponentially, and the corresponding growth rate satisfies*

$$\liminf_{t \to \infty} \frac{\log \mathcal{K}_t}{t} \overset{\text{a.s.}}{\geq} \frac{\mathbb{E}[f^\star(Z_1, Z_2)]}{4} \cdot \left(\frac{\mathbb{E}[f^\star(Z_1, Z_2)]}{\mathbb{E}[(f^\star(Z_1, Z_2))^2]} \wedge 1\right), \tag{15}$$

where $f^\star$ is the oracle payoff defined in (12).

Remark B.7. While it will be clear from the proof that the i.i.d. assumption is sufficient but not necessary for asymptotic power one, the more relaxed sufficient conditions are slightly technical to state and thus omitted.

Proof. 1. First, let us show that the predictable estimates of the oracle payoff function are bounded when the scaling factor $s = 1/2$ is used. Recall that:

$$\begin{aligned} f_t((x', y'), (x, y)) &= \frac{1}{2}\left(\hat{g}_t(x', y') - \hat{g}_t(x', y) + \hat{g}_t(x, y) - \hat{g}_t(x, y')\right) \\ &= \frac{1}{2}\left\langle \hat{g}_t, \varphi(x') \otimes \psi(y') - \varphi(x') \otimes \psi(y) + \varphi(x) \otimes \psi(y) - \varphi(x) \otimes \psi(y')\right\rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &= \frac{1}{2}\left\langle \hat{g}_t, (\varphi(x') - \varphi(x)) \otimes (\psi(y') - \psi(y))\right\rangle_{\mathcal{G}\otimes\mathcal{H}}. \end{aligned} \tag{37}$$

Note that:

$$\begin{aligned} |f_t((x', y'), (x, y))| &\leq \frac{1}{2}\|\hat{g}_t\|_{\mathcal{G}\otimes\mathcal{H}} \|(\varphi(x') - \varphi(x)) \otimes (\psi(y') - \psi(y))\|_{\mathcal{G}\otimes\mathcal{H}} \\ &\leq \frac{1}{2}\|(\varphi(x') - \varphi(x)) \otimes (\psi(y') - \psi(y))\|_{\mathcal{G}\otimes\mathcal{H}} \\ &= \frac{1}{2}\|\varphi(x') - \varphi(x)\|_{\mathcal{G}} \cdot \|\psi(y') - \psi(y)\|_{\mathcal{H}} \\ &= \frac{1}{2}\sqrt{2(1 - k(x', x))} \cdot \sqrt{2(1 - l(y', y))} \\ &= 1. \end{aligned}$$

and hence, $f_t((x', y'), (x, y)) \leq [-1, 1]$. Next, we show that constructed payoff function yields a fair bet. Indeed, we have that:

$$\mathbb{E}\left[f_t(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] = \left\langle \hat{g}_t, \mu_{XY} - \mu_X \otimes \mu_Y\right\rangle_{\mathcal{G}\otimes\mathcal{H}},$$

and in particular, the above implies that $\mathbb{E}_{H_0}\left[f_t(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] = 0$ for H_0 in (1a). For H_0 in (2a), it is easy to see that the result holds using the form (37). We use that $X_{2t-1} \perp\!\!\!\perp Y_{2t-1}$, $X_{2t} \perp\!\!\!\perp Y_{2t}$, $X_{2t} \perp\!\!\!\perp Y_{2t-1}$, $X_{2t-1} \perp\!\!\!\perp Y_{2t}$, and the fact that at least one of the marginal distributions does not change. Next, we show that for all strategies for selecting betting fractions that are considered in this work, the resulting wealth process is a nonnegative martingale. In case aGRAPA/ONS strategies are used, the resulting wealth process is clearly a nonnegative martingale since betting fractions are predictable. The mixed wealth process $\left(\mathcal{K}_t^{\text{mixed}}\right)_{t\geq 1}$ is a nonnegative martingale under the null H_0, and hence

$$\begin{aligned} \mathbb{E}_{H_0}\left[\mathcal{K}_t^{\text{mixed}} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\int_0^1 \mathcal{K}_{t-1}^{\lambda}(1 + \lambda f_t(Z_{2t-1}, Z_{2t}))\nu(\lambda)d\lambda \mid \mathcal{F}_{t-1}\right] \\ &= \int_0^1 \mathcal{K}_{t-1}^{\lambda}\mathbb{E}_{H_0}\left[1 + \lambda f_t(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right]\nu(\lambda)d\lambda \\ &= \int_0^1 \mathcal{K}_{t-1}^{\lambda}\nu(\lambda)d\lambda \\ &= \mathcal{K}_{t-1}^{\text{mixed}}, \end{aligned}$$

where the interchange of conditional expectation and integration is justified by the conditional monotone convergence theorem. The assertion of the Theorem then follows directly from Ville's inequality (Proposition B.1) when $a = 1/\alpha$.

2. Next, we establish the consistency of HSIC-based SKIT with ONS betting strategy. Under the ONS betting strategy, for any sequence of outcomes $(f_i)_{i\geq 1}$, $f_i \in [-1, 1]$, $i \geq 1$, it holds that (see the proof of Theorem 1 in (Cutkosky & Orabona, 2018)):

$$\log \mathcal{K}_t(\lambda_0) - \log \mathcal{K}_t = O\left(\log\left(\sum_{i=1}^t f_i^2\right)\right), \tag{38}$$

where $\mathcal{K}_t(\lambda_0)$ is the wealth of any constant betting strategy $\lambda_0 \in [-1/2, 1/2]$ and $\mathcal{K}_t$ is the wealth corresponding to the ONS betting strategy. It follows that the wealth process corresponding to the ONS betting strategy satisfies

$$\frac{\log \mathcal{K}_t}{t} \geq \frac{\log \mathcal{K}_t(\lambda_0)}{t} - C \cdot \frac{\log t}{t}, \tag{39}$$

for some absolute constant $C > 0$. Next, let us consider:

$$\lambda_0 = \frac{1}{2}\left(\left(\frac{\sum_{i=1}^t f_i}{\sum_{i=1}^t f_i^2} \wedge 1\right) \vee 0\right).$$

We obtain:

$$\begin{aligned}\frac{\log \mathcal{K}_t(\lambda_0)}{t} &= \frac{1}{t}\sum_{i=1}^t \log(1+\lambda_0 f_i) \\ &\overset{(a)}{\geq} \frac{1}{t}\sum_{i=1}^t (\lambda_0 f_i - \lambda_0^2 f_i^2) \\ &= \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{4} \vee 0\right) \cdot \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{\frac{1}{t}\sum_{i=1}^t f_i^2} \wedge 1\right),\end{aligned} \tag{40}$$

where in (a) we used[5] that $\log(1+x) \geq x - x^2$ for $x \in [-1/2, 1/2]$. From Lemma B.6, it follows for $f_i = f_i(Z_{2i-1}, Z_{2i})$ that:

$$\frac{\frac{1}{t}\sum_{i=1}^t f_i(Z_{2i-1}, Z_{2i})}{4} \cdot \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i(Z_{2i-1}, Z_{2i})}{\frac{1}{t}\sum_{i=1}^t (f_i(Z_{2i-1}, Z_{2i}))^2} \wedge 1\right) \xrightarrow{\text{a.s.}} \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{4} \cdot \left(\frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1\right). \tag{41}$$

Further, note that:

$$\mathbb{E}\left[f^\star(Z_1, Z_2)\right] = \|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}} = \sqrt{\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})},$$

which is positive if the H_1 is true. Hence, using (39), we deduce that the growth rate of the ONS wealth process satisfies

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \geq \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{4} \cdot \left(\frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1\right). \tag{42}$$

We conclude that the test is consistent, that is, if H_1 is true, then $\mathbb{P}(\tau < \infty) = 1$.

□

Proposition 2.5. *The optimal log-wealth $S^\star := \mathbb{E}\left[\log(1 + \lambda^\star f^\star(Z_1, Z_2))\right]$ — that can be achieved by an oracle betting scheme (16) which knows $f^\star$ from (12) and the underlying distribution — satisfies:*

$$S^\star \leq \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{2}\left(\frac{8\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{3\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1\right). \tag{17}$$

Proof. We start by establishing the upper bound in (17). The fact that $S^\star \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$ trivially follows from $\mathbb{E}\left[\log(1 + \lambda f^\star(Z_1, Z_2))\right] \leq \lambda\mathbb{E}\left[f^\star(Z_1, Z_2)\right] \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$. Since for any $x \in [-0.5, 0.5]$, it holds that: $\log(1 + x) \leq x - 3x^2/8$, we know that:

$$S^\star \leq \max_{\lambda\in[-0.5,0.5]} \left(\lambda\mathbb{E}\left[f^\star(Z_1, Z_2)\right] - \frac{3}{8}\lambda^2\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]\right), \tag{43}$$

and by solving the maximization problem, we get the upper bound:

$$S^\star \leq \frac{2}{3}\frac{(\mathbb{E}\left[f^\star(Z_1, Z_2)\right])^2}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]}, \tag{44}$$

assuming $(\mathbb{E}\left[f^\star(Z_1, Z_2)\right])^2/\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right] \leq 3/8$. On the other hand, it always holds that: $S^\star \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$. To obtain the claimed bound, we multiply the RHS of (44) by two, which completes the proof of (17).

