Adaptive test of independence based on HSIC measures

Anouar Meynaoui, Mélisande Albert, Béatrice Laurent, Amandine Marrel

Adaptive test of independence based on HSIC measures

Anouar Meynaoui[1,2], Mélisande Albert[2], Béatrice Laurent-Bonneau[2], and Amandine Marrel[1]

[1]CEA, DEN, DER, SESI, LEMS, 13108 Saint-Paul-lez-Durance, France
[2]Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse
INSA de Toulouse,
135, avenue de Rangueil,
31077 Toulouse Cedex 4, France.

August 21, 2019

Abstract

Dependence measures based on reproducing kernel Hilbert spaces, also known as Hilbert-Schmidt Independence Criterion and denoted HSIC, are widely used to statistically decide whether or not two random vectors are dependent. Recently, non-parametric HSIC-based statistical tests of independence have been performed. However, these tests lead to the question of the choice of the kernels associated to the HSIC. In particular, there is as yet no method to objectively select specific kernels with theoretical guarantees in terms of first and second kind errors. One of the main contributions of this work is to develop a new HSIC-based aggregated procedure which avoids such a kernel choice, and to provide theoretical guarantees for this procedure. To achieve this, we first introduce non-asymptotic single tests based on Gaussian kernels with a given bandwidth, which are of prescribed level $\alpha \in (0,1)$. From a theoretical point of view, we upper-bound their uniform separation rate of testing over Sobolev and Nikol'skii balls. Then, we aggregate several single tests, and obtain similar upper-bounds for the uniform separation rate of the aggregated procedure over the same regularity spaces. Another main contribution is that we provide a lower-bound for the non-asymptotic minimax separation rate of testing over Sobolev balls, and deduce that the aggregated procedure is adaptive in the minimax sense over such regularity spaces. Finally, from a practical point of view, we perform numerical studies in order to assess the efficiency of our aggregated procedure and compare it to existing independence tests in the literature.

1 Introduction

We study here the problem of testing the independence of two real random vectors $X = \left(X^{(1)}, \ldots, X^{(p)}\right)$ and $Y = \left(Y^{(1)}, \ldots, Y^{(q)}\right)$. Let us first introduce some notations and assumptions. The couple (X, Y) is assumed to have a joint density f w.r.t. Lebesgue measure on $\mathbb{R}^p \times \mathbb{R}^q$. The marginal densities of X and Y are respectively denoted f_1 and f_2. We also denote by $f_1 \otimes f_2$, the product of the marginal densities f_1 and f_2, defined as follows:

$$f_1 \otimes f_2 : (x, y) \in \mathbb{R}^p \times \mathbb{R}^q \mapsto f_1(x) f_2(y).$$

The density f is assumed to be unknown as well as the marginals f_1 and f_2. We also assume that we have an n-sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent and identically distributed (i.i.d.) random variables with common density f. The probability measure associated to this n-sample is denoted P_f. By analogy, $P_{f_1 \otimes f_2}$ designates the probability measure associated to a n-sample with common density $f_1 \otimes f_2$.

We address here the question of testing the null hypothesis $(\mathcal{H}_0)$: "X and Y are independent" against the alternative $(\mathcal{H}_1)$: "X and Y are dependent". That is equivalent to test

$$(\mathcal{H}_0)\text{: } \text{“}f = f_1 \otimes f_2\text{”} \quad \text{against} \quad (\mathcal{H}_1)\text{: } \text{“}f \neq f_1 \otimes f_2\text{”}.$$

Throughout this document, the densities f, f_1 and f_2 are assumed to be bounded and M_f denotes the maximum of their infinity norms: $M_f = \max\{\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\}$.

Non parametric tests of independence. To test independence between X and Y, many approaches have been explored in the last few decades. Among them, [Hoeffding, 1948] introduces an independence test based on the difference between the distribution function of (X, Y) and the product of the marginal distribution functions. This test has good properties in the asymptotic framework: consistent and distribution-free under the null hypothesis. But, it is only designated to univariate continuous random variables. The authors of [Bergsma and Dassios, 2014] suggest an improvement of Hoeffding's test, which is applicable to discrete, continuous or mixture of discrete and continuous distributions. Lately, [Weihs et al., 2018] extend Hoeffding's test to the case of multivariate random variables. Another classical method for testing independence between X and Y is based on comparing the joint density f and the product of the marginals $f_1 \otimes f_2$ [Rosenblatt, 1975, Ahmad and Li, 1997]. For this, an intermediate step is to estimate these densities using the kernel-based method of Parzen-Rosenblatt [Parzen, 1962]. The major drawback of this method is that the convergence is slow for high dimensions i.e. when $p + q$ is large (this fact is also called the *curse of dimensionality*, see e.g. [Scott, 2012]). This approach is therefore not feasible in the case of high dimensions with limited sample size. More recently, many approaches based on Reproducing Kernel Hilbert Spaces (RKHS, see [Aronszajn, 1950] for more details) have been developed. In particular, several RKHS-based dependence measures have been considered. By definition, all these measures characterize the independence once the involved RKHS kernels belong to a class called *universal kernels* [Micchelli et al., 2006]. It has been established that the universality of kernels is a very restricting condition [Steinwart, 2001]. A largest class called *characteristic kernels* detecting the independence has been later introduced [Fukumizu et al., 2008, Sriperumbudur et al., 2010]. One of the earlier RKHS measures is the Kernel Canonical Correlation (KCC), introduced in [Bach and Jordan, 2002]. Unfortunately, the estimation of the KCC is not practical, it requires an extra regularization which has to be adjusted. Other dependence measures, easier to estimate have been studied later. For instance, the Kernel Mutual Information (KMI) [Gretton et al., 2003, Gretton et al., 2005b] and the Constrained covariance (COCO) [Gretton et al., 2005c, Gretton et al., 2005b], which are relatively easy to interpret and implement, are widely used. Last but not least, one of the most interesting kernel dependence measure is the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005a]. The HSIC has a very low computational cost and seems to numerically outperform all previous RKHS measures [Gretton et al., 2005a]. Furthermore, beyond the good quality of a given dependence measure, a straightforward interpretation of its estimated value, may not be enough to discern the dependence from the independence. To further study the independence between X and Y, tests based on these measures can be used. A first RKHS-based asymptotic independence test is performed using general large deviation inequalities [Gretton et al., 2005a]. A more optimal asymptotic independence test based on a Gamma approximation of the distributions of HSIC estimators under $(\mathcal{H}_0)$ is developed by [Gretton et al., 2008]. This last statistical test remains by far the most commonly used kernel-based test for independence. A generalization of this test for the joint and mutual independence of several random variables is presented in [Pfister et al., 2018]. We also mention the RKHS-based test [Póczos et al., 2012], based on a new dependence measure called Copula-based kernel dependency measure. Yet, this test is more conservative than the test of [Gretton et al., 2008], since it is based on large deviation inequalities rather than the asymptotic distributions of the estimators under $(\mathcal{H}_0)$. Lately, the distance covariance which is based on the difference between the characteristic function of (X, Y) and the product of the marginal characteristic functions has been introduced in [Székely et al., 2007]. The distance covariance has good properties and has been used to study the independence between random variables of high dimensions [Székely and Rizzo, 2013, Yao et al., 2018]. Furthermore, it has been shown that the distance covariance coincides with the HSIC using specific choice of kernels. We also mention the statistical test of independence based on the kernel mutual information recently introduced by [Berrett and Samworth, 2017]. This new statistical test seems to achieve comparable results with the classical tests based on HSIC. In this paper, we focus on HSIC measures to test independence between X and Y.

Review on HSIC measures. The definition of the HSIC is derived from the notion of cross-covariance operator [Baker, 1973, Fukumizu et al., 2004], which can be seen as a generalization of the

classical covariance, measuring many forms of dependence between X and Y (not only linear ones). For this, [Gretton et al., 2005a] associate to X a RKHS $\mathcal{F}$ composed of functions mapping from $\mathbb{R}^p$ to $\mathbb{R}$ ($\mathcal{F}$ is a set of transformations for X), and characterized by a scalar product $\langle .,. \rangle_{\mathcal{F}}$. The same operation is carried out for Y, considering a RKHS denoted $\mathcal{G}$ and a scalar product $\langle .,. \rangle_{\mathcal{G}}$. The cross-covariance operator $C_{X,Y}$ associated to RKHS $\mathcal{F}$ and $\mathcal{G}$ is the operator mapping from $\mathcal{G}$ to $\mathcal{F}$ and verifying for all $(F, G) \in \mathcal{F} \times \mathcal{G}$,

$$\langle F, C_{X,Y}(G) \rangle_{\mathcal{F}} = \mathrm{Cov}\left(F(X), G(Y)\right).$$

Designating by $(u_i)_i$ and $(v_j)_j$ respectively orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$, the HSIC between X and Y is the square of the operator's $C_{X,Y}$ Hilbert-Schmidt norm [Gretton et al., 2005a] defined as

$$\mathrm{HSIC}(X, Y) = \|C_{X,Y}\|^2_{\mathrm{HS}} = \sum_{i,j} \langle u_i, C_{X,Y}(v_j) \rangle^2_{\mathcal{F}} = \sum_{i,j} \mathrm{Cov}\left(u_i(X), v_j(Y)\right)^2.$$

The fundamental idea behind this definition is that $\mathrm{HSIC}(X, Y)$ is zero if and only if $\mathrm{Cov}\left(F(X), G(Y)\right) = 0$ for all $(F, G) \in \mathcal{F} \times \mathcal{G}$. Furthermore, we already know (see e.g. [Jacod and Protter, 2012]) that X and Y are independent if and only if $\mathrm{Cov}\left(F(X), G(Y)\right) = 0$ for all bounded and continuous functions F and G. It follows that, for well chosen RKHS, the nullity of the HSIC characterizes independence. Before giving such a condition, we recall that [Gretton et al., 2005a] express the $\mathrm{HSIC}(X, Y)$ in a very convenient form, using kernels k and l respectively associated to $\mathcal{F}$ and $\mathcal{G}$,

$$\begin{aligned} \mathrm{HSIC}(X, Y) = {} & \mathbb{E}\left[k(X, X')l(Y, Y')\right] + \mathbb{E}\left[k\left(X, X'\right)\right]\mathbb{E}\left[l\left(Y, Y'\right)\right] \\ & - 2\mathbb{E}\left[\mathbb{E}\left[k\left(X, X'\right) \mid X\right]\mathbb{E}\left[l\left(Y, Y'\right) \mid Y\right]\right], \end{aligned} \tag{1}$$

where (X', Y') is an independent and identically distributed copy of (X, Y). Note that $\mathrm{HSIC}(X, Y)$ only depends on the density f of (X, Y). We thus denote it $\mathrm{HSIC}(f)$ in the following.

Authors of [Gretton et al., 2005d] show that a sufficient condition so that the nullity of the associated HSIC is characteristic of independence is that the RKHS $\mathcal{F}$ (resp. $\mathcal{G}$) induced by k and (resp. l) is dense in the space of bounded and continuous functions mapping from $\mathbb{R}^p$ (resp. $\mathbb{R}^q$) to $\mathbb{R}$. These kernels are called universal [Micchelli et al., 2006]. However, the universality is a very limiting condition and only adapted to compact domains. Recently, a wider class of kernels called *characteristic kernels* has been introduced in [Fukumizu et al., 2008, Sriperumbudur et al., 2010]. These kernels characterize independence on compact as well as non-compact sets. Among them, the most commonly used are Gaussian kernels [Steinwart, 2001]. We consider in this paper Gaussian kernels. Let us introduce some notations. We denote by g_s the density of the standard Gaussian distribution on $\mathbb{R}^s$ defined for all $x \in \mathbb{R}^s$ by

$$g_s(x) = \frac{1}{(2\pi)^{s/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{s} x_i^2\right). \tag{2}$$

For any bandwidths $\lambda = (\lambda_1, ..., \lambda_p) \in (0, +\infty)^p$ and $\mu = (\mu_1, ..., \mu_q) \in (0, +\infty)^q$, we define for any $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$,

$$\varphi_\lambda(x) = \frac{1}{\lambda_1 \dots \lambda_p} g_p\left(\frac{x_1}{\lambda_1}, \dots, \frac{x_p}{\lambda_p}\right), \tag{3}$$

$$\phi_\mu(y) = \frac{1}{\mu_1 \dots \mu_q} g_q\left(\frac{y_1}{\mu_1}, \dots, \frac{y_q}{\mu_q}\right). \tag{4}$$

Finally, we define the Gaussian kernels, for $x, x' \in \mathbb{R}^p$ and $y, y' \in \mathbb{R}^q$,

$$k_\lambda(x, x') = \varphi_\lambda(x - x'), \quad l_\mu(y, y') = \phi_\mu(y - y').$$

We denote by $\mathrm{HSIC}_{\lambda,\mu}(f)$ the HSIC measure defined in (1), where the kernels k and l are respectively the Gaussian kernels k_λ and l_μ.

In practice, the computation of $\mathrm{HSIC}_{\lambda,\mu}(f)$ is not feasible, since it depends on the unknown density f. Given an i.i.d. n-sample $(X_i, Y_i)_{1 \le i \le n}$ with common density f, $\mathrm{HSIC}_{\lambda,\mu}(f)$ can be estimated by

estimating each expectation of Equation (1). For this, we introduce the following U-statistics, respectively with order 2, 3 and 4,

$$\widehat{\mathrm{HSIC}}^{(2)}_{\lambda,\mu} = \frac{1}{n(n-1)} \sum_{(i,j)\in \mathbf{i}^n_2} k_\lambda\left(X_i, X_j\right) l_\mu\left(Y_i, Y_j\right),$$

$$\widehat{\mathrm{HSIC}}^{(3)}_{\lambda,\mu} = \frac{1}{n(n-1)(n-2)} \sum_{(i,j,r)\in \mathbf{i}^n_3} k_\lambda\left(X_i, X_j\right) l_\mu\left(Y_j, Y_r\right),$$

and

$$\widehat{\mathrm{HSIC}}^{(4)}_{\lambda,\mu} = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{(i,j,q,r)\in \mathbf{i}^n_4} k_\lambda\left(X_i, X_j\right) l_\mu\left(Y_q, Y_r\right),$$

where $\mathbf{i}^n_r$ is the set of all r-tuples drawn without replacement from the set $\{1, ..., n\}$. We estimate $\mathrm{HSIC}_{\lambda,\mu}\left(f\right)$ by the U-statistic

$$\widehat{\mathrm{HSIC}}_{\lambda,\mu} = \widehat{\mathrm{HSIC}}^{(2)}_{\lambda,\mu} + \widehat{\mathrm{HSIC}}^{(4)}_{\lambda,\mu} - 2\widehat{\mathrm{HSIC}}^{(3)}_{\lambda,\mu}. \tag{5}$$

Such estimators of the HSIC have been used to construct independence tests. A first asymptotic test of level $\alpha \in (0,1)$ has been introduced by [Gretton et al., 2008]. For this, the authors show that under $(\mathcal{H}_0)$, the asymptotic distribution of the HSIC estimator can be approximated by a Gamma distribution with parameters which are easy to estimate. Furthermore, [Gretton and Györfi, 2010] also show the asymptotic consistency of the test (the convergence to one of the power under any reasonable alternative). However, there are two main disadvantages of this testing procedure. Firstly, it is purely asymptotic in the sense that the critical value of the test is obtained from an approximation of the asymptotic distribution under $(\mathcal{H}_0)$. In particular, the first kind error is controlled only in the asymptotic framework. Secondly, only an heuristic choice of the bandwidths λ and μ is considered with no theoretical guarantees. One of the main contributions of this work is to introduce a testing procedure which avoids such an arbitrary choice with theoretical guarantees in terms of first and second kind errors.

Adaptive HSIC-based independence tests. To avoid the unjustified choice of the bandwidths λ and μ, a first step is to define a criterion allowing to compare the performances of the HSIC-tests associated to different bandwidths. For this, we consider the *uniform separation rate* as defined in [Baraud, 2002]. For any level-α test Δ_α with values in $\{0,1\}$, rejecting independence when $\Delta_\alpha = 1$, the uniform separation rate $\rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right)$ of the test Δ_α, over a class $\mathcal{C}_\delta$ of alternatives f such that the difference between f and its marginals, namely $f - f_1 \otimes f_2$, satisfies smoothness assumptions, with respect to the $\mathbb{L}_2$-norm, is defined for all β in $(0,1)$ by

$$\rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right) = \inf\left\{\rho > 0;\ \sup_{f\in\mathcal{F}_\rho(\mathcal{C}_\delta)} \mathbb{P}_f\left(\Delta_\alpha = 0\right) \leq \beta\right\}, \tag{6}$$

where $\mathcal{F}_\rho(\mathcal{C}_\delta) = \{f; f - f_1 \otimes f_2 \in \mathcal{C}_\delta, \|f - f_1 \otimes f_2\|_2 > \rho\}$ and $\|\cdot\|_2$ designates the usual $\mathbb{L}_2$-norm. To avoid any misunderstanding, let us highlight that f_1 and f_2 always denote the marginals of the function f and are not fixed *a priori*.

The uniform separation rate is then the smallest value in the sense of the $\mathbb{L}_2$-norm of $f - f_1 \otimes f_1$ (the difference between the joint density and the product of marginals) allowing to control the second kind error of the test by β. This definition extends to the non-asymptotic framework, the notion of *critical radius* introduced and studied for several examples in a series of Ingster papers (see e.g. [Ingster, 1993a, Ingster, 1993b]). A test of level α having the optimal performances, should then have the smallest possible uniform separation rate (up to a multiplicative constant) over $\mathcal{C}_\delta$. To quantify this, let us introduce, as in [Baraud, 2002], the *non-asymptotic minimax rate of testing*, defined by

$$\rho\left(\mathcal{C}_\delta, \alpha, \beta\right) = \inf_{\Delta_\alpha} \rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right), \tag{7}$$

where the infimum is taken over all α-level tests of $(\mathcal{H}_0)$ against $(\mathcal{H}_1)$. If the uniform separation rate of a test is upper-bounded up to a constant by the non-asymptotic minimax rate of testing, then this

test is said to be *optimal in the minimax sense*. The problem of non-asymptotic minimax rate of testing was raised in many papers over the past years. Among them, we mention for example [Ingster and Suslina, 1998, Laurent et al., 2012] for minimax signal detection testing. However, only few works exist for the problem of minimax independence testing. The notable works are those of Ingster [Ingster, 1989, Ingster, 1993b] and of Yodé [Yodé, 2004, Yodé, 2011] in the asymptotic framework, and the one of Albert [Albert, 2015] in the non-asymptotic framework considered in this paper. As far as we know, no lower-bound for the minimax rate of testing independence was yet proved in the non-asymptotic framework. Furthermore, beyond the problem of minimax rate, the straightforward practical construction of a minimax test is impossible. Indeed, this construction depends on the unknown smoothness parameters defining the regularity space $\mathcal{C}_\delta$. The objective is then to construct a minimax test which does not need any smoothness property to be implemented. These tests are called *minimax adaptive* (or assumption free). It has been shown that a standard logarithmic price is sometimes inevitable for adaptivity [Spokoiny, 1996]. The problem of adaptivity has received a good attention in the literature. We mention for instance [Baraud et al., 2003] for linear regression model testing with normal noise and [Butucea and Tribouley, 2006] for testing the equality of two samples densities. For the specific case of testing independence, the adaptive testing procedure introduced in [Yodé, 2011] seems to be the only currently existing. As mentioned above, this test is purely asymptotic, but we are interested here in the non-asymptotic framework. Recently, an interesting approach of testing developed in [Fromont et al., 2013], consists in testing the equality of two poisson processes intensities by aggregating several kernels in a unique testing procedure. It has been shown in [Fromont et al., 2013] that this testing procedure is adaptive over several regularity spaces. Inspired by these works, and following the work of [Gretton et al., 2008, Gretton and Györfi, 2010], we consider in this paper a procedure of testing independence based on HSIC measures and aggregating a collection of Gaussian-kernel HSIC tests. Another main contribution of this paper is to prove that this aggregated procedure is adaptive over Sobolev balls by proving both an upper-bound of its uniform separation rate, and a lower bound of the non-asymptotic minimax rate of testing in this setting. Moreover, the upper bound of its uniform separation rate over Nikol'skii-Besov balls seems optimal compared to "classical" testing rates in other frameworks. This suggests that this test may also be adaptive over Nikol'skii-Besov spaces.

The structure of the paper is as follows. In Section 2, we first present a theoretical non-asymptotic HSIC-based test of level α as well as its practical implementation. We then provide theoretical conditions based on concentration inequalities for U-statistics, allowing to control the second kind error by a given β. This last step leads us to sharp upper bounds of the uniform separation rate over two classes of alternatives, namely *Sobolev* and *Nikol'skii-Besov* balls. In Section 3, we introduce the general aggregated procedure of testing, taking into account various single HSIC-based tests. Thereafter, we show a non-asymptotic oracle upper bound of uniform separation rate over Sobolev and Nikol'skii-Besov balls. These upper bounds are shown to be optimal over Sobolev spaces in Section 4. Finally, we illustrate the procedure in Section 5 on simulated data. Methodological choices are tested and compared to other existing independence tests.

All along the paper, the generic notation $C(a, b, \ldots)$ denotes a positive constant depending only on its arguments $(a, b, \ldots)$ and that may vary from line to line.

2 Single kernel-based tests of independence

The target of this section is to sharply upper bound the uniform separation rate of non-asymptotic HSIC-based tests over some "classical" regularity spaces, specifically Sobolev and Nikol'skii-Besov spaces. For this, theoretical conditions allowing to control the second kind error are given in terms of $\mathrm{HSIC}_{\lambda,\mu}(f)$ value and then in terms of the $\mathbb{L}_2$-norm of $f - f_1 \otimes f_2$.

2.1 The testing procedures

A first theoretical test. Since Gaussian kernels are characteristic, testing the independence between X and Y is equivalent to testing

$$(\mathcal{H}_0) : \mathrm{HSIC}_{\lambda,\mu}(f) \;=\; 0 \qquad \text{against} \qquad (\mathcal{H}_1) : \mathrm{HSIC}_{\lambda,\mu}(f) \;>\; 0.$$

The statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined in Equation (5) is then a natural choice to test independence between X and Y, since it is an unbiased estimator of $\mathrm{HSIC}_{\lambda,\mu}(f)$. The corresponding test rejects independence if $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ is significantly large. Specifically, for $\alpha \in (0,1)$, we consider the statistical test which rejects $(\mathcal{H}_0)$ if $\widehat{\mathrm{HSIC}}_{\lambda,\mu} > q^{\lambda,\mu}_{1-\alpha}$, where $q^{\lambda,\mu}_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ under $P_{f_1\otimes f_2}$. Note that the analytical computation of the quantile $q^{\lambda,\mu}_{1-\alpha}$ is not possible since its value depends on the unknown marginals f_1 and f_2 of the couple (X,Y). In practice, a permutation approach is used as described in the following paragraph.

The associated test function is defined by

$$\Delta^{\lambda,\mu}_{\alpha} = \mathbb{1}_{\widehat{\mathrm{HSIC}}_{\lambda,\mu} > q^{\lambda,\mu}_{1-\alpha}}. \tag{8}$$

Then, the null hypothesis is rejected if and only if $\Delta^{\lambda,\mu}_{\alpha} = 1$. By definition of the quantile, this theoretical test is of non-asymptotic level α, that is

$$P_{f_1\otimes f_2}\left(\Delta^{\lambda,\mu}_{\alpha} = 1\right) \leq \alpha.$$

A permutation-based test of independence. Let us now describe the permutation method with Monte Carlo approximation applied to approach the unknown theoretical quantile $q^{\lambda,\mu}_{1-\alpha}$. The practical validity of this approach is illustrated on simulated examples in Section 5.2.1.

We denote $\mathbb{Z}_n = (X_i, Y_i)_{1\leq i\leq n}$ the original sample and we compute the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}(\mathbb{Z}_n)$ defined by Equation (5). Then, we consider B independent and uniformly distributed random permutations of $\{1,...,n\}$, denoted $\tau_1,...,\tau_B$, independent of $\mathbb{Z}_n$. We define for each permutation τ_b the corresponding permuted sample $\mathbb{Z}_n^{\tau_b} = (X_i, Y_{\tau_b(i)})_{1\leq i\leq n}$ and compute the permuted test statistic

$$\widehat{H}^{\star b}_{\lambda,\mu} = \widehat{\mathrm{HSIC}}_{\lambda,\mu}(\mathbb{Z}_n^{\tau_b})$$

on this new sample.

Under $P_{f_1\otimes f_2}$, each permuted sample $\mathbb{Z}_n^{\tau_b}$ has the same distribution than the original sample $\mathbb{Z}_n$. Hence, the random variables $\widehat{H}^{\star b}_{\lambda,\mu}$, $1\leq b\leq B$, have the same distribution as $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$. We apply a trick, based on [Romano and Wolf, 2005, Lemma 1], which consists in adding the original sample to the Monte Carlo sample in order to obtain a test of non-asymptotic level α. To do so, denote

$$\widehat{H}^{\star B+1}_{\lambda,\mu} = \widehat{\mathrm{HSIC}}_{\lambda,\mu}, \quad \text{and} \quad \widehat{H}^{\star(1)}_{\lambda,\mu} \leq \widehat{H}^{\star(2)}_{\lambda,\mu} \leq \ldots \leq \widehat{H}^{\star(B+1)}_{\lambda,\mu}$$

the order statistic. Then, the permuted quantile with Monte Carlo approximation $\widehat{q}^{\lambda,\mu}_{1-\alpha}$ is thus defined as

$$\widehat{q}^{\lambda,\mu}_{1-\alpha} = \widehat{H}^{\star(\lceil (B+1)(1-\alpha)\rceil)}_{\lambda,\mu}. \tag{9}$$

where $\lceil\cdot\rceil$ denotes the ceiling function. The permuted test with Monte Carlo approximation $\widehat{\Delta}^{\lambda,\mu}_{\alpha}$ performed in practice is then defined as

$$\widehat{\Delta}^{\lambda,\mu}_{\alpha} = \mathbb{1}_{\widehat{\mathrm{HSIC}}_{\lambda,\mu} > \widehat{q}^{\lambda,\mu}_{1-\alpha}}. \tag{10}$$

Proposition 1. *Let α be in $(0,1)$ and $\widehat{\Delta}^{\lambda,\mu}_{\alpha}$ the test defined by Equation (10). Then,*

$$P_{f_1\otimes f_2}\left(\widehat{\Delta}^{\lambda,\mu}_{\alpha} = 1\right) \leq \alpha, \tag{11}$$

that is, this permuted test with Monte Carlo approximation is of prescribed non-asymptotic level α.

2.2 Control of the second kind error in terms of HSIC

For an arbitrarily small β given in $(0,1)$, Lemma 1 provides a first non-asymptotic condition on the alternative f ensuring that the probability of second kind error of the theoretical test under P_f is at most equal to β. This condition is given for the value of $\mathrm{HSIC}_{\lambda,\mu}(f)$. It involves the variance of the estimator $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ which is finite since this estimator is a bounded random variable.

Lemma 1. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Let α, β in $(0,1)$, and $q_{1-\alpha}^{\lambda,\mu}$ be the $(1-\alpha)$-quantile of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ under $P_{f_1\otimes f_2}$. Then $P_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q_{1-\alpha}^{\lambda,\mu}) \leq \beta$ as soon as*

$$\mathrm{HSIC}_{\lambda,\mu}(f) \geq \sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}} + q_{1-\alpha}^{\lambda,\mu}.$$

Lemma 1 gives a threshold for $\mathrm{HSIC}_{\lambda,\mu}(f)$ from which the dependence between X and Y is detectable with probability at least $1-\beta$ using Gaussian kernels k_λ and l_μ with given bandwidths λ and μ. Furthermore, it would be useful to give more explicit conditions w.r.t. the bandwidths λ and μ and the sample size n. The objective of this section is to provide a condition w.r.t. λ, μ and n on the theoretical value $\mathrm{HSIC}_{\lambda,\mu}$, so that the test $\Delta_\alpha^{\lambda,\mu}$ has a second kind error controlled by $\beta \in (0,1)$. For this, Lemma 1 already provides a condition involving $\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})$ and $q_{1-\alpha}^{\lambda,\mu}$. It is therefore necessary to establish sharp upper bounds for these two quantities w.r.t. λ, μ and n. Propositions 2 and 3 give these upper bounds.

Proposition 2. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Assume that the densities f, f_1 and f_2 are bounded. Then,*

$$\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu}) \leq C\left(M_f, p, q\right)\left\{\frac{1}{n} + \frac{1}{\lambda_1 ... \lambda_p \mu_1 ... \mu_q n^2}\right\},$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$.

Proposition 3. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Let α in $(0,1)$ and $q_{1-\alpha}^{\lambda,\mu}$ be the $(1-\alpha)$-quantile of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ under $P_{f_1\otimes f_2}$. Assuming that the densities f_1, f_2 are bounded,*

$$\max\left(\lambda_1 ... \lambda_p\ ,\ \mu_1 ... \mu_q\right) < 1 \ \text{ and } \ n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}\ > \log\left(\frac{1}{\alpha}\right)\ > 1.$$

Then,

$$q_{1-\alpha}^{\lambda,\mu} \leq \frac{C\left(\|f_1\|_\infty, \|f_2\|_\infty, p, q\right)}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \log\left(\frac{1}{\alpha}\right).$$

Combining Lemma 1, Propositions 2 and 3, we can then give a sufficient condition on $\mathrm{HSIC}_{\lambda,\mu}$ depending on the parameters λ, μ and the sample size n in order to control the second kind error by β. This result is presented in Corollary 1.

Corollary 1. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Let α, β in $(0,1)$, and $q_{1-\alpha}^{\lambda,\mu}$ be the $(1-\alpha)$-quantile of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ under $P_{f_1\otimes f_2}$ as defined in Section 2.1. Assume that the densities f, f_1 and f_2 are bounded, and that*

$$\max\left(\lambda_1 ... \lambda_p\ ,\ \mu_1 ... \mu_q\right) < 1 \ \text{ and } \ n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}\ > \log\left(\frac{1}{\alpha}\right)\ > 1.$$

Then, one has $P_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q_{1-\alpha}^{\lambda,\mu}) \leq \beta$ as soon as

$$\mathrm{HSIC}_{\lambda,\mu}(f) > C\left(M_f, p, q, \beta\right)\left\{\frac{1}{\sqrt{n}} + \frac{1}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \log\left(\frac{1}{\alpha}\right)\right\},$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$.

Note that the right hand term given in Corollary 1 is not computable in practice since it depends on the unknown density f. However, this dependence is weak since it only depends on the infinite norm of f and its marginals. For a given $\beta \in (0,1)$, Corollary 1 provides a condition on the value of $\mathrm{HSIC}_{\lambda,\mu}(f)$ ensuring that the probability of second kind error of the theoretical test under such f is at most equal to β. We now want to express such conditions in terms of the $\mathbb{L}_2$-norm of the function $f - f_1 \otimes f_2$, for the sake of interpretation, and in order to be able to determine separation rates with respect to this $\mathbb{L}_2$-norm for our test.

2.3 Control of the second kind error in terms of $\mathbb{L}_2$-norm

In order to express a condition on the $\mathbb{L}_2$-norm of the difference $f - f_1 \otimes f_2$ ensuring a probability of second kind error controlled by β, we first give in Lemma 2 a link between $\mathrm{HSIC}_{\lambda,\mu}$ and $\|f - f_1 \otimes f_2\|_2^2$.