□

[5] A slightly better constant for the growth rate (0.3 in place of 1/4) can be obtained by using the inequality: $\log(1+x) \geq x - \frac{5}{6}x^2$, that holds $\forall x \in [-0.5, 0.5]$.

Theorem 2.8. *Suppose that H_0 in (18a) is true. Further, assume that Assumption 2.7 holds. Then HSIC-based SKIT (Algorithm 2) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Proof. Recall that at round t, the payoff function has form:

$$f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) = \frac{1}{2}\langle \hat{g}_t, (\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1}))\rangle_{\mathcal{G}\otimes\mathcal{H}}.$$

Let $\mathcal{D}_t = \{(X_i, Y_i)\}_{i \leq 2(t-1)}$. To establish validity, we need to show that under H_0 in (18a),

$$\mathbb{E}\left[f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) \mid \mathcal{D}_t\right] = 0, \tag{45}$$

and hence it suffices to show that:

$$\mathbb{E}\left[(\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1})) \mid \mathcal{D}_t\right] = 0.$$

Due to independence under the null H_0, we have:

$$\begin{aligned}
\mathbb{E}\left[\varphi(X_{2t-1}) \otimes \psi(Y_{2t-1}) \mid \mathcal{D}_t\right] &= \mathbb{E}\left[\varphi(X_{2t-1}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] =: \mu_X^{2t-1} \otimes \mu_Y^{2t-1}, \\
\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t}) \mid \mathcal{D}_t\right] &= \mathbb{E}\left[\varphi(X_{2t}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t}) \mid \mathcal{D}_t\right] =: \mu_X^{2t} \otimes \mu_Y^{2t},
\end{aligned}$$

Consider one of the cross-terms $\varphi(X_{2t}) \otimes \psi(Y_{2t-1})$. We have the following:

$$\begin{aligned}
\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t-1}) \mid \mathcal{D}_t\right] &\overset{a}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t-1}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{b}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{c}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{d}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \\
&\overset{e}{=} \mathbb{E}\left[\varphi(X_{2t}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \\
&\overset{f}{=} \mu_X^{2t} \otimes \mu_Y^{2t-1}.
\end{aligned}$$

In the above, (a) uses the law of iterated expectations and conditioning on X_{2t-1}, (b) uses the assumption (19) about conditional independence, (c) uses the independence null assumption (1a), (d) uses that $\mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right]$ is $\sigma(\mathcal{D}_t)$-measurable, (e) uses the law of iterated expectations, and (f) uses the definitions of the mean embeddings of conditional distributions. An analogous argument can be used to deduce:

$$\mathbb{E}\left[\varphi(X_{2t-1}) \otimes \psi(Y_{2t}) \mid \mathcal{D}_t\right] = \mu_X^{2t-1} \otimes \mu_Y^{2t}.$$

We get that:

$$\begin{aligned}
\mathbb{E}\left[(\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1})) \mid \mathcal{D}_t\right] &= \mu_X^{2t-1} \otimes \mu_Y^{2t-1} + \mu_X^{2t} \otimes \mu_Y^{2t} - \mu_X^{2t-1} \otimes \mu_Y^{2t} - \mu_X^{2t} \otimes \mu_Y^{2t-1} \\
&= (\mu_X^{2t} - \mu_X^{2t-1}) \otimes (\mu_Y^{2t} - \mu_Y^{2t-1}),
\end{aligned}$$

and hence, if either (X_{2t-1}, X_{2t}) or (Y_{2t-1}, X_{2t}) are exchangeable conditional on $\mathcal{D}_t$, it follows that either $\mu_X^{2t} = \mu_X^{2t-1}$ or $\mu_Y^{2t} = \mu_Y^{2t-1}$ respectively. This, in turn, implies that (45) holds, and hence, the result follows.

□

B.3. Proofs for Section 3

Theorem 3.1. *Suppose that (A1) in Assumption 2.3 is satisfied. Then, under H_0 in (1a) and (18a), COCO/KCC-based SKIT (Algorithm 3) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Proof. It suffices to show that the proposed payoff functions are bounded. The rest of the proof follows will follow the same steps as the proof of Theorem 2.4 (for a stream of independent observations) or Theorem 2.8 (for time-varying independence

null), and we omit the details. Note that:

$$\begin{aligned}\left|\hat{h}_t(y') - \hat{h}_t(y)\right| &= \left|\langle \hat{h}_t, \psi(y')\rangle_{\mathcal{H}} - \langle \hat{h}_t, \psi(y)\rangle_{\mathcal{H}}\right| \\ &= \left|\langle \hat{h}_t, \psi(y') - \psi(y)\rangle_{\mathcal{H}}\right| \\ &\leq \left\|\hat{h}_t\right\|_{\mathcal{H}} \|\psi(y') - \psi(y)\|_{\mathcal{H}} \\ &\leq \|\psi(y') - \psi(y)\|_{\mathcal{H}} \\ &= \sqrt{2(1 - l(y, y'))} \\ &\leq \sqrt{2},\end{aligned}$$

where we used that $\left\|\hat{h}_t\right\|_{\mathcal{H}} \leq 1$ due to normalization. Analogous bound holds for $|\hat{g}_t(x') - \hat{g}_t(x)|$. We conclude that any predictable estimate of the oracle payoff function for COCO (or KCC) satisfies

$$|f_t((x', y'), (x, y))| \leq 1,$$

as proposed. The fact that the payoff function is fair trivially follows from the definition. Regarding the existence of the oracle payoff, whose mean is positive under H_1 in (1b), note that if k and l are characteristic kernels, then COCO and KCC satisfy the characteristic condition (6); see Jordan & Bach (2001); Gretton et al. (2005c;b). Hence, the result follows from Theorem 2.1. This completes the proof.

□

B.4. Proofs for Section 4

Theorem 4.2. *Under H_0 in (1a) and (18a), the symmetry-based SKIT (Algorithm 4) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Proof. For any $t \geq 1$, we have that the payoffs defined in (26), (27), and (28) are bounded: $f_t(w) \in [-1, 1], \forall w \in \mathbb{R}$. Due to Proposition 4.1, we know that, under the null, W_t is a random variable that is symmetric around zero (conditional on $\mathcal{F}_{t-1}$). Hence, for the composition approach, it trivially follows that $\mathbb{E}_{H_0}\left[f_t^{\text{odd}}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$ since a composition with an odd function is used. For the rank and predictive approaches, we use the fact that, under the null, $\text{sign}(W_t) \perp\!\!\!\perp |W_t| \mid \mathcal{F}_{t-1}$. Since, $\mathbb{E}_{H_0}\left[\text{sign}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$, it then follows that $\mathbb{E}_{H_0}\left[f_t^{\text{rank}}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$. Using that $\text{sign}(W_t) \perp\!\!\!\perp |W_t| \mid \mathcal{F}_{t-1}$ and by conditioning on the sign of W_t, we get:

$$\mathbb{E}_{H_0}\left[\ell_t(W_t) \mid \mathcal{F}_{t-1}\right] = \frac{1}{2}\mathbb{P}_{H_0}\left(p_t(|W_t|) \geq 1/2\right) + \frac{1}{2}\mathbb{P}_{H_0}\left(p_t(|W_t|) < 1/2\right) = \frac{1}{2}.$$

Hence $\mathbb{E}_{H_0}\left[1 - 2\ell_t(W_t) \mid \mathcal{F}_{t-1}\right] = 0$. The rest of the proof regarding the validity of the symmetry-based SKITs follows the same steps as the proof of Theorem 2.4, and we omit the details.

□

C. Selecting Betting Fractions

As alluded to in Remark 2.2, sticking to a single fixed betting fraction, $\lambda_t = \lambda \in [0, 1]$, $t \geq 1$, may result in a wealth process that either has a sub-optimal growth rate under the alternative or tends to zero almost surely (see Figure 9). *Mixing* over different betting fractions is a simple approach that often works well in practice. Given a fine grid of values: $\Lambda = \{\lambda^{(1)}, \ldots, \lambda^{(J)}\}$, e.g., uniformly spaced values on the unit interval, consider

$$\mathcal{K}_t^{\text{mixed}} = \frac{1}{|\Lambda|} \sum_{\lambda^{(j)} \in \Lambda} \mathcal{K}_t(\lambda^{(j)}), \tag{46}$$

where $\left(\mathcal{K}_t(\lambda^{(j)})\right)_{t \geq 0}$ is a wealth process corresponding to a constant-betting strategy with betting fraction $\lambda^{(j)}$ [6].

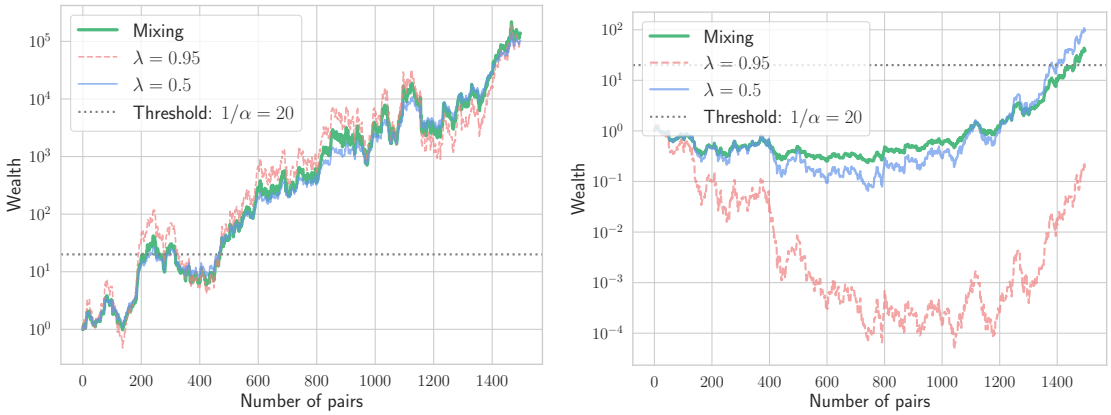


Figure 9. SKIT with HSIC payoff function on two particular realizations of streams of dependent data: $Y_t = 0.1 \cdot X_t + \varepsilon_t$, $X_t, \varepsilon_t \sim \mathcal{N}(0, 1)$. For both cases, we consider a mixed wealth process for $\Lambda = \{0.05, 0.1, \ldots, 0.95\}$. We observe that the mixed wealth process follows closely the best of constant-betting strategies with $\lambda \in \{0.5, 0.95\}$.

While mixing often works well in practice, it introduces additional tuning hyperparameters, e.g., grid size. We consider two compelling approaches for the selection of betting fractions in a predictable way, meaning that λ_t depends only on $\{(X_i, Y_i)\}_{i \leq 2(t-1)}$. In addition to the ONS strategy (Algorithm 1), we also consider *aGRAPA* strategy (Algorithm 5). The idea that effective betting strategies are ones that maximize a gambler's expected log capital dates back to early works of Kelly (1956) and Breiman (1962). Assuming that the same betting fraction is used, the log capital after round $(t-1)$ is

$$\log \mathcal{K}_{t-1}(\lambda) = \sum_{i=1}^{t-1} \log\left(1 + \lambda f_i(Z_{2i-1}, Z_{2i})\right).$$

Algorithm 5 aGRAPA strategy for selecting betting fractions

Input: sequence of payoffs $(f_t(Z_{2t-1}, Z_{2t}))_{t \geq 1}$, $\lambda_1^{\text{aGRAPA}} = 0$, $\mu_0^{(1)} = 0$, $\mu_0^{(2)} = 1$, $c = 0.9$.
for $t = 1, 2, \ldots$ **do**
  Set $\mu_t^{(1)} = \mu_{t-1}^{(1)} + f_t(Z_{2t-1}, Z_{2t})$;
  Set $\mu_t^{(2)} = \mu_{t-1}^{(2)} + (f_t(Z_{2t-1}, Z_{2t}))^2$;
  Set $\lambda_{t+1}^{\text{aGRAPA}} = c \wedge \left(0 \vee \left(\mu_t^{(1)} / \mu_t^{(2)}\right)\right)$;
end for

[6] Practically, it is advisable to start with a coarse grid (small J) at small t and occasionally add another grid point, so that the grid becomes finer over time. Whenever a grid point is added, it is like adding another stock to a portfolio, and the wealth must be appropriately redistributed; we omit the details for brevity.

Following Waudby-Smith & Ramdas (2023), we set the derivative to zero and use Taylor's expansion to get

$$\lambda_t^{\mathrm{aGRAPA}} = \left(\left(\frac{\sum_{i=1}^{t-1} f_i(Z_{2i-1}, Z_{2i})}{\sum_{i=1}^{t-1} \left(f_i(Z_{2i-1}, Z_{2i})\right)^2} \right) \vee 0 \right) \wedge c.$$

Truncation at zero is inspired by the fact that $\mathbb{E}_{H_1}\left[f^\star(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] > 0$, whereas truncation at $c \in (0, 1]$ (e.g., $c = 0.9$) is necessary to guarantee that the wealth process is indeed nonnegative.

D. Omitted Details for Sections 2 and 3

In this section, we complement the material presented in the main paper by deriving the forms of the witness functions for the dependence criteria considered in this work.

Oracle Witness Function for HSIC. Let us derive the form of the oracle witness function for HSIC. Note that:

$$\begin{aligned}
&\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\mathbb{E}_{P_{XY}}\left[g(X,Y)\right] - \mathbb{E}_{P_X\times P_Y}\left[g(X',Y')\right]\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\mathbb{E}_{P_{XY}}\left[\langle g, \varphi(X)\otimes\psi(Y)\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] - \mathbb{E}_{P_X\times P_Y}\left[\langle g, \varphi(X')\otimes\psi(Y')\rangle_{\mathcal{G}\otimes\mathcal{H}}\right]\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\langle g, \mathbb{E}_{P_{XY}}\left[\varphi(X)\otimes\psi(Y)\right]\rangle_{\mathcal{G}\otimes\mathcal{H}} - \langle g, \mathbb{E}_{P_X\times P_Y}\left[\varphi(X')\otimes\psi(Y')\right]\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\langle g, \mu_{XY}\rangle_{\mathcal{G}} - \langle g, \mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \langle g, \mu_{XY} - \mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}},
\end{aligned}$$

from which it is easy to derive the oracle witness function for HSIC.

Remark D.1. Note that in (13) the witness function is defined as an operator: $\hat{g}_t : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. To clarify, for any $z=(x,y)\in\mathcal{X}\times\mathcal{Y}$, we have

$$(\hat{\mu}_{XY} - \hat{\mu}_X\otimes\hat{\mu}_Y)(z) = \frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} k(X_i,x)l(Y_i,y) - \left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} k(X_i,x)\right)\cdot\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} l(Y_i,y)\right),$$

and the denominator in (13) can be expressed in terms of kernel matrices $K, L\in\mathbb{R}^{2(t-1)\times 2(t-1)}$ with entries $K_{ij} = k(X_i,X_j)$, $L_{ij}=l(Y_i,Y_j)$, $i,j\in\{1,\dots,2(t-1)\}$, as:

$$\|\hat{\mu}_{XY} - \hat{\mu}_X\otimes\hat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}} = \frac{1}{2(t-1)}\sqrt{\mathrm{tr}(KHLH)},$$

where $H = \mathbf{I}_{2(t-1)} - (1/(2(t-1)))\mathbf{1}\mathbf{1}^\top$ is the centering projection matrix.

Remark D.2. While the empirical witness functions for COCO/KCC (21) are defined as operators, we use those as functions in the definition of the corresponding payoff function. To clarify, for any $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, we have

$$\begin{aligned}
\hat{g}_t(x) &= \sum_{i=1}^{2(t-1)}\alpha_i\left(k(X_i,x) - \frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)} k(X_j,x)\right), \\
\hat{h}_t(y) &= \sum_{i=1}^{2(t-1)}\beta_i\left(l(Y_i,y) - \frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)} l(Y_j,y)\right).
\end{aligned}$$

Minibatched Payoff Function for HSIC. The minibatched payoff function at round t has the following form:

$$f_t(Z_{b(t-1)+1},\dots,Z_{bt}) = \frac{1}{b}\sum_{i=1}^{b}\hat{g}_t(X_{b(t-1)+i},Y_{b(t-1)+i}) - \frac{1}{b(b-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{b}\hat{g}_t(X_{b(t-1)+i},Y_{b(t-1)+j}).$$

Note that:

$$\begin{aligned}
f_t(Z_{b(t-1)+1},\dots,Z_{bt}) = &\frac{1}{b}\sum_{i=1}^{b}\langle\hat{g}_t, \varphi(X_{b(t-1)+i})\otimes\psi(Y_{b(t-1)+i})\rangle_{\mathcal{G}\otimes\mathcal{H}} \\
&- \frac{1}{b(b-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{b}\langle\hat{g}_t, \varphi(X_{b(t-1)+i})\otimes\psi(Y_{b(t-1)+j})\rangle_{\mathcal{G}\otimes\mathcal{H}}
\end{aligned}$$

$$= \left\langle \hat{g}_t, \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{b} \left(\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\right) \otimes \left(\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\right) \right\rangle_{\mathcal{G}\otimes\mathcal{H}}.$$

Let $\mathcal{F}'_{t-1} = \sigma(\{(X_i, Y_i)\}_{i \leq b(t-1)})$. We have that:

$$\mathbb{E}\left[f_t(Z_{b(t-1)+1}, \ldots, Z_{bt}) \mid \mathcal{F}'_{t-1}\right] = \langle \hat{g}_t, \mu_{XY} - \mu_X \otimes \mu_Y \rangle_{\mathcal{G}\otimes\mathcal{H}},$$

and in particular, $\mathbb{E}_{H_0}\left[f_t(Z_{b(t-1)+1}, \ldots, Z_{bt}) \mid \mathcal{F}'_{t-1}\right] = 0$ if the null H_0 in (1a) is true. It suffices to show that the payoff is bounded. Since $\|\hat{g}_t\|_{\mathcal{G}\otimes\mathcal{H}} = 1$, we can easily deduce that:

$$\begin{aligned}
\left|f_t(Z_{b(t-1)+1}, \ldots, Z_{bt})\right| &\leq \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{b} \left\|\left(\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\right) \otimes \left(\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\right)\right\|_{\mathcal{G}\otimes\mathcal{H}} \\
&= \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{b} \left\|\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\right\|_{\mathcal{G}} \left\|\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\right\|_{\mathcal{H}} \\
&= \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{b} \sqrt{2(1 - k(X_{b(t-1)+i}, X_{b(t-1)+j}))} \sqrt{2(1 - l(Y_{b(t-1)+i}, Y_{b(t-1)+j}))} \\
&\leq 1.
\end{aligned}$$

Hence, we conclude that the wealth process constructed using a minibatched version of the payoff function is also a nonnegative martingale.