Lemma 2. *Let $\psi = f - f_1 \otimes f_2$. The* HSIC *measure $\mathrm{HSIC}_{\lambda,\mu}(f)$ associated to kernels k_λ and l_μ and defined in Equation (1) can be written as*

$$\mathrm{HSIC}_{\lambda,\mu}(f) = \langle \psi, \psi * (\varphi_\lambda \otimes \phi_\mu) \rangle_2,$$

where φ_λ and ϕ_μ are the functions respectively defined in Equations (3) and (4). Moreover, the notation $\langle ., . \rangle_2$ designates the usual scalar product in the space $\mathbb{L}_2$. One can easily deduce that

$$\mathrm{HSIC}_{\lambda,\mu}(f) = \frac{1}{2}\Big(\|\psi\|_2^2 + \|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 - \|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 \Big). \tag{12}$$

The Theorem 1 gives a sufficient condition on $\|f - f_1 \otimes f_2\|_2^2$, for the second kind error of the test $\Delta_\alpha^{\lambda,\mu}$ to be upper bounded by β.

Theorem 1. *Let $(X_i, Y_i)_{1 \le i \le n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Denote $\psi = f - f_1 \otimes f_2$. Let α, β in $(0,1)$, and $q_{1-\alpha}^{\lambda,\mu}$ be the $(1-\alpha)$-quantile of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ under $P_{f_1 \otimes f_2}$. Assume that the densities f, f_1 and f_2 are bounded, and that*

$$\max(\lambda_1 ... \lambda_p \ , \ \mu_1 ... \mu_q) < 1 \quad \text{and} \quad n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q} \ > \log\left(\frac{1}{\alpha}\right) \ > 1.$$

One has $P_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \le q_{1-\alpha}^{\lambda,\mu}) \le \beta$ as soon as

$$\|\psi\|_2^2 > \|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 + \frac{C(M_f, p, q, \beta)}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \log\left(\frac{1}{\alpha}\right).$$

where $M_f = \max(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty)$, and $C(M_f, p, q)$ denotes a positive constant depending only on its arguments.

In the condition given in Theorem 1, appears a compromise between a bias term $\|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2$ and a term induced by the square-root of the variance of the estimator $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$. Comparing the conditions on the HSIC given in Corollary 1 and on $\|f - f_1 \otimes f_2\|_2^2$ given in Theorem 1, the meticulous reader may notice that the term in $1/\sqrt{n}$ has been removed. This suppression seems to be necessary to obtain optimal separation rates according to the literature in other testing frameworks. This derives from quite tricky computations that we point out here and that directly prove Theorem 1. By combining Lemmas 1 and 2, direct computations lead to the condition

$$\|\psi\|_2^2 > \|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 - \|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 + 2\sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}} + 2q_{1-\alpha}^{\lambda,\mu}. \tag{13}$$

If one directly considers the upper bound of the variance $\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})$ given in Proposition 2, one would get the unwanted $1/\sqrt{n}$ term. The idea is to take advantage of the negative term $-\|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2$ to compensate such term. To do so, we need a more refined control of the variance given in the technical Proposition 4.

Proposition 4. *Let $(X_i, Y_i)_{1 \le i \le n}$ be an i.i.d. sample with distribution P_f and consider the test statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ defined by (5). Assume that the densities f, f_1 and f_2 are bounded. Then,*

$$\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu}) \le C(M_f)\frac{\|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2}{n} + \frac{C(M_f, p, q)}{\lambda_1 ... \lambda_p \mu_1 ... \mu_q n^2},$$

where $M_f = \max(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty)$.

Finally, using standard inequalities such as $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq \delta a + b/\delta$ for all positive a, b and δ, one can prove

$$2\sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}} \leq \|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 + \frac{C(M_f,\beta)}{n} + \frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1 \ldots \lambda_p \mu_1 \ldots \mu_q}},$$

which leads to Theorem 1 when combined with Equation (13) and Proposition 3. Notice that such trick is already present in [Fromont et al., 2013].

2.4 Uniform separation rate

The bias term in Theorem 1 comes from the fact that we do not estimate $\|f - f_1 \otimes f_2\|_2^2$ but $\mathrm{HSIC}_{\lambda,\mu}(f)$. In order to have a control of the bias term w.r.t λ and μ, we assume that $f - f_1 \otimes f_2$ belongs some class of regular functions. We introduce the two following classes: Sobolev balls (isotropic case) and Nikol'skii-Besov balls (anisotropic case).

2.4.1 Case Sobolev balls

For $d \in \mathbb{N}^*$, $\delta > 0$ and $R > 0$, the Sobolev ball $\mathcal{S}_d^\delta(R)$ is the set defined by

$$\mathcal{S}_d^\delta(R) = \left\{ s : \mathbb{R}^d \to \mathbb{R} / s \in \mathbb{L}^1(\mathbb{R}^d) \cap \mathbb{L}^2(\mathbb{R}^d), \int_{\mathbb{R}^d} \|u\|^{2\delta} |\hat{s}(u)|^2 du \leq (2\pi)^d R^2 \right\}, \tag{14}$$

where $\hat{s}$ denotes the Fourier transform of s defined by $\hat{s}(u) = \int_{\mathbb{R}^d} s(x) e^{i\langle x,u\rangle}\, \mathrm{d}x$, $\langle .,. \rangle$ denotes the usual scalar product in $\mathbb{R}^d$ and $\|.\|$ the Euclidean norm in $\mathbb{R}^d$.

Lemma 3 gives an upper bound for the bias term in the case when $f - f_1 \otimes f_2$ belongs to particular Sobolev balls.

Lemma 3. *Let $\psi = f - f_1 \otimes f_2$. We assume that $\psi \in \mathcal{S}_{p+q}^\delta(R)$, where $\delta \in (0,2]$ and $\mathcal{S}_d^\delta(R)$ is defined by (14). Let φ_λ and ϕ_μ be the functions respectively defined in Equations (3) and (4). Then we have the following inequality,*

$$\|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 \leq C(p,q,\delta,R) \left[\sum_{i=1}^p \lambda_i^{2\delta} + \sum_{j=1}^q \mu_j^{2\delta} \right].$$

One can deduce from Theorem 1 upper bounds for the uniform separation rates (defined in (6)) of the test $\Delta_\alpha^{\lambda,\mu}$ over Sobolev balls.

Theorem 2. *Let $\alpha, \beta \in (0,1)$ and consider the same notation and assumptions as in Theorem 1. Let $\delta \in (0,2]$ and $R > 0$. Then, the uniform separation rate $\rho\left(\Delta_\alpha^{\lambda,\mu}, \mathcal{S}_{p+q}^\delta(R), \beta\right)$ defined in (6) over the Sobolev ball $\mathcal{S}_{p+q}^\delta(R)$ can be upper bounded as follows*

$$\left[\rho\left(\Delta_\alpha^{\lambda,\mu}, \mathcal{S}_{p+q}^\delta(R), \beta\right)\right]^2 \leq C(p,q,\delta,R) \left[\sum_{i=1}^p \lambda_i^{2\delta} + \sum_{j=1}^q \mu_j^{2\delta} \right] + \frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1 \ldots \lambda_p \mu_1 \ldots \mu_q}} \log\left(\frac{1}{\alpha}\right), \tag{15}$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$, $C\left(M_f,p,q,\beta\right)$ and $C(R,\delta)$ are positive constants depending only on their arguments.

One can now determine optimal bandwidths (λ^*, μ^*) in order to minimize the right-hand side of Equation (15). To do so, the idea is to find for which (λ,μ) both terms in the right hand side of (15) are of the same order w.r.t. n. We also provide an upper bound for the uniform separation rate of the optimized test $\Delta_\alpha^{\lambda^*,\mu^*}$ on Sobolev balls.

Corollary 2. *Consider the assumptions of Theorem 2, and define for all i in $\{1,\dots,p\}$ and for all j in $\{1,\dots,q\}$,*

$$\lambda_i^* = \mu_j^* = n^{-2/(4\delta+p+q)}.$$

The uniform separation rate of the test $\Delta_\alpha^{\lambda^*,\mu^*}$ over the Sobolev ball $\mathcal{S}_{p+q}^\delta(R)$ is controlled as follows

$$\rho\left(\Delta_\alpha^{\lambda^*,\mu^*}, \mathcal{S}_{p+q}^\delta(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta, R\right) n^{-2\delta/(4\delta+p+q)}. \tag{16}$$

Note that, in the definition of the Sobolev ball $\mathcal{S}_{p+q}^\delta(R)$, we have the same regularity parameter $\delta > 0$ for all the directions in $\mathbb{R}^{p+q}$. This corresponds to isotropic regularity conditions. We now introduce other classes of functions allowing to take into account possible anisotropic regularity properties.

2.4.2 Case of Nikol'skii-Besov balls

For $d \in \mathbb{N}^*$, $\delta = (\delta_1, ..., \delta_d) \in (0, +\infty)^d$ and $R > 0$, we consider the anisotropic Nikol'skii-Besov ball $\mathcal{N}_{2,d}^\delta(R)$ defined by

$$\mathcal{N}_{2,d}^\delta(R) = \Bigg\{ s : \mathbb{R}^d \to \mathbb{R} / s \text{ has continuous partial derivatives } D_i^{\lfloor \delta_i \rfloor} \text{ of order } \lfloor \delta_i \rfloor \text{ w.r.t } u_i, \text{and } \forall i = 1, ..., d,$$
$$u_1, ..., u_d, v \in \mathbb{R}, \|D_i^{\lfloor \delta_i \rfloor} s(u_1, ..., u_i + v, ..., u_d) - D_i^{\lfloor \delta_i \rfloor} s(u_1, ..., u_d)\|_2 \leq R|v|^{\delta_i - \lfloor \delta_i \rfloor} \Bigg\}, \tag{17}$$

where $\lfloor \delta_i \rfloor$ denotes the floor function of δ_i if δ_i is not integer and $\lfloor \delta_i \rfloor = \delta_i - 1$ if δ_i is an integer. We give in the following proposition an upper bound of the bias term, similar to that of Lemma 3, in the case when $f - f_1 \otimes f_2$ belongs to particular Nikol'skii-Besov balls.

Lemma 4. *We assume that $\psi \in \mathcal{N}_{2,p+q}^\delta(R)$, where $\delta = (\nu_1, ..., \nu_p, \gamma_1, ..., \gamma_q) \in (0, 2]^{p+q}$. Then, we have the following inequality,*

$$\|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2 \leq C(R, \delta) \left[\sum_{i=1}^p \lambda_i^{2\nu_i} + \sum_{j=1}^q \mu_j^{2\gamma_j} \right].$$

As in Section 2.4.1, one can deduce from Theorem 1 upper bounds for the uniform separation rates of the test $\Delta_\alpha^{\lambda,\mu}$ over Nikol'skii-Besov balls.

Theorem 3. *Let $\alpha, \beta \in (0, 1)$ and consider the same notation and assumptions as in Theorem 1. Let $\delta = (\nu_1, ..., \nu_p, \gamma_1, ..., \gamma_q) \in (0, 2]^{p+q}$ and $R > 0$. Then, the uniform separation rate $\rho\left(\Delta_\alpha^{\lambda,\mu}, \mathcal{N}_{2,p+q}^\delta(R), \beta\right)$ defined in (6) over the Nikol'skii-Besov ball $\mathcal{N}_{2,p+q}^\delta(R)$ can be upper bounded as follows*

$$\left[\rho\left(\Delta_\alpha^{\lambda,\mu}, \mathcal{N}_{2,p+q}^\delta(R), \beta\right)\right]^2 \leq C(R, \delta) \left[\sum_{i=1}^p \lambda_i^{2\nu_i} + \sum_{j=1}^q \mu_j^{2\gamma_j} \right] + \frac{C\left(M_f, p, q, \beta\right)}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \log\left(\frac{1}{\alpha}\right). \tag{18}$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$, $C\left(M_f, p, q, \beta\right)$ and $C(R, \delta)$ are positive constants depending only on their arguments.

As in Section 2.4.1, we now determine optimal bandwidths (λ^*, μ^*) which minimize the right-hand side of Equation (18) and compute an upper bound for the uniform separation rate of the optimized test $\Delta_\alpha^{\lambda^*,\mu^*}$ on Nikol'skii-Besov balls.

Corollary 3. *Consider the assumptions of Theorem 3, and define for all i in $\{1,\dots,p\}$ and for all j in $\{1,\dots,q\}$,*

$$\lambda_i^* = n^{-2\eta/[\nu_i(1+4\eta)]} \quad \text{and} \quad \mu_j^* = n^{-2\eta/[\gamma_j(1+4\eta)]},$$

where η is defined by $\dfrac{1}{\eta} = \displaystyle\sum_{i=1}^{p} \frac{1}{\nu_i} + \sum_{j=1}^{q} \frac{1}{\gamma_j}$,

The uniform separation rate of the test $\Delta_\alpha^{\lambda^*,\mu^*}$ over the Nikol'skii-Besov ball $\mathcal{N}_{2,p+q}^{\delta}(R)$ is controlled as follows

$$\rho\left(\Delta_\alpha^{\lambda^*,\mu^*}, \mathcal{N}_{2,p+q}^{\delta}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right) n^{-2\eta/(1+4\eta)}. \tag{19}$$

Notice that the upper bound obtained for Nikol'skii-Besov balls in Corollary 3 is analogue to that obtained for Sobolev balls in Corollary 2. Indeed, if we consider the same regularities in all directions in the case of Nikol'skii-Besov balls: $\nu_1 = \ldots = \nu_p = \gamma_1 = \ldots = \gamma_q$, we obtain a similar upper bound. These upper bounds obtained in Corollaries 2 and 3 remind the asymptotic minimax separation rate of testing independence w.r.t. the $\mathbb{L}_2$-norm over Hölder spaces [Ingster, 1989, Yodé, 2004]. However, the test having a rate with the smallest upper bound is not adaptive, it depends on the regularity parameter δ. In the next section, for the purpose of adaptivity, we build an aggregated testing procedure taking into account a collection of bandwidths $(\lambda,\mu) \in \Lambda \times U$. In particular, this avoids the delicate choice of arbitrary bandwidths. We then prove that the uniform separation rate of this aggregated procedure is of the same order as the smallest uniform separation rate of the chosen collection, up to a logarithmic term.

3 Aggregated non-asymptotic kernel-based test

In Section 2, we consider single tests based on Gaussian kernels associated to a particular choice of the bandwidths (λ,μ). However, applying such a procedure leads to the question of the choice of these parameters. There is as yet no justified method to choose λ and μ. In many cases, authors choose these parameters w.r.t the available data $(X_i,Y_i)_{1\leq i\leq n}$ by taking for example λ (resp. μ) as the empirical median or standard deviation of the X_i's (resp. the Y_i's), which is not necessarily an optimal choice. To avoid this delicate choice, we consider in this section an aggregated testing procedure combining a collection of single tests based on different bandwithds.

3.1 The aggregated testing procedure

Consider now a finite or countable collection $\Lambda \times U$ of bandwidths in $(0,+\infty)^p \times (0,+\infty)^q$. Consider a collection of positive weights $\left\{\omega_{\lambda,\mu} \,/\, (\lambda,\mu) \in \Lambda \times U\right\}$ such that $\sum_{(\lambda,\mu)\in\Lambda\times U} e^{-\omega_{\lambda,\mu}} \leq 1$.

For a given $\alpha \in (0,1)$, we define the aggregated test which rejects $(\mathcal{H}_0)$ if there is at least one $(\lambda,\mu) \in \Lambda \times U$ such that

$$\widehat{\mathrm{HSIC}}_{\lambda,\mu} \;>\; q^{\lambda,\mu}_{1-u_\alpha e^{-\omega_{\lambda,\mu}}},$$

where u_α is the less conservative value such that the test is of level α, and is defined by

$$u_\alpha = \sup\left\{u>0 \;;\; P_{f_1\otimes f_2}\left(\sup_{(\lambda,\mu)\in\Lambda\times U}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - q^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}\right) > 0\right) \leq \alpha\right\}. \tag{20}$$

We should mention here that the supremum in Equation (20) exists. Indeed, one may notice that the function

$$u \mapsto P_{f_1\otimes f_2}\left(\sup_{(\lambda,\mu)\in\Lambda\times U}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - q^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}\right) > 0\right)$$

is well defined for u in the interval $(0, \inf\{\exp(\omega_{\lambda,\mu}); \lambda\in\Lambda, \mu\in U\})$, non-decreasing, and converges to 0 and 1 respectively at the boundaries of this interval.

The test function Δ_α associated to this aggregated test, takes values in $\{0,1\}$ and is defined by

$$\Delta_\alpha = 1 \quad \Longleftrightarrow \quad \sup_{(\lambda,\mu)\in\Lambda\times U} \left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - q^{\lambda,\mu}_{1-u_\alpha e^{-\omega_{\lambda,\mu}}} \right) > 0. \tag{21}$$

It is easy to check that the test Δ_α is of level α, this is directly derived from the definition of u_α. Note that in order to guarantee the level of the aggregated procedure, the level α of each single test $\Delta^{\lambda,\mu}_\alpha$ is here replaced by the *corrected level* $u_\alpha \exp(-\omega_{\lambda,\mu})$.

For computational limitations, the collections Λ and U are finite in practice. Moreover, note that, as for the quantile, the correction u_α of the level is not analytically computable since it depends on the unknown marginals f_1 and f_2. In practice, it can also be approached by a permutation method with Monte Carlo approximation, as done in [Albert, 2015]. More precisely, consider the notations of Section 2.1. First, generate B_1 independent and uniformly distributed random permutations of $\{1,...,n\}$, denoted $\tau_1,...,\tau_{B_1}$, independent of $\mathbb{Z}_n$ and compute for each $(\lambda,\mu)\in\Lambda\times U$ and each $u>0$ the permuted quantile with Monte Carlo approximation $\hat{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}$ as defined in (9).
Second, in order to estimate the probabilities under $P_{f_1\otimes f_2}$ in Equation (20), generate B_2 independent and uniformly distributed random permutations of $\{1,...,n\}$, denoted $\kappa_1,...,\kappa_{B_2}$, independent of $\mathbb{Z}_n$ and of $\tau_1,...,\tau_{B_1}$. Denote for all permutation κ_b, the corresponding permuted statistic

$$\widehat{H}^{\kappa_b}_{\lambda,\mu} = \widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}^{\kappa_b}_n\right)$$

Then, the correction u_α is approached by

$$\hat{u}_\alpha = \sup\left\{ u>0 \ ; \ \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}_{\max_{(\lambda,\mu)\in\Lambda\times U}\left\{\widehat{H}^{\kappa_b}_{\lambda,\mu} - \hat{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}\right\}>0} \leq \alpha \right\}. \tag{22}$$

In the end, the aggregated testing procedure with permutation approach rejects the null hypothesis if

$$\max_{(\lambda,\mu)\in\Lambda\times U}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - \hat{q}^{\lambda,\mu}_{1-\hat{u}_\alpha e^{-\omega_{\lambda,\mu}}} \right) > 0.$$

In the next section, we will provide a uniform separation rate similar to that of Corollaries 2 and 3 for the aggregated test Δ_α. This uniform separation rate will be given in the two cases mentioned earlier in Section 2.4 where $f - f_1\otimes f_2$ belongs to isotropic Sobolev balls or to anisotropic Nikol'skii-Besov balls.

3.2 Oracle type conditions for the second kind error

As a reminder, our goal is to construct a testing procedure with a uniform separation rate as small as possible and whose implementation does not require any information about the regularity of the difference $f - f_1\otimes f_2$.

The main advantage of the aggregated procedure is that its second kind error is as small as the one of the single test corresponding to the best bandwidths in the collection $\Lambda\times U$ with a corrected level. The main argument is highlighted in Lemma 5.

Lemma 5. *Let α, β in $(0,1)$, and consider the aggregated test Δ_α defined in (21). Then, $u_\alpha \geq \alpha$ and*

$$P_f\left(\Delta_\alpha = 0\right) \leq \inf_{(\lambda,\mu)\in\Lambda\times U}\left\{ P_f\left(\Delta^{\lambda,\mu}_{\alpha e^{-\omega_{\lambda,\mu}}} = 0\right)\right\}.$$

According to Lemma 5, if there exists at least one $(\lambda,\mu)\in\Lambda\times U$ such that the associated single test $\Delta^{\lambda,\mu}_{\alpha e^{-\omega_{\lambda,\mu}}}$ has a probability of second kind error at most equal to β, then the probability of the second kind error of the aggregated test Δ_α is at most equal to β.

We now give an oracle inequality for the uniform separation rate of the aggregation procedure Δ_α. This inequality given in the following theorem shows the interest of this testing procedure.

Theorem 4. *Let $\alpha, \beta \in (0,1)$, $\left\{(k_\lambda, l_\mu) \,/\, (\lambda,\mu) \in \Lambda \times U\right\}$ a collection of Gaussian kernels and $\left\{\omega_{\lambda,\mu} \,/\, (\lambda,\mu) \in \Lambda \times U\right\}$ a collection of positive weights, such that $\sum_{(\lambda,\mu)\in\Lambda\times U} e^{-\omega_{\lambda,\mu}} \leq 1$. We also assume that all bandwidths (λ,μ) in $\Lambda \times U$ verify the conditions given in Theorem 1, and that f, f_1 and f_2 are bounded. Then, the test Δ_α of level α defined in Equation (21) has a uniform separation rate $\rho(\Delta_\alpha, C_\delta, \beta)$ which can be upper bounded as follows*

- *If $C_\delta = \mathcal{S}^\delta_{p+q}(R)$, where $\delta \in (0,2]$ and $R > 0$, then*

$$\left[\rho\left(\Delta_\alpha, \mathcal{S}^\delta_{p+q}(R), \beta\right)\right]^2 \leq C\left(M_f, p, q, \beta, \delta\right) \inf_{(\lambda,\mu)\in\Lambda\times U} \left\{ \frac{1}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \left(\log\left(\frac{1}{\alpha}\right) + \omega_{\lambda,\mu}\right) + \left[\sum_{i=1}^p \lambda_i^{2\delta} + \sum_{j=1}^q \mu_j^{2\delta}\right] \right\},$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$ and $C\left(M_f, p, q, \beta, \delta\right)$ is a positive constant depending only on its arguments.

- *If $C_\delta = \mathcal{N}^\delta_{2,p+q}(R)$, where $\delta = (\nu_1, ..., \nu_p, \gamma_1, ..., \gamma_q) \in (0,2]^{p+q}$ and $R > 0$, then*

$$\left[\rho\left(\Delta_\alpha, \mathcal{N}^\delta_{2,p+q}(R), \beta\right)\right]^2 \leq C\left(M_f, p, q, \beta, \delta\right) \inf_{(\lambda,\mu)\in\Lambda\times U} \left\{ \frac{1}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \left(\log\left(\frac{1}{\alpha}\right) + \omega_{\lambda,\mu}\right) + \left[\sum_{i=1}^p \lambda_i^{2\nu_i} + \sum_{j=1}^q \mu_j^{2\gamma_j}\right] \right\},$$

where $C\left(M_f, p, q, \beta, \delta\right)$ is a positive constant depending only on its arguments.

According to Theorem 4, the uniform separation rate of the aggregated testing procedure Δ_α is the infimum of all $(\lambda,\mu) \in \Lambda \times U$, up to the additional term $\omega_{\lambda,\mu}$. This theorem can also be interpreted as an oracle type condition for the second kind error of the test Δ_α. Indeed, without knowing the regularity of $f - f_1 \otimes f_2$, we prove that the uniform separation rate of Δ_α is of the same order as the smallest uniform separation rate over $(\lambda,\mu) \in \Lambda \times U$, up to $\omega_{\lambda,\mu}$.

3.3 Uniform separation rate over Sobolev balls and Nikol'skii-Besov balls

In this section, we provide an upper bound of the uniform separation rate $\rho(\Delta_\alpha, C_\delta, \beta)$ of the multiple testing procedure Δ_α over the classes of Sobolev balls and Nikol'skii-Besov balls. For this, we consider the collections Λ and U of parameters λ and μ respectively, defined by

$$\Lambda = \{(2^{-m_{1,1}}, \ldots, 2^{-m_{1,p}}) \;;\; (m_{1,1}, \ldots, m_{1,p}) \in (\mathbb{N}^*)^p\}, \tag{23}$$

and

$$U = \{(2^{-m_{2,1}}, \ldots, 2^{-m_{2,q}}) \;;\; (m_{2,1}, \ldots, m_{2,q}) \in (\mathbb{N}^*)^q\}. \tag{24}$$

In addition, we associate to every $\lambda = (2^{-m_{1,1}}, \ldots, 2^{-m_{1,p}})$ in Λ and $\mu = (2^{-m_{2,1}}, \ldots, 2^{-m_{2,q}})$ in U the positive weights

$$\omega_{\lambda,\mu} = 2\sum_{i=1}^p \log\left(m_{1,i} \times \frac{\pi}{\sqrt{6}}\right) + 2\sum_{j=1}^q \log\left(m_{2,j} \times \frac{\pi}{\sqrt{6}}\right), \tag{25}$$

so that $\sum_{(\lambda,\mu)\in\Lambda\times U} e^{-\omega_{\lambda,\mu}} = 1$. Corollary 4 provides these upper bounds.

Corollary 4. *Assuming that $\log\log(n) > 1$, $\alpha, \beta \in (0,1)$ and Δ_α the test defined in (21), with the particular choice of Λ, U and the weights $(\omega_{\lambda,\mu})_{(\lambda,\mu)\in\Lambda\times U}$ defined in (23), (24) and (25). Then, the uniform separation rate $\rho(\Delta_\alpha, C_\delta, \beta)$ of the aggregated test Δ_α can be upper bounded as follows.*

- *If* $C_\delta = \mathcal{S}^\delta_{p+q}(R)$*, where* $\delta \in (0,2]$ *and* $R > 0$*, then,*

$$\rho\left(\Delta_\alpha, \mathcal{S}^\delta_{p+q}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right)\left(\frac{\log\log(n)}{n}\right)^{2\delta/(4\delta+p+q)}, \tag{26}$$

where $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$.

- *If* $C_\delta = \mathcal{N}^\delta_{2,p+q}(R)$*, where* $\delta = (\nu_1, \ldots, \nu_p, \gamma_1, \ldots, \gamma_q) \in (0,2]^{p+q}$ *and* $R > 0$*, then,*

$$\rho\left(\Delta_\alpha, \mathcal{N}^\delta_{2,p+q}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right)\left(\frac{\log\log(n)}{n}\right)^{2\eta/(1+4\eta)}, \tag{27}$$

where $\dfrac{1}{\eta} = \displaystyle\sum_{i=1}^{p} \frac{1}{\nu_i} + \sum_{j=1}^{q} \frac{1}{\gamma_j}$ *and* $M_f = \max\left(\|f\|_\infty, \|f_1\|_\infty, \|f_2\|_\infty\right)$.

According to Corollary 4, the rate of the aggregation procedure over the classes of Sobolev balls and Nikol'skii-Besov balls is in the same order of the best rate of single tests (given in Theorem 1), up to a $\log\log(n)$ factor.

4 Lower bound for the non-asymptotic minimax separation rates over Sobolev balls.

In this section, we study the optimality of the single test introduced in Corollary 2 and of the aggregated testing procedure defined in Corollary 4 over Sobolev balls. For this, we first present a general method based on a Bayesian approach to lower bound the non-asymptotic minimax separation rate of testing as defined in (7). The general idea of this method is due to [Ingster, 1993a] and relies on Lemma 6.

Lemma 6. *Let* α, β *in* $(0,1)$ *such that* $\alpha + \beta < 1$*. Let* $\rho_* > 0$ *and* $\mathcal{C}_\delta$ *some regularity space. We recall that the set* $\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)$ *is the set defined as*

$$\mathcal{F}_{\rho_*}(\mathcal{C}_\delta) = \{f; f - f_1 \otimes f_2 \in \mathcal{C}_\delta, \|f - f_1 \otimes f_2\|_2 \geq \rho_*\}.$$

Let ν_{ρ_*} *be a probability measure on* $\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)$ *and* $P_{\nu_{\rho_*}}$ *the probability measure defined for all measurable set* A *in* $\mathbb{R}^{p+q}$ *by*

$$P_{\nu_{\rho_*}}(A) = \int P_f(A)\, \mathrm{d}\nu_{\rho_*}(f).$$

Denote for all $\rho > 0$*,*

$$\beta\left[\mathcal{F}_\rho(\mathcal{C}_\delta)\right] = \inf_{\Delta_\alpha} \sup_{f \in \mathcal{F}_\rho(\mathcal{C}_\delta)} P_f\left(\Delta_\alpha = 0\right),$$

where the infimum is taken over all α*-level tests of* $(\mathcal{H}_0)$ *against* $(\mathcal{H}_1)$*. Assume there exists a distribution* f_0 *that satisfies* $(\mathcal{H}_0)$ *such that the probability measure* $P_{\nu_{\rho_*}}$ *is absolutely continuous w.r.t.* P_{f_0} *and verifies*

$$\mathbb{E}_{P_{f_0}}\left[L^2_{\nu_{\rho_*}}(\mathbb{Z}_n)\right] < 1 + 4(1-\alpha-\beta)^2, \tag{28}$$

where the likelihood ratio $L_{\nu_{\rho_*}}$ *is defined by* $L_{\nu_{\rho_*}} = \dfrac{\mathrm{d}P_{\nu_{\rho_*}}}{\mathrm{d}P_{f_0}}$*. Then, for all* $\rho \leq \rho_*$ *we have that*

$$\beta\left[\mathcal{F}_\rho(\mathcal{C}_\delta)\right] > \beta.$$

It follows that

$$\rho\left(\mathcal{C}_\delta, \alpha, \beta\right) = \inf_{\Delta_\alpha} \rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right) \geq \rho_*.$$

We aim at proving that

$$\rho_n^* = C n^{-2\delta/(4\delta+p+q)}$$

is a lower bound for the non-asymptotic minimax rate of testing over Sobolev balls $\mathcal{S}^\delta_{p+q}(R)$, for some positive constant C, that is, $\rho\left(\mathcal{S}^\delta_{p+q}(R), \alpha, \beta\right) \geq \rho_n^*$. According to Lemma 6, it is sufficient to find a probability distribution $\nu_{\rho_n^*}$ over $\mathcal{F}_{\rho_n^*}(\mathcal{S}^\delta_{p+q}(R))$ so that the condition (28) holds.

To do so, we generalize the construction of [Butucea, 2007] to our multidimensional framework. The idea is to construct a finite set of alternatives $(f_\theta)_\theta$ by perturbing the uniform density on $[0,1]^p \times [0,1]^q$, and define $\nu_{\rho_n^*}$ as a uniform mixture of these alternatives. For this, consider the function G defined for all t in $\mathbb{R}$ by

$$G(t) = \exp\left(-\frac{1}{1-(4t+3)^2}\right)\mathbb{1}_{(-1,-1/2)}(t) - \exp\left(-\frac{1}{1-(4t+1)^2}\right)\mathbb{1}_{(-1/2,0)}(t). \tag{29}$$

One may notice that G is continuous, with support in $[-1,0]$ and that $\int_{\mathbb{R}} G(t)\mathrm{d}t = 0$. The function G together with its Fourier transform has valuable properties for our study.

Let $(h_n)_n$ be a sequence of positive numbers to be specified later, and consider an integer M_n such the $M_n h_n = 1$ (possibly rounded to the nearest integer). Denote $I_{n,p,q} = \{1,\dots,M_n\}^p \times \{1,\dots,M_n\}^q$. For all $\theta = (\theta_{(j,l)})_{(j,l)\in I_{n,p,q}}$ in $\{-1,1\}^{M_n^{p+q}}$, define for all (x,y) in $\mathbb{R}^p \times \mathbb{R}^q$,

$$f_\theta(x,y) = \mathbb{1}_{[0,1]^{p+q}}(x,y) + h_n^{\delta+(p+q)} \sum_{(j,l)\in I_{n,p,q}} \theta_{(j,l)} \prod_{r=1}^{p} G_{h_n}(x_r - j_r h_n) \prod_{s=1}^{q} G_{h_n}(y_s - l_s h_n), \tag{30}$$

where for all $h > 0$, $G_h(\cdot) = (1/h)\, G(\cdot/h)$. One may notice that for all θ, the alternative f_θ is supported in $[0,1]^{p+q}$. Moreover, since the integral of G over $\mathbb{R}$ equals 0, the marginals $f_{\theta,1}$ and $f_{\theta,2}$ of f_θ are respectively the uniform densities on $[0,1]^p$ and $[0,1]^q$. Proposition 5 justifies the choice of these alternatives.