Example 3. *For $t \geq 1$, consider*

$$(X_t, Y_t) = \left(\frac{V_t + 1 - 1/t}{2}, \frac{V'_t + 1 - 1/t}{2}\right),$$

where $V_t, V'_t \overset{iid}{\sim} \mathrm{Ber}(1/2)$. Note that $\mathcal{X} = \mathcal{Y} \subseteq [0,1]$, which means that a pair of linear kernels, $k(x, x') = xx'$ and $l(y, y') = yy'$ are nonnegative and bounded by one on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Note that for a linear kernel,

$$\hat{g}_t(x, y) = \hat{g}_t \cdot x \cdot y.$$

Hence,

$$\begin{aligned}
f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) &= \frac{\hat{g}_t}{2} (X_{2t} - X_{2t-1}) (Y_{2t} - Y_{2t-1}) \\
&= \frac{\hat{g}_t}{8} \left(V_{2t} - V_{2t-1} + \frac{1}{2t(2t-1)}\right) \left(V'_{2t} - V'_{2t-1} + \frac{1}{2t(2t-1)}\right).
\end{aligned}$$

In particular, $\mathbb{E}\left[f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) \mid \mathcal{F}_{t-1}\right] \neq 0$, implying that the wealth process $(\mathcal{K}_t)_{t \geq 0}$ is no longer a nonnegative martingale.

Witness Functions for COCO. Let Φ and Ψ be a pair of matrices whose columns represent embeddings of $X_1, \ldots, X_{2(t-1)}$ and $Y_1, \ldots, Y_{2(t-1)}$, that is, $\varphi(X_i) = k(X_i, \cdot)$ and $\psi(Y_i) = l(Y_i, \cdot)$ for $i = 1, \ldots, 2(t-1)$. Recall that

$$\begin{aligned}
\hat{g} &= \sum_{i=1}^{2(t-1)} \alpha_i \left(\varphi(X_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \varphi(X_j)\right) = \Phi H \alpha, \\
\hat{h} &= \sum_{i=1}^{2(t-1)} \beta_i \left(\psi(Y_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \psi(Y_j)\right) = \Psi H \beta,
\end{aligned}$$

where $H = \mathbf{I}_{2(t-1)} - \frac{1}{2(t-1)} 1 1^\top$ is the centering projection matrix. We have

$$\begin{aligned}
\langle g, \hat{C}_{XY} h \rangle_{\mathcal{G}} &= \frac{1}{2(t-1)} (\alpha^\top H \Phi^\top)(\Phi H \Psi^\top)(\Psi H \beta) = \frac{1}{2(t-1)} \alpha^\top H K H L H \beta = \frac{1}{2(t-1)} \alpha^\top \tilde{K} \tilde{L} \beta, \\
\|g\|_{\mathcal{G}}^2 &= \alpha^\top \tilde{K} \alpha, \\
\|h\|_{\mathcal{H}}^2 &= \beta^\top \tilde{L} \beta,
\end{aligned}$$

where $\tilde{K} := HKH$ and $\tilde{L} := HLH$ are centered kernel matrices. Hence, the maximization problem in (20) can be expressed as:

$$\begin{aligned}
\max_{\alpha, \beta} \quad & \frac{1}{2(t-1)} \alpha^\top \tilde{K} \tilde{L} \beta \\
\text{subject to} \quad & \alpha^\top \tilde{K} \alpha = 1, \quad \beta^\top \tilde{L} \beta = 1.
\end{aligned} \tag{47}$$

After introducing Lagrange multipliers, it can then be shown that α and β, which solve (47), exactly correspond to the generalized eigenvalue problem (22).

Witness Functions for KCC. Introduce empirical covariance operators:

$$\begin{aligned}
\hat{C}_X &= \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \otimes \varphi(X_i) - \left(\frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \right) \otimes \left(\frac{1}{2(t-1)} \sum_{i=1}^{n} \varphi(X_i) \right) = \frac{1}{2(t-1)} \Phi H \Phi^\top, \\
\hat{C}_Y &= \frac{1}{n} \sum_{i=1}^{2(t-1)} \psi(Y_i) \otimes \psi(Y_i) - \left(\frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i) \right) \otimes \left(\frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i) \right) = \frac{1}{2(t-1)} \Psi H \Psi^\top.
\end{aligned}$$

Then the empirical variance terms can be expressed as:

$$\begin{aligned}
\hat{\mathbb{V}}\left[g(X)\right] &= \langle g, \hat{C}_X g \rangle_{\mathcal{G}} = \frac{1}{2(t-1)} (\alpha^\top H \Phi^\top)(\Phi H \Phi^\top)(\Phi H \alpha) = \frac{1}{2(t-1)} \alpha^\top \tilde{K}^2 \alpha, \\
\hat{\mathbb{V}}\left[h(Y)\right] &= \langle h, \hat{C}_Y h \rangle_{\mathcal{H}} = \frac{1}{2(t-1)} (\beta^\top H \Psi^\top)(\Psi H \Psi^\top)(\Psi H \beta) = \frac{1}{2(t-1)} \beta^\top \tilde{L}^2 \beta.
\end{aligned}$$

Thus, an empirical estimator of the kernel canonical correlation (23) can be obtained by solving:

$$\begin{aligned}
\max_{\alpha, \beta} \quad & \frac{1}{2(t-1)} \alpha^\top \tilde{K} \tilde{L} \beta \\
\text{subject to} \quad & \frac{1}{2(t-1)} \alpha^\top \tilde{K}^2 \alpha + \kappa_1 \alpha^\top \tilde{K} \alpha = 1, \\
& \frac{1}{2(t-1)} \beta^\top \tilde{L}^2 \beta + \kappa_2 \beta^\top \tilde{L} \beta = 1.
\end{aligned}$$

After introducing Lagrange multipliers, it can then be shown that α and β, which solve (23), correspond to the generalized eigenvalue problem:

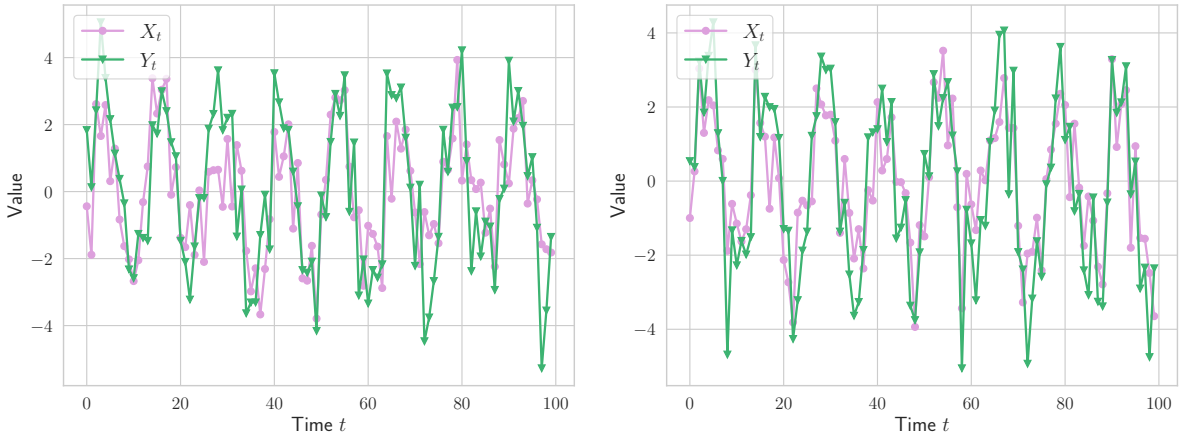
$$\begin{pmatrix} 0 & \frac{1}{2(t-1)} \tilde{K} \tilde{L} \\ \frac{1}{2(t-1)} \tilde{L} \tilde{K} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \gamma \begin{pmatrix} \kappa_1 \tilde{K} + \frac{1}{2(t-1)} \tilde{K}^2 & 0 \\ 0 & \kappa_2 \tilde{L} + \frac{1}{2(t-1)} \tilde{L}^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

E. Additional Simulations

This section contains: (a) additional experiments on synthetic dataset and (b) data visualizations of the datasets used in this paper.

E.1. Test of Instantaneous Dependence

In Figure 10, we demonstrate it is hard to visually tell the difference between independence and dependence under distribution drift setting (2). See Example 1 for details.



(a) Independent noise.

(b) Dependent noise.

Figure 10. Sample of independent (subplot (a)) and dependent ($\rho = 0.5$, subplot(b)) data according to (3). The purpose of visualizing raw data is to demonstrate that dependence is hard to detect visually, and dependence refers to more than temporal correlation which may be present due to cyclical trends.

E.2. Distribution Drift

In this section, we consider the linear Gaussian model with an underlying distribution drift:

$$Y_t = X_t \beta_t + \varepsilon_t, \quad X_t, \varepsilon_t \sim \mathcal{N}(0,1), \quad t \geq 1,$$

that is, in contrast to the Gaussian linear model (Section 3), β_t changes over time. We gradually increase it from $\beta_t = 0$ to $\beta_t = 0.1$ in increments of 0.02, that is:

$$\underbrace{\beta_0, \ \ldots, \ \beta_{b-1}}_{=0}, \ \underbrace{\beta_b, \ \ldots, \ \beta_{2b-1}}_{=0.02}, \ \ldots, \ \underbrace{\beta_{5b-1}, \ \ldots}_{=0.1}$$

and, starting with β_{5b}, we keep it equal to 0.1. We consider $b \in \{100, 200, 400\}$ as possible block sizes. Note that there is a transition from independence (first b datapoints in a stream) to dependence. In Figure 11, we show that our test performs well under the distribution drift setting and consistently detects dependence.

E.3. Symmetry-based Payoff Functions

In this section, we complement the comparison presented in Section 4 between the rank- and composition-based betting strategies (since those require minimal tuning) used with ONS or aGRAPA criteria for selecting betting fractions. We also increase the monitoring horizon to 20000 datapoints. In Figure 12(a), we consider the Gaussian linear model, but in contrast to the setting considered in Section 4, we focus on harder testing settings by considering $\beta \in [0, 0.3]$. In Figure 12(b), we compare composition- and rank-based approaches when data are sampled from the spherical model. In both cases, composition and rank-based approaches are similar; none of the payoffs uniformly dominates the other. We also observe that selecting betting fractions via aGRAPA criterion tends to result in a bit more powerful testing procedure.

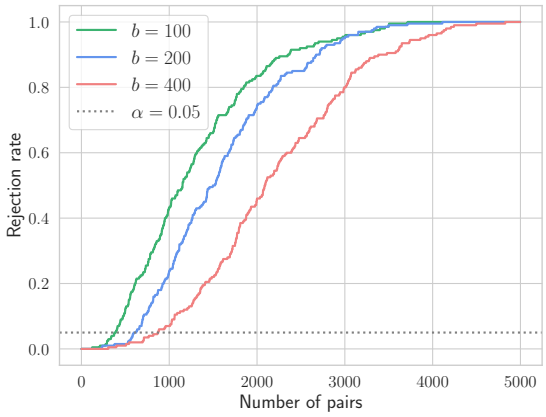
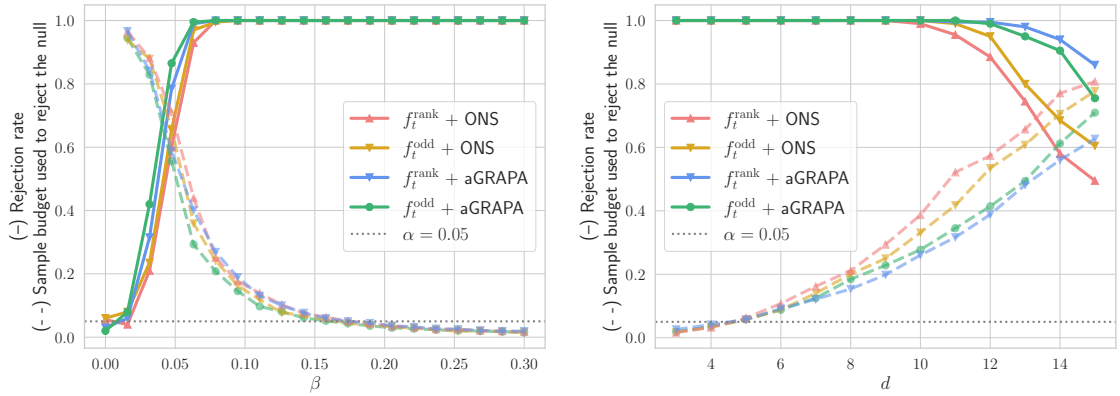


Figure 11. Rejection rate of sequential independence test under distribution drift setting. Focusing on the non-i.i.d. time-varying setting, we confirm that our test has high power under the alternative.



(a) (b)

Figure 12. (a) Comparison of symmetry-based betting strategies under the Gaussian model. The betting strategy based on composition with an odd function performs only slightly better than the rank-based strategy. (b) SKIT with composition- and rank-based betting strategies under the spherical model. None of the betting strategies uniformly dominates the other. aGRAPA criterion for selecting betting fractions tends to result in a bit more powerful testing procedure.

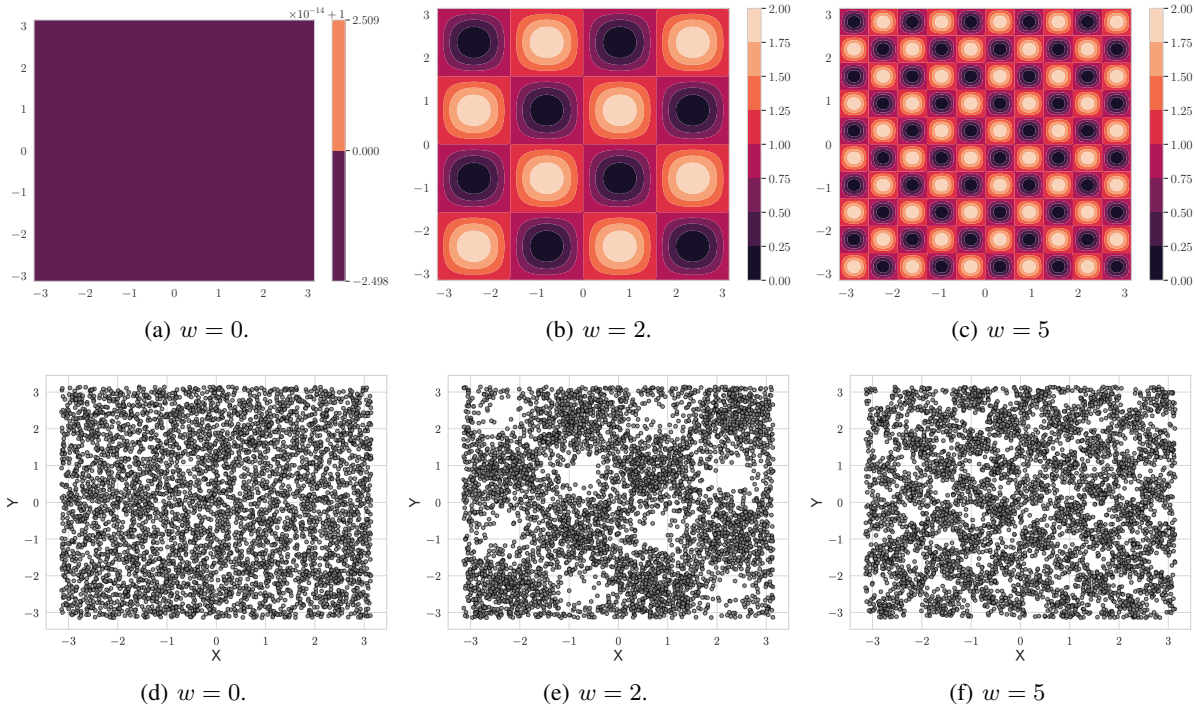
E.4. Hard-to-detect Dependence

Hard-to-detect dependence. Consider the joint density $p(x, y)$ of the form:

$$\frac{1}{4\pi^2}\left(1 + \sin(wx)\sin(wy)\right) \cdot \mathbb{1}\left\{(x, y) \in [-\pi, \pi]^2\right\}. \tag{48}$$

With the null case corresponding to $w = 0$, the testing problem becomes harder with growing w. In Figure 13, we illustrate the densities and a data sample for the hard-to-detect setting (48).

We use $\lambda_X = \lambda_Y = 3/(4\pi^2)$ as RBF kernel hyperparameters. For visualization purposes, we stop monitoring after observing 20000 datapoints from P_{XY}, and if a SKIT does not reject H_0 by that time, we assume that the null is retained. The results are aggregated over 200 runs for each value of w. In Figure 14, where the null case corresponds to $w = 0$, we confirm that SKITs have time-uniform type I error control. The average rejection rate starts to drop for $w \geq 3$, meaning that observing 20000 points from P_{XY} does not suffice to detect dependence.