Proposition 5. *Let $\delta > 0$ and $R > 0$. Fix a sequence $(h_n)_n$ of positive numbers and consider an integer M_n such the $M_n h_n = 1$. Then, for all $\theta = (\theta_{(j,l)})_{(j,l)\in I_{n,p,q}}$ in $\{-1,1\}^{M_n^{p+q}}$, the function f_θ defined in Equation (30) satisfies the following properties.*

1. *The function f_θ is a density function for h_n small enough.*
2. *The function $f_\theta - f_{\theta,1} \otimes f_{\theta,2}$ belongs to the Sobolev ball $\mathcal{S}^\delta_{p+q}(R)$ for n large enough.*
3. *The function f_θ is such that $\|f_\theta - f_{\theta,1} \otimes f_{\theta,2}\|_2 = C(p,q)h_n^\delta$.*

Let us now consider a uniform mixture $\nu_{\rho_n^*}$ of the alternatives (f_θ), for θ in $\{-1,1\}^{M_n^{p+q}}$. Note that this is equivalent to considering a random alternative f_Θ where $\Theta = (\Theta_{(j,l)})_{(j,l)\in I_{n,p,q}}$ with i.i.d. Rademacher components $\Theta_{(j,l)}$.
Following Lemma 6, let P_ν be the probability measure defined for all measurable set A in $\mathbb{R}^{p+q}$ by

$$P_\nu(A) = \int_{\{-1,1\}^{M_n^{p+q}}} P_{f_\theta}(A)\pi(\mathrm{d}\theta) = \frac{1}{2^{M_n^{p+q}}} \sum_{\theta\in\{-1,1\}^{M_n^{p+q}}} P_{f_\theta}(A), \tag{31}$$

where π is the distribution of a (M_n^{p+q})-sample of i.i.d. Rademacher random variables. Proposition 6 justifies the use of these alternatives and this probability measure to prove the lower bound.

Proposition 6. *Let α, β in $(0,1)$ such that $\alpha + \beta < 1$ and let $\delta > 0$. Denote f_0 the uniform density on $[0,1]^{p+q}$. There exists some positive constant $C(p,q,\alpha,\beta,\delta)$ such that, if we set*

$$h_n = C(p,q,\alpha,\beta,\delta) n^{-2/(4\delta+p+q)}, \tag{32}$$

and define P_ν by Equations (30) and (31), then we have

$$\mathbb{E}_{P_{f_0}}\left[\left(\frac{\mathrm{d}P_\nu}{\mathrm{d}P_{f_0}}(\mathbb{Z}_n)\right)^2\right] < 1 + 4(1-\alpha-\beta)^2,$$

for n large enough.

Finally, combining Lemma 6 with Propositions 5 and 6, we obtain a lower bound for the non-asymptotic minimax separation rate of testing in Theorem 5.

Theorem 5. *Consider α, β in $(0,1)$ such that $\alpha + \beta < 1$. Let $\delta > 0$ and $R > 0$. Then, there exists a positive constant $C(p, q, \alpha, \beta, \delta, R)$ such that*

$$\rho\left(\mathcal{S}^{\delta}_{p+q}(R), \alpha, \beta\right) \geq C(p, q, \alpha, \beta, \delta, R)\ n^{-2\delta/(4\delta+p+q)},$$

for n large enough.

Theorem 5 proves that each single test introduced in Corollary 2 is optimal in the minimax sense over Sobolev balls $\mathcal{S}^{\delta}_{p+q}(R)$ for δ in $(0,2]$ since the upper and lower bounds coincide up to constants. Moreover, the aggregated testing procedure defined in Corollary 4 is optimal up to a logarithmic term over Sobolev balls. Since it does not depend on the prior knowledge of the regularity parameter δ, it is adaptive.

5 Numerical simulations

In this section, numerical simulations are performed in order to study the practical validity of our testing procedures. More precisely, after describing the different simulated alternatives, we first compare the "theoretical" single tests (for which we have proved optimality) and the permutation-based single tests (which are applied in practice) in terms of power. A similar verification is also carried out for the aggregated procedure. Then, we compare different possible strategies of aggregation based on different choices of bandwidth collections and of weights. Finally, a comparison with the nonparametric independence testing procedure based on the mutual information of [Berrett and Samworth, 2017] is done for different data generating mechanisms provided by the authors.

5.1 Data generating mechanisms

In this simulation study, different data generating mechanisms are used.

1. First, in order to compare the theoretical and the permuted tests, we rely on the following data generating mechanism inspired from the Ishigami function in [Ishigami and Homma, 1990].

   Let U, V and W be independent uniform random variables on $[0,1]$, we define the variables (X, Y) by

   $$X = U \quad \text{and} \quad Y = \sin(U) + 4\sin^2(V) + 0.5W^4 \sin(U). \tag{33}$$

2. Then, in order to study the performance of the permuted tests in terms of power and compare it to existing procedures, we consider the same data generating mechanisms as in [Berrett and Samworth, 2017], defined bellow.

   (i) Define the joint density $f_{[l]}$ of the couple (X, Y) for all (x, y) in $[-\pi, \pi]$ by

   $$f_{[l]}(x, y) = \frac{1}{4\pi^2}\left\{1 + \sin(lx)\sin(ly)\right\}.$$

   Densities $f_{[l]}$ with $l = 1 \dots 10$ are considered here.

   (ii) Define X and Y as

   $$X = L\cos\Theta + \frac{\varepsilon_1}{4}, \quad Y = L\sin\Theta + \frac{\varepsilon_2}{4},$$

   where L, Θ, ε_1 and ε_2 are independent, with L is uniformly distributed on $\{1, \dots, l\}$ for some l in $\{1, \dots, 10\}$, Θ is uniformly distributed on $[0, 2\pi]$ and ε_1, ε_2 are standard normal random variables.

   (iii) Let X be a uniform random variable on $[-1, 1]$. For a given $\rho \geq 0$, Y is defined as

   $$Y = |X|^{\rho}\varepsilon,$$

   where ε is a standard normal random variable independent with X. The considered values of ρ are $0.1, 0.2, \dots, 1$.

   In addition, we also consider the following bi-variate case: $X = (X^{(1)}, X^{(2)})$ and $Y = (Y^{(1)}, Y^{(2)})$ where $(X^{(1)}, Y^{(1)})$ is generated according to any described mechanism above, while $X^{(2)}$, $Y^{(2)}$ are independent uniform random variables on $[0,1]$ and independent from $(X^{(1)}, Y^{(1)})$.

5.2 Power comparison between the theoretical and the permuted tests

In this section, we first numerically illustrate that the power of the permuted single HSIC tests approximate very well the power of the theoretical tests, as soon as enough permutations are used for the estimation of the critical value (that is the quantile under the null hypothesis). In particular, the level of the permuted tests being guaranteed by Proposition 1, the permutation approach does not affect the quality of these tests. Thereafter, a similar study is conducted for the aggregated testing procedure introduced in Section 3.1.

All along this section, we rely on the data generating mechanism inspired from the Ishigami function, defined in Equation (33).

5.2.1 Single tests comparison

In order to evaluate the accuracy of permuted single HSIC tests, we choose the kernel bandwidth associated to X (resp. Y) to be the empirical standard deviation of X (resp. Y), which is a usual choice in the literature on single HSIC-test [Zhang et al., 2012],

In the following, we illustrate the power for three sample sizes n in $\{50, 100, 200\}$ and two test levels α in $\{0.05, 0.001\}$. For each sample size n and test level α, we first estimate the power of the theoretical test. To achieve this, we simulate 500.000 n-samples under the null hypothesis[1] and estimate the theoretical $(1-\alpha)$-quantile (denoted $q_{1-\alpha}$) by Monte Carlo. Then, we generate 1000 different n-samples of (X, Y) under the alternative (according to (33)) and we estimate the "theoretical" power $\beta_{th}(n, \alpha)$ as being the ratio of times that the observed test-statistic $\widehat{\mathrm{HSIC}}$ exceeds the quantile $q_{1-\alpha}$.

The second step consists in estimating the power of permuted tests for several values of the number of permutations B. The chosen values of B are $\{10, 20, \ldots, 100, 200, \ldots, 2500\}$. For each value of n, α and B, we generate 1000 n-sample of (X, Y) according to (33). For each n-sample, we compute the permuted quantile $\widehat{q}_{1-\alpha}$ defined in Equation (9) using B random permutations of this sample. Thereafter, we compute the power $\beta(n, \alpha, B)$ of the permuted test, as being the ratio of times the value of $\widehat{\mathrm{HSIC}}$ exceeds the permuted quantile $\widehat{q}_{1-\alpha}$ (computed on the corresponding sample).

To compare the powers of theoretical and permuted tests (resp. $\beta_{th}(n, \alpha)$ and $\beta(n, \alpha, B)$), we consider the relative absolute error $Err(n, \alpha, B)$ defined as

$$Err(n, \alpha, B) = \frac{|\beta(n, \alpha, B) - \beta_{th}(n, \alpha)|}{\beta_{th}(n, \alpha)}.$$

The results obtained for $\alpha = 0.05$ and different n values are given by Figure 1. We can see that the accuracy of the permuted approach tends to increase as n increases. This is probably due to the fact that the power of the theoretical test increases as the sample size increases. Another explanation may be that, on the one hand, the power of the theoretical test is more difficult to estimate for small sizes, which explains the fluctuations observed for $n = 50$. On the other hand, as n increases, the approximation of the distribution of $\widehat{\mathrm{HSIC}}$ under the null hypothesis thanks to B permutations becomes more accurate, and this for any value of B larger than 500. Hence, the approximation of the quantile by permutation becomes more accurate, and thus, there are less fluctuations for larger sample sizes.

Generally, the permutation approach allows to obtain the theoretical test power with an acceptable level of precision, even for small values of B. In particular, we observe for $n = 50$ that aside from very small values of B and two outliers, the absolute relative error is always lower than 10%. Moreover, from $n = 100$ this error is mostly less than 10% and no observed error is more than 5% for $n = 200$.

In order to study the impact of the level on the accuracy of the permutation approximation, we show in Figure 2 the relative absolute error of the power w.r.t. n and B for the extreme level value $\alpha = 0.001$. Contrary to the case $\alpha = 0.05$, we observe here much less precision of the power approximation. In particular, for $n = 50$, $B = 2000$ permutations are required to obtain satisfactory accuracy (against $B = 30$ for $\alpha = 0.05$). Similar observations are done for $n = 100$ and 200 with respectively $B = 1200$ and $B = 500$ permutations required (against $B = 30$ and $B = 10$ for $\alpha = 0.05$). This slow convergence results from the difficulty of estimating the extreme quantile $q_{1-\alpha}$, this phenomenon being more significant for

[1] To generate an independent n-sample of (X, Y) under the null hypothesis, we first generate an independent $2n$-sample of X. Only the first n elements are used to compute the marginal sample of Y and the remaining n elements are considered to be the marginal sample of X.

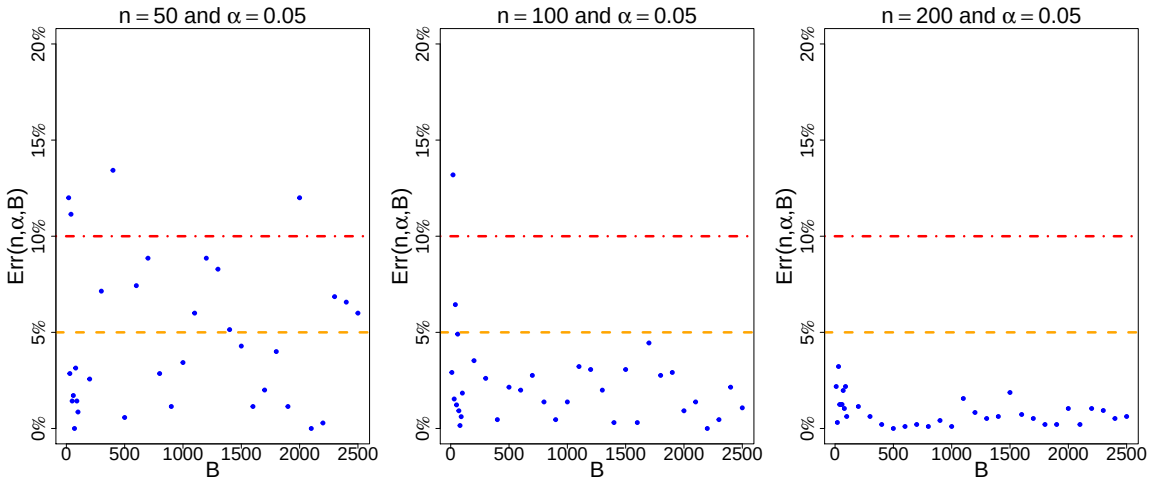

Figure 1: Absolute relative error between the powers of the theoretical and permuted HSIC-tests, w.r.t the number B of permutations, for sample sizes $n = 50, 100$ and 200. The presumed level of tests is $\alpha = 0.05$. The red (resp. orange) dashed line represents the error threshold of 10% (resp. 5%).

small-size samples. Indeed, as in the previous case, the lowest the power of the test, the biggest its sensitivity to the quantile estimation error.

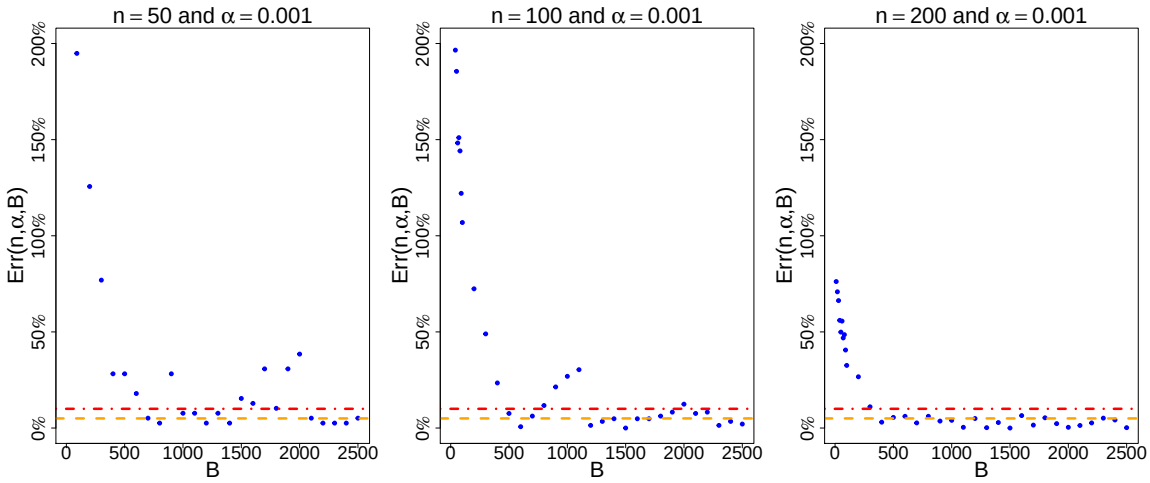

Figure 2: Absolute relative error between the powers of the theoretical and the permuted HSIC-tests, w.r.t the number B of permutations, for sample sizes $n = 50, 100$ and 200. The presumed level of tests is $\alpha = 0.001$. The red (resp. orange) dashed line represents the error threshold of 10% (resp. 5%).

5.2.2 Aggregated testing procedure comparison

Similarly to the previous section, the objective here is to check that the permutation approach does not impact the quality of the aggregated HSIC procedure. The practical implementation of the theoretical and permuted aggregated testing procedures are described in Algorithms 1 and 2. They both require the estimation of the value of u_α defined in Equation (20). A very straightforward approach to do so is to proceed by dichotomy on the interval $[\alpha, M]$, where $M = \inf_{(\lambda,\mu)\in\Lambda\times U}\{e^{\omega_{\lambda,\mu}}\}$ (u_α belonging to this interval as proved in Section 3.1). More precisely, we need to estimate for different values of u, the probability

$$P(u) = P_{f_1\otimes f_2}\left(\sup_{(\lambda,\mu)\in\Lambda\times U}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - q^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}\right) > 0\right). \tag{34}$$

Algorithm 1 *Theoretical aggregated procedure*

Input: We are given an n-sample (observation) and a prescribed level α. We are also given a collection of bandwidths $\Lambda \times U$ and a family of weights $(\omega_{\lambda,\mu})_{(\lambda,\mu)\in\Lambda\times U}$.

1. Simulate a first set, denoted set `(A)`, of 500.000 n-samples under the null hypothesis (to estimate the quantiles) and a second set, denoted `(B)` of 1000 n-samples also under the null hypothesis (to estimate the probabilities $P(u)$ defined in Equation (34) for different values of u).
2. Set $u_{min} = \alpha$ and $u_{max} = M$, where $M = \inf_{(\lambda,\mu)\in\Lambda\times U}\{e^{\omega_{\lambda,\mu}}\}$.
3. While $(u_{max} - u_{min}) > 10^{-3}u_{min}$, repeat the following steps.
   (a) Set $u = (u_{min} + u_{max})/2$.
   (b) For all (λ,μ) in $\Lambda\times U$, compute the estimation $\tilde{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}$ of the quantile $q^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}$ by Monte Carlo using the 500.000 samples of set `(A)`.
   (c) Estimate the probability$P(u)$ by Monte Carlo using the 1000 samples of set `(B)`. More precisely, consider $\hat{P}_u$ as the ratio of times at least one $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ is greater than $\tilde{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}$.
   (d) If $\hat{P}_u \leq \alpha$, then set $u_{min} = u$. Else set $u_{max} = u$ and repeat Step 3.
4. Set $\tilde{u}_\alpha = u$ and the quantiles with corrected levels $(\tilde{q}^{\lambda,\mu}_{1-\tilde{u}_\alpha e^{-\omega_{\lambda,\mu}}})_{(\lambda,\mu)\in\Lambda\times U}$.
5. Finally, compute the observed statistics $(\widehat{\mathrm{HSIC}}_{\lambda,\mu})_{(\lambda,\mu)\in\Lambda\times U}$ (on the given observation) and reject the null hypothesis if there is at least one (λ,μ) such that

$$\widehat{\mathrm{HSIC}}_{\lambda,\mu} > \tilde{q}^{\lambda,\mu}_{1-\tilde{u}_\alpha e^{-\omega_{\lambda,\mu}}}.$$

One may notice that in the theoretical case, this probability is approached by Monte Carlo independently on the observation (provided that we can simulate under the null hypothesis) whereas in the permuted case, it is based on samples obtained by permuting the observation.

Theoretical power. From a given sample of size n and for a given collection of bandwidths $\Lambda \times U$, we estimate the theoretical power of the aggregated test as follows.

Since the approximation of the value of u_α and of the quantiles can be done independently of the observation, it can be done once and for all by running Steps 1 to 4 of Algorithm 1.

Then, to estimate the theoretical power, we then generate 1000 i.i.d samples (observations) and for each one, we apply Step 5 of Algorithm 1 and estimate the theoretical power as the proportion of times the aggregated procedure rejects the null hypothesis.

Permuted power. In this paragraph, we describe the aggregated testing procedure with the permutation approach. In particular, we do not assume we are able to simulate under the null hypothesis to compute the correction u_α.

Note that here, Step 3 depends on the observation and needs to be done for each new observation.

Finally, the power of the aggregated test with permutation is then the ratio of times that the null hypothesis is rejected when applied to 1000 n-samples.

Numerical results. In all the following, the prescribed level of the tests is set to $\alpha = 0.05$ and we consider sample sizes n in $\{50, 100, 200\}$.

For our numerical tests, we consider 6 different collections of bandwidths $(\Lambda_r \times U_r)_{2\leq r\leq 7}$, defined for all r by

$$\Lambda_r \times U_r = \left\{1, 1/2, \ldots, 1/2^{r-1}\right\}^2.$$

Moreover, we choose uniform weights defined for all (λ,μ) by

$$\omega_{\lambda,\mu} = \log(r^2). \tag{35}$$

Algorithm 2 *Permuted aggregated procedure*

Input: We are given an n-sample (observation) denote $\mathbb{Z}_n$ and a prescribed level α. We are also given a collection of bandwidths $\Lambda \times U$ and a family of weights $(\omega_{\lambda,\mu})_{(\lambda,\mu)\in\Lambda\times U}$.

1. Generate a first set, say `(A')`, of B_1 i.i.d. random permutations of $\{1, ..., n\}$ (to estimate the quantiles), and independently generate a second set, denoted `(B')`, of B_2 i.i.d. random permutations of $\{1, ..., n\}$ (to estimate the probabilities $P(u)$ defined in Equation (34)), all independent of $\mathbb{Z}_n$.
2. Set $u_{min} = \alpha$ and $u_{max} = M$, where $M = \inf_{(\lambda,\mu)\in\Lambda\times U}\{e^{\omega_{\lambda,\mu}}\}$.
3. While $(u_{max} - u_{min}) > 10^{-3}u_{min}$, repeat the following steps.
   (a) Set $u = (u_{min} + u_{max})/2$.
   (b) For all (λ, μ) in $\Lambda \times U$, compute the permuted quantile with Monte Carlo approximation $\hat{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}$ as defined in (9) using the set `(A')`.
   (c) Estimate $P(u)$ by permutation with Monte Carlo approximation using the set `(B')`. More precisely, consider
   $$\hat{P}^{\star}_u(\mathbb{Z}_n) = \frac{1}{B_2}\sum_{b=1}^{B_2} \mathbb{1}_{\max_{(\lambda,\mu)\in\Lambda\times U}\left\{\widehat{H}^{\kappa_b}_{\lambda,\mu} - \hat{q}^{\lambda,\mu}_{1-ue^{-\omega_{\lambda,\mu}}}\right\}>0},$$
   where $(\kappa_b)_{1\le b\le B_2}$ denote the permutations of set `(B')` and $\widehat{H}^{\kappa_b}_{\lambda,\mu} = \widehat{\mathrm{HSIC}}_{\lambda,\mu}(\mathbb{Z}^{\kappa_b}_n)$ is the statistic computed on the bth permuted sample $\mathbb{Z}^{\kappa_b}_n$.
   (d) If $\hat{P}^{\star}_u(\mathbb{Z}_n) \le \alpha$, then set $u_{min} = u$. Else set $u_{max} = u$ and repeat Step 3.
4. Set $\hat{u}_\alpha = u$ and the quantiles with corrected levels $(\hat{q}^{\lambda,\mu}_{1-\hat{u}_\alpha e^{-\omega_{\lambda,\mu}}})_{(\lambda,\mu)\in\Lambda\times U}$.
5. Finally, compute the observed statistics $(\widehat{\mathrm{HSIC}}_{\lambda,\mu})_{(\lambda,\mu)\in\Lambda\times U}$ (on the given observation) and reject the null hypothesis if there is at least one (λ, μ) such that
   $$\widehat{\mathrm{HSIC}}_{\lambda,\mu} > \hat{q}^{\lambda,\mu}_{1-\hat{u}_\alpha e^{-\omega_{\lambda,\mu}}}.$$

Note that, the case $r = 1$ corresponds to the single test with $\lambda = \mu = 1$.

For the permuted aggregated procedure, the number B_1 of permutations to estimate the quantiles varies in $\{100, 200, 500, 1000, ..., 5000\}$ and the number of permutations used to estimate the probabilities $P(u)$ is set to $B_2 = 500$.

For each triplet (r, n, B_1), the power of both the theoretical and permuted aggregated testing procedures are estimated from 1000 different samples as described above. Results in terms of absolute relative error on power are given by Figure 3. Notice that, regardless of the n value, the required value of B_1 to well approximate the theoretical power increases with r. In fact, the supremum in Equation (22) becomes more difficult to estimate as the number r^2 of aggregated tests increases. Unsurprisingly, for a given B_1 value, the accuracy of the power estimation increases with n as in the case of single tests. In particular, we observe that for $n = 50$, the biggest error becomes less than 10% from $B_1 = 3500$, while this threshold seems to be achieved from $B_1 = 3000$ for $r = 4, 5, 6$ and from $B_1 = 500$ for $r = 2, 3$. For bigger sample sizes $n = 100$ and 200, a good approximation of the theoretical test seems to be achieved from small values of B_1, even for a relatively large number of aggregated tests. In particular, for $n = 200$, an error smallest than 10% is reached for all tested B_1 values.

All these results show that both theoretical and permuted tests have comparable powers provided that the sample size and the number of permutations are large enough. In the following, we study the power of the permuted tests, which are used in practice.

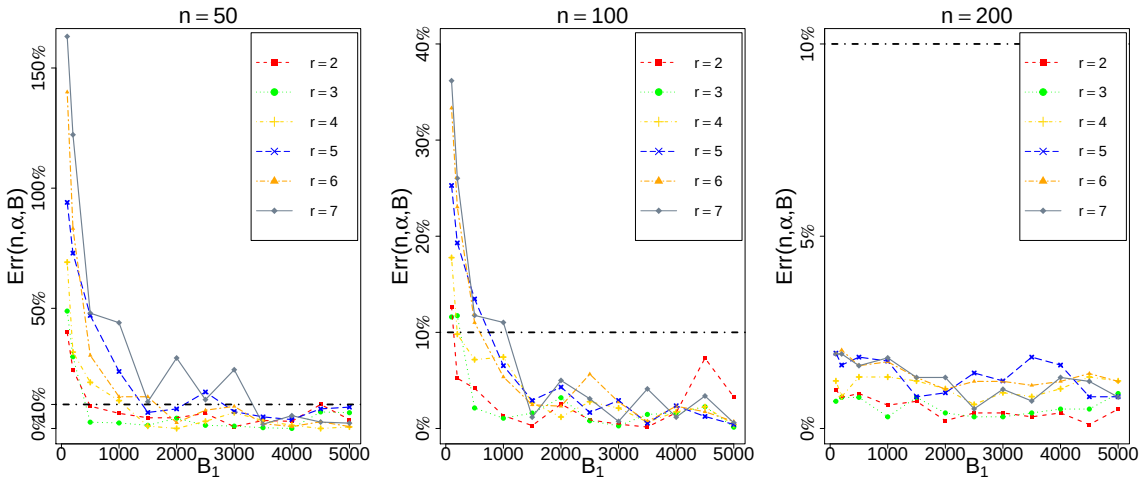

Figure 3: Absolute relative error between the powers of the theoretical and permuted aggregated HSIC procedures w.r.t. the number B_1 of permutations, for the sample sizes $n = 50, 100$ and 200 and the collection Λ_r and U_r sizes r. The prescribed level of tests is $\alpha = 0.05$. The black dashed line represents the error threshold of 10%.

5.3 How to choose the collections of bandwidths and the weights?

In our aggregated procedure, collections of bandwidths Λ and U, together with weights have to be chosen. There is no universal best collection that would ensure optimal test power. To determine the collection, we first study the impact of the bandwidth choice on single HSIC-based tests. This leads us to particular forms of collections. Then, we investigate different choices of the collections Λ and U and together with different weights. A comparison with the single HSIC test is also carried out.

5.3.1 Impact of the bandwidths choice on the power of the single tests

The optimal bandwidth depends on the intrinsic characteristics of X and Y and their dependence structure. Consequently, it seems relevant to consider the possible bandwidths relatively to the standard deviations of X and Y. Moreover, as already mentioned, the standard deviation is a usual choice for the bandwidth in the literature on single HSIC-test [Zhang et al., 2012]. We assume here that the exact values of standard deviations of X and Y, respectively denoted s and s', are known. In such a way, we are able to construct collections which do not depend on the observation. In practice, when only a n-sample of (X, Y) is available, we estimate these standard deviations by the usual empirical estimators. Note that, practice shows that the effect of this estimation does not significantly impact the single tests performance. Indeed, standard deviation estimators converge in most cases rapidly w.r.t. n. More particularly, this estimation error is small compared to the estimation error of the quantiles.

For this, we consider the uni-variate mechanism of dependence (ii) with $l = 2$. Moreover, we consider, as possible bandwidths λ and μ respectively associated to X and Y, multiple or dyadic fractions of s and s' respectively. For each couple (λ, μ), the power of the single HSIC tests is computed. Figure 4 shows the obtained power maps w.r.t. (λ, μ), for different sample sizes. First, we can observe that the bandwidths significantly impact the power: in this case, there is an optimal area around $(\lambda, \mu) = (s/4, s'/4)$ with a power close to one for $n = 200$. The power decreases progressively as we move away from this area, until being null for very high and very low values of bandwidths. We can also see that the regularity of the maps increases with the sample-size (just like the power for each point, obviously). Similar conclusions can be observed for other values of l and the other data generating mechanisms (i) and (iii) with one or several best-power areas, but are not presented here. These results illustrate that an arbitrary choice of bandwidths is not relevant and justify the interest of considering several bandwidths through an aggregation strategy. Moreover, according to our experience, it is usually not appropriate to consider bandwidths higher than standard deviations. Consequently, in the following, we consider aggregating

procedures based on collections Λ and U of types

$$\Lambda_r^s = \left\{s, s/2, \ldots, s/2^{r-1}\right\}, \quad U_r^{s'} = \left\{s', s'/2, \ldots, s'/2^{r-1}\right\}, \tag{36}$$

where r belongs to $\mathbb{N}^*$.

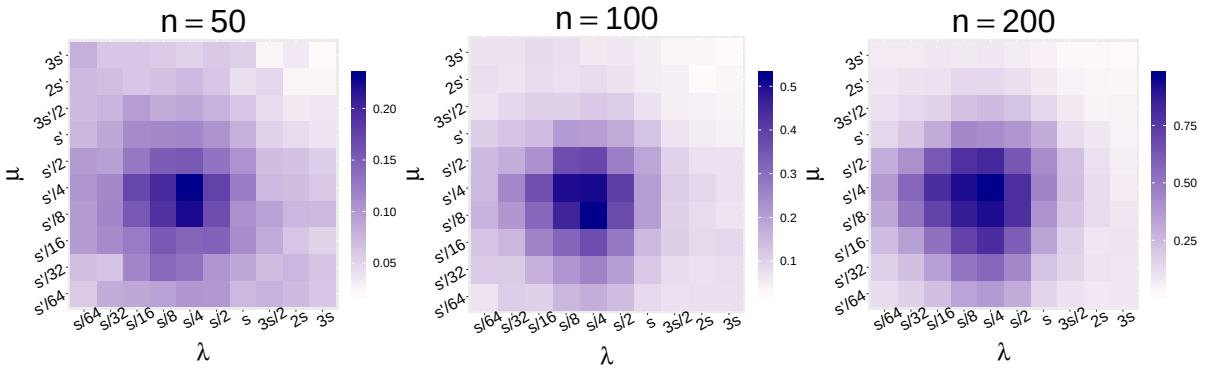

Figure 4: Power map of single HSIC test w.r.t. to kernel bandwidths λ and μ respectively associated to X and Y, for sample sizes $n = 50, 100$ and 200.

5.3.2 Impact of the weights choice on the power of the aggregated procedure

Following the results of Section 5.3.1, we consider bandwidth collections $\Lambda_r^s \times U_r^{s'}$ as defined in Equation (36), where s and s' are respectively the empirical standard deviations of the X_i's and the Y_i's. By now, let us compare two possible choices of weights. Likewise, without prior knowledge, one should not consider bandwidths too small, which would lean in favor of decreasing weights with bandwidth. We study in particular two types of weights : uniform and exponential weights. On the one hand, we recall that uniform weights depend only on the cardinals of Λ and U, and are defined in Equation (35) for all (λ, μ) by

$$\omega_{\lambda,\mu} = \log(r^2).$$

On the second hand, the exponential weights are adapted to the collections Λ_r^s and $U_r^{s'}$. They are defined, by analogy with Equation (25), for all bandwidths $(s/2^i, s'/2^j)$ as

$$\omega_{s/2^i, s'/2^j} = 2\log(i+1) + 2\log(j+1) + \log\left(\sum_{1\leq u,v\leq r} \frac{1}{u^2 v^2}\right). \tag{37}$$

Note that the last term in the right hand side of (37) ensures that $\sum_{(\lambda,\mu)\in\Lambda\times U} e^{-\omega_{\lambda,\mu}} = 1$.

The results obtained with the two types of weights are given in Figure 5, for different collection sizes r and sample sizes n. In this case, the uniform weights seem to give a better power than the exponential ones. However, we can observe a different behavior w.r.t. r. For the uniform weights, the power increases until a specific r ($r = 3$ or 4 w.r.t n), before decreasing with r, to being lower than power with exponential weights. On the contrary, the power with exponential weights has a more robust behavior, since it increases with r until it stabilizes. This is a crucial advantage in favor of exponential weights, as the optimal r is unknown in practice. It prevents deterioration of the quality of the test, when too large collection sizes have been chosen. We can also observe that the two aggregated strategies yield a greater power than the single test (which corresponds to the case $r = 1$), as soon as the collection $\Lambda \times U$ is large enough.

Similar conclusions have been drawn from the other analytical examples, which are not presented here for the sake of brevity. Thus, from our experience, we recommend in practice the use of the aggregated procedure with exponential weights with $r = 5$ or 6.

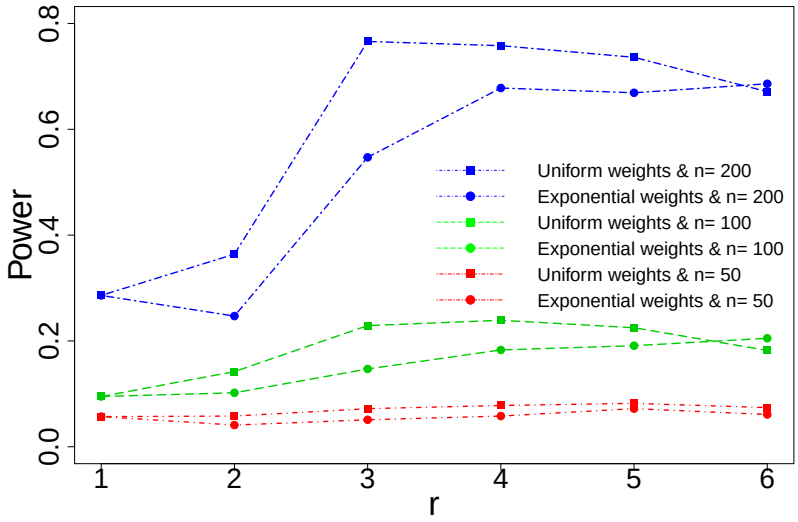

Figure 5: Power of aggregated procedures with uniform and exponential weights, w.r.t. the number r of aggregated bandwidths in each direction, for sample sizes $n = 50, 100$ and 200.

5.4 Comparison with Mutual Information Test

To complete this simulation study, we compare our aggregated procedure with some existing reference tests of independence. For this, we simulate accordingly to the data generating mechanisms (i), (ii) and (iii) of [Berrett and Samworth, 2017] presented in Section 5.1 and consider a sample size $n = 200$. In their paper, [Berrett and Samworth, 2017] numerically compare the powers of several independence tests. More precisely, they compare their Mutual Information Test (called MINT) with the tests based on the copula defined in [Kojadinovic and Holmes, 2009], on the distance covariance [Székely et al., 2007], on the Kendall's tau introduced in [Bergsma and Dassios, 2014] and on the single HSIC using the permutation method [Pfister et al., 2018] with $B = 1000$ permutations. For this last test, the kernel bandwidths defined for multivariate variables are chosen for X and for Y by

$$\lambda^2 = \frac{1}{2} median\left\{\|X_i - X_j\|_2^2 : i < j\right\} \quad \text{and} \quad \mu^2 = \frac{1}{2} median\left\{\|Y_i - Y_j\|_2^2 : i < j\right\},$$

where p (resp. q) is the dimension of X (resp. Y) and $\|.\|_p$ (resp. $\|.\|_p$) is the Euclidean norm in dimension p (resp. q). For the sake of consistency with Section 5.2, we make here slightly different choices for these bandwidths, by taking λ and μ such that

$$\widetilde{\lambda}^2 = \frac{1}{2} mean\left\{\|X_i - X_j\|_2^2 : i < j\right\} \quad \text{and} \quad \widetilde{\mu}^2 = \frac{1}{2} mean\left\{\|Y_i - Y_j\|_2^2 : i < j\right\}. \tag{38}$$

Notice that, when X (resp. Y) is one-dimensional, then $\widetilde{\lambda}$ (resp. $\widetilde{\mu}$) is the empirical estimator of the standard deviation of X (resp. Y). In [Berrett and Samworth, 2017], the most powerful tests on the univariate and bivariate examples (i), (ii) and (iii) are the MINT and single HSIC test. Let us compare the performances of these two tests with our aggregated procedure with the following methodological choices.

- In the univariate case, the bandwidth collections Λ and U defined as

$$\Lambda = \left\{\widetilde{\lambda}, \widetilde{\lambda}/2, \ldots, \widetilde{\lambda}/2^6\right\} \quad \text{and} \quad U = \left\{\widetilde{\mu}, \widetilde{\mu}/2, \ldots, \widetilde{\mu}/2^6\right\},$$

  where $\widetilde{\lambda}$ and $\widetilde{\mu}$ are the bandwidths introduced in Equation (38). Similarly, in the bivariate case, the bandwidth are defined by

$$\Lambda = \left\{(\widetilde{\lambda}, \widetilde{\lambda}), (\widetilde{\lambda}/2, \widetilde{\lambda}/2), \ldots, (\widetilde{\lambda}/2^6, \widetilde{\lambda}/2^6)\right\} \quad \text{and} \quad U = \left\{(\widetilde{\mu}, \widetilde{\mu}), (\widetilde{\mu}/2, \widetilde{\mu}/2), \ldots, (\widetilde{\mu}/2^6, \widetilde{\mu}/2^6)\right\}.$$

- Exponential weights of Equation (37) are chosen.
- Algorithm 2 is used with $B_1 = 3000$ and $B_2 = 500$.

For each example, the power of the different testing procedures is estimated using 1000 different samples of (X, Y). The obtained power curves are given in Figure 6, w.r.t parameters l for simulated data from (i) and (ii) and w.r.t ρ for (iii). No procedure of testing constantly yields the best performances. For the case (i), the MINT and the HSIC aggregated procedure have competitive results, much better than single HSIC. For the mechanism (ii), the MINT is the most powerful method, then comes the aggregated procedure, single HSIC giving the worst results. For the last example (iii), results are opposite.

Thus, the HSIC aggregated procedure seems to yield intermediate results between MINT and single HSIC: it provides better results on average, regardless of the mechanism of dependence between the variables. Moreover, in most examples of Section 5.3 and 5.4, the HSIC aggregated procedure performs better than the single HISC test.

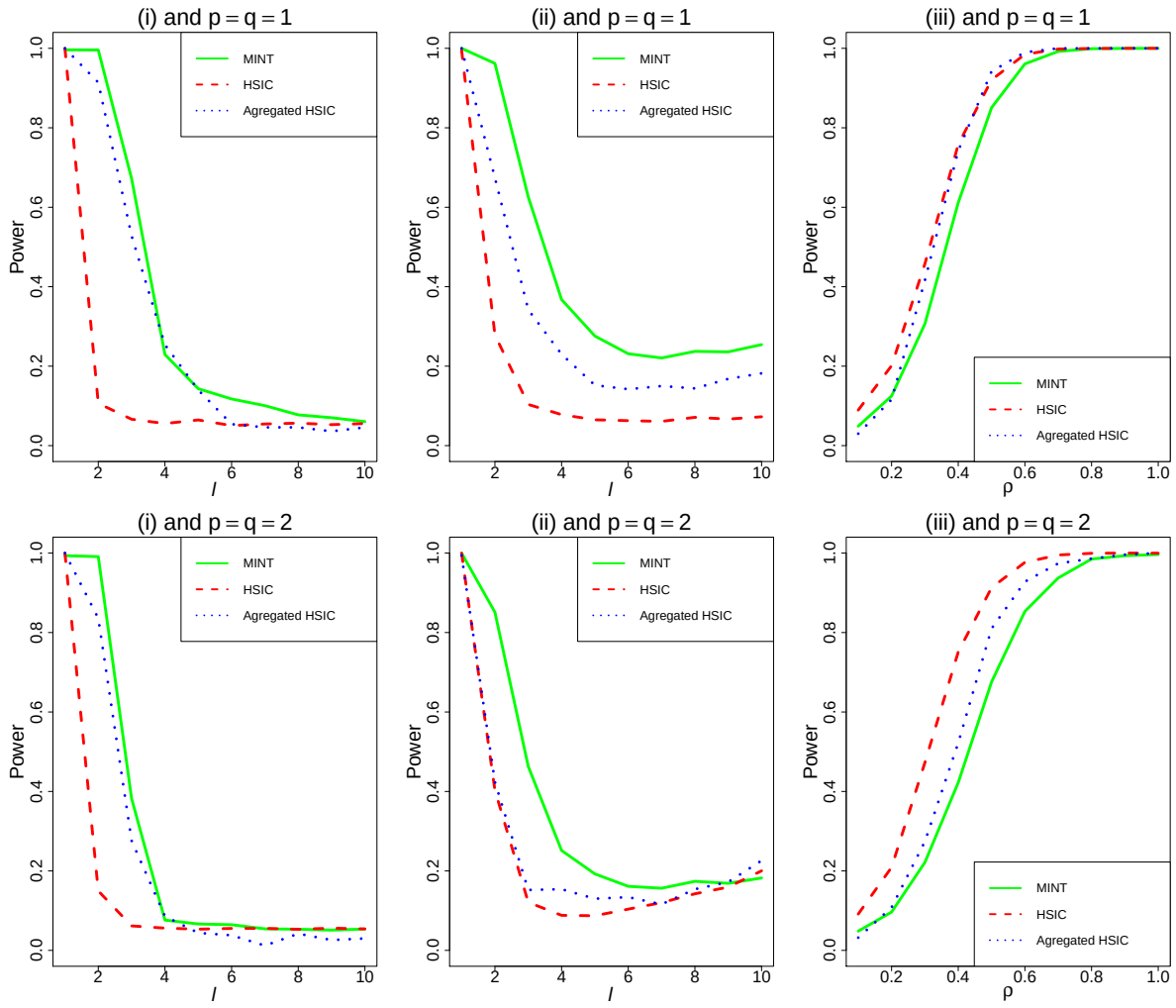

Figure 6: Power curves of MINT, single HSIC test and aggregated procedure for the mechanisms of dependence (i), (ii) and (iii) in the uni-variate ($p = q = 1$) and the bi-variate ($p = q = 2$) cases.

6 Proofs

All along the proofs, we set $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$ for all i in $\{1, \dots, n\}$. We also denote by A, B and C positive universal constants whose values may change from line to line.

6.1 Proof of Proposition 1

Let α be in $(0,1)$. In order to prove that the permuted test with Monte Carlo approximation $\widehat{\Delta}_\alpha^{\lambda,\mu}$ is of prescribed level α, we use the following lemma of [Romano and Wolf, 2005].

Lemma 7 ([Romano and Wolf, 2005, Lemma 1]). *Let $R_1, \ldots, R_{B+1}$ be $(B+1)$ exchangeable random variables. Then, for all u in $(0,1)$*

$$\mathbb{P}\left(\frac{1}{B+1}\left(1+\sum_{b=1}^{B}\mathbb{1}_{R_b\geq R_{B+1}}\right)\leq u\right)\leq u.$$

Recall that for all $1\leq b\leq B$,

$$\widehat{H}_{\lambda,\mu}^{\star b}=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n^{\tau_b}\right)\quad\text{and}\quad\widehat{H}_{\lambda,\mu}^{\star B+1}=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n\right)=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n^{\tau_{B+1}}\right),$$

where $\tau_{B+1}=\mathrm{id}$ is the identity permutation of $\{1,\ldots,B+1\}$ (deterministic).

Assume that $f=f_1\otimes f_2$. Then the random variables $\widehat{H}_{\lambda,\mu}^{\star 1},\ldots,\widehat{H}_{\lambda,\mu}^{\star B}$ and $\widehat{H}_{\lambda,\mu}^{\star B+1}$ are exchangeable. Indeed, let π be a (deterministic) permutation of $\{1,\ldots,B+1\}$ and let us prove that

$$\left(\widehat{H}_{\lambda,\mu}^{\star 1},\ldots,\widehat{H}_{\lambda,\mu}^{\star B},\widehat{H}_{\lambda,\mu}^{\star B+1}\right)\quad\text{and}\quad\left(\widehat{H}_{\lambda,\mu}^{\star\pi(1)},\ldots,\widehat{H}_{\lambda,\mu}^{\star\pi(B+1)}\right)\quad\text{have the same distribution.}\tag{39}$$

1st case: if $\pi(B+1)=B+1$. Then, since the permutations $(\tau_b)_{1\leq b\leq B}$ are i.i.d., they are exchangeable. Hence, $(\tau_{\pi(1)},\ldots,\tau_{\pi(B)})$ is an i.i.d. sample of uniform permutations of $\{1,\ldots,n\}$, independent of $\mathbb{Z}_n$ and (39) holds by construction.

2nd case: if $\pi(B+1)\neq B+1$. Then,

$$\widehat{H}_{\lambda,\mu}^{\star\pi(B+1)}=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n^{\tau_{\pi(B+1)}}\right)=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\tilde{\mathbb{Z}}_n\right),\quad\text{where}\quad\tilde{\mathbb{Z}}_n=\mathbb{Z}_n^{\tau_{\pi(B+1)}}.$$

In particular, for all b in $\{1,\ldots,B\}$,

$$\begin{cases}\widehat{H}_{\lambda,\mu}^{\star\pi(b)}=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n^{\tau_{\pi(b)}}\right)=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\tilde{\mathbb{Z}}_n^{\tau_{\pi(b)}\circ\tau_{\pi(B+1)}^{-1}}\right) & \text{if } \pi(b)\neq B+1,\\ \widehat{H}_{\lambda,\mu}^{\star\pi(b)}=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\mathbb{Z}_n\right)=\widehat{\mathrm{HSIC}}_{\lambda,\mu}\left(\tilde{\mathbb{Z}}_n^{\mathrm{id}\circ\tau_{\pi(B+1)}^{-1}}\right) & \text{if } \pi(b)=B+1.\end{cases}$$

Therefore, in order to prove (39), it is sufficient to prove that $\{\tau_{\pi(1)}\circ\tau_{\pi(B+1)}^{-1},\ldots,\tau_{\pi(B)}\circ\tau_{\pi(B+1)}^{-1}\}$ is an i.i.d. sample of uniform permutations of $\{1,\ldots,n\}$ independent of $\tilde{\mathbb{Z}}_n$.

Let A be a mesurable set, and $\sigma_1,\ldots,\sigma_B$ be (fixed) permutations of $\{1,\ldots,n\}$. Then

$$\begin{aligned}
&\mathbb{P}\big(\tilde{\mathbb{Z}}_n\in A,\ \tau_{\pi(1)}\circ\tau_{\pi(B+1)}^{-1}=\sigma_1,\ \ldots,\ \tau_{\pi(B)}\circ\tau_{\pi(B+1)}^{-1}=\sigma_B\big)\\
&=\mathbb{P}\big(\mathbb{Z}_n^{\tau_{\pi(B+1)}}\in A,\tau_{\pi(1)}=\sigma_1\circ\tau_{\pi(B+1)},\ldots,\tau_{\pi(B)}=\sigma_B\circ\tau_{\pi(B+1)}\big)\\
&=\mathbb{E}\big[\mathbb{P}\big(\mathbb{Z}_n^{\tau_{\pi(B+1)}}\in A,\tau_{\pi(1)}=\sigma_1\circ\tau_{\pi(B+1)},\ldots,\tau_{\pi(B)}=\sigma_B\circ\tau_{\pi(B+1)}\big|\tau_{\pi(B+1)}\big)\big]\\
&=\mathbb{E}\left[\mathbb{P}(\mathbb{Z}_n\in A)\times\left(\prod_{\substack{b=1\\ b\neq\pi^{-1}(B+1)}}^{B}\mathbb{P}\big(\tau_{\pi(b)}=\sigma_b\circ\tau_{\pi(B+1)}\big|\tau_{\pi(B+1)}\big)\right)\times\mathbb{P}\big(\mathrm{id}=\sigma_{\pi^{-1}(B+1)}\circ\tau_{\pi(B+1)}\big|\tau_{\pi(B+1)}\big)\right],\qquad(40)\\
&=\mathbb{E}\left[\mathbb{P}(\mathbb{Z}_n\in A)\left(\frac{1}{n!}\right)^{B-1}\mathbb{P}\big(\mathrm{id}=\sigma_{\pi^{-1}(B+1)}\circ\tau_{\pi(B+1)}\big|\tau_{\pi(B+1)}\big)\right],\\
&=\mathbb{P}(\mathbb{Z}_n\in A)\left(\frac{1}{n!}\right)^{B-1}\mathbb{P}\left(\tau_{\pi(B+1)}=\sigma_{\pi^{-1}(B+1)}^{-1}\right),\\
&=\mathbb{P}(\mathbb{Z}_n\in A)\left(\frac{1}{n!}\right)^{B},
\end{aligned}$$

where (40) holds by independence of all permutations τ_b and of $\mathbb{Z}_n$ and since, if $f = f_1 \otimes f_2$, $\mathbb{Z}_n^{\tau_\pi(B+1)}$ and $\mathbb{Z}_n$ have the same distribution. This ends the proof of the exchangeability of the $(\widehat{H}^{\star b}_{\lambda,\mu})_{1\leq b\leq B+1}$.

Then, by applying Lemma 7 to the $(\widehat{H}^{\star b}_{\lambda,\mu})_{1\leq b\leq B+1}$, we obtain

$$
\begin{aligned}
P_{f_1\otimes f_2}\left(\widehat{\Delta}^{\lambda,\mu}_{\alpha}=1\right) &= P_{f_1\otimes f_2}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} > \widehat{q}^{\lambda,\mu}_{1-\alpha}\right)\\
&= P_{f_1\otimes f_2}\left(\widehat{H}^{\star B+1}_{\lambda,\mu} > \widehat{H}^{\star(\lceil (B+1)(1-\alpha)\rceil)}_{\lambda,\mu}\right)\\
&= P_{f_1\otimes f_2}\left(\sum_{b=1}^{B+1} \mathbb{1}_{\widehat{H}^{\star b}_{\lambda,\mu} < \widehat{H}^{\star B+1}_{\lambda,\mu}} \geq \lceil (B+1)(1-\alpha)\rceil\right)\\
&= P_{f_1\otimes f_2}\left(\sum_{b=1}^{B+1} \mathbb{1}_{\widehat{H}^{\star b}_{\lambda,\mu} \geq \widehat{H}^{\star B+1}_{\lambda,\mu}} \leq \lfloor \alpha(B+1)\rfloor\right) \qquad (41)\\
&= P_{f_1\otimes f_2}\left(\sum_{b=1}^{B+1} \mathbb{1}_{\widehat{H}^{\star b}_{\lambda,\mu} \geq \widehat{H}^{\star B+1}_{\lambda,\mu}} \leq \alpha(B+1)\right)\\
&= P_{f_1\otimes f_2}\left(\frac{1}{B+1}\left(1+\sum_{b=1}^{B} \mathbb{1}_{\widehat{H}^{\star b}_{\lambda,\mu} \geq \widehat{H}^{\star B+1}_{\lambda,\mu}}\right) \leq \alpha\right)\\
&\leq \alpha, \qquad (42)
\end{aligned}
$$

where (41) comes from the fact that

$$B+1-\lceil (B+1)(1-\alpha)\rceil = \lfloor \alpha(B+1)\rfloor,$$

and (42) is obtained from Lemma 7.

6.2 Proof of Lemma 1

Let α and β be in $(0,1)$. We aim here to give a condition on $\mathrm{HSIC}_{\lambda,\mu}(f)$ w.r.t. the variance $\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})$ and the quantile $q^{\lambda,\mu}_{1-\alpha}$, so that the statistical test $\Delta^{\lambda,\mu}_{\alpha}$ has a second kind error controlled by β. For this, we use Chebyshev's inequality. Since $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ is an unbiased estimator of $\mathrm{HSIC}_{\lambda,\mu}(f)$,

$$P_f\left(\left|\widehat{\mathrm{HSIC}}_{\lambda,\mu} - \mathrm{HSIC}_{\lambda,\mu}(f)\right| \geq \sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}}\right) \leq \beta.$$

We then have the following inequality:

$$P_f\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq \mathrm{HSIC}_{\lambda,\mu}(f) - \sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}}\right) \leq \beta.$$

Consequently, one has $P_f\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q^{\lambda,\mu}_{1-\alpha}\right) \leq \beta$, as soon as

$$\mathrm{HSIC}_{\lambda,\mu}(f) \geq \sqrt{\frac{\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})}{\beta}} + q^{\lambda,\mu}_{1-\alpha}.$$

6.3 Proof of Proposition 2

In order to provide an upper bound of the variance $\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})$ w.r.t. the bandwidths λ, μ and the sample-size n, let us first give the following lemma for a general U-statistic of any order r in $\{1,\ldots,n\}$.

Lemma 8. *Let h be a symmetric function with $r \leq n$ inputs, $V_1, \ldots, V_n$ be independent and identically distributed random variables and U_n be the U-statistic defined by*

$$U_n = \frac{(n-r)!}{n!} \sum_{(i_1,\ldots,i_r)\in \mathbf{i}_r^n} h(V_{i_1}, \ldots, V_{i_r}).$$

The following inequality gives an upper bound of the variance of U_n,

$$\mathrm{Var}(U_n) \leq C(r) \left(\frac{\sigma^2}{n} + \frac{s^2}{n^2} \right), \tag{43}$$

where $\sigma^2 = \mathrm{Var}\left(\mathbb{E}[h(V_1, \ldots, V_r) \mid V_1]\right)$ and $s^2 = \mathrm{Var}\left(h(V_1, \ldots, V_r)\right)$.

Proof of Lemma 8. First, using Hoeffding's decomposition (see e.g. [Serfling, 2009, Lemma A, p. 183]), the variance of U_n can be decomposed as

$$\mathrm{Var}(U_n) = \binom{n}{r}^{-1} \sum_{c=1}^{r} \binom{r}{c} \binom{n-r}{r-c} \zeta_c,$$

where $\zeta_c = \mathrm{Var}(\mathbb{E}[h(V_1, \ldots, V_r) \mid V_1, \ldots, V_c])$.

Let us now prove that, for all $n \in \mathbb{N}^*$, $r \in \{1, \ldots, n\}$ and $c \in \{1, \ldots, r\}$,

$$\binom{n}{r}^{-1} \binom{r}{c} \binom{n-r}{r-c} \leq \frac{C(r,c)}{n^c}. \tag{44}$$

We first write

$$\begin{aligned} \binom{n}{r}^{-1} \binom{r}{c} \binom{n-r}{r-c} &= \binom{r}{c} \times \frac{(n-r)!}{(r-c)!(n+c-2r)!} \times \frac{r!(n-r)!}{n!} \\ &= \binom{r}{c} \times \frac{r!}{(r-c)!} \times \frac{(n-r)!}{(n+c-2r)!} \times \frac{(n-r)!}{n!}. \end{aligned} \tag{45}$$

Moreover,

$$\begin{aligned} n! &= (n-r)! \times (n-r+1) \times \ldots \times (n-r+r) \\ &\geq (n-r)! \times (n-r+1)^r, \end{aligned}$$

and

$$\begin{aligned} (n-r)! &= (n-2r+c)! \times (n-2r+c+1) \times \ldots \times (n-2r+c+r-c) \\ &\leq (n-2r+c)! \times (n-r+1)^{r-c}. \end{aligned}$$

Then, we have

$$\frac{(n-r)!}{(n+c-2r)!} \times \frac{(n-r)!}{n!} \leq \frac{1}{(n-r+1)^c}.$$

Furthemore, using that $n \geq r$, one can write

$$\begin{aligned} \frac{n-r+1}{n} &= 1 - \frac{r-1}{n} \\ &\geq 1 - \frac{r-1}{r} \\ &= \frac{1}{r}. \end{aligned}$$

This leads to, $\frac{1}{n-r+1} \leq \frac{r}{n}$. Finally, Equation (45) leads to Equation (44).

By upper bounding each term in Hoeffding's decomposition of the variance of U_n according to Equation (44), we obtain the following inequality:

$$\mathrm{Var}(U_n) \leq C(r) \sum_{c=1}^{r} \frac{\zeta_c}{n^c}. \tag{46}$$

On the one hand, $\zeta_1 = \sigma^2$. On the other hand, using the law of total variance (see e.g. [Weiss, 2006]), for all c in $\{2, .., r\}$: $\zeta_c \leq s^2$. By injecting this last inequality in Equation (46), we obtain for all n in $\mathbb{N}^*$:

$$\mathrm{Var}(U_n) \leq C(r) \left(\frac{\sigma^2}{n} + \frac{s^2}{n^2} \right),$$

which achieves the proof of Lemma 8. □

Let us now apply Lemma 8 in order to control the variance of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ w.r.t λ, μ and n. For this, we first recall that $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ can be written as a single U-statistic of order 4 [Gretton et al., 2008] as

$$\widehat{\mathrm{HSIC}}_{\lambda,\mu} = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{(i,j,q,r) \in \mathbf{i}_4^n} h_{i,j,q,r},$$

where the general term $h_{i,j,q,r}$ of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ is defined as

$$h_{i,j,q,r} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} \left(k_{t,u} l_{t,u} + k_{t,u} l_{v,w} - 2 k_{t,u} l_{t,v} \right). \tag{47}$$

where $k_{t,u}$ (resp. $l_{t,u}$) is defined for all t, u in $\{1, \ldots, n\}$ as $k_{t,u} = k(X_t, X_u)$ (resp. $l_{t,u} = l(Y_t, Y_u)$) and the sum represents all ordered quadruples (t, u, v, w) drawn without replacement from (i, j, q, r).

Thus, using Lemma 8, the variance of $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ can be upper bounded as follows:

$$\mathrm{Var}_f \left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \right) \leq C \left(\frac{\sigma^2(\lambda,\mu)}{n} + \frac{s^2(\lambda,\mu)}{n^2} \right), \tag{48}$$

where, recalling that $Z_i = (X_i, Y_i)$ for all i in $\{1, \ldots, n\}$, $\sigma^2(\lambda,\mu) = \mathrm{Var}_f \left(\mathbb{E}[h_{1,2,3,4} \mid Z_1] \right)$ and $s^2(\lambda,\mu) = \mathrm{Var}_f \left(h_{1,2,3,4} \right)$.