(a) $w = 0$. (b) $w = 2$. (c) $w = 5$

(d) $w = 0$. (e) $w = 2$. (f) $w = 5$

Figure 13. Visualization of the densities (top) and a dataset of size 5000 (bottom) sampled from the corresponding distribution.

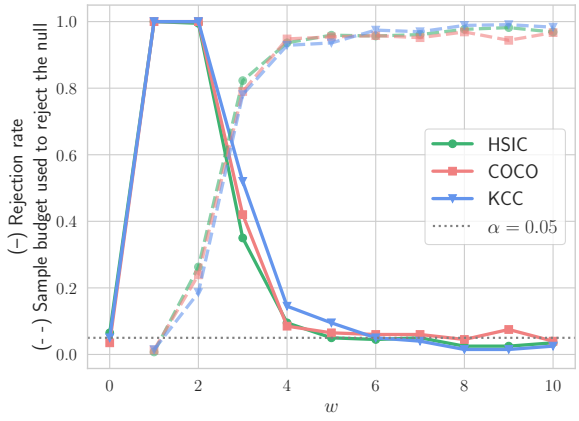


Figure 14. Rejection rate (solid) and fraction of samples used before the null hypothesis was rejected (dashed) for hard-to-detect dependence model. By inspecting the rejection rate for $w = 0$ (independence holds), we confirm that the type I error is controlled. Further, SKIT is adaptive to the complexity of a problem (larger w corresponds to a harder setting).

E.5. Additional Results for Real Data

In Figure 15, we illustrate that the average daily temperature in selected cities share similar seasonal patterns. We repeat the same experiment as in Section 4, but for four cities in South Africa: Cape Town (CT), Port Elizabeth (PE), Durban (DRN), and Bloemfontein (BFN). In Figures 15(d) and 15(e), we illustrate the resulting wealth processes for each pair of cities and for each region. Finally, we illustrate the pairs of cities for which the null has been rejected in Figure 15(c).

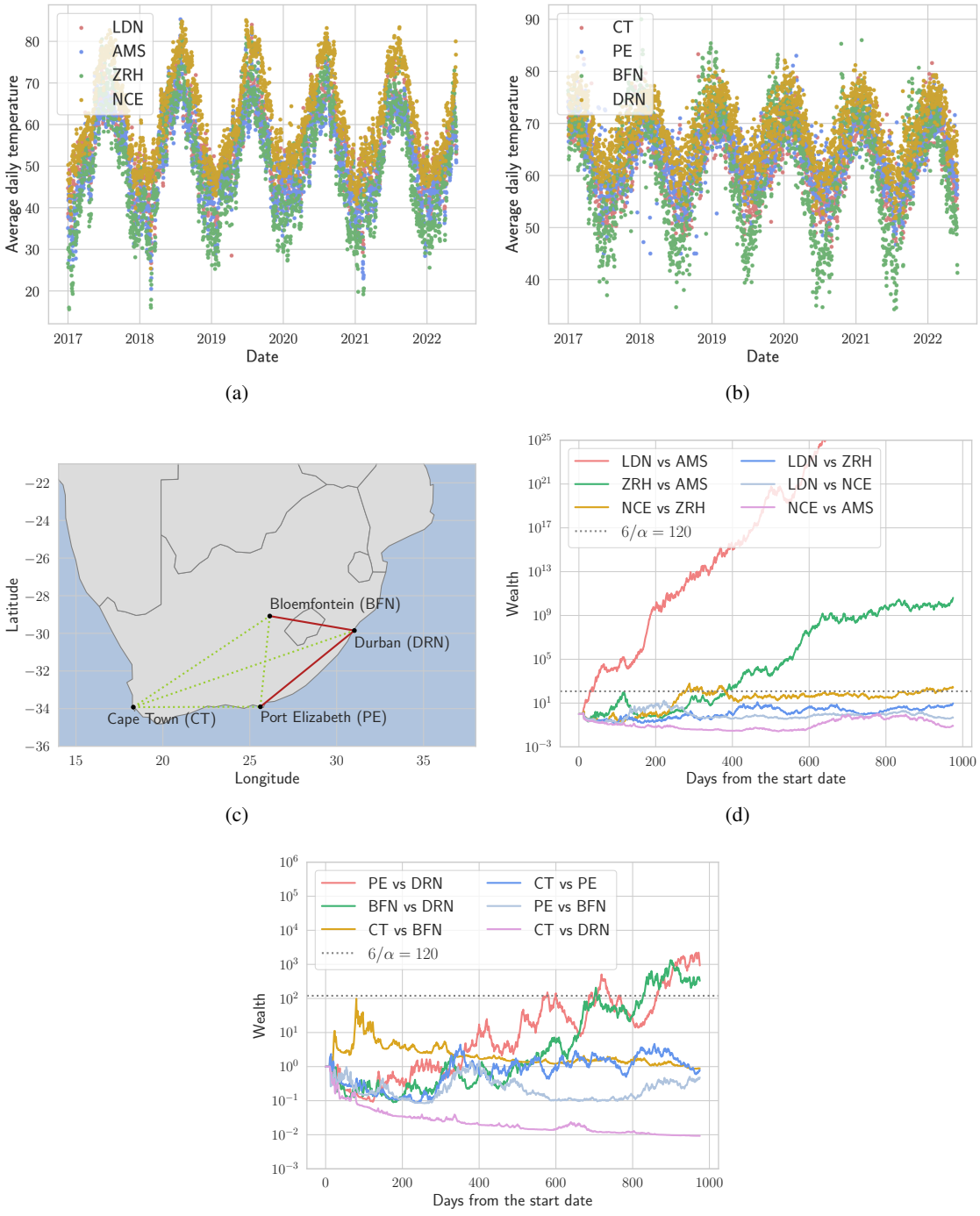


Figure 15. Temperatures for selected cities in Europe (subplot (a)) and South Africa (subplot (b)) share similar seasonal patterns. Map (subplot (c)) where solid red lines connect those cities for which the null is rejected. SKIT supports our conjecture about dependent temperature fluctuations for geographically close cities. For completeness, we also plot wealth processes for SKIT used on weather data for Europe (subplot (d)) and South Africa (subplot (e)).

E.6. Experiment with MNIST data

In this section, we analyze the performance of SKIT on high-dimensional real data. This experiment is based on MNIST dataset (LeCun et al., 1998) where pairs of digits are observed at each step; under the null one sees digits (a, b) where a and b are uniformly randomly chosen, but under the alternative one sees (a, a'), i.e., two different images of the same digit. To estimate kernel hyperparameters, we deploy the median heuristic using 20 pairs of images.

We illustrate the results in Figure 16. Under the null, our test does not reject more often than the required 5%, but its power increases with sample size under the alternative, reaching power one after processing ≈ 500 pairs of digits (points from P_{XY}) on average.

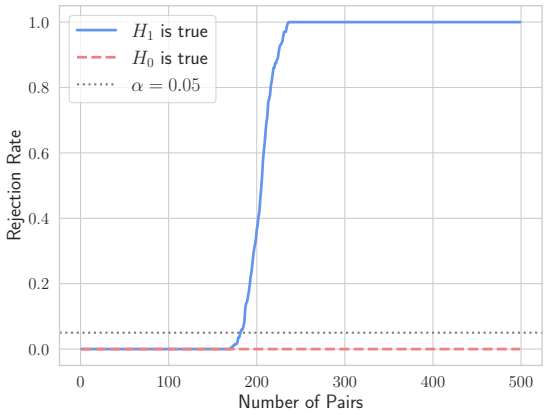


Figure 16. Rejection rate for SKIT on MNIST data. Under the null (red dashed line), our test does not reject more often than the required 5%, but its power increases with sample size under the alternative (blue solid line). Each pair corresponds to two points from P_{XY}, and hence, SKIT reaches power one after processing ≈ 500 pairs of images on average.

F. Scaling Sequential Testing Procedures

Updating the wealth process at each round requires evaluating the payoff function at a new pair of observations (and hence computing the witness function corresponding to a chosen dependence criterion). In this section, we provide details about the ways of reducing the computational complexity of this step, which are necessary to scale the proposed sequential testing frameworks to moderately large sample sizes. Note that the proposed implementation of COCO allows updating kernel hyperparameters on the fly. In contrast, linear-time updates for HSIC require fixing kernel hyperparameters in advance.