6.3.1 Upper bound of $\sigma^2(\lambda,\mu)$

By now, we upper bound $\sigma^2(\lambda,\mu)$ defined in Equation (48) w.r.t. λ and μ. For this, we first notice that in the cases when $k_\lambda(X_a, X_b) l_\mu(Y_c, Y_d)$ is independent from Z_1, the variance of its expectation conditionally on Z_1 equals 0. That are the cases when a, b, c and d are all different from 1. We then have the following inequality:

$$\sigma^2(\lambda,\mu) \leq C \sum_{i=1}^{6} \sigma_i^2(\lambda,\mu),$$

where

$$\begin{aligned}
\sigma_1^2(\lambda,\mu) &= \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_1, X_2) l_\mu(Y_1, Y_2) \mid Z_1] \right), & \sigma_2^2(\lambda,\mu) &= \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_1, X_2) l_\mu(Y_3, Y_4) \mid X_1] \right), \\
\sigma_3^2(\lambda,\mu) &= \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_3, X_4) l_\mu(Y_1, Y_2) \mid Y_1] \right), & \sigma_4^2(\lambda,\mu) &= \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_1, X_2) l_\mu(Y_1, Y_3) \mid Z_1] \right), \\
\sigma_5^2(\lambda,\mu) &= \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_2, X_1) l_\mu(Y_2, Y_3) \mid X_1] \right), & \sigma_6^2(\lambda,\mu) &=, \mathrm{Var}_f \left(\mathbb{E}[k_\lambda(X_2, X_3) l_\mu(Y_2, Y_1) \mid Y_1] \right).
\end{aligned}$$

Case 1. Upper bound of $\sigma_1^2(\lambda,\mu)$

$$\begin{aligned}\sigma_1^2(\lambda,\mu) \leq & \mathbb{E}\left[\left(\mathbb{E}\left[k_\lambda(X_1,X_2)l_\mu(Y_1,Y_2) \mid Z_1\right]\right)^2\right] \\ \leq & \mathbb{E}\left[k_\lambda(X_1,X_2)l_\mu(Y_1,Y_2)k_\lambda(X_1,X_3)l_\mu(Y_1,Y_3)\right].\end{aligned}$$

Moreover, we have

$$\begin{aligned}&\mathbb{E}\left[k_\lambda(X_1,X_2)k_\lambda(X_1,X_3)l_\mu(Y_1,Y_2)l_\mu(Y_1,Y_3)\right] \\ =&\int_{(\mathbb{R}^p\times\mathbb{R}^q)^3} k_\lambda(x_1,x_2)k_\lambda(x_1,x_3)l_\mu(y_1,y_2)l_\mu(y_1,y_3)\prod_{k=1}^3 f(x_k,y_k)\mathrm{d}x_k\mathrm{d}y_k.\end{aligned}$$

By upper bounding $f(x_2,y_2)$ and $f(x_3,y_3)$ by $\|f\|_\infty$, we have

$$\begin{aligned}\sigma_1^2(\lambda,\mu) &\leq \|f\|_\infty^2\int_{(\mathbb{R}^p\times\mathbb{R}^q)^3} k_\lambda(x_1,x_2)k_\lambda(x_1,x_3)l_\mu(y_1,y_2)l_\mu(y_1,y_3)\, f(x_1,y_1)\prod_{k=1}^3 \mathrm{d}x_k\mathrm{d}y_k \\ &= \|f\|_\infty^2\int_{\mathbb{R}^p\times\mathbb{R}^q}\left[\int_{\mathbb{R}^p} k_\lambda(x_1,x_2)\mathrm{d}x_2\right]\left[\int_{\mathbb{R}^p} k_\lambda(x_1,x_3)\mathrm{d}x_3\right]\left[\int_{\mathbb{R}^q} l_\mu(y_1,y_2)\mathrm{d}y_2\right]\left[\int_{\mathbb{R}^q} l_\mu(y_1,y_3)\mathrm{d}y_3\right] f(x_1,y_1)\mathrm{d}x_1\mathrm{d}y_1.\end{aligned}$$

Finally, using that $\int_{\mathbb{R}^p} k_\lambda(.,x)\mathrm{d}x = \int_{\mathbb{R}^q} l_\mu(.,y)\mathrm{d}y = 1$, we write

$$\sigma_1^2(\lambda,\mu) \leq \|f\|_\infty^2. \tag{49}$$

Case 2. Upper bound of $\sigma_2^2(\lambda,\mu)$

$$\begin{aligned}\sigma_2^2(\lambda,\mu) \leq & \mathbb{E}\left[\left(\mathbb{E}\left[k_\lambda(X_1,X_2)l_\mu(Y_3,Y_4) \mid X_1\right]\right)^2\right] \\ \leq & \mathbb{E}\left[\left(\mathbb{E}\left[k_\lambda(X_1,X_2) \mid X_1\right]\right)^2\right]\left(\mathbb{E}\left[l_\mu(Y_3,Y_4)\right]\right)^2 \\ \leq & \mathbb{E}\left[k_\lambda(X_1,X_2)k_\lambda(X_1,X_3)\right]\left(\mathbb{E}\left[l_\mu(Y_3,Y_4)\right]\right)^2.\end{aligned}$$

Moreover, it is easy to see that by upper bounding $f_1(x_2)$ and $f_1(x_3)$ by $\|f_1\|_\infty$, and recalling that $\int_{\mathbb{R}^p} k_\lambda(x_1,x)\mathrm{d}x = 1$, we have,

$$\begin{aligned}\mathbb{E}\left[k_\lambda(X_1,X_2)k_\lambda(X_1,X_3)\right] &= \int_{\mathbb{R}^p}\left[\int_{\mathbb{R}^p} k_\lambda(x_1,x_2)f_1(x_2)\mathrm{d}x_2\right]\left[\int_{\mathbb{R}^p} k_\lambda(x_1,x_3)f_1(x_3)\mathrm{d}x_3\right] f_1(x_1)\mathrm{d}x_1 \\ &\leq \|f_1\|_\infty^2.\end{aligned}$$

Besides, upper bounding $f_2(y_3)$ by $\|f_2\|_\infty$ in the integral form of $\mathbb{E}\left[l_\mu(Y_3,Y_4)\right]$ gives

$$\mathbb{E}\left[l_\mu(Y_3,Y_4)\right] \leq \|f_2\|_\infty.$$

By combining these inequalities, we obtain

$$\sigma_2^2(\lambda,\mu) \leq \|f_1\|_\infty^2\|f_2\|_\infty^2. \tag{50}$$

Case 3. Upper bound of $\sigma_3^2(\lambda,\mu)$
This case is similar to case 2 by exchanging X by Y and k_λ by l_μ. Thus, we have the inequality

$$\sigma_3^2(\lambda,\mu) \leq \|f_1\|_\infty^2\|f_2\|_\infty^2. \tag{51}$$

Case 4. Upper bound of $\sigma_4^2(\lambda,\mu)$

$$\begin{aligned}\sigma_4^2(\lambda,\mu) \leq & \mathbb{E}\left[\left(\mathbb{E}\left[k_\lambda(X_1,X_2)l_\mu(Y_1,Y_3) \mid Z_1\right]\right)^2\right] \\ \leq & \mathbb{E}\left[k_\lambda(X_1,X_2)k_\lambda(X_1,X_4)l_\mu(Y_1,Y_3)l_\mu(Y_1,Y_5)\right].\end{aligned}$$

By upper bounding $f_1(x_2)$, $f_1(x_4)$ by $\|f_1\|_\infty$ and $f_2(y_3)$, $f_2(y_5)$ by $\|f_2\|_\infty$ in the integral form of $\mathbb{E}\left[k_\lambda(X_1,X_2)k_\lambda(X_1,X_4)l_\mu(Y_1,Y_3)l_\mu(Y_1,Y_5)\right]$, we obtain

$$\sigma_4^2(\lambda,\mu) \leq \|f_1\|_\infty^2 \|f_2\|_\infty^2. \tag{52}$$

Case 5. Upper bound of $\sigma_5^2(\lambda,\mu)$

$$\begin{aligned}\sigma_5^2(\lambda,\mu) \leq & \mathbb{E}\left[\left(\mathbb{E}\left[k_\lambda(X_2,X_1)l_\mu(Y_2,Y_3) \mid X_1\right]\right)^2\right] \\ \leq & \mathbb{E}\left[k_\lambda(X_2,X_1)k_\lambda(X_4,X_1)l_\mu(Y_2,Y_3)l_\mu(Y_4,Y_5)\right].\end{aligned}$$

By upper bounding $f(x_2,y_2)$ and $f(x_4,y_4)$ by $\|f\|_\infty$ in the integral form of the last expectation, we have

$$\sigma_5^2(\lambda,\mu) \leq \|f\|_\infty^2. \tag{53}$$

Case 6. Upper bound of $\sigma_6^2(\lambda,\mu)$
This case is similar to case 5 by exchanging X by Y and k_λ by l_μ. We have then the inequality

$$\sigma_6^2(\lambda,\mu) \leq \|f\|_\infty^2. \tag{54}$$

Finally, by combining inequalities (49), (50), (51), (52), (53) and (54), we have the following inequality

$$\sigma^2(\lambda,\mu) \leq C(M_f). \tag{55}$$

6.3.2 Upper bound of $s^2(\lambda,\mu)$

Let us first recall that the general term of the U-statistic $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ is written as

$$h_{1,2,3,4}(Z_1,Z_2,Z_3,Z_4) = \frac{1}{4!}\sum_{(u,v,w,t)}^{(1,2,3,4)} k_\lambda(X_u,X_v)\left[l_\mu(Y_u,Y_v)+l_\mu(Y_w,Y_t)-2l_\mu(Y_u,Y_w)\right].$$

Moreover, all the terms of the last sum have the same distribution. We then have

$$\begin{aligned}s^2(\lambda,\mu) &= \mathrm{Var}_f\left(h_{1,2,3,4}(Z_1,Z_2,Z_3,Z_4)\right) \\ &\leq C\,\mathrm{Var}_f\left(k_\lambda(X_1,X_2)\left[l_\mu(Y_1,Y_2)+l_\mu(Y_3,Y_4)-2l_\mu(Y_1,Y_3)\right]\right),\end{aligned}$$

It follows that,

$$\begin{aligned}\mathrm{Var}_f\left(h_{1,2,3,4}(Z_1,Z_2,Z_3,Z_4)\right) \leq & C\left[\mathrm{Var}_f\left(k_\lambda(X_1,X_2)l_\mu(Y_1,Y_2)\right)+\mathrm{Var}_f\left(k_\lambda(X_1,X_2)l_\mu(Y_3,Y_4)\right)\right. \\ & + \left.\mathrm{Var}_f\left(k_\lambda(X_1,X_2)l_\mu(Y_1,Y_3)\right)\right] \\ \leq & C\left(\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_2)\right]+\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_3,Y_4)\right]\right. \\ & + \left.\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_3)\right]\right),\end{aligned}$$

In order to bring back to multivariate normal densities, we express k_λ^2 and l_μ^2 as

$$k_\lambda^2 = \frac{k_{\lambda'}}{(4\pi)^{\frac{p}{2}}\lambda_1\ldots\lambda_p} \text{ and } l_\mu^2 = \frac{l_{\mu'}}{(4\pi)^{\frac{q}{2}}\mu_1\ldots\mu_q},$$

where $\lambda' = \dfrac{\lambda}{\sqrt{2}}$ and $\mu' = \dfrac{\mu}{\sqrt{2}}$.

Consequently, the expectation $\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_2)\right]$ can be expressed as

$$\begin{aligned}\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_2)\right] =& \frac{1}{(4\pi)^{\frac{p+q}{2}}\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}\,\mathbb{E}\left[k_{\lambda'}(X_1,X_2)l_{\mu'}(Y_1,Y_2)\right] \\ =& \frac{1}{(4\pi)^{\frac{p+q}{2}}\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}\int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_{\lambda'}(x_1,x_2)l_{\mu'}(y_1,y_2)f(x_1,y_1)f(x_2,y_2)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}y_1\mathrm{d}y_2.\end{aligned}$$

By upper bounding $f(x_2, y_2)$ by $\|f\|_\infty$ in the last integral, we have

$$\begin{aligned}
&\int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_{\lambda'}(x_1,x_2)l_{\mu'}(y_1,y_2)f(x_1,y_1)f(x_2,y_2)\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}y_1\mathrm{d}y_2\\
\leq&\|f\|_\infty \int_{\mathbb{R}^p\times\mathbb{R}^q}\left[\int_{\mathbb{R}^p} k_{\lambda'}(x_1,x_2)\mathrm{d}x_2\right]\left[\int_{\mathbb{R}^q} l_{\mu'}(y_1,y_2)\mathrm{d}y_2\right] f(x_1,y_1)\mathrm{d}x_1\mathrm{d}y_1\\
=&\|f\|_\infty.
\end{aligned}$$

This leads to,

$$\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_2)\right] \leq \frac{\|f\|_\infty}{(4\pi)^{\frac{p+q}{2}}\lambda_1...\lambda_p\mu_1...\mu_q}. \tag{56}$$

We can easily show by similar argument that

$$\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_3,Y_4)\right] \leq \frac{\|f_1\|_\infty\|f_2\|_\infty}{(4\pi)^{\frac{p+q}{2}}\lambda_1...\lambda_p\mu_1...\mu_q}. \tag{57}$$

and

$$\mathbb{E}\left[k_\lambda^2(X_1,X_2)l_\mu^2(Y_1,Y_3)\right] \leq \frac{\|f\|_\infty}{(4\pi)^{\frac{p+q}{2}}\lambda_1...\lambda_p\mu_1...\mu_q}. \tag{58}$$

From Equations (56), (57) and (58), we have

$$s^2(\lambda,\mu) \leq \frac{C(M_f)}{(4\pi)^{\frac{p+q}{2}}\lambda_1...\lambda_p\mu_1...\mu_q}. \tag{59}$$

From Equations (55) and (59) we obtain the following inequality for $\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu})$

$$\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu}) \leq C\left(M_f,p,q\right)\left\{\frac{1}{n}+\frac{1}{\lambda_1...\lambda_p\mu_1...\mu_q n^2}\right\}.$$

6.4 Proof of Proposition 3

To give an upper bound for the quantile $q_{1-\alpha}^{\lambda,\mu}$ w.r.t λ and μ, we use concentration inequalities for general U-statistics. However, sharp upper bounds are obtained only for degenerate U-statistics (see e.g. [Houdré and Reynaud-Bouret, 2003]). We recall that, a U-statistic $U_n = U_n(V_1,...,V_r)$ is degenerate if $\mathbb{E}[U_n \mid V_1,...,V_i] = 0$ for all i in $\{1,...,r-1\}$. The first step to upper bound $q_{1-\alpha}^{\lambda,\mu}$ is then to write $\widehat{\mathrm{HSIC}}_{\lambda,\mu}$ as a sum of degenerate U-statistics. For this, we rely on the ANOVA-decomposition (ANOVA for ANalyse Of VAriance, see e.g. [Sobol, 2001]) of the symmetrical function $h_{i,j,q,r}$ introduced in Equation (47). We then write:

$$h_{i,j,q,r} = \frac{1}{2}\sum_{(t,u)}^{(i,j,q,r)} h_{t,u} + \frac{1}{6}\sum_{(t,u,v)}^{(i,j,q,r)} h_{t,u,v} + \widetilde{h}_{i,j,q,r}, \tag{60}$$

where the first (resp. the second) sum represents all ordered pairs (t,u) (resp. triplets (t,u,v)) drawn without replacement from (i,j,q,r) and the terms $h_{t,u}$, $h_{t,u,v}$ and $\widetilde{h}_{i,j,q,r}$ are defined as

$$\begin{aligned}
h_{t,u} &= \mathbb{E}\left[h_{i,j,q,r} \mid Z_t, Z_u\right],\\
h_{t,u,v} &= \mathbb{E}\left[h_{i,j,q,r} \mid Z_t, Z_u, Z_v\right] - \frac{1}{2}\sum_{(t',u')}^{(t,u,v)} h_{t',u'},\\
\widetilde{h}_{i,j,q,r} &= h_{i,j,q,r} - \frac{1}{6}\sum_{(t,u,v)}^{(i,j,q,r)} h_{t,u,v} - \frac{1}{2}\sum_{(t,u)}^{(i,j,q,r)} h_{t,u}.
\end{aligned}$$

Hence, by summing all terms $h_{i,j,q,r}$ for (i,j,q,r) in $\mathbf{i}_4^n$ and then dividing by $n(n-1)(n-2)(n-3)$, we have

$$\widehat{\mathrm{HSIC}}_{\lambda,\mu} = 6\widehat{\mathrm{HSIC}}_{\lambda,\mu}^{(2,D)} + 4\widehat{\mathrm{HSIC}}_{\lambda,\mu}^{(3,D)} + \widehat{\mathrm{HSIC}}_{\lambda,\mu}^{(4,D)}, \tag{61}$$

where

$$\widehat{\mathrm{HSIC}}^{(2,D)}_{\lambda,\mu} = \frac{1}{n(n-1)} \sum_{(i,j)\in \mathbf{i}^n_2} h_{i,j} \quad , \quad \widehat{\mathrm{HSIC}}^{(3,D)}_{\lambda,\mu} = \frac{1}{n(n-1)(n-2)} \sum_{(i,j,q)\in \mathbf{i}^n_3} h_{i,j,q}$$

$$\widehat{\mathrm{HSIC}}^{(4,D)}_{\lambda,\mu} = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{(i,j,q,r)\in \mathbf{i}^n_4} \widetilde{h}_{i,j,q,r}.$$

Lemma 9. *Let us assume that $f = f_1 \otimes f_2$. Then, the U-statistics $\widehat{\mathrm{HSIC}}^{(2,D)}_{\lambda,\mu}$, $\widehat{\mathrm{HSIC}}^{(3,D)}_{\lambda,\mu}$ and $\widehat{\mathrm{HSIC}}^{(4,D)}_{\lambda,\mu}$ are degenerated.*

Proof. According to Theorem 2 of [Gretton et al., 2008], if $f = f_1 \otimes f_2$, we have:

$$\mathbb{E}[h_{i,j,q,r} \mid Z_i] = 0.$$

We then easily show that $\widehat{\mathrm{HSIC}}^{(2,D)}_{\lambda,\mu}$ is degenerated by writing

$$\mathbb{E}[h_{i,j} \mid Z_i] = \mathbb{E}[h_{i,j,q,r} \mid Z_i] = 0. \tag{62}$$

Moreover, to prove that $\widehat{\mathrm{HSIC}}^{(3,D)}_{\lambda,\mu}$ is degenerated, we have

$$\begin{aligned}\mathbb{E}[h_{i,j,q} \mid Z_i, Z_j] =& \mathbb{E}[h_{i,j,q,r} \mid Z_i, Z_j] - \mathbb{E}[h_{i,j} \mid Z_i, Z_j] - \mathbb{E}[h_{i,q} \mid Z_i] - \mathbb{E}[h_{j,q} \mid Z_j] \\ =& h_{i,j} - h_{i,j} \quad \text{(by definition of } h_{i,j} \text{ and Equation (62))} \\ =& 0.\end{aligned} \tag{63}$$

Finally, to show that $\widehat{\mathrm{HSIC}}^{(4,D)}_{\lambda,\mu}$ is degenerated, we write

$$\begin{aligned}\mathbb{E}[\widetilde{h}_{i,j,q,r} \mid Z_i, Z_j, Z_q] =& \mathbb{E}[h_{i,j,q,r} \mid Z_i, Z_j, Z_q] - h_{i,j,q} - h_{i,j} - h_{i,q} - h_{j,q} \\ =& 0.\end{aligned} \tag{64}$$

□

Once we have upper bounds of the $(1-\alpha)$-quantiles of $\widehat{\mathrm{HSIC}}^{(r,D)}_{\lambda,\mu}$ with r in $\{2,3,4\}$ under the assumption $f = f_1 \otimes f_2$, an upper bound of the quantile $q^{\lambda,\mu}_{1-\alpha}$ is naturally obtained. In fact, we can easily show that,

$$q^{\lambda,\mu}_{1-\alpha} \leq 6q^{\lambda,\mu}_{1-\alpha/3,2} + 4q^{\lambda,\mu}_{1-\alpha/3,3} + q^{\lambda,\mu}_{1-\alpha/3,4}$$

where $q^{\lambda,\mu}_{1-\alpha,r}$ is the $(1-\alpha)$-quantiles of $\widehat{\mathrm{HSIC}}^{(r,D)}_{\lambda,\mu}$ under the assumption $f = f_1 \otimes f_2$.

6.4.1 Upper bound of $q^{\lambda,\mu}_{1-\alpha,2}$

In this part, we give an upper bound of $q^{\lambda,\mu}_{1-\alpha,2}$. For this, we use the concentration Inequality 3.5, page 15 of [Giné et al., 2000], given for degenerated U-statistics of order 2. We write for all $t > 0$:

$$\mathbb{P}\left(|\sum_{i,j} h_{i,j}| > t\right) \leq A \exp\left(-\frac{1}{A}\min\left[\frac{t}{M}, \left(\frac{t}{L}\right)^{2/3}, \left(\frac{t}{K}\right)^{1/2}\right]\right), \tag{65}$$

where

$$K = \max_{i,j} \|h_{i,j}\|_\infty, \; M^2 = \sum_{i,j} \mathbb{E}[h^2_{i,j}]$$

$$L^2 = \max\left[\| \sum_i \mathbb{E}\left[h^2_{i,j}\left(Z_i, y\right)\right] \|_\infty, \| \sum_j \mathbb{E}\left[h^2_{i,j}\left(x, Z^{(j)}\right)\right] \|_\infty\right].$$

By setting $\varepsilon = \dfrac{t}{n^2}$, and using Equation (65), we obtain

$$\mathbb{P}\left(\frac{1}{n^2}|\sum_{i,j} h_{i,j}| > \varepsilon\right) \leq A\exp\left(-\frac{1}{A}\min\left[\frac{n^2\varepsilon}{M}, \left(\frac{n^2\varepsilon}{L}\right)^{2/3}, \left(\frac{n^2\varepsilon}{K}\right)^{1/2}\right]\right).$$

Therefore, we have for all $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n^2}|\sum_{i,j} h_{i,j}| > \varepsilon\right) \leq & A\exp\left(-\frac{1}{A}\min\left[\frac{n^2\varepsilon}{M}, \left(\frac{n^2\varepsilon}{L}\right)^{2/3}, \left(\frac{n^2\varepsilon}{K}\right)^{1/2}\right]\right) \\ = & A\max\left[\exp\left(-\frac{n^2\varepsilon}{AM}\right), \exp\left(-\frac{n^{4/3}\varepsilon^{2/3}}{AL^{2/3}}\right), \exp\left(-\frac{n\varepsilon^{1/2}}{AK^{1/2}}\right)\right].\end{aligned}$$

By adjusting the constant A, we can replace in the last inequality $\dfrac{1}{n^2}\sum_{i,j} h_{i,j}$ by $\widehat{\mathrm{HSIC}}^{(2,D)}_{\lambda,\mu}$,

$$\mathbb{P}\left(|\widehat{\mathrm{HSIC}}^{(2,D)}_{\lambda,\mu}| > \varepsilon\right) \leq A\max\left[\exp\left(-\frac{n^2\varepsilon}{AM}\right), \exp\left(-\frac{n^{4/3}\varepsilon^{2/3}}{AL^{2/3}}\right), \exp\left(-\frac{n\varepsilon^{1/2}}{AK^{1/2}}\right)\right].$$

Furthermore, if ε_α is a positif number verifying

$$\alpha = A\max\left[\exp\left(-\frac{n^2\varepsilon_\alpha}{AM}\right), \exp\left(-\frac{n^{4/3}\varepsilon_\alpha^{2/3}}{AL^{2/3}}\right), \exp\left(-\frac{n\varepsilon_\alpha^{1/2}}{AK^{1/2}}\right)\right].$$

Then, we can easily show the following inequality

$$q^{\lambda,\mu}_{1-\alpha,2} \leq \varepsilon_\alpha. \tag{66}$$

By now, we upper bound ε_α (and consequently $q^{\lambda,\mu}_{1-\alpha,2}$), in the 3 following cases.

Case 1. $\alpha = A\exp\left(-\dfrac{n^2\varepsilon_\alpha}{AM}\right)$

In this case, ε_α is expressed as

$$\varepsilon_\alpha = \frac{AM}{n^2}\left(\log\left(\frac{1}{\alpha}\right) + \log(A)\right).$$

We can then upper bound ε_α as

$$\varepsilon_\alpha \leq \frac{CM}{n^2}\left(\log\left(\frac{1}{\alpha}\right) + 1\right).$$

Furthermore, considering the values of α such that $\log\left(\frac{1}{\alpha}\right) > 1$ and by changing constant C value, we obtain

$$\varepsilon_\alpha \leq \frac{CM}{n^2}\log\left(\frac{1}{\alpha}\right). \tag{67}$$

Let us upper bound M w.r.t λ, μ and n. For this, we first write

$$M^2 = \sum_{i,j}\mathbb{E}[h_{i,j}^2] = n^2\,\mathbb{E}[h_{1,2}^2].$$

Moreover, using the law of total variance, we have under the hypothesis $f = f_1 \otimes f_2$,

$$\begin{aligned}\mathbb{E}[h_{1,2}^2] &= \mathrm{Var}\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1, Z_2]\right) \\ &\leq \mathrm{Var}\left(h_{1,2,3,4}\right).\end{aligned}$$

Furthermore, we have shown in Annexe 6.3.2 that,

$$\mathrm{Var}\left(h_{1,2,3,4}\right) \leq \frac{C\left(M_f, p, q\right)}{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}.$$

Hence, we can upper bound M as follows,

$$M \leq \frac{C\left(M_f, p, q\right) n}{\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}}. \tag{68}$$

Consequently, by combining Equations (67) and (68), we obtain

$$q^{\lambda,\mu}_{1-\alpha,2} \leq \frac{C\left(M_f, p, q\right)}{n\sqrt{\lambda_1 ... \lambda_p \mu_1 ... \mu_q}} \log\left(\frac{1}{\alpha}\right). \tag{69}$$

Case 2. $\alpha = A \exp\left(-\frac{n^{4/3}\varepsilon_\alpha^{2/3}}{AL^{2/3}}\right)$

In this case, ε_α verify that,

$$\varepsilon_\alpha^{2/3} = \frac{AL^{2/3}}{n^{4/3}}\left(\log\left(\frac{1}{\alpha}\right) + \log\left(A\right)\right).$$

Thus, ε_α can be upper bounded as

$$\varepsilon_\alpha \leq \frac{CL}{n^2} \log\left(\frac{1}{\alpha}\right)^{3/2}, \tag{70}$$

Let us upper bound L w.r.t n, λ and μ. For this, knowing that $h_{i,j}$ is symmetrical we write

$$L^2 = \| \sum_i \mathbb{E}[h^2_{i,j}(Z_i, y)] \|_\infty.$$

Moreover, according to [Gretton et al., 2008] page 10, we have under the hypothesis $f = f_1 \otimes f_2$,

$$h_{i,j}(Z_i, Z_j) = \frac{1}{6}\left[k_\lambda(X_i, X_j) + (k_\lambda)_{.,.} - (k_\lambda)_{i,.} - (k_\lambda)_{.,j}\right]\left[l_\mu(Y_i, Y_j) + (l_\mu)_{.,.} - (l_\mu)_{i,.} - (l_\mu)_{.,j}\right],$$

where $(k_\lambda)_{.,.} = \mathbb{E}[k_\lambda(X_i, X_j)]$, $(k_\lambda)_{i,.} = \mathbb{E}[k_\lambda(X_i, X_j) \mid X_i]$, $(k_\lambda)_{.,j} = \mathbb{E}[k_\lambda(X_i, X_j) \mid X_j]$ and $(l_\mu)_{.,.}$, $(l_\mu)_{i,.}$, $(l_\mu)_{.,j}$ are defined in a similar way.

Hence, we write for all $y = (y_1, y_2) \in \mathbb{R}^2$,

$$\begin{aligned} h^2_{i,j}(Z_i, y) = \frac{1}{36}&\left[k_\lambda(X_i, y_1) + (k_\lambda)_{.,.} - \mathbb{E}[k_\lambda(X_i, X_j) \mid X_i] - \mathbb{E}[k_\lambda(X_i, y_1)]\right]^2 \\ &\times \left[l_\mu(Y_i, y_2) + (l_\mu)_{.,.} - \mathbb{E}[l_\mu(Y_i, Y_j) \mid Y_i] - \mathbb{E}[l_\mu(Y_i, y_2)]\right]^2. \end{aligned}$$

Therefore, we have the following inequality for $h^2_{i,j}(Z_i, y)$,

$$\begin{aligned} h^2_{i,j}(Z_i, y) \leq C&\left[k_\lambda(X_i, y_1)^2 + (k_\lambda)^2_{.,.} + \mathbb{E}[k_\lambda(X_i, X_j) \mid X_i]^2 + \mathbb{E}[k_\lambda(X_i, y_1]^2\right] \\ &\times \left[l_\mu(Y_i, y_2)^2 + (l_\mu)^2_{.,.} + \mathbb{E}[l_\mu(Y_i, Y_j) \mid Y_i]^2 + \mathbb{E}[l_\mu(Y_i, y_2]^2\right]. \end{aligned}$$

Using that $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are independent, we write

$$\begin{aligned} L^2 \leq C\left(M_f\right) n\mathbb{E}&\left[k_\lambda(X_i, y_1)^2 + (k_\lambda)^2_{.,.} + \mathbb{E}[k_\lambda(X_i, X_j) \mid X_i]^2 + \mathbb{E}[k_\lambda(X_i, y_1]^2\right] \\ \times\mathbb{E}&\left[l_\mu(Y_i, y_2)^2 + (l_\mu)^2_{.,.} + \mathbb{E}[l_\mu(Y_i, Y_j) \mid Y_i]^2 + \mathbb{E}[l_\mu(Y_i, y_2]^2\right]. \end{aligned}$$

Each term can be upper bounded by similar arguments as 6.3.2, we then have

$$L^2 \leq C\left(M_f\right) n \left(1 + \frac{1}{\lambda_1 ... \lambda_p}\right)\left(1 + \frac{1}{\mu_1 ... \mu_q}\right).$$

Thus, using that $\lambda_1...\lambda_p < 1$ and $\mu_1...\mu_q < 1$, we obtain:

$$L \leq \frac{C\left(M_f\right)\sqrt{n}}{\sqrt{\lambda_1...\lambda_p\mu_1...\mu_q}}. \tag{71}$$

By combining Equations (70) and (71), we have

$$\varepsilon_\alpha \leq \frac{C\left(M_f\right)}{\sqrt{\lambda_1...\lambda_p\mu_1...\mu_q}n^{3/2}}\left[\log\left(\frac{1}{\alpha}\right)\right]^{3/2}.$$

Moreover, knowing that $\lambda_1...\lambda_p\mu_1...\mu_q < 1$, we obtain

$$\varepsilon_\alpha \leq \frac{C\left(M_f\right)}{(n\sqrt{\lambda_1...\lambda_p\mu_1...\mu_q})^{3/2}}\left[\log\left(\frac{1}{\alpha}\right)\right]^{3/2}. \tag{72}$$

Case 3. $\alpha = A\exp\left(-\frac{n\varepsilon_\alpha^{1/2}}{AK^{1/2}}\right)$

In this case, ε_α is expressed as

$$\varepsilon_\alpha^{1/2} = \frac{AK^{1/2}}{n}\left(\log\left(\frac{1}{\alpha}\right) + \log\left(A\right)\right).$$

Using that $\log\left(\frac{1}{\alpha}\right) > 1$ and by adjusting the value of A, we upper bound ε_α as

$$\varepsilon_\alpha \leq \frac{AK}{n^2}\left[log\left(\frac{1}{\alpha}\right)\right]^2. \tag{73}$$

Morover, we can easily show that

$$K \leq \frac{4}{\lambda_1...\lambda_p\mu_1...\mu_q}. \tag{74}$$

By combining Equations (73) and (74), we obtain:

$$q_{1-\alpha,2}^{\lambda,\mu} \leq \frac{C}{\lambda_1...\lambda_p\mu_1...\mu_q n^2}\left[\log\left(\frac{1}{\alpha}\right)\right]^2. \tag{75}$$

using (69), (72) and (75) and the fact that $\frac{1}{\sqrt{\lambda_1...\lambda_p\mu_1...\mu_q}n}\log\left(\frac{1}{\alpha}\right) < 1$, we have the following inequality

$$q_{1-\alpha,2}^{\lambda,\mu} \leq \frac{C\left(\|f_1\|_\infty, \|f_2\|_\infty, p, q\right)}{n\sqrt{\lambda_1...\lambda_p\mu_1...\mu_q}}\log\left(\frac{1}{\alpha}\right). \tag{76}$$

6.4.2 Upper bound of $q_{1-\alpha,3}^{\lambda,\mu}$

In this part, we give an upper bound for the $(1-\alpha)$-quantile of $\widehat{\text{HSIC}}_{\lambda,\mu}^{(3,D)}$. For this, we use the concentration inequality (c), page 1501 of [Arcones and Gine, 1993]. We write for all $t > 0$,

$$\mathbb{P}\left(n^{-3/2}|\sum_{i,j,q}h_{i,j,q}| > t\right) \leq A\exp\left[-\frac{Bt^{2/3}}{M^{2/3} + K^{1/2}t^{1/6}n^{-1/4}}\right], \tag{77}$$

where $K = \|h_{i,j,q}\|_\infty$, $M^2 = \mathbb{E}[h_{1,2,3}^2]$ and B an absolute postitive constant.

By setting $\varepsilon = \frac{t}{n^{3/2}}$ and using Equation (77), we have

$$\mathbb{P}\left(\frac{1}{n^3}|\sum_{i,j,q}h_{i,j,q}| > \varepsilon\right) \leq A\exp\left[-\frac{Bn\varepsilon^{2/3}}{M^{2/3} + K^{1/2}\varepsilon^{1/6}}\right].$$

Moreover, by adjusting the value of B, we can write

$$\mathbb{P}\left(|\widehat{\mathrm{HSIC}}_{\lambda,\mu}^{(3,D)}| > \varepsilon\right) \leq A\exp\left[-\frac{Bn\varepsilon^{2/3}}{M^{2/3}+K^{1/2}\varepsilon^{1/6}}\right]. \tag{78}$$

Furthermore, if ε_α is a positive number verifying

$$A\exp\left[-\frac{Bn\varepsilon_\alpha^{2/3}}{M^{2/3}+K^{1/2}\varepsilon_\alpha^{1/6}}\right] = \alpha, \tag{79}$$

then, we have the following inequality

$$q_{1-\alpha,3}^{\lambda,\mu} \leq \varepsilon_\alpha.$$

In order to upper bound ε_α in (79), we set $\gamma_\alpha = \varepsilon_\alpha^{1/6}$ and we obtain

$$Bn\gamma_\alpha^4 = K^{1/2}\log\left(\frac{A}{\alpha}\right)\gamma_\alpha + M^{2/3}\log\left(\frac{A}{\alpha}\right). \tag{80}$$

The polynomial Equation (80) is not resolvable. However, it's possible to give an upper bound of its roots. Indeed,

$$Bn\gamma_\alpha^4 \leq 2\max\left[K^{1/2}\gamma_\alpha + M^{2/3}\right]\log\left(\frac{A}{\alpha}\right).$$

Case 1. $\max\left[K^{1/2}\gamma_\alpha + M^{2/3}\right] = K^{1/2}\gamma_\alpha$
In this case, γ_α verify the following inequality,

$$\gamma_\alpha^3 \leq \frac{BK^{1/2}}{n}\left(\log\left(\frac{1}{\alpha}\right) + \log(A)\right).$$

Hence,

$$\varepsilon_\alpha \leq \frac{BK}{n^2}\left(\log\left(\frac{A}{\alpha}\right)\right)^2.$$

Since $K \leq \frac{4}{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}$, $\frac{1}{\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q n}}\log\left(\frac{1}{\alpha}\right) < 1$ and $\log\left(\frac{1}{\alpha}\right) > 1$, we write

$$\varepsilon_\alpha \leq \frac{C}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

Case 2. $\max\left[K^{1/2}\gamma_\alpha + M^{2/3}\right] = M^{2/3}$
In this case,

$$\gamma_\alpha^4 \leq \frac{BM^{2/3}}{n}\left[\log\left(\frac{A}{\alpha}\right)\right].$$

Therefore, ε_α can be upper bounded as

$$\varepsilon_\alpha \leq \frac{BM}{n^{3/2}}\left[\log\left(\frac{A}{\alpha}\right)\right]^2.$$

Moreover, using the law of total variance, it's easy to see that under the hypothesis $f = f_1 \otimes f_2$,

$$M^2 = \mathrm{Var}\left(h_{1,2,3}\right) \leq C\,\mathrm{Var}\left(h_{1,2,3,4}\right). \tag{81}$$

Hence, according to Annexe 6.3.2, M can be upper bounded as

$$M \leq \frac{C(M_f,p,q)}{\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}.$$

To conclude, in all cases we have the following inequality for $q_{1-\alpha,3}^{\lambda,\mu}$

$$q_{1-\alpha,3}^{\lambda,\mu} \leq \frac{C(\|f_1\|_\infty,\|f_2\|_\infty,p,q)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

6.4.3 Upper bound of $q^{\lambda,\mu}_{1-\alpha,4}$

In this part, we give an upper bound for the $(1-\alpha)$-quantile of $\widehat{\text{HSIC}}^{(4,D)}_{\lambda,\mu}$. For this, we use the concentration inequality (d), page 1501 of [Arcones and Gine, 1993]. We have for all $t > 0$:

$$\mathbb{P}\left(\frac{1}{n^2}|\sum_{i,j,q,r}\widetilde{h}_{i,j,q,r}| > t\right) \leq A\exp\left(-B\sqrt{\frac{t}{K}}\right),$$

where, $K = \|\widetilde{h}_{1,2,3,4}\|_\infty$.