F.1. Incomplete/Pivoted Cholesky Decomposition for COCO and KCC

Suppose that we want to evaluate COCO payoff function on the next pair of points $(X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})$. In order to do so, we need to compute $g_{1,t}$ and $g_{2,t}$, that is solve the generalized eigenvalue problem. Note that solving generalized eigenvalue problem at each iteration could be computationally prohibitive. One simple way is to use a random subsample of datapoints when performing witness function estimation, e.g., once the sample size n exceeds n_s, e.g., $n_s = 25$, we randomly subsample (without replacement) a sample of size n_s to estimate witness functions. Alternatively, a common approach is to reduce computational burden through incomplete Cholesky decomposition. The idea is to use the fact that kernel matrices tend to demonstrate rapid spectrum decay, and thus low-rank approximations can be used to scale the procedures. Suppose that $K \approx G_1 G_1^T$ and $L \approx G_2 G_2^T$ where G_i's are lower triangular matrices of size $n \times M$ (M depends on the preset approximation error level). After computing Cholesky decomposition, we center both matrices via left multiplication by H and compute SVDs of HG_1 and HG_2, that is, $HG_1 = U_1 \Lambda_1 V_1^\top$ and $HG_2 = U_2 \Lambda_2 V_2^\top$. We have:

$$\tilde{K} \approx U_1 \Lambda_1^2 U_1^\top, \quad \tilde{L} \approx U_2 \Lambda_2^2 U_2^\top.$$

Our goal is to find the largest eigenvalue/eigenvector pair for $Ax = \gamma Bx$ for a PD matrix B. Since:

$$Ax = \gamma Bx \iff B^{-1/2} A B^{-1/2} (B^{1/2} x) = \gamma (B^{1/2} x),$$

it suffices to leading eigenvalue/eigenvector pair for:

$$B^{-1/2} A B^{-1/2} y = \gamma y.$$

Then $x = B^{-1/2} y$ is a generalized eigenvector for the initial problem.

COCO. For COCO, we have:

$$B = \begin{pmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{pmatrix} \approx \begin{pmatrix} U_1 \Lambda_1^2 U_1^\top & 0 \\ 0 & U_2 \Lambda_2^2 U_2^\top \end{pmatrix} = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1^2 & 0 \\ 0 & \Lambda_2^2 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top$$

$$\implies B^{-1/2} \approx \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & \Lambda_2^{-1} \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top =: \mathcal{U} \Lambda^{-1} \mathcal{U}^\top.$$

We also have:

$$A \approx \begin{pmatrix} 0 & \frac{1}{n} U_1 \Lambda_1^2 U_1^\top U_2 \Lambda_2^2 U_2^\top \\ \frac{1}{n} U_2 \Lambda_2^2 U_2^\top U_1 \Lambda_1^2 U_1^\top & 0 \end{pmatrix}$$

$$= \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{n} \Lambda_1^2 U_1^\top U_2 \Lambda_2^2 \\ \frac{1}{n} \Lambda_2^2 U_2^\top U_1 \Lambda_1^2 & 0 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top.$$

Thus we have:

$$B^{-1/2} A B^{-1/2} \approx \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{n} \Lambda_1 U_1^\top U_2 \Lambda_2 \\ \frac{1}{n} \Lambda_2 U_2^\top U_1 \Lambda_1 & 0 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top.$$

Hence, we only need to compute the leading eigenvector (say, z^*) for:

$$\begin{pmatrix} 0 & \frac{1}{n} \Lambda_1 U_1^\top U_2 \Lambda_2 \\ \frac{1}{n} \Lambda_2 U_2^\top U_1 \Lambda_1 & 0 \end{pmatrix} \in \mathbb{R}^{(M_1+M_2)\times(M_1+M_2)}.$$

It implies that the leading eigenvector for $B^{-1/2} A B^{-1/2}$ is then $\mathcal{U} z^*$, and the solution for the generalized eigenvalue problem is given by:

$$\mathcal{U} \Lambda^{-1} z^* = \begin{pmatrix} U_1 \Lambda_1^{-1} z_1^* \\ U_2 \Lambda_2^{-1} z_2^* \end{pmatrix} =: \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

Next, we need to normalize this vector of coefficients appropriately, i.e., we need to guarantee that $\left\|\tilde{K}^{1/2}\alpha\right\|_2 = 1$ and $\left\|\tilde{L}^{1/2}\beta\right\|_2 = 1$, and thus re-normalizing naively is quadratic in n. Instead, note that in order to compute incomplete Cholesky decomposition, we choose a tolerance parameter δ so that: $\left\|PKP^\top - G_1G_1^\top\right\|_* = \left\|K - G_1G_1^\top\right\|_* \leq \delta$ (nuclear norm). Let $\Delta = K - G_1G_1^\top$. We know that:

$$\alpha^\top \tilde{K}\alpha = \alpha^\top HKH\alpha = \alpha^\top H(\Delta + G_1G_1^\top)H\alpha = \alpha^\top H\Delta H\alpha + \alpha^\top HG_1G_1^\top H\alpha$$

First, note that $\alpha^\top H\Delta H\alpha \leq \delta\|H\alpha\|_2^2$. Next,

$$G_1^\top H = V_1\Lambda_1U_1^\top.$$

Given an initial vector of parameters α_0 and β_0, vectors of coefficients can be normalized in linear time using

$$\alpha = \frac{\alpha_0}{\sqrt{\left\|G_1^\top H\alpha_0\right\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\Lambda_1^{-1}z_1^*}{\sqrt{\left\|V_1z_1^*\right\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\Lambda_1^{-1}z_1^*}{\sqrt{\left\|z_1^*\right\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}},$$
$$\beta = \frac{\beta_0}{\sqrt{\left\|G_2^\top H\beta_0\right\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\Lambda_2^{-1}z_2^*}{\sqrt{\left\|V_2z_2^*\right\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\Lambda_2^{-1}z_2^*}{\sqrt{\left\|z_2^*\right\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}}.$$

For small δ, we essentially normalize by $\alpha_0^\top\tilde{K}\alpha_0$ and $\beta_0^\top\tilde{L}\beta_0$ as expected. It also makes sense to use $\delta = n\cdot\delta_0$. Still, re-estimating the witness functions after processing $2t$, $t \geq 1$ points is computationally intensive. In contrast to HSIC, for which there are no clear benefits of skipping certain estimation steps, for COCO we estimate the witness functions after processing $2t^2$, $t \geq 1$ points.

KCC. For KCC, we have:

$$\begin{aligned}
B &= \begin{pmatrix} \kappa_1\tilde{K} + \frac{1}{n}\tilde{K}^2 & 0 \\ 0 & \kappa_2\tilde{L} + \frac{1}{n}\tilde{L}^2 \end{pmatrix} \\
&\approx \begin{pmatrix} \kappa_1U_1\Lambda_1^2U_1^\top + \frac{1}{n}U_1\Lambda_1^4U_1^\top & 0 \\ 0 & \kappa_2U_2\Lambda_2^2U_2^\top + \frac{1}{n}U_2\Lambda_2^4U_2^\top \end{pmatrix} \\
&= \begin{pmatrix} U_1\Lambda_1^2\left(\kappa_1\mathbf{I}_n + \Lambda_1^2\frac{1}{n}\right)U_1^\top & 0 \\ 0 & U_2\Lambda_2^2\left(\kappa_2\mathbf{I}_n + \Lambda_2^2\frac{1}{n}\right)U_2^\top \end{pmatrix} \\
&= \mathcal{U}\begin{pmatrix} \Lambda_1^2\left(\kappa_1\mathbf{I}_n + \Lambda_1^2\frac{1}{n}\right) & 0 \\ 0 & \Lambda_2^2\left(\kappa_2\mathbf{I}_n + \Lambda_2^2\frac{1}{n}\right) \end{pmatrix}\mathcal{U}^\top,
\end{aligned}$$

which implies that:

$$B^{-1/2} \approx \mathcal{U}\begin{pmatrix} \left(\kappa_1\mathbf{I}_n + \Lambda_1^2\frac{1}{n}\right)^{-1/2}\Lambda_1^{-1} & 0 \\ 0 & \left(\kappa_2\mathbf{I}_n + \Lambda_2^2\frac{1}{n}\right)^{-1/2}\Lambda_2^{-1} \end{pmatrix}\mathcal{U}^\top.$$

Recall that:

$$A \approx \mathcal{U}\begin{pmatrix} 0 & \frac{1}{n}\Lambda_1^2U_1^\top U_2\Lambda_2^2 \\ \frac{1}{n}\Lambda_2^2U_2^\top U_1\Lambda_1^2 & 0 \end{pmatrix}\mathcal{U}^\top.$$

Thus,

$$B^{-1/2}AB^{-1/2} \approx \mathcal{U}\begin{pmatrix} 0 & M_* \\ M_*^\top & 0 \end{pmatrix}\mathcal{U}^\top,$$
$$\text{where}\quad M_* = \frac{1}{n}\left(\kappa_1\mathbf{I}_n + \Lambda_1^2\frac{1}{n}\right)^{-1/2}\Lambda_1U_1^\top U_2\Lambda_2\left(\kappa_2\mathbf{I}_n + \Lambda_2^2\frac{1}{n}\right)^{-1/2}.$$

Equivalently,

$$M_* = \frac{1}{n}\rho_{\kappa_1}(\Lambda_1)\Lambda_1U_1^\top U_2\Lambda_2\rho_{\kappa_2}(\Lambda_2),\quad \text{where}\quad \rho_\kappa(x) = \frac{1}{\sqrt{x^2/n + \kappa}}.$$