By setting $\varepsilon = \dfrac{t}{n^2}$, we have

$$\mathbb{P}\left(\frac{1}{n^4}|\sum_{i,j,q,r}\widetilde{h}_{i,j,q,r}| > \varepsilon\right) \leq A\exp\left(-Bn\sqrt{\frac{\varepsilon}{K}}\right).$$

Furthermore, by adjusting the constant B, we replace $\dfrac{1}{n^4}\sum_{i,j,q,r}\widetilde{h}_{i,j,q,r}$ by $\widehat{\text{HSIC}}^{(4,D)}_{\lambda,\mu}$. We write

$$\mathbb{P}\left(|\widehat{\text{HSIC}}^{(4,D)}_{\lambda,\mu}| > \varepsilon\right) \leq A\exp\left(-Bn\sqrt{\frac{\varepsilon}{K}}\right). \tag{82}$$

Moreover, if ε_α is a positive number verifying

$$A\exp\left(-Bn\sqrt{\frac{\varepsilon_\alpha}{K}}\right) = \alpha, \tag{83}$$

then,

$$q^{\lambda,\mu}_{1-\alpha,4} \leq \varepsilon_\alpha.$$

By resolving Equation (83), we obtain the following equality

$$\varepsilon_\alpha = \frac{BK}{n^2}\left[\log\left(\frac{A}{\alpha}\right)\right]^2.$$

Therefore, we can easily show that

$$\varepsilon_\alpha \leq \frac{CK}{n^2}\left[\log\left(\frac{1}{\alpha}\right)\right]^2.$$

Moreover, by using the Inequality $K \leq \dfrac{4}{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}$ we have

$$q^{\lambda,\mu}_{1-\alpha,4} \leq \frac{C}{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q n^2}\log\left(\frac{1}{\alpha}\right)^2.$$

Consequently,

$$q^{\lambda,\mu}_{1-\alpha,4} \leq \frac{C}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}n}\log\left(\frac{1}{\alpha}\right). \tag{84}$$

To conclude, the quantile $q^{\lambda,\mu}_{1-\alpha}$ can be upper bounded under the hypothesis $f = f_1 \otimes f_2$ as follows,

$$q^{\lambda,\mu}_{1-\alpha} \leq \frac{C\left(\|f_1\|_\infty, \|f_2\|_\infty, p, q\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right). \tag{85}$$

6.5 Proof of Corollary 1

The proof of this corollary is immediately obtained from Lemma 1, Proposition 2 and Proposition 3.

6.6 Proof of Lemma 2

Recalling the formulation of $\mathrm{HSIC}_{\lambda,\mu}(f)$ given in Equation (1) with $k = k_\lambda$ and $l = l_\mu$, we obtain

$$\begin{aligned}\mathrm{HSIC}_{\lambda,\mu}(f) = &\int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_\lambda(x,x')l_\mu(y,y')f(x,y)f(x',y')\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y' \\ &-2\int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_\lambda(x,x')l_\mu(y,y')f(x,y)f_1(x')f_2(y')\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y' \\ &+\int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_\lambda(x,x')l_\mu(y,y')f_1(x)f_2(y)f_1(x')f_2(y')\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y'.\end{aligned}$$

This expression can be factorized using the symmetry of the kernels k_λ and l_μ:

$$\begin{aligned}\mathrm{HSIC}_{\lambda,\mu}(f) &= \int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_\lambda(x,x')l_\mu(y,y')\Big[f(x,y)-f_1(x)f_2(y)\Big]\Big[f(x',y')-f_1(x')f_2(y')\Big]\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y' \\ &= \int_{(\mathbb{R}^p\times\mathbb{R}^q)^2} k_\lambda(x,x')l_\mu(y,y')\psi(x,y)\psi(x',y')\mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y',\end{aligned}$$

where $\psi(x,y) = f(x,y) - f_1(x)f_2(y)$.

Thereafter, we reformulate this equation by replacing $k_\lambda(x,x')$ with $\varphi_\lambda(x-x')$ and replacing $l_\lambda(y,y')$ with $\phi_\mu(y-y')$, where φ_λ and ϕ_μ are respectively the functions defined in Equations (3) and (4):

$$\begin{aligned}\mathrm{HSIC}_{\lambda,\mu}(f) &= \int_{\mathbb{R}^p\times\mathbb{R}^q} \psi(x,y)\left[\int_{\mathbb{R}^p\times\mathbb{R}^q}\psi(x',y')\varphi_\lambda(x-x')\phi_\mu(y-y')\mathrm{d}x'\mathrm{d}y'\right]\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathbb{R}^p\times\mathbb{R}^q} \psi(x,y)\left[\psi*(\varphi_\lambda\otimes\phi_\mu)\right](x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \langle\psi,\psi*(\varphi_\lambda\otimes\phi_\mu)\rangle_2.\end{aligned}$$

6.7 Proof of Proposition 4

First notice that according to Equations (48) and (59), one can write:

$$\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu}) \leq \frac{C}{n}\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4}\mid Z_1]\right) + \frac{C\left(M_f,p,q\right)}{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q n^2}, \tag{86}$$

where $h_{1,2,3,4}$ is defined in Equation (47).

To prove the intended result from the last equation, we aim now to upper bound $\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4}\mid Z_1]\right)$ by $\|\psi*(\varphi_\lambda\otimes\phi_\mu)\|_2^2$ up to a positive constant which depends only on M_f. The following lemma gives such an upper bound.

Lemma 10. *For all λ in $(0,+\infty)^p$ and μ in $(0,+\infty)^q$, we have*

$$\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4}\mid Z_1]\right) \leq C(M_f)\,\|\psi*(\varphi_\lambda\otimes\phi_\mu)\|_2^2.$$

Proof. The first step to upper bound $\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4}\mid Z_1]\right)$ is to rewrite $h_{1,2,3,4}$ by isolating all the terms depending on Z_1.

$$\begin{aligned}h_{1,2,3,4} &= \frac{1}{4!}\sum_{(t,u,v,w)}^{(1,2,3,4)}\left[k_{t,u}l_{t,u}+k_{t,u}l_{v,w}-2k_{t,u}l_{t,v}\right] \\ &= \frac{2}{4!}\sum_{(u,v,w)}^{(2,3,4)}\left[k_{1,u}l_{1,u}+k_{1,u}l_{v,w}+k_{u,v}l_{1,w}-k_{w,v}l_{w,1}-k_{u,1}l_{u,v}-k_{1,u}l_{1,v}\right]+R(Z_2,Z_3,Z_4),\end{aligned}$$

where the last sum represents all triplets (u, v, w) drawn without replacement from $(2, 3, 4)$ and $R(Z_2, Z_3, Z_4)$ is a random variable depending only on Z_2, Z_3 and Z_4.

Then,

$$h_{1,2,3,4} = R(Z_2, Z_3, Z_4) + \frac{1}{12} \sum_{(u,v,w)}^{(2,3,4)} \left[k_{1,u}(l_{1,u} - l_{1,v}) - k_{u,1}(l_{u,v} - l_{v,w}) - (k_{w,v} - k_{u,v}) l_{1,w} \right].$$

The random variable $R(Z_2, Z_3, Z_4)$ being independent from Z_1, the variance of its expectation conditionally to Z_1 is equal to 0. It is then easy to see that $\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right)$ can be upper bounded as follows:

$$\begin{aligned} \mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right) \leq C \big[& \mathrm{Var}_f\left(\mathbb{E}[k_{1,2}(l_{1,2} - l_{1,3}) \mid Z_1]\right) + \mathrm{Var}_f\left(\mathbb{E}[k_{2,1}(l_{2,3} - l_{3,4}) \mid X_1]\right) \\ & + \mathrm{Var}_f\left(\mathbb{E}[(k_{2,3} - k_{4,3}) l_{1,2} \mid Y_1]\right) \big], \end{aligned} \tag{87}$$

By now, we reformulate the function $\psi * (\varphi_\lambda \otimes \phi_\mu)$ in a simpler form in order to link its $\mathbb{L}_2$-norm with the upper bound given in Equation (87). For notational convenience, we denote $G_{\lambda,\mu} = \psi * (\varphi_\lambda \otimes \phi_\mu)$. We then write

$$\begin{aligned} G_{\lambda,\mu}(x, y) &= \int_{\mathbb{R}^p \times \mathbb{R}^q} \psi(x', y') k_\lambda(x, x') l_\mu(y, y') \, \mathrm{d}x' \mathrm{d}y' \\ &= \int_{\mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^q} k_\lambda(x, x') \Big(l_\mu(y, y') - l_\mu(y, y'') \Big) \, f(x', y') f_2(y'') \mathrm{d}x' \mathrm{d}y' \mathrm{d}y'' \\ &= \mathbb{E}\Big[k_\lambda(x, X') \Big(l_\mu(y, Y') - l_\mu(y, Y'') \Big) \Big], \end{aligned}$$

where (X', Y') and Y'' are independent random variables with respective densities f and f_2.

Thereafter, the conditional expectations in Equation (87) can all be expressed as follows:

$$\begin{aligned} \mathbb{E}[k_{1,2}(l_{1,2} - l_{1,3}) \mid Z_1] &= G_{\lambda,\mu}(X_1, Y_1), \\ \mathbb{E}[k_{2,1}(l_{2,3} - l_{3,4}) \mid X_1] &= \mathbb{E}\left[G_{\lambda,\mu}(X_1, Y_3) \mid X_1 \right], \\ \mathbb{E}[(k_{2,3} - k_{4,3}) l_{1,2} \mid Y_1] &= \mathbb{E}\left[G_{\lambda,\mu}(X_3, Y_1) \mid Y_1 \right]. \end{aligned}$$

Thus, using the law of total variance [Weiss, 2006], we have the following upper bound for $\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right)$:

$$\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right) \leq C \Big[\mathrm{Var}_f\left(G_{\lambda,\mu}(X_1, Y_1)\right) + \mathrm{Var}_f\left(G_{\lambda,\mu}(X_1, Y_3)\right) + \mathrm{Var}_f\left(G_{\lambda,\mu}(X_3, Y_1)\right) \Big].$$

On the other hand, it is straightforward to upper bound the three variances in the last equation as

$$\begin{aligned} \mathrm{Var}_f\left(G_{\lambda,\mu}(X_1, Y_1)\right) &\leq \|f\|_\infty \, \|G_{\lambda,\mu}\|_2^2, \\ \mathrm{Var}_f\left(G_{\lambda,\mu}(X_1, Y_3)\right) &\leq \|f_1 \otimes f_2\|_\infty \, \|G_{\lambda,\mu}\|_2^2, \\ \mathrm{Var}_f\left(G_{\lambda,\mu}(X_3, Y_1)\right) &\leq \|f_1 \otimes f_2\|_\infty \, \|G_{\lambda,\mu}\|_2^2. \end{aligned}$$

Consequently, combining the three last Equations with Equation (87) gives us the following upper bound of $\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right)$:

$$\mathrm{Var}_f\left(\mathbb{E}[h_{1,2,3,4} \mid Z_1]\right) \leq C(M_f) \, \|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2.$$

□

We then obtain as a result of Equation (86) and Lemma 10:

$$\mathrm{Var}_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu}) \leq \frac{C(M_f) \, \|\psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2}{n} + \frac{C\left(M_f, p, q\right)}{\lambda_1 \dots \lambda_p \mu_1 \dots \mu_q n^2}.$$

6.8 Proof of Lemma 3

The objective here is the provide an upper bound of the bias term $\|\psi - \psi * (\varphi_\lambda \otimes \phi_\mu)\|_2^2$ w.r.t λ and μ, when $\psi \in \mathcal{S}^\delta_{p+q}(R)$, where $\delta \in (0,2]$. We first set $b = \psi*(\varphi_\lambda \otimes \phi_\mu) - \psi$, using that $b \in \mathbb{L}^1(\mathbb{R}^{p+q}) \cap \mathbb{L}^2(\mathbb{R}^{p+q})$, Plancherel's theorem gives that

$$\begin{aligned}(2\pi)^{p+q}\,\|b\|_2^2 =& \|\hat{b}\|_2^2 \\ =& \|(1 - \widehat{\varphi_\lambda \otimes \phi_\mu})\hat{\psi}\|_2^2. \end{aligned} \tag{88}$$

Let us denote g_1 as in Equation (2), the real function defined for all $z \in \mathbb{R}$ as $g_1(z) = \dfrac{1}{\sqrt{2\pi}} \exp(-z^2/2)$. We then obviously have the following equation

$$\varphi_\lambda \otimes \phi_\mu(x,y) = \prod_{i=1}^{p} \left[\frac{1}{\lambda_i} g_1\left(\frac{x_i}{\lambda_i}\right)\right] \prod_{j=1}^{q} \left[\frac{1}{\mu_j} g_1\left(\frac{y_j}{\mu_j}\right)\right].$$

Moreover, it is known that $\hat{g}_1 = \sqrt{2\pi} g_1$, and that the Fourier transform of a tensor product of functions is the product of Fourier transform of each of these functions. We also recall that if G is a real function and $a > 0$ then, the Fourier transform of $z \mapsto 1/a \cdot G(z/a)$ is $u \mapsto \widehat{G}(au)$. We then obtain

$$\begin{aligned}\widehat{\varphi_\lambda \otimes \phi_\mu}(\xi,\zeta) =& (2\pi)^{\frac{p+q}{2}} \prod_{i=1}^{p} g_1(\lambda_i \xi_i) \prod_{j=1}^{q} g_1(\mu_j \zeta_j) \\ =& \exp\left(-(\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2)/2\right).\end{aligned}$$

Thereafter, using Equation (88), the bias term $\|b\|_2^2$ can then expressed as follows

$$\begin{aligned}\|b\|_2^2 =& \frac{1}{(2\pi)^{p+q}} \int \left(1 - \exp\left(-(\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2)/2\right)\right)^2 |\hat{\psi}(\xi,\zeta)|^2 \, d\xi d\zeta \\ =& \frac{1}{(2\pi)^{p+q}} \int \left[\frac{\left(1 - \exp\left(-(\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2)/2\right)\right)}{\|(\xi,\zeta)\|^\delta}\right]^2 \|(\xi,\zeta)\|^{2\delta} |\hat{\psi}(\xi,\zeta)|^2 \, d\xi d\zeta.\end{aligned} \tag{89}$$

In order to upper bound the last integral, one can first notice that for all λ, ξ in $(0,+\infty)^p$ and μ, ζ in $(0,+\infty)^p$, we have: $\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2 \leq \|(\lambda,\mu)\|^2 \|(\xi,\zeta)\|^2$. We then obtain for all (ξ,ζ) in $\mathbb{R}^{p+q} \setminus \{0\}$,

$$\begin{aligned}\frac{1 - \exp\left(-(\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2)/2\right)}{\|(\xi,\zeta)\|^\delta} &\leq \frac{1 - \exp\left(-\|(\lambda,\mu)\|^2\|(\xi,\zeta)\|^2/2\right)}{\|(\xi,\zeta)\|^\delta} \\ &\leq \|(\lambda,\mu)\|^\delta \sup_{H>0} \frac{1 - \exp(-H/2)}{H^{\delta/2}}.\end{aligned}$$

For δ in $(0,2]$, the function $H \mapsto \dfrac{1 - \exp(-H/2)}{H^{\delta/2}}$ is bounded in $(0,+\infty)$. Indeed, it is continuous on $(0,+\infty)$, tends to 0 in $+\infty$ and has a finite limit at 0 (1/2 if $\delta = 2$ and 0 otherwise). Hence, we thus obtain for all (ξ,ζ) in $\mathbb{R}^{p+q} \setminus \{0\}$,

$$\left[\frac{\left(1 - \exp\left(-(\lambda_1^2\xi_1^2 + \ldots + \lambda_p^2\xi_p^2 + \mu_1^2\zeta_1^2 + \ldots + \mu_q^2\zeta_q^2)/2\right)\right)}{\|(\xi,\zeta)\|^\delta}\right]^2 \leq C(\delta)\|(\lambda,\mu)\|^{2\delta}.$$

Thereafter, using Hölder's inequality if $\delta \geq 1$ and the concavity of $t \mapsto t^\delta$ on $\mathbb{R}_+$ if $\delta < 1$, it is straightforward to see that

$$\|(\lambda,\mu)\|^{2\delta} \leq C(p,q,\delta) \left[\sum_{i=1}^{p} \lambda_i^{2\delta} + \sum_{j=1}^{q} \mu_j^{2\delta}\right]. \tag{90}$$

Hence, combining the two last inequalities gives

$$\|b\|_2^2 \leq C(p,q,\delta)\left[\sum_{i=1}^{p}\lambda_i^{2\delta}+\sum_{j=1}^{q}\mu_j^{2\delta}\right]\int \|(\xi,\zeta)\|^{2\delta}|\hat{\psi}(\xi,\zeta)|^2\,d\xi d\zeta.$$

Recalling that ψ belongs to the Sobolev ball $\mathcal{S}_{p+q}^{\delta}(R)$, we obtain

$$\|\psi-\psi*(\varphi_\lambda\otimes\phi_\mu)\|_2^2 \leq C(p,q,\delta,R)\left[\sum_{i=1}^{p}\lambda_i^{2\delta}+\sum_{j=1}^{q}\mu_j^{2\delta}\right].$$

6.9 Proof of Theorem 2

We easily deduce from Theorem 1 and Lemma 3 that if ψ belongs to the Sobolev balls $\mathcal{S}_{p+q}^{\delta}(R)$ with δ in $(0,2]$, $P_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q_{1-\alpha}^{\lambda,\mu}) \leq \beta$ as soon as

$$\|\psi\|_2^2 > C(p,q,\delta,R)\left[\sum_{i=1}^{p}\lambda_i^{2\delta}+\sum_{j=1}^{q}\mu_j^{2\delta}\right]+\frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

It now follows from the definition (6) of the uniform separation rate that

$$\left[\rho\left(\Delta_\alpha^{\lambda,\mu},\mathcal{S}_{p+q}^{\delta}(R),\beta\right)\right]^2 \leq C(p,q,\delta,R)\left[\sum_{i=1}^{p}\lambda_i^{2\delta}+\sum_{j=1}^{q}\mu_j^{2\delta}\right]+\frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

6.10 Proof of Corollary 2

The objective here is to give the uniform separation rate having the smallest upper bound w.r.t. the sample-size n, when ψ belongs to a Sobolev ball $\mathcal{S}_{p+q}^{\delta}(R)$ with δ in $(0,2]$. For this, we recall that according to Theorem 2, we have

$$\left[\rho\left(\Delta_\alpha^{\lambda,\mu},\mathcal{S}_{p+q}^{\delta}(R),\beta\right)\right]^2 \leq C(p,q,\delta,R)\left[\sum_{i=1}^{p}\lambda_i^{2\delta}+\sum_{j=1}^{q}\mu_j^{2\delta}\right]+\frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

In order to have the smallest behavior of the right side of the last inequality w.r.t. n, one has then to choose bandwidths $\lambda^*=(\lambda_1^*,\ldots,\lambda_p^*)$ and $\mu^*=(\mu_1^*,\ldots,\mu_q^*)$ w.r.t. n in such a way that

$$\sum_{i=1}^{p}\lambda_i^{*2\delta}+\sum_{j=1}^{q}\mu_j^{*2\delta} \quad\text{and}\quad \frac{1}{n\sqrt{\lambda_1^*\ldots\lambda_p^*\mu_1^*\ldots\mu_q^*}}$$

have the same behavior in n. Thereafter, it is clear that all λ_i^*'s and μ_j^*'s have the same behavior w.r.t. n. It obviously follows that for all i in $\{1,\ldots,p\}$ and all j in $\{1,\ldots,q\}$, we have

$$\lambda_i^*=\mu_j^*=n^{-2/(4\delta+p+q)}.$$

Consequently, the separation rate $\rho\left(\Delta_\alpha^{\lambda^*,\mu^*},\mathcal{S}_{p+q}^{\delta}(R),\beta\right)$ can be upper bounded as

$$\rho\left(\Delta_\alpha^{\lambda^*,\mu^*},\mathcal{S}_{p+q}^{\delta}(R),\beta\right) \leq C\left(M_f,p,q,\alpha,\beta,\delta\right)n^{-2\delta/(4\delta+p+q)}.$$

6.11 Proof of Lemma 4

The objective here is to give an upper bound of the bias term $\|\psi-\psi*(\varphi_\lambda\otimes\phi_\mu)\|_2^2$ w.r.t. λ and μ, when ψ belongs to a Nikol'skii-Besov ball $\mathcal{N}_{2,p+q}^{\delta}(R)$, with $\delta=(\nu_1,\ldots,\nu_p,\gamma_1,\ldots,\gamma_q)$ in $(0,2]^{p+q}$. We first set $b=\psi*(\varphi_\lambda\otimes\phi_\mu)-\psi$ and we write

$$\begin{aligned} b(x,y) =& \psi*(\varphi_\lambda\otimes\phi_\mu)(x,y)-\psi(x,y)\\ =& \int\psi(x',y')\varphi_\lambda(x-x')\phi_\mu(y-y')\mathrm{d}x'\mathrm{d}y'-\psi(x,y).\end{aligned}$$

Moreover, using Equations (3) and (4), the fonction b can be written in terms of the functions g_p and g_q defined in Equation (2) as

$$\begin{aligned} b(x,y) =& \frac{1}{\lambda_1 \dots \lambda_p \mu_1 \dots \mu_q} \int \psi(x',y') g_p\left(\frac{x_1 - x_1'}{\lambda_1}, \dots, \frac{x_p - x_p'}{\lambda_p}\right) g_q\left(\frac{y_1 - y_1'}{\mu_1}, \dots, \frac{y_q - y_q'}{\mu_p}\right) \mathrm{d}x' \mathrm{d}y' - \psi(x,y) \\ =& \int \psi(x_1 + \lambda_1 u_1, \dots, x_p + \lambda_p u_p, y_1 + \mu_1 v_1, \dots, y_q + \mu_q v_q) g_p(u_1, \dots, u_p) g_q(v_1, \dots, v_q) \, \mathrm{d}u\mathrm{d}v - \psi(x,y). \end{aligned}$$

Thereafter, using that $\int_{\mathbb{R}^p} g_p = \int_{\mathbb{R}^q} g_q = 1$, the function b can be expressed as

$$b(x,y) = \int g_p(u_1, \dots, u_p) g_q(v_1, \dots, v_q) \Big[\psi(x_1 + \lambda_1 u_1, \dots, x_p + \lambda_p u_p, y_1 + \mu_1 v_1, \dots, y_q + \mu_q v_q) - \psi(x,y) \Big] \, \mathrm{d}u\mathrm{d}v.$$

Let us from now define for all i in $\{1, \dots, p\}$ and j in $\{1, \dots, q\}$, the functions $b_{1,i}$ and $b_{2,j}$ by

$$\begin{aligned} b_{1,i}(x,y) &= \int g_p(u_1, \dots, u_p) g_q(v_1, \dots, v_q) \omega_{1,i}(x, y, u_1, \dots, u_i) \, \mathrm{d}u\mathrm{d}v, \\ b_{2,j}(x,y) &= \int g_p(u_1, \dots, u_p) g_q(v_1, \dots, v_q) \omega_{2,j}(x, y, u_1, \dots, u_p, v_1, \dots, v_j) \, \mathrm{d}u\mathrm{d}v, \end{aligned}$$

where the function $\omega_{1,i}$ is defined as

$$\omega_{1,i}(x, y, u_1, \dots, u_i) = \psi(x_1 + \lambda_1 u_1, \dots, x_i + \lambda_i u_i, x_{i+1}, \dots, x_p, y) - \psi(x_1 + \lambda_1 u_1, \dots, x_{i-1} + \lambda_{i-1} u_{i-1}, x_i, \dots, x_p, y),$$

while the function $\omega_{2,j}$ is defined as

$$\begin{aligned} \omega_{2,j}(x, y, u_1, \dots, u_p, v_1, \dots, v_j) =& \psi(x_1 + \lambda_1 u_1, \dots, x_p + \lambda_p u_p, y_1 + \mu_1 v_1, \dots, y_j + \mu_j v_j, y_{j+1}, \dots, y_q) \\ & - \psi(x_1 + \lambda_1 u_1, \dots, x_p + \lambda_p u_p, y_1 + \mu_1 v_1, \dots, y_{j-1} + \mu_{j-1} v_{j-1}, y_j, \dots, y_q). \end{aligned}$$

It is then easy to see that the function b is the sum of all the functions $b_{1,i}$ and $b_{2,j}$

$$b(x,y) = \sum_{i=1}^{p} b_{1,i}(x,y) + \sum_{j=1}^{q} b_{2,j}(x,y).$$

One can then deduce that it would be sufficient for the control of the $\mathbb{L}_2$-norm of b, to control the $\mathbb{L}_2$-normes of all the functions $b_{1,i}$ and $b_{2,j}$. Using the triangular inequality, we have

$$\|b\|_2 \leq \sum_{i=1}^{p} \|b_{1,i}\|_2 + \sum_{j=1}^{q} \|b_{2,j}\|_2. \tag{91}$$

By now, let us upper bound $\|b_{1,i}\|_2^2$ and $\|b_{2,j}\|_2^2$ for all i in $\{1, \dots, p\}$ and j in $\{1, \dots, q\}$. We distinguish two cases

Case 1. $0 < \nu_i \leq 1$

We first recall that $\|b_{1,i}\|_2^2$ can be written as

$$\|b_{1,i}\|_2^2 = \int \left[\int g_p(u_1, \dots, u_p) g_q(v_1, \dots, v_q) \omega_{1,i}(x, y, u_1, \dots, u_i) \, \mathrm{d}u\mathrm{d}v \right]^2 \mathrm{d}x\mathrm{d}y.$$

We use the following lemma from page 13 of [Tsybakov, 2009].

Lemma 11. *Let $\rho : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ be a Borel function, then we have the following inequality:*

$$\int \left(\int \rho(\theta, z) d\theta \right)^2 dz \leq \left[\int \left(\int \rho^2(\theta, z) dz \right)^{1/2} d\theta \right]^2.$$

By applying Lemma 11 to the function $((u,v),(x,y)) \mapsto g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\omega_{1,i}(x,y,u_1,\dots,u_i)$, we obtain:

$$\begin{aligned}\|b_{1,i}\|_2^2 &\leq \left[\int\left(\int g_p^2(u_1,\dots,u_p)g_q^2(v_1,\dots,v_q)\omega_{1,i}^2(x,y,u_1,\dots,u_i)\,\mathrm{d}x\mathrm{d}y\right)^{1/2}\mathrm{d}u\mathrm{d}v\right]^2 \\ &= \left[\int g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left(\int \omega_{1,i}^2(x,y,u_1,\dots,u_i)\,\mathrm{d}x\mathrm{d}y\right)^{1/2}\mathrm{d}u\mathrm{d}v\right]^2. \qquad (92)\end{aligned}$$

On the other hand, since ψ belongs to the Nikol'skii-Besov ball $\mathcal{N}_{2,p+q}^{\delta}(R)$, we have:

$$\left(\int \omega_{1,i}^2(x,y,u_1,\dots,u_i)\,\mathrm{d}x\mathrm{d}y\right)^{1/2} \leq R\lambda_i^{\nu_i}|u_i|^{\nu_i}.$$

We then have by injecting this last inequation in Equation (92), that

$$\|b_{1,i}\|_2^2 \leq C(R,\nu_i)\lambda_i^{2\nu_i}.$$

Case 2. $1 < \nu_i \leq 2$

In this case the function ψ has continuous first-order partial derivatives. Using Taylor expansion with integral form of the remainder w.r.t. the i^{th} variable of ψ, we have:

$$\omega_{1,i}(x,y,u_1,\dots,u_i) = \lambda_i u_i \int_0^1 (1-\tau)D_i^1\psi(x_1+\lambda_1u_1,\dots,x_i+\tau\lambda_iu_i,x_{i+1},\dots,y)d\tau.$$

where D_i^1 denotes the first-order partial derivative of ψ w.r.t. the i^{th} variable.