Hence, we only need to compute the leading eigenvector (say, z^*) for:

$$\begin{pmatrix} 0 & M_* \\ M_*^\top & 0 \end{pmatrix}$$

It implies that the leading eigenvector for $B^{-1/2}AB^{-1/2}$ is then $\mathcal{U}z^*$. For the initial generalized eigenvalue problem, an approximate solution (due to using low-rank approximations of kernel matrices) is given by:

$$B^{-1/2}\mathcal{U}z^* = \begin{pmatrix} U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^* \\ U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^* \end{pmatrix} = \begin{pmatrix} U_1\Lambda_1^{-1}\rho_{\kappa_1}(\Lambda_1)z_1^* \\ U_2\Lambda_2^{-1}\rho_{\kappa_2}(\Lambda_2)z_2^* \end{pmatrix} =: \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

Next, we need to normalize this vector of coefficients appropriately, i.e., we need to guarantee that $\left\|\tilde{K}^{1/2}\alpha\right\|_2 = 1$ and $\left\|\tilde{L}^{1/2}\beta\right\|_2 = 1$, and thus re-normalizing naively is quadratic in n. Instead, note that in order to compute incomplete Cholesky decomposition, we choose a tolerance parameter δ so that: $\left\|PKP^\top - G_1G_1^\top\right\|_* = \left\|K - G_1G_1^\top\right\|_* \leq \delta$ (nuclear norm). Let $\Delta = K - G_1G_1^\top$. We know that:

$$\alpha^\top\tilde{K}\alpha = \alpha^\top HKH\alpha = \alpha^\top H(\Delta + G_1G_1^\top)H\alpha = \alpha^\top H\Delta H\alpha + \alpha^\top HG_1G_1^\top H\alpha$$

First, note that $\alpha^\top H\Delta H\alpha \leq \delta\|H\alpha\|_2^2$. Next,

$$G_1^\top H = V_1\Lambda_1U_1^\top.$$

Given an initial vector of parameters α_0 and β_0, vectors of coefficients can be normalized in linear time using

$$\alpha = \frac{\alpha_0}{\sqrt{\left\|G_1^\top H\alpha_0\right\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^*}{\sqrt{\|V_1\rho_{\kappa_1}(\Lambda_1)z_1^*\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^*}{\sqrt{\|\rho_{\kappa_1}(\Lambda_1)z_1^*\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}},$$
$$\beta = \frac{\beta_0}{\sqrt{\left\|G_2^\top H\beta_0\right\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^*}{\sqrt{\|V_2\rho_{\kappa_2}(\Lambda_2)z_2^*\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^*}{\sqrt{\|\rho_{\kappa_2}(\Lambda_2)z_2^*\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}}.$$

F.2. Linear-time Updates of the HSIC Payoff Function

Suppose that we want to evaluate HSIC payoff function on the next pair of points $(X_{2t+1}, Y_{2t+1}), (X_{2t+2}, Y_{2t+2})$. In order to do so, we need to compute: $\hat{g}_t(X_{2t+2}, Y_{2t+2})$. It is clear that the computational of evaluating $\hat{\mu}_{XY}(x, y)$ and $(\hat{\mu}_X \otimes \hat{\mu}_Y)(x, y)$ on a given pair (x, y) is linear in t. However, we also need to compute the normalization constant:

$$\|\hat{\mu}_{XY} - \hat{\mu}_X \otimes \hat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}}. \tag{49}$$

Recall that:

$$\left\|\hat{\mu}_{XY}^{(t)} - \hat{\mu}_X^{(t)} \otimes \hat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 = \frac{1}{(2t)^2}\mathrm{tr}\left(K^{(t)}H^{(t)}L^{(t)}H^{(t)}\right),$$

where $K^{(t)}$ and $L^{(t)}$ are kernel matrices corresponding to the first $2t$ pairs, $H^{(t)} := \mathbf{I}_{2t} - \frac{1}{2t}\mathbf{1}_{2t}\mathbf{1}_{2t}^\top$. Instead of computing the normalization constant naively, we next establish a more efficient way of computing (49) in time linear in t by caching certain values. Introduce:

$$\Delta_1^{(t)} = \sum_{i,j=1}^{2t} K_{ij}L_{ij} = \mathrm{tr}\left(K^{(t)}L^{(t)}\right),$$
$$\Delta_2^{(t)} = \sum_{i,j}^{2t} K_{ij} = \mathbf{1}_{2t}^\top K^{(t)}\mathbf{1}_{2t},$$
$$\Delta_3^{(t)} = \sum_{i,j}^{2t} L_{ij} = \mathbf{1}_{2t}^\top L^{(t)}\mathbf{1}_{2t},$$
$$\Delta_4^{(t)} = \sum_{i=1}^{2t}\sum_{j,q=1}^{2t} K_{ij}L_{iq} = \mathbf{1}_{2t}^\top K^{(t)}L^{(t)}\mathbf{1}_{2t}.$$

We have:

$$\left\|\hat{\mu}_{XY}^{(t+1)} - \hat{\mu}_X^{(t+1)} \otimes \hat{\mu}_Y^{(t+1)}\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 = \frac{1}{(2t+2)^2}\Delta_1^{(t+1)} + \frac{1}{(2t+2)^4}\Delta_2^{(t+1)}\cdot\Delta_3^{(t)} - \frac{2}{(2t+2)^3}\Delta_4^{(t+1)}.$$

Next, we show how to speed up computations via caching certain intermediate values. Kernel matrices have the following structure:

$$K^{(t+1)} = \begin{pmatrix} K^{(t)} & K_{\cdot,2t+1} & K_{\cdot,2t+2} \\ K_{\cdot,2t+1}^\top & K_{2t+1,2t+1} & K_{2t+1,2t+2} \\ K_{\cdot,2t+2}^\top & K_{2t+2,2t+1} & K_{2t+2,2t+2} \end{pmatrix}, \quad L^{(t+1)} = \begin{pmatrix} L^{(t)} & L_{\cdot,2t+1} & L_{\cdot,2t+2} \\ L_{\cdot,2t+1}^\top & L_{2t+1,2t+1} & L_{2t+1,2t+2} \\ L_{\cdot,2t+2}^\top & L_{2t+2,2t+1} & L_{2t+2,2t+2} \end{pmatrix},$$

where $K_{\cdot,2t+1}, K_{\cdot,2t+2}, L_{\cdot,2t+1}, L_{\cdot,2t+2} \in \mathbb{R}^{2t}$ contain kernel function evaluations:

$$K_{\cdot,m} = \begin{pmatrix} k(X_1, X_m) \\ \vdots \\ k(X_{2t}, X_m) \end{pmatrix}, \quad L_{\cdot,m} = \begin{pmatrix} l(Y_1, Y_m) \\ \vdots \\ l(Y_{2t}, Y_m) \end{pmatrix}, \quad m \in \{2t+1, 2t+2\}.$$

First, it is easy to derive that:

$$\begin{aligned} \text{tr}\left(K^{(t+1)}L^{(t+1)}\right) = \text{tr}\left(K^{(t)}L^{(t)}\right) &+ 2(L_{\cdot,2t+1}^\top K_{\cdot,2t+1}) + 2(L_{\cdot,2t+2}^\top K_{\cdot,2t+2}) + \\ &+ K_{2t+1,2t+1}L_{2t+1,2t+1} + K_{2t+2,2t+2}L_{2t+2,2t+2} \\ &+ K_{2t+1,2t+2}L_{2t+2,2t+1} + K_{2t+2,2t+1}L_{2t+1,2t+2}. \end{aligned}$$

Thus, if the value $\text{tr}\left(K^{(t)}L^{(t)}\right)$ is cached, then $\text{tr}\left(K^{(t+1)}L^{(t+1)}\right)$ can be computed in linear time. Note that:

$$K^{(t+1)}\mathbf{1}_{2t+2} = \begin{pmatrix} K^{(t)}\mathbf{1}_{2t} + k_{\cdot,2t+1} + k_{\cdot,2t+2} \\ K_{\cdot,2t+1}^\top\mathbf{1}_{2t} + K_{2t+1,2t+1} + K_{2t+1,2t+2} \\ K_{\cdot,2t+2}^\top\mathbf{1}_{2t} + K_{2t+2,2t+1} + K_{2t+2,2t+2} \end{pmatrix},$$

which can be computed in linear time if $K^{(t)}\mathbf{1}_{2t}$ is stored (similar result holds for $L^{(t+1)}\mathbf{1}_{2t+2}$). It thus follows that $\mathbf{1}_{2t+2}^\top K^{(t+1)}\mathbf{1}_{2t+2}$, $\mathbf{1}_{2t+2}^\top L^{(t+1)}\mathbf{1}_{2t+2}$ and $\mathbf{1}_{2t+2}^\top K^{(t+1)}L^{(t+1)}\mathbf{1}_{2t+2}$ can all be computed in linear time. To sum up, we need to cache $\text{tr}\left(K^{(t)}L^{(t)}\right)$, $K^{(t)}\mathbf{1}_{2t}$, $L^{(t)}\mathbf{1}_{2t}$ to compute the normalization constant in linear time.