Thereafter, by injecting the last equation in the expression of $b_{1,i}$, we obtain:

$$b_{1,i}(x,y) = \int \lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left[\int_0^1 (1-\tau)D_i^1\psi(x_1+\lambda_1u_1,\dots,x_i+\tau\lambda_iu_i,x_{i+1},\dots,y)d\tau\right]\mathrm{d}u\mathrm{d}v.$$

Furthermore, using the fact that g_p is the density function of the multivariate normal distribution with mean 0 and covariance matrix equals identity, we have that $\int u_i g_p(u_1,\dots,u_p)\mathrm{d}u_i = 0$. The function $b_{1,i}$ can then be written as

$$b_{1,i}(x,y) = \int \lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left[\int_0^1 (1-\tau)D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)\,d\tau\right]\mathrm{d}u\mathrm{d}v.$$

We have then the following equation for the $\mathbb{L}_2$-norm of $b_{1,i}$:

$$\|b_{1,i}\|_2^2 = \int\left[\int \lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left(\int_0^1 (1-\tau)D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)\,d\tau\right)\mathrm{d}u\mathrm{d}v\right]^2\mathrm{d}x\mathrm{d}y.$$

By now, we use as in Case 1 of Lemma 11 in order to upper bound $\|b_{1,i}\|_2^2$. We then obtain:

$$\begin{aligned}\|b_{1,i}\|_2^2 &\leq \left(\int\left[\int\left(\lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\int_0^1 (1-\tau)D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)\,d\tau\right)^2\mathrm{d}x\mathrm{d}y\right]^{1/2}\mathrm{d}u\mathrm{d}v\right)^2 \\ &= \left(\int \lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left[\int\left(\int_0^1 (1-\tau)D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)\,d\tau\right)^2\mathrm{d}x\mathrm{d}y\right]^{1/2}\mathrm{d}u\mathrm{d}v\right)^2.\end{aligned}$$

We apply a second time Lemma 11. For this, consider the function $\rho\left((x,y),\tau\right) = (1-\tau)D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)$, we then have:

$$\|b_{1,i}\|_2^2 \leq \left(\int \lambda_i u_i g_p(u_1,\dots,u_p)g_q(v_1,\dots,v_q)\left[\int_0^1 (1-\tau)\left(\int \left(D_i^1\omega_{1,i}(x,y,u_1,\dots,\tau u_i)\right)^2\mathrm{d}x\mathrm{d}y\right)^{1/2}d\tau\right]\mathrm{d}u\mathrm{d}v\right)^2. \qquad (93)$$

On the other hand, using that ψ belongs to the Nikol'skii-Besov ball $\mathcal{N}^{\delta}_{2,p+q}(R)$:

$$\left(\int \left(D_i^1 \omega_{1,i}(x,y,u_1,\ldots,\tau u_i)\right)^2 \mathrm{d}x\mathrm{d}y\right)^{1/2} \leq R\lambda_i^{\nu_i-1}|\tau u_i|^{\nu_i-1}.$$

We then obtain by injecting this last inequation in Equation (93), that

$$\|b_{1,i}\|_2^2 \leq C(R,\nu_i)\lambda_i^{2\nu_i}.$$

Besides, for all j in $\{1,\ldots,q\}$, by similar arguments:

$$\|b_{2,j}\|_2^2 \leq C(R,\gamma_j)\mu_j^{2\gamma_j}.$$

Consequently, according to Equation (91), we have the following upper bound of $\|b\|_2^2$

$$\|b\|_2^2 \leq C(R,\delta)\left[\sum_{i=1}^{p}\lambda_i^{2\nu_i} + \sum_{j=1}^{q}\mu_j^{2\gamma_j}\right].$$

6.12 Proof of Theorem 3

The proof of this theorem is similar to that of Theorem 2. Indeed, assuming the conditions of Theorem 1, we have according to this theorem and Lemma 4 that if ψ belongs to $\mathcal{N}^{\delta}_{2,p+q}(R)$, with $\delta = (\nu_1,\ldots,\nu_p,\gamma_1,\ldots,\gamma_q)$ in $(0,2]^{p+q}$, then $P_f(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q^{\lambda,\mu}_{1-\alpha}) \leq \beta$ as soon as

$$\|\psi\|_2^2 > C(R,\delta)\left[\sum_{i=1}^{p}\lambda_i^{2\nu_i} + \sum_{j=1}^{q}\mu_j^{2\gamma_j}\right] + \frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

One can then conclude from the definition (6) of the uniform separation rate that

$$\left[\rho\left(\Delta_{\alpha}^{\lambda,\mu},\mathcal{N}^{\delta}_{2,p+q}(R),\beta\right)\right]^2 \leq C(R,\delta)\left[\sum_{i=1}^{p}\lambda_i^{2\nu_i} + \sum_{j=1}^{q}\mu_j^{2\gamma_j}\right] + \frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

6.13 Proof of Corollary 3

We aim here to give the uniform separation rate having the smallest upper bound w.r.t. the sample-size n, when ψ belongs to a Nikol'skii-Besov ball $\mathcal{N}^{\delta}_{2,p+q}(R)$, with $\delta = (\nu_1,\ldots,\nu_p,\gamma_1,\ldots,\gamma_q)$ in $(0,2]^{p+q}$. We first recall that Theorem 3 shows that:

$$\left[\rho\left(\Delta_{\alpha}^{\lambda,\mu},\mathcal{N}^{\delta}_{2,p+q}(R),\beta\right)\right]^2 \leq C(R,\delta)\left[\sum_{i=1}^{p}\lambda_i^{2\nu_i} + \sum_{j=1}^{q}\mu_j^{2\gamma_j}\right] + \frac{C\left(M_f,p,q,\beta\right)}{n\sqrt{\lambda_1\ldots\lambda_p\mu_1\ldots\mu_q}}\log\left(\frac{1}{\alpha}\right).$$

So as to minimize the right side of the last inequality w.r.t. n, we have to choose bandwidths $\lambda^* = (\lambda_1^*,\ldots,\lambda_p^*)$ and $\mu^* = (\mu_1^*,\ldots,\mu_q^*)$ w.r.t. n such as

$$\left[\sum_{i=1}^{p}\lambda_i^{*2\nu_i} + \sum_{j=1}^{q}\mu_j^{*2\gamma_j}\right] \quad \text{and} \quad \frac{1}{n\sqrt{\lambda_1^*\ldots\lambda_p^*\mu_1^*\ldots\mu_q^*}}$$

have the same behavior in n. Let us set for all i in $\{1,\ldots,p\}$ and all j in $\{1,\ldots,q\}$, $\lambda_i^* = n^{a_i}$ and $\mu_j^* = n^{b_j}$. It is than clear that for all i and all j:

$$2a_i\nu_i = 2b_j\gamma_j = -\frac{1}{2}\left[\sum_{r=1}^{p}a_r + \sum_{s=1}^{q}b_s\right] - 1. \tag{94}$$

One can first express all a_i's and all b_j's w.r.t a_1 as

$$a_i = a_1 \frac{\nu_1}{\nu_i} \quad \text{and} \quad b_j = a_1 \frac{\nu_1}{\gamma_j}.$$

Thereafter, using Equation (94) we have the following:

$$2a_1\nu_1 = \frac{-a_1\nu_1}{2\eta} - 1.$$

We then first write that $a_1 = \dfrac{-2\eta}{\nu_1(4\eta+1)}$. We next obtain for all i and for all j that:

$$a_i = \frac{-2\eta}{\nu_i(4\eta+1)} \quad \text{and} \quad b_j = \frac{-2\eta}{\gamma_j(4\eta+1)}.$$

Consequently, the separation rate $\rho\left(\Delta_\alpha^{\lambda^*,\mu^*}, \mathcal{N}_{2,p+q}^{\delta}(R), \beta\right)$ can be upper bound as

$$\rho\left(\Delta_\alpha^{\lambda^*,\mu^*}, \mathcal{N}_{2,p+q}^{\delta}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right) n^{-\frac{2\eta}{(1+4\eta)}}.$$

6.14 Proof of Lemma 5

Let α be in $(0,1)$, we first prove that $u_\alpha \geq \alpha$. For this, we apply Bonferroni's Inequality:

$$\begin{aligned}
&P_{f_1\otimes f_2}\left(\sup_{(\lambda,\mu)\in\Lambda\times U}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} - q^{\lambda,\mu}_{1-\alpha e^{-\omega_{\lambda,\mu}}}\right) > 0\right)\\
=&P_{f_1\otimes f_2}\left(\bigcup_{(\lambda,\mu)\in\Lambda\times U}\left\{\widehat{\mathrm{HSIC}}_{\lambda,\mu} > q^{\lambda,\mu}_{1-\alpha e^{-\omega_{\lambda,\mu}}}\right\}\right)\\
\leq& \sum_{(\lambda,\mu)\in\Lambda\times U} P_{f_1\otimes f_2}\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} > q^{\lambda,\mu}_{1-\alpha e^{-\omega_{\lambda,\mu}}}\right)\\
\leq& \sum_{(\lambda,\mu)\in\Lambda\times U} \alpha e^{-\omega_{\lambda,\mu}}\\
\leq& \ \alpha.
\end{aligned}$$

Then, by definition of u_α we have: $u_\alpha \geq \alpha$. Thereafter, we obtain:

$$\begin{aligned}
P_f\left(\Delta_\alpha = 0\right) &= P_f\left(\bigcap_{(\lambda,\mu)\in\Lambda\times U}\left\{\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q^{\lambda,\mu}_{1-u_\alpha e^{-\omega_{\lambda,\mu}}}\right\}\right)\\
&\leq \inf_{(\lambda,\mu)\in\Lambda\times U} P_f\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q^{\lambda,\mu}_{1-u_\alpha e^{-\omega_{\lambda,\mu}}}\right)\\
&\leq \inf_{(\lambda,\mu)\in\Lambda\times U} P_f\left(\widehat{\mathrm{HSIC}}_{\lambda,\mu} \leq q^{\lambda,\mu}_{1-\alpha e^{-\omega_{\lambda,\mu}}}\right)\\
&= \inf_{(\lambda,\mu)\in\Lambda\times U}\left\{P_f\left(\Delta^{\lambda,\mu}_{\alpha e^{-\omega_{\lambda,\mu}}} = 0\right)\right\},
\end{aligned}$$

which concludes the proof.

6.15 Proof of Theorem 4

Let α and β be in $(0,1)$. According to Lemma 5, $P_f\left(\Delta_\alpha = 0\right) \leq \beta$ as soon as there exists (λ,μ) in $\Lambda\times U$ such that

$$P_f\left(\Delta^{\lambda,\mu}_{\alpha e^{-\omega_{\lambda,\mu}}} = 0\right) \leq \beta.$$

Then, according to Theorem 2 (resp. Theorem 3) if ψ belongs to $\mathcal{N}_{2,p+q}^{\delta}(R)$ (resp. ψ belongs to $\mathcal{S}_{p+q}^{\delta}(R)$): we take the infimum of the upper bounds for the uniform separation rates of the single tests over $\Lambda\times U$ while replacing $\log\left(\dfrac{1}{\alpha}\right)$ by $\log\left(\dfrac{1}{\alpha}\right) + \omega_{\lambda,\mu}$.

6.16 Proof of Corollary 4

Let us start with the case where ψ belongs to $\mathcal{N}^{\delta}_{2,p+q}(R)$. In this case, using Theorem 4, we have the following inequality for $\rho\left(\Delta_\alpha, \mathcal{N}^{\delta}_{2,p+q}(R), \beta\right)$,

$$\left[\rho\left(\Delta_\alpha, \mathcal{N}^{\delta}_{2,p+q}(R), \beta\right)\right]^2 \leq C\left(M_f, p, q, \beta, \delta\right) \inf_{(\lambda,\mu)\in\Lambda\times U} \left\{ \frac{1}{\sqrt{\lambda_1 \dots \lambda_p \mu_1 \dots \mu_q n}} \left(\log\left(\frac{1}{\alpha}\right) + \omega_{\lambda,\mu} \right) + \left[\sum_{i=1}^{p} \lambda_i^{2\nu_i} + \sum_{j=1}^{q} \mu_j^{2\gamma_j} \right] \right\}$$

Let us take $\lambda^* = (2^{-m^*_{1,1}}, \dots, 2^{-m^*_{1,p}})$ and $\mu^* = (2^{-m^*_{2,1}}, \dots, 2^{-m^*_{2,q}})$, where the integers $m^*_{1,1}, \dots, m^*_{1,p}, m^*_{2,1}, \dots, m^*_{2,q}$ are defined as follows

$$m^*_{1,i} = \left\lfloor \log_2\left(\left(\frac{n}{\log\log(n)} \right)^{\frac{2\eta}{\nu_i(1+4\eta)}} \right) \right\rfloor \quad \text{and} \quad m^*_{2,j} = \left\lfloor \log_2\left(\left(\frac{n}{\log\log(n)} \right)^{\frac{2\eta}{\gamma_j(1+4\eta)}} \right) \right\rfloor.$$

where $\dfrac{1}{\eta} = \displaystyle\sum_{i=1}^{p} \frac{1}{\nu_i} + \sum_{j=1}^{q} \frac{1}{\gamma_j}$.

Then, we obviously have

$$\left[\rho\left(\Delta_\alpha, \mathcal{N}^{\delta}_{2,p+q}(R), \beta\right)\right]^2 \leq C\left(M_f, p, q, \beta, \delta\right) \left[\frac{1}{\sqrt{\lambda^*_1 \dots \lambda^*_p \mu^*_1 \dots \mu^*_q n}} \left(\log\left(\frac{1}{\alpha}\right) + \omega_{\lambda^*,\mu^*} \right) + \sum_{i=1}^{p} (\lambda^*_i)^{2\nu_i} + \sum_{j=1}^{q} (\mu^*_j)^{2\gamma_j} \right].$$

Besides, using the inequalities

$$m^*_{1,i} \leq \log_2\left(\left(\frac{n}{\log\log(n)} \right)^{\frac{2\eta}{\nu_i(1+4\eta)}} \right) \quad \text{and} \quad m^*_{2,j} \leq \log_2\left(\left(\frac{n}{\log\log(n)} \right)^{\frac{2\eta}{\gamma_j(1+4\eta)}} \right),$$

we upper bound $(\lambda^*_i)^{-1/2}$ and $(\mu^*_j)^{-1/2}$ by

$$(\lambda^*_i)^{-1/2} = 2^{m^*_{1,i}/2} \leq \left(\frac{n}{\log\log(n)} \right)^{\frac{\eta}{\nu_i(1+4\eta)}} \quad \text{and} \quad \mu^*_j = 2^{m^*_{2,j}/2} \leq \left(\frac{n}{\log\log(n)} \right)^{\frac{\eta}{\gamma_j(1+4\eta)}}.$$

Therefore, we obtain

$$(\lambda^*_1 \dots \lambda^*_p \mu^*_1 \dots \mu^*_q)^{-1/2} \leq \left(\frac{n}{\log\log(n)} \right)^{\frac{1}{(1+4\eta)}}. \tag{95}$$

Let us now upper bound ω_{λ^*,μ^*}, we first write

$$\begin{aligned} \omega_{\lambda^*,\mu^*} &= 2\sum_{i=1}^{p} \log(m^*_{1,i} \times \pi/\sqrt{6}) + 2\sum_{j=1}^{q} \log(m^*_{2,j} \times \pi/\sqrt{6}) \\ &= 2\log\left(m^*_{1,1} \dots m^*_{1,p} m^*_{2,1} \dots m^*_{2,q}\right) + 2(p+q)\log(\pi/\sqrt{6}). \end{aligned}$$

Moreover, it is easy to see that

$$m^*_{1,i} \leq \frac{2\eta}{\nu_i(1+4\eta)} \log(n) \quad \text{and} \quad \mu^*_j \leq \frac{2\eta}{\gamma_j(1+4\eta)} \log(n).$$

Then,

$$\log(m^*_{1,1} \dots m^*_{1,p} m^*_{2,1} \dots m^*_{2,q}) \leq C(\delta) \log\log(n).$$

Thereafter, ω_{λ^*,μ^*} can be upper bound as

$$\omega_{\lambda^*,\mu^*} \leq C(\delta) \log\log(n). \tag{96}$$

From Equations (95) and (96), we have

$$\frac{1}{n\sqrt{\lambda^*_1 \dots \lambda^*_p \mu^*_1 \dots \mu^*_q}} \left(\log\left(\frac{1}{\alpha}\right) + \omega_{\lambda^*,\mu^*} \right) \leq C(\alpha, \delta) \left(\frac{\log\log(n)}{n} \right)^{\frac{4\eta}{(1+4\eta)}}. \tag{97}$$

We aim now to upper bound $\sum_{i=1}^{p}(\lambda_i^*)^{2\nu_i} + \sum_{j=1}^{q}(\mu_j^*)^{2\gamma_j}$. For this, we first write

$$m_{1,i}^* \geq \log_2\left(\left(\frac{n}{\log\log(n)}\right)^{\frac{2\eta}{\nu_i(1+4\eta)}}\right) - 1 \quad \text{and} \quad m_{2,j}^* \geq \log_2\left(\left(\frac{n}{\log\log(n)}\right)^{\frac{2\eta}{\gamma_j(1+4\eta)}}\right) - 1.$$

We then have the following inequalities for $(\lambda_i^*)^{2\nu_i}$ and $(\mu_j^*)^{2\gamma_j}$,

$$(\lambda_i^*)^{2\nu_i} \leq 2^{2\nu_i}\left(\frac{\log\log(n)}{n}\right)^{\frac{4\eta}{(1+4\eta)}} \quad \text{and} \quad (\mu_j^*)^{2\gamma_j} \leq 2^{2\gamma_j}\left(\frac{\log\log(n)}{n}\right)^{\frac{4\eta}{(1+4\eta)}}.$$

Therefore, we obtain

$$\sum_{i=1}^{p}(\lambda_i^*)^{2\nu_i} + \sum_{j=1}^{q}(\mu_j^*)^{2\gamma_j} \leq C(\delta)\left(\frac{\log\log(n)}{n}\right)^{\frac{4\eta}{(1+4\eta)}}. \tag{98}$$

Consequently, from Equations (97) and (98),

$$\rho\left(\Delta_\alpha, \mathcal{N}_{2,p+q}^{\delta}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right)\left(\frac{\log\log(n)}{n}\right)^{\frac{2\eta}{(1+4\eta)}}.$$

In the case where ψ belongs to $\mathcal{S}_{p+q}^{\delta}(R)$, the same arguments above is applied by taking $\delta_1 = \ldots = \delta_p = \gamma_1 = \ldots = \gamma_q = \delta$, lead to

$$\rho\left(\Delta_\alpha, \mathcal{S}_{p+q}^{\delta}(R), \beta\right) \leq C\left(M_f, p, q, \alpha, \beta, \delta\right)\left(\frac{\log\log(n)}{n}\right)^{\frac{2\eta}{(1+4\eta)}},$$

where $\dfrac{1}{\eta} = (p+q)\dfrac{1}{\delta}$.

6.17 Proof of Lemma 6

Assume there exists a distribution f_0 that satisfies $(\mathcal{H}_0)$ such that the probability measure $P_{\nu_{\rho_*}}$ is absolutely continuous w.r.t. P_{f_0} and verifies Equation (28).

Let us first lower bound $\beta\left[\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)\right]$ w.r.t. the distributions $P_{\nu_{\rho_*}}$ and P_{f_0},

$$\begin{aligned}
\beta\left[\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)\right] &\geq \inf_{\Delta_\alpha} P_{\nu_{\rho_*}}\left(\Delta_\alpha = 0\right) \\
&= 1 - \sup_{\Delta_\alpha} P_{\nu_{\rho_*}}\left(\Delta_\alpha = 1\right) \\
&\geq 1 - \alpha - \sup_{\Delta_\alpha}\left|P_{\nu_{\rho_*}}\left(\Delta_\alpha = 1\right) - P_{f_0}\left(\Delta_\alpha = 1\right)\right|.
\end{aligned}$$

We denote by $\left\|P_{\nu_{\rho_*}} - P_{f_0}\right\|_{TV}$ the total variation distance between the distributions $P_{\nu_{\rho_*}}$ and P_{f_0}. We recall that,

$$\left\|P_{\nu_{\rho_*}} - P_{f_0}\right\|_{TV} = \sup_{E\in\mathcal{E}}\left|P_{\nu_{\rho_*}}(E) - P_{f_0}(E)\right|,$$

where $\mathcal{E}$ is the space of measurable sets. We then obtain

$$\beta\left[\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)\right] \geq 1 - \alpha - \left\|P_{\nu_{\rho_*}} - P_{f_0}\right\|_{TV}.$$

Notice that,

$$\left\|P_{\nu_{\rho_*}} - P_{f_0}\right\|_{TV} = \sup_{E\in\mathcal{E}}\left[P_{\nu_{\rho_*}}(E) - P_{f_0}(E)\right] = \sup_{E\in\mathcal{E}}\left[P_{f_0}(E) - P_{\nu_{\rho_*}}(E)\right].$$

It is then straightforward to show that

$$\begin{aligned}
\left\|P_{\nu_{\rho_*}} - P_{f_0}\right\|_{TV} &= \frac{1}{2}\int\left|L_{\nu_{\rho_*}} - 1\right| \mathrm{d}P_{f_0} \\
&= \frac{1}{2}\mathbb{E}_{P_{f_0}}\left[\left|L_{\nu_{\rho_*}}(\mathbb{Z}_n) - 1\right|\right] \\
&\leq \frac{1}{2}\left(\mathbb{E}_{P_{f_0}}\left[L_{\nu_{\rho_*}}^2(\mathbb{Z}_n)\right] - 1\right)^{1/2},
\end{aligned}$$

where the last inequality holds by applying Cauchy-Schwarz and the fact that $\mathbb{E}_{P_{f_0}}\left[L_{\nu_{\rho_*}}(\mathbb{Z}_n)\right] = 1$. Thus,

$$\beta\big[\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)\big] \geq 1 - \alpha - \frac{1}{2}\left(\mathbb{E}_{P_{f_0}}\left[L^2_{\nu_{\rho_*}}(\mathbb{Z}_n)\right] - 1\right)^{1/2}.$$

If the condition (28) holds, we then obtain

$$\beta\big[\mathcal{F}_{\rho_*}(\mathcal{C}_\delta)\big] > \beta.$$

Furthermore, using that $\mathcal{F}_{\rho_*}(\mathcal{C}_\delta) \subset \mathcal{F}_\rho(\mathcal{C}_\delta)$ for all $\rho \leq \rho_*$, we have

$$\beta\big[\mathcal{F}_{\rho}(\mathcal{C}_\delta)\big] > \beta.$$

Let us now prove that this implies the lower bound

$$\rho\left(\mathcal{C}_\delta, \alpha, \beta\right) = \inf_{\Delta_\alpha} \rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right) \geq \rho_*. \tag{99}$$

Assume $\beta\big[\mathcal{F}_{\rho_n^*}(\mathcal{C}_\delta)\big] > \beta$, then

$$\forall \Delta_\alpha, \qquad \sup_{f \in \mathcal{F}_{\rho_n^*}(\mathcal{C}_\delta)} P_f\left(\Delta_\alpha = 0\right) > \beta.$$

In particular, since the family $\left\{\mathcal{F}_\rho\left(\mathcal{C}_\delta\right)\right\}_{\rho>0}$ is non increasing for the inclusion,

$$\forall \Delta_\alpha, \quad \rho\left(\Delta_\alpha, \mathcal{C}_\delta, \beta\right) = \inf\left\{\rho > 0 \ ; \ \sup_{f \in \mathcal{F}_\rho(\mathcal{C}_\delta)} P_f\left(\Delta_\alpha = 0\right) \leq \beta\right\} > \rho_n^*,$$

which directly implies (99).

6.18 Proof of Proposition 5

Proof of 1. Let us prove that the functions f_θ are probability density functions for n large enough. First, it is obvious from Equation (30) that

$$\int_{\mathbb{R}^{p+q}} f_\theta(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1,$$

since $f_1 \otimes f_2$ is a probability density function and that $\displaystyle\int_{\mathbb{R}} G(x)\,\mathrm{d}x = 0$. It remains to check that f_θ is a non-negative function for n large enough.

Let $j = (j_1, \ldots, j_p)$ in $\{1, \ldots, M_n\}^p$ and $l = (l_1, \ldots, l_1)$ in $\{1, \ldots, M_n\}^q$. Knowing that for all $1 \leq r \leq p$ and all $1 \leq s \leq q$, the supports of the functions $G_{h_n}(\cdot - j_r h_n)$ and $G_{h_n}(\cdot - l_s h_n)$ are respectively the intervals $\Big((j_r - 1)h_n, j_r h_n\Big]$ and $\Big((l_s - 1)h_n, l_s h_n\Big]$, the support of the function

$$g_{n,j,l} : (x,y) \mapsto \prod_{r=1}^{p} G_{h_n}(x_r - j_r h_n) \prod_{s=1}^{q} G_{h_n}(y_s - l_s h_n) \tag{100}$$

is the set

$$D_{(j,l)} = \prod_{r=1}^{p} \Big((j_r - 1)h_n, j_r h_n\Big] \times \prod_{s=1}^{q} \Big((l_s - 1)h_n, l_s h_n\Big]. \tag{101}$$

These supports are then disjoint for different multi-indexes (j,l) in $I_{n,p,q}$ and have as union set $(0,1]^{p+q}$ (since $M_n h_n = 1$). In particular, for all (x,y) in $(0,1]^{p+q}$,

$$\begin{aligned}\left|\sum_{(j,l)\in I_{n,p,q}} \theta_{(j,l)} \prod_{r=1}^{p} G_{h_n}(x_r - j_r h_n) \prod_{s=1}^{q} G_{h_n}(y_s - l_s h_n)\right| &\leq \frac{1}{h_n^{p+q}}\left(\sup_{t\in[-1,0]} |G(t)|\right)^{p+q} \\ &= \frac{1}{(eh_n)^{p+q}}.\end{aligned}$$

Hence, if (x, y) belongs to $[0,1]^{p+q}$, then

$$f_\theta(x,y) \geq 1 - \frac{h_n^\delta}{e^{p+q}},$$

which is non negative for h_n small enough. Otherwise, $f_\theta(x,y) = 0$. In particular, for all (x,y) in $\mathbb{R}^{p+q}$, $f_\theta(x,y) \geq 0$ which ends the proof of this first point.

Proof of 2. Let us prove that the functions $f_\theta - f_1 \otimes f_2$ belong to the Sobolev ball $\mathcal{S}^\delta_{p+q}(R)$ for n large enough. This point relies on Lemma [Butucea, 2007, Lemma 2] recalled bellow.

Lemma 12 ([Butucea, 2007]). *Let G be the function defined in Equation (29). Then G is an infinitely differentiable function such that $\int_{\mathbb{R}} G(x)\,\mathrm{d}x = 0$. Its Fourier transform verifies*

$$|\widehat{G}(u)| \leq C \exp(-a\sqrt{|u|}) \quad as \quad |u| \to \infty,$$

for some positive constants C and a. Moreover, $\widehat{G}$ is an infinitely differentiable and bounded function.

According to the Fourier transform properties, we write

$$\widehat{f}_\theta(u,v) = \widehat{f}_1 \otimes \widehat{f}_2(u,v) + h_n^{\delta+(p+q)} \sum_{(j,l)\in I_{n,p,q}} \theta_{j,l} \prod_{r=1}^{p} \exp(iu_r j_r h_n)\widehat{G}(h_n u_r) \prod_{s=1}^{q} \exp(iv_s l_s h_n)\widehat{G}(h_n v_s).$$

Then,

$$\left|\widehat{f}_\theta(u,v) - \widehat{f}_1 \otimes \widehat{f}_2(u,v)\right|^2 = H_{1,n}(u,v) + H_{2,n}(u,v), \tag{102}$$

where the functions $H_{1,n}$ and $H_{2,n}$ are respectively defined by

$$H_{1,n}(u,v) = M_n^{p+q} h_n^{2\delta+2(p+q)} \prod_{r=1}^{p} |\widehat{G}(h_n u_r)|^2 \prod_{s=1}^{q} |\widehat{G}(h_n v_s)|^2, \tag{103}$$

$$H_{2,n}(u,v) = h_n^{2\delta+2(p+q)} \sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \theta_{j_1,l_1}\theta_{j_2,l_2}\mathcal{G}_{j_1,l_1,j_2,l_2}(h_n u, h_n v), \tag{104}$$

where the vectors j_1, j_2, l_1 and l_2 in the last sum are defined as $j_1 = (j_{1,1}, \ldots, j_{1,p})$, $j_2 = (j_{2,1}, \ldots, j_{2,p})$, $l_1 = (l_{1,1}, \ldots, l_{1,q})$ and $l_2 = (l_{2,1}, \ldots, l_{2,q})$. In addition, the function $\mathcal{G}_{j_1,l_1,j_2,l_2}$ is defined as

$$\mathcal{G}_{j_1,l_1,j_2,l_2} : (u,v) \mapsto \prod_{r=1}^{p} \exp\left(iu_r(j_{1,r} - j_{2,r})\right) |\widehat{G}(u_r)|^2 \prod_{s=1}^{q} \exp\left(iv_s(l_{1,s} - l_{1,s})\right) |\widehat{G}(v_s)|^2. \tag{105}$$

By now, we aim to upper bound the integral of the function $(u,v) \mapsto \|(u,v)\|^{2\delta} \left|\widehat{f}_\theta(u,v) - \widehat{f}_1 \otimes \widehat{f}_2(u,v)\right|^2$. First, one easily show, as in Equation (90), that

$$\|(u,v)\|^{2\delta} \leq C(p,q,\delta) \left[\sum_{i=1}^{p} |u_i|^{2\delta} + \sum_{j=1}^{q} |v_j|^{2\delta}\right]. \tag{106}$$

We then obtain from Equations (103) and (106) the following result,

$$\begin{aligned}\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{1,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v &\leq C(p,q,\delta) M_n^{p+q} h_n^{2\delta+2(p+q)} \left(\int_{\mathbb{R}} |t|^{2\delta}|\widehat{G}(h_n t)|^2\,\mathrm{d}t\right)\left(\int_{\mathbb{R}} |\widehat{G}(h_n z)|^2\,\mathrm{d}z\right)^{p+q-1} \\ &= C(p,q,\delta)(M_n h_n)^{p+q} \left(\int_{\mathbb{R}} |t|^{2\delta}|\widehat{G}(t)|^2\,\mathrm{d}t\right)\left(\int_{\mathbb{R}} |\widehat{G}(z)|^2\,\mathrm{d}z\right)^{p+q-1}.\end{aligned}$$

The functions $t \mapsto |t|^{2\delta}|\widehat{G}(t)|^2$ and $z \mapsto |\widehat{G}(z)|^2$ being integrable according to Lemma 12, we have

$$\begin{aligned}\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{1,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v &\leq C(p,q,\delta)(M_n h_n)^{p+q} \\ &\leq C(p,q,\delta). \end{aligned} \tag{107}$$

To complete this point of the proof, we demonstrate a similar equation for $H_{2,n}$. Starting from the expression of $H_{2,n}$ in (104), we write

$$\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{2,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v = h_n^{p+q} \sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \theta_{j_1,l_1}\theta_{j_2,l_2} \int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

Therefore, according to the triangular inequality, we obtain

$$\left|\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{2,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v\right| \leq h_n^{p+q} \sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \left|\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v\right|. \tag{108}$$

We first assume that all the components of the vectors (j_1,l_1) and (j_2,l_2) are different. Knowing that $\widehat{G}$ is differentiable, the function $|\widehat{G}|^2 = \widehat{G}\overline{\widehat{G}}$ is also differentiable. We apply an integration by parts to the function $\mathcal{G}_{j_1,l_1,j_2,l_2}$ w.r.t. the variable u_1,

$$\begin{aligned}\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v &= \frac{1}{i(j_{1,1}-j_{2,1})} \int_{\mathbb{R}^{p+q-1}} \left[\|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\right]_{u_1=-\infty}^{u_1=+\infty} \frac{\mathrm{d}u}{\mathrm{d}u_1}\,\mathrm{d}v \\ &\quad - \frac{1}{i(j_{1,1}-j_{2,1})} \int_{\mathbb{R}^{p+q}} \frac{\partial}{\partial u_1}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2\right\} \frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2}\,\mathrm{d}u\,\mathrm{d}v. \end{aligned} \tag{109}$$

Moreover, it is straightforward to see that the first term on the right side of Equation (109) is equal to zero. Indeed, for all a in $\mathbb{R}$ and $(u_2,\ldots,u_p,v)$ in $\mathbb{R}^{p+q-1}$, we have the following

$$\left|\left[\|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\right]_{u_1=-a}^{u_1=+a}\right| \leq \|(a,u_2,\ldots,u_p,v)\|^{2\delta} \left(|\widehat{G}(a)|^2 + |\widehat{G}(-a)|^2\right) \prod_{r=2}^{p} |\widehat{G}(u_r)|^2 \prod_{s=1}^{q} |\widehat{G}(v_s)|^2.$$

In addition, according to Lemma 12 we obviously obtain for all $u_2,\ldots,u_p,v$ the following

$$\lim_{|a|\to+\infty} \|(a,u_2,\ldots,u_p,v)\|^{2\delta}|\widehat{G}(a)|^2 = \lim_{|a|\to+\infty} \|(a,u_2,\ldots,u_p,v)\|^{2\delta}|\widehat{G}(-a)|^2 = 0. \tag{110}$$

It follows from Equation (109) that

$$\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v = \frac{i}{(j_{1,1}-j_{2,1})} \int_{\mathbb{R}^{p+q}} \frac{\partial}{\partial u_1}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2\right\} \frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2}\,\mathrm{d}u\,\mathrm{d}v. \tag{111}$$

From Equation (111), we perform an integration by parts w.r.t. u_2,

$$\begin{aligned}&(j_{1,1}-j_{2,1})(j_{1,2}-j_{2,2}) \int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\ &\quad = \int_{\mathbb{R}^{p+q-1}} \left[\frac{\partial}{\partial u_1}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2\right\} \frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2}\right]_{u_2=-\infty}^{u_2=+\infty} \frac{\mathrm{d}u}{\mathrm{d}u_2}\,\mathrm{d}v \\ &\qquad - \int_{\mathbb{R}^{p+q}} \frac{\partial^2}{\partial u_1 \partial u_2}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2|\widehat{G}(u_2)|^2\right\} \frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2|\widehat{G}(u_2)|^2}\,\mathrm{d}u\,\mathrm{d}v. \end{aligned}$$

Moreover, by analogy with Equation (110), we can easily show using Lemma 12 that

$$\left[\frac{\partial}{\partial u_1}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2\right\}\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2}\right]_{u_2=-\infty}^{u_2=+\infty}=0.$$

This leads to

$$\begin{aligned}(j_{1,1}-j_{2,1})(j_{1,2}-j_{2,2})\int_{\mathbb{R}^{p+q}}\|(u,v)\|^{2\delta}\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v\\ =-\int_{\mathbb{R}^{p+q}}\frac{\partial^2}{\partial u_1\partial u_2}\left\{\|(u,v)\|^{2\delta}|\widehat{G}(u_1)|^2|\widehat{G}(u_2)|^2\right\}\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{|\widehat{G}(u_1)|^2|\widehat{G}(u_2)|^2}\,\mathrm{d}u\,\mathrm{d}v.\end{aligned}$$

By repeating this last process for the variables $u_3,\ldots,u_p,v_1,\ldots,v_q$, we get

$$\begin{aligned}&\prod_{r=1}^{p}(j_{1,r}-j_{2,r})\prod_{s=1}^{q}(l_{1,s}-l_{1,s})\int_{\mathbb{R}^{p+q}}\|(u,v)\|_2^{2\delta}\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v\\ &=(i)^{p+q}\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q}}{\partial u_1\ldots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2}\,\mathrm{d}u\,\mathrm{d}v.\end{aligned}\tag{112}$$

Starting from Equation (112), we perform a second integration by part w.r.t. u_1. We write

$$\begin{aligned}&\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q}W}{\partial u_1\ldots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v\\ &\quad=\frac{1}{i(j_{1,1}-j_{2,1})}\int_{\mathbb{R}^{p+q-1}}\left[\frac{\partial^{p+q}W}{\partial u_1\ldots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\right]_{u_1=-\infty}^{u_1=+\infty}\frac{\mathrm{d}u}{\mathrm{d}u_1}\,\mathrm{d}v\\ &\qquad-\frac{1}{i(j_{1,1}-j_{2,1})}\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q+1}W}{\partial^2 u_1\partial u_2\ldots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v,\end{aligned}\tag{113}$$

where the functions W and R_{j_1,l_1,j_2,l_2} are defined as

$$\begin{aligned}W:(u,v)&\mapsto\|(u,v)\|^{2\delta}\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2,\\ R_{j_1,l_1,j_2,l_2}:(u,v)&\mapsto\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2}.\end{aligned}$$

Let us now show that the first term in the right side of Equation (113) is equal to zero. Using the differentiability of $\widehat{G}$, we write

$$\begin{aligned}&\frac{\partial^{p+q}}{\partial u_1\ldots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}\\ &\qquad=|\widehat{G}(u_1)|^2\frac{\partial^{p+q}}{\partial u_1\ldots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}\\ &\qquad\quad+\frac{\partial}{\partial u_1}|\widehat{G}(u_1)|^2\frac{\partial^{p+q}}{\partial u_2\ldots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}.\end{aligned}\tag{114}$$

Furthermore, according to Lemma 12 we obviously have

$$\lim_{|u_1|\to+\infty}|\widehat{G}(u_1)|^2\frac{\partial^{p+q}}{\partial u_1\ldots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}=0.$$

In addition,

$$\frac{\partial}{\partial u_1}|\widehat{G}(u_1)|^2\frac{\partial^{p+q}}{\partial u_2\dots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}$$
$$=\left\{\widehat{G}(u_1)\frac{\partial\overline{\widehat{G}}(u_1)}{\partial u_1}+\overline{\widehat{G}}(u_1)\frac{\partial\widehat{G}(u_1)}{\partial u_1}\right\}\frac{\partial^{p+q}}{\partial u_2\dots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}.$$

Moreover, $\dfrac{\partial\widehat{G}(u_1)}{\partial u_1}$ is the Fourier transform of $x\mapsto ixG(x)$ which is $\mathbb{L}_1$-integrable as it is continuous with a bounded support. Thus, we deduce according to Riemann–Lebesgue lemma [Bochner and Chandrasekharan, 1949] that

$$\lim_{|u_1|\to+\infty}\frac{\partial\widehat{G}(u_1)}{\partial u_1}=0.$$

It follows that

$$\lim_{|u_1|\to+\infty}\left|\frac{\partial\widehat{G}(u_1)}{\partial u_1}\right||\widehat{G}(u_1)|\left|\frac{\partial^{p+q}}{\partial u_2\dots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=2}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}\right|=0.$$

We then obtain from Equation (114) that

$$\left[\frac{\partial^{p+q}}{\partial u_1\dots\partial v_q}\left\{\|(u,v)\|^{2\delta}\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2\right\}\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{\prod_{r=1}^{p}|\widehat{G}(u_r)|^2\prod_{s=1}^{q}|\widehat{G}(v_s)|^2}\right]_{u_1=-\infty}^{u_1=+\infty}=0.$$

Therefore, we have according to Equation (113) that

$$\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q}W}{\partial u_1\dots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v$$
$$=-\frac{1}{i(j_{1,1}-j_{2,1})}\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q+1}W}{\partial^2u_1\partial u_2\dots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

Repeating this last process for the variables $u_2,\dots,u_p,v_1,\dots,v_q$ gives

$$\int_{\mathbb{R}^{p+q}}\frac{\partial^{p+q}W}{\partial u_1\dots\partial v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v$$
$$=\frac{(i)^{p+q}}{\prod_{r=1}^{p}(j_{1,r}-j_{2,r})\prod_{s=1}^{q}(l_{1,s}-l_{1,s})}\int_{\mathbb{R}^{p+q}}\frac{\partial^{2(p+q)}W}{\partial^2u_1\dots\partial^2v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v.\tag{115}$$

By injecting this last Equation (115) in Equation (112), we obtain

$$\int_{\mathbb{R}^{p+q}}\|(u,v)\|^{2\delta}\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v$$
$$=\frac{(-1)^{p+q}}{\prod_{r=1}^{p}(j_{1,r}-j_{2,r})^2\prod_{s=1}^{q}(l_{1,s}-l_{1,s})^2}\int_{\mathbb{R}^{p+q}}\frac{\partial^{2(p+q)}W}{\partial^2u_1\dots\partial^2v_q}(u,v)R_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

By analogy with the last equation, we can show that in the general case for two different vectors (j_1,l_1) and (j_2,l_2) that

$$\int_{\mathbb{R}^{p+q}}\|(u,v)\|^{2\delta}\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v$$
$$=\frac{(-1)^{|\mathcal{S}_1|+|\mathcal{S}_2|}}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2\prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2}\int_{\mathbb{R}^{p+q}}\frac{\partial^{2(p+q)}W}{\prod_{r\in\mathcal{S}_1}\partial^2u_r\prod_{s\in\mathcal{S}_2}\partial^2v_s}(u,v)T_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v,$$

where $\mathcal{S}_1$ (resp. $\mathcal{S}_2$) is the set of indices r (resp. s) such that $j_{1,r}\neq j_{2,r}$ (resp. $l_{1,s}\neq l_{1,s}$), while the notation $|\cdot|$ designates the cardinal. In addition, the function T is defined as

$$T_{j_1,l_1,j_2,l_2}:(u,v)\mapsto\frac{\mathcal{G}_{j_1,l_1,j_2,l_2}(u,v)}{\prod_{r\in\mathcal{S}_1}|\widehat{G}(u_r)|^2\prod_{s\in\mathcal{S}_2}|\widehat{G}(v_s)|^2}.$$

Moreover, one can easily show using Lemma 12 that independently from the values of j_1, l_1 and j_2, l_2, we have

$$\left| \int_{\mathbb{R}^{p+q}} \frac{\partial^{2(p+q)} W}{\prod_{r\in\mathcal{S}_1} \partial^2 u_r \prod_{s\in\mathcal{S}_2} \partial^2 v_s}(u,v) T_{j_1,l_1,j_2,l_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v \right| \leq C(p,q).$$

We deduce using Equation (108) the following

$$\left| \int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{2,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v \right| \leq C(p,q,\delta) h_n^{p+q} \sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \frac{1}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2 \prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2}. \tag{116}$$

Furthermore,

$$\begin{aligned}
&\sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \frac{1}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2 \prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2} \\
&\qquad = \sum_{\vartheta=1}^{p+q} \sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \frac{1}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2 \prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2} \\
&\qquad = \sum_{\vartheta=1}^{p+q} \binom{p+q}{\vartheta} M_n^{p+q-\vartheta} \left[\sum_{j\neq k=1}^{M_n} \frac{1}{(j-k)^2} \right]^{\vartheta}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \frac{1}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2 \prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2} &= \sum_{\vartheta=1}^{p+q} \binom{p+q}{\vartheta} M_n^{p+q-\vartheta} \left[\sum_{l=1}^{M_n} \frac{M_n - l}{l^2} \right]^{\vartheta} \\
&\leq \sum_{\vartheta=1}^{p+q} \binom{p+q}{\vartheta} M_n^{p+q} \left[\sum_{l=1}^{M_n} \frac{1}{l^2} \right]^{\vartheta}.
\end{aligned}$$

Thus, using the convergence of the sum $\sum_{l>0} 1/l^2$, we obtain

$$\sum_{\substack{(j_1,l_1)\in I_{n,p,q}\\(j_2,l_2)\in I_{n,p,q}\\(j_1,l_1)\neq(j_2,l_2)}} \frac{1}{\prod_{r\in\mathcal{S}_1}(j_{1,r}-j_{2,r})^2 \prod_{s\in\mathcal{S}_2}(l_{1,s}-l_{1,s})^2} \quad \leq \quad C(p,q) M_n^{p+q}.$$

We deduce from Equation (116) that

$$\left| \int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} H_{2,n}(u,v)\,\mathrm{d}u\,\mathrm{d}v \right| \leq C(p,q,\delta). \tag{117}$$

Consequently, combining Equations (102), (107) and (117) we have the following

$$\int_{\mathbb{R}^{p+q}} \|(u,v)\|^{2\delta} \left| \widehat{f}_\theta(u,v) - \widehat{f}_1 \otimes \widehat{f}_2(u,v) \right|^2 \mathrm{d}u\,\mathrm{d}v \leq C(p,q,\delta).$$

By well choosing the constant of the sequence $(h_n)_n$ w.r.t. p, q and δ, the constant $C(p,q,\delta)$ in the last equation can be replaced by R, which achieves this point of the proof.

Proof of 3. Let us prove that, for all $\theta \in \{-1,1\}^{M_n^{p+q}}$, f_θ satisfies $\|f_\theta - f_{\theta,1} \otimes f_{\theta,2}\|_2 = Cn^{-2\delta/(4\delta+p+q)}$. Since, $\int_{\mathbb{R}} G(t)\mathrm{d}t = 0$, we know that $f_{\theta,1} = \mathbb{1}_{[0,1]^p}$ and $f_{\theta,2} = \mathbb{1}_{[0,1]^q}$, thus $f_{\theta,1} \otimes f_{\theta,2} = \mathbb{1}_{[0,1]^{p+q}}$ and

$$f_\theta - f_{\theta,1} \otimes f_{\theta,2} = h_n^{\delta+(p+q)} \sum_{(j,l)\in I_{n,p,q}} \theta_{(j,l)} g_{n,j,l}(x,y),$$

where the functions $g_{n,j,l}$ are defined in (100), with disjoint supports.
In particular,

$$\|f_\theta - f_{\theta,1} \otimes f_{\theta,2}\|_2^2 = h_n^{2\delta+2(p+q)} \sum_{(j,l)\in I_{n,p,q}} \|g_{n,j,l}\|_2^2 .$$

Moreover, for all $(j,l) \in I_{n,p,q}$,

$$\begin{aligned}
\|g_{n,j,l}\|_2^2 &= \int_{\mathbb{R}^{p+q}} \left[\prod_{r=1}^{p} G_{h_n}^2(x_r - j_r h_n) \prod_{s=1}^{q} G_{h_n}^2(y_s - l_s h_n)\right] \mathrm{d}x_1 \dots \mathrm{d}x_p \mathrm{d}y_1 \dots \mathrm{d}y_q \\
&= \left[\prod_{r=1}^{p} \left(\int_{\mathbb{R}} G_{h_n}^2(x_r - j_r h_n)\mathrm{d}x_r\right)\right] \times \left[\prod_{s=1}^{q} \left(\int_{\mathbb{R}} G_{h_n}^2(y_s - l_s h_n)\mathrm{d}y_s\right)\right],
\end{aligned}$$

and for all k in $\{1,\dots,M_n\}$, a simple change of variables implies that

$$\int_{\mathbb{R}} G_{h_n}^2(t - kh_n)\mathrm{d}t = \frac{1}{h_n^2}\int_{\mathbb{R}} G^2\left(\frac{t-kh_n}{h_n}\right)\mathrm{d}t = \frac{1}{h_n}\int_{\mathbb{R}} G^2(t)\mathrm{d}t = \frac{C}{h_n},$$

since G belongs to $\mathbb{L}_2(\mathbb{R})$.
We thus deduce that

$$\|g_{n,j,l}\|_2^2 = \frac{C(p,q)}{h_n^{p+q}} \tag{118}$$

and that, since the cardinality of $I_{n,p,q}$ equals M_n^{p+q},

$$\|f_\theta - f_{\theta,1} \otimes f_{\theta,2}\|_2^2 = h_n^{2\delta+2(p+q)} \times M_n^{p+q} \times \frac{C(p,q)}{h_n^{p+q}} = C(p,q)h_n^{2\delta}.$$

6.19 Proof of Proposition 6

Let $\mathbb{Z}_n = (X_i, Y_i)_{1\le i\le n}$ be an i.i.d sample with uniform distribution P_{f_0}.

For simplicity, denote for all $1 \le i \le n$ and all (j,l) in $I_{n,p,q}$,

$$a_{i,j,l} = h_n^{\delta+(p+q)} g_{n,j,l}(X_i, Y_i) = h_n^{\delta+(p+q)} \prod_{r=1}^{p} G_{h_n}(X_i^{(r)} - j_r h_n) \prod_{s=1}^{q} G_{h_n}(Y_i^{(s)} - l_s h_n),$$

where $g_{n,j,l}$ is defined in Equation (100), such that $f_\theta(X_i,Y_i) = 1 + \sum_{(j,l)\in I_{n,p,q}} \theta_{(j,l)} a_{i,j,l}$. Note that $a_{i,j,l} \neq 0$ if and only if (X_i,Y_i) belongs to the set $D_{(j,l)}$ defined in Equation (101).

Then, since $f_0 = \mathbb{1}_{[0,1]^{p+q}}$, the likelihood ratio equals

$$\begin{aligned}
L_\nu(\mathbb{Z}_n) = \frac{\mathrm{d}P_\nu}{\mathrm{d}P_{f_0}}(\mathbb{Z}_n) &= \int \prod_{i=1}^{n} \frac{f_\theta}{f_0}(X_i,Y_i)\pi(\mathrm{d}\theta) \\
&= \mathbb{E}_\Theta\left[\prod_{i=1}^{n}\left(1 + \sum_{(j,l)\in I_{n,p,q}} \Theta_{(j,l)} a_{i,j,l}\right)\right],
\end{aligned}$$

where $\Theta = (\Theta_{(j,l)})_{(j,l)\in I_{n,p,q}}$ has i.i.d. Rademacher components $\Theta_{(j,l)}$, and $\mathbb{E}_\Theta[\cdot]$ denotes the expectation w.r.t. Θ.

Noticing that for all $1 \leq i \leq n$, there exists a unique (j,l) in $I_{n,p,q}$ such that $a_{i,j,l} \neq 0$, we obtain

$$1 + \sum_{(j,l)\in I_{n,p,q}} \Theta_{(j,l)} a_{i,j,l} = \prod_{(j,l)\in I_{n,p,q}} \left(1 + \Theta_{(j,l)} a_{i,j,l}\right).$$

Thus,

$$\begin{aligned} L_\nu(\mathbb{Z}_n) &= \mathbb{E}_\Theta\left[\prod_{(j,l)\in I_{n,p,q}} \prod_{i=1}^{n} \left(1 + \Theta_{(j,l)} a_{i,j,l}\right)\right] \\ &= \prod_{(j,l)\in I_{n,p,q}} \left[\frac{1}{2}\prod_{i=1}^{n}(1 - a_{i,j,l}) + \frac{1}{2}\prod_{i=1}^{n}(1 + a_{i,j,l})\right]. \end{aligned}$$

Moreover, for ε in $\{-1, 1\}$,

$$\prod_{i=1}^{n}(1 + \varepsilon a_{i,j,l}) = 1 + \sum_{k=1}^{n} \varepsilon^k \left(\sum_{1\leq i_1<\ldots<i_k\leq n} a_{i_1,j,l} \ldots a_{i_k,j,l}\right).$$

Hence, by cancelling the odd terms, we obtain

$$\begin{aligned} \frac{1}{2}\prod_{i=1}^{n}(1 - a_{i,j,l}) + \frac{1}{2}\prod_{i=1}^{n}(1 + a_{i,j,l}) &= 1 + \sum_{k=1}^{[n/2]} \sum_{1\leq i_1<\ldots<i_{2k}\leq n} a_{i_1,j,l} \ldots a_{i_{2k},j,l} \\ &= 1 + \sum_{k=1}^{[n/2]} A_{k,j,l}, \end{aligned}$$

where $[\cdot]$ denotes the integer part, and

$$A_{k,j,l} = \sum_{1\leq i_1<\ldots<i_{2k}\leq n} a_{i_1,j,l} \ldots a_{i_{2k},j,l}. \tag{119}$$

Thus,

$$[L_\nu(\mathbb{Z}_n)]^2 = \prod_{(j,l)\in I_{n,p,q}} \left(1 + \sum_{k=1}^{[n/2]} A_{k,j,l}\right)^2 = \prod_{(j,l)\in I_{n,p,q}} \left(1 + B_{j,l}\right),$$

where

$$B_{j,l} = 2\sum_{k=1}^{[n/2]} A_{k,j,l} + \sum_{k,k'=1}^{[n/2]} A_{k,j,l} A_{k',j,l}. \tag{120}$$

Then,

$$[L_\nu(\mathbb{Z}_n)]^2 = 1 + \sum_{m=1}^{M_n^{p+q}} \frac{1}{m!} \sum_{(j_1,l_1),\ldots,(j_m,l_m)}^{\neq} B_{j_1,l_1} \ldots B_{j_m,l_m},$$

where $\sum^{\neq}$ means that the indexes are all distinct.

From Equation (119), one may easily see that for all k, $A_{k,j,l}$ only depends on the (X_i, Y_i) for i in $D_{(j,l)}$ (since $a_{i,j,l} = 0$ for all $i \notin D_{(j,l)}$). Thus, the same holds for $B_{j,l}$ and this for all (j,l) in $I_{n,p,q}$. Hence, since the set $D_{(j,l)}$ are all disjoints, and by independence between the $(X_i, Y_i)_{1\leq i\leq n}$, we deduce that the $(B_{j,l})_{(j,l)\in I_{n,p,q}}$ are independent. Therefore,

$$\mathbb{E}_{f_0}\left[\left(L_\nu(\mathbb{Z}_n)\right)^2\right] = 1 + \sum_{m=1}^{M_n^{p+q}} \frac{1}{m!} \sum_{(j_1,l_1),\ldots,(j_m,l_m)}^{\neq} \mathbb{E}_{f_0}[B_{j_1,l_1}] \ldots \mathbb{E}_{f_0}[B_{j_m,l_m}] \tag{121}$$

Now fix (j,l) in $I_{n,p,q}$. By Equation (120),

$$\mathbb{E}_{f_0}[B_{j,l}] = 2\sum_{k=1}^{[n/2]} \mathbb{E}_{f_0}[A_{k,j,l}] + \sum_{k,k'=1}^{[n/2]} \mathbb{E}_{f_0}[A_{k,j,l}A_{k',j,l}]\,.$$

On the one hand, by definition of $A_{k,j,l}$,

$$\mathbb{E}_{f_0}[A_{k,j,l}] = \sum_{1\le i_1<\ldots<i_{2k}\le n} \mathbb{E}_{f_0}[a_{i_1,j,l}\ldots a_{i_{2k},j,l}]\,.$$

Yet, for all i, $(a_{i,j,l})$ only depends on (X_i, Y_i). Thus, the $(a_{i,j,l})_{1\le i\le n}$ are i.i.d. and centered under f_0 since $\int_{\mathbb{R}} G(t)dt = 0$ which directly leads to

$$\mathbb{E}_{f_0}[A_{k,j,l}] = 0.$$

On the other hand, for all $1 \le k, k' \le [n/2]$,

$$\mathbb{E}_{f_0}[A_{k,j,l}A_{k',j,l}] = \sum_{1\le i_1<\ldots<i_{2k}\le n}\ \sum_{1\le i'_1<\ldots<i'_{2k'}\le n} \mathbb{E}_{f_0}\left[a_{i_1,j,l}\ldots a_{i_{2k},j,l}a_{i'_1,j,l}\ldots a_{i'_{2k'},j,l}\right].$$

Yet, if there exists at least one i (or one i') such that $a_{i,j,l}$ appears only once in the product, by independence and since $\mathbb{E}_{f_0}[a_{i,j,l}] = 0$,

$$\mathbb{E}_{f_0}\left[a_{i_1,j,l}\ldots a_{i_{2k},j,l}a_{i'_1,j,l}\ldots a_{i'_{2k'},j,l}\right] = 0.$$

Thus, if $k \ne k'$, $\mathbb{E}_{f_0}[A_{k,j,l}A_{k',j,l}] = 0$ (since there are at least $2|k-k'|$ isolated terms). Moreover, if $k = k'$, we obtain

$$\begin{aligned}\mathbb{E}_{f_0}[A_{k,j,l}A_{k',j,l}] = \mathbb{E}_{f_0}\left[A^2_{k,j,l}\right] &= \sum_{1\le i_1<\ldots<i_{2k}\le n} \mathbb{E}_{f_0}\left[a^2_{i_1,j,l}\ldots a^2_{i_{2k},j,l}\right] \\ &= \binom{n}{2k}\left(\mathbb{E}_{f_0}\left[a^2_{1,j,l}\right]\right)^{2k}\end{aligned}$$

Besides, $\mathbb{E}_{f_0}\left[a^2_{1,j,l}\right] = h_n^{2\delta+2p+2q}\left\|g_{n,j,l}\right\|_2^2 = C(p,q)h_n^{2\delta+p+q}$ by Equation (118). Thus,

$$\mathbb{E}_{f_0}\left[A^2_{k,j,l}\right] = \binom{n}{2k}\left[C(p,q)h_n^{2\delta+p+q}\right]^{2k} \le \left[n\times C(p,q)h_n^{2\delta+p+q}\right]^{2k},$$

since $\binom{n}{2k} \le n^{2k}$.

Therefore, we obtain

$$\mathbb{E}_{f_0}[B_{j,l}] = \sum_{k=1}^{[n/2]} \mathbb{E}_{f_0}\left[A^2_{k,j,l}\right] \le \sum_{k=1}^{[n/2]} \left\{\left[C(p,q)\ n\times h_n^{2\delta+p+q}\right]^2\right\}^k.$$

Furthermore, for h_n defined in (32) (for any constant $C(p,q,\alpha,\beta,\delta)$),

$$C(p,q)\ n\times h_n^{2\delta+p+q} = C(p,q,\alpha,\beta,\delta)n^{-(p+q)/(4\delta+p+q)} < 1/2$$

for n large enough, and thus, by property of geometric series, we get

$$\mathbb{E}_{f_0}[B_{j,l}] \le \frac{\left[C(p,q)\ n\times h_n^{2\delta+p+q}\right]^2}{1-\left[C(p,q)\ n\times h_n^{2\delta+p+q}\right]^2} \le \left[C(p,q)\ n\times h_n^{2\delta+p+q}\right]^2.$$

This being true for all $(j,l) \in I_{n,p,q}$, from Equation (121), we deduce that

$$\begin{aligned}
\mathbb{E}_{f_0}\left[\{L_\nu(\mathbb{Z}_n)\}^2\right] &\leq 1 + \sum_{m=1}^{M_n^{p+q}} \frac{1}{m!}\binom{M_n^{p+q}}{m}\left[C(p,q)\ n \times h_n^{2\delta+p+q}\right]^{2m} \\
&\leq 1 + \sum_{m=1}^{M_n^{p+q}} \left\{M_n^{p+q}\left[C(p,q)\ n \times h_n^{2\delta+p+q}\right]^2\right\}^m \\
&\leq 1 + \sum_{m=1}^{M_n^{p+q}} \left[C(p,q)\ n^2 \times h_n^{4\delta+p+q}\right]^m,
\end{aligned}$$

since $1/m! \leq 1$, $\binom{M_n^{p+q}}{m} \leq [M_n^{p+q}]^m$ and $M_n h_n = 1$.

Finally, for h_n defined in (32), with

$$C(p,q,\alpha,\beta,\delta) = \left(\frac{1}{C(p,q)} \times \frac{2(1-\alpha-\beta)^2}{1+4(1-\alpha-\beta)^2}\right)^{1/(4\delta+p+q)},$$

we directly obtain that

$$C(p,q)\ n^2 \times h_n^{4\delta+p+q} = \frac{2(1-\alpha-\beta)^2}{1+4(1-\alpha-\beta)^2} < 1.$$

Hence, by property of the geometric series we obtain,

$$\begin{aligned}
\mathbb{E}_{f_0}\left[(L_\nu(\mathbb{Z}_n))^2\right] &\leq 1 + \frac{\left[C(p,q)\ n^2 \times h_n^{4\delta+p+q}\right]}{1-\left[C(p,q)\ n^2 \times h_n^{4\delta+p+q}\right]} \\
&< 1 + 4(1-\alpha-\beta)^2,
\end{aligned}$$

which ends the proof of Proposition 6.

References

[Ahmad and Li, 1997] Ahmad, I. A. and Li, Q. (1997). Testing independence by nonparametric kernel method. *Statistics & probability letters*, 34(2):201–210.

[Albert, 2015] Albert, M. (2015). *Tests of independence by bootstrap and permutation: an asymptotic and non-asymptotic study. Application to neurosciences.* PhD thesis, Université Nice Sophia Antipolis.

[Arcones and Gine, 1993] Arcones, M. A. and Gine, E. (1993). Limit theorems for u-processes. *The Annals of Probability*, pages 1494–1542.

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

[Bach and Jordan, 2002] Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

[Baker, 1973] Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289.

[Baraud, 2002] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

[Baraud et al., 2003] Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of linear hypotheses by model selection. *The Annals of Statistics*, 31(1):225–251.

[Bergsma and Dassios, 2014] Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli*, 20(2):1006–1028.

[Berrett and Samworth, 2017] Berrett, T. B. and Samworth, R. J. (2017). Nonparametric independence testing via mutual information. *arXiv preprint arXiv:1711.06642*.

[Bochner and Chandrasekharan, 1949] Bochner, S. and Chandrasekharan, K. (1949). *Fourier transforms*. Princeton University Press.

[Butucea, 2007] Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930.

[Butucea and Tribouley, 2006] Butucea, C. and Tribouley, K. (2006). Nonparametric homogeneity tests. *Journal of statistical planning and inference*, 136(3):597–639.

[Fromont et al., 2013] Fromont, M., Laurent, B., and Reynaud-Bouret, P. (2013). The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461.

[Fukumizu et al., 2004] Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99.

[Fukumizu et al., 2008] Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496.

[Giné et al., 2000] Giné, E., Latała, R., and Zinn, J. (2000). Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Springer.

[Gretton et al., 2005a] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

[Gretton et al., 2008] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592.

[Gretton and Györfi, 2010] Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(Apr):1391–1423.

[Gretton et al., 2005b] Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129.

[Gretton et al., 2003] Gretton, A., Herbrich, R., and Smola, A. J. (2003). The kernel mutual information. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, pages IV–880. IEEE.

[Gretton et al., 2005c] Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005c). Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119.

[Gretton et al., 2005d] Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005d). Kernel constrained covariance for dependence measurement. In *AISTATS*, volume 10, pages 112–119.

[Hoeffding, 1948] Hoeffding, W. (1948). A non-parametric test of independence. *The annals of mathematical statistics*, pages 546–557.

[Houdré and Reynaud-Bouret, 2003] Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer.

[Ingster, 1989] Ingster, Y. I. (1989). An asymptotically minimax test of the hypothesis of independence. *J. Soviet Math*, 44:466–476.

[Ingster, 1993a] Ingster, Y. I. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114.

[Ingster, 1993b] Ingster, Y. I. (1993b). Minimax testing of the hypothesis of independence for ellipsoids in l_p. *Zapiski Nauchnykh Seminarov POMI*, 207:77–97.

[Ingster and Suslina, 1998] Ingster, Y. I. and Suslina, I. A. (1998). Minimax detection of a signal for besov bodies and balls. *Problemy Peredachi Informatsii*, 34(1):56–68.

[Ishigami and Homma, 1990] Ishigami, T. and Homma, T. (1990). An importance quantification technique in uncertainty analysis for computer models. In *[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, pages 398–403. IEEE.

[Jacod and Protter, 2012] Jacod, J. and Protter, P. (2012). *Probability essentials*. Springer Science & Business Media.

[Kojadinovic and Holmes, 2009] Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on cramér–von mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154.

[Laurent et al., 2012] Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics*, 6:91–122.

[Micchelli et al., 2006] Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667.

[Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

[Pfister et al., 2018] Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31.

[Póczos et al., 2012] Póczos, B., Ghahramani, Z., and Schneider, J. (2012). Copula-based kernel dependency measures. *arXiv preprint arXiv:1206.4682*.

[Romano and Wolf, 2005] Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

[Rosenblatt, 1975] Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, pages 1–14.

[Scott, 2012] Scott, D. W. (2012). Multivariate density estimation and visualization. In *Handbook of computational statistics*, pages 549–569. Springer.

[Serfling, 2009] Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

[Sobol, 2001] Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.

[Spokoiny, 1996] Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24(6):2477–2498.

[Sriperumbudur et al., 2010] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

[Steinwart, 2001] Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93.

[Székely and Rizzo, 2013] Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

[Székely et al., 2007] Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

[Tsybakov, 2009] Tsybakov, A. B. (2009). Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats.

[Weihs et al., 2018] Weihs, L., Drton, M., and Meinshausen, N. (2018). Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika*, 105(3):547–562.

[Weiss, 2006] Weiss, N. A. (2006). *A course in probability*. Addison-Wesley.

[Yao et al., 2018] Yao, S., Zhang, X., and Shao, X. (2018). Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):455–480.

[Yodé, 2004] Yodé, A. (2004). Asymptotically minimax test of independence. *Mathematical Methods of Statistics*, 13(2):201–234.

[Yodé, 2011] Yodé, A. (2011). Adaptive minimax test of independence. *Mathematical Methods of Statistics*, 20(3):246.

[Zhang et al., 2012] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.