Max–Planck–Institut für biologische Kybernetik
Max Planck Institute for Biological Cybernetics

Technical Report No. 130

Behaviour and Convergence of the Constrained Covariance

Arthur Gretton,[1] Alexander Smola,[2] Olivier Bousquet,[1] Ralf Herbrich,[3] Bernhard Schölkopf,[1] and Nikos Logothetis[4]

October 2004

[1] Department Schölkopf, email: firstname.lastname@tuebingen.mpg.de; [2] NICTA, Canberra, Australia, email: alex.smola@anu.edu.au; [3] Microsoft Research, Cambridge, UK, email: rherb@microsoft.com; [4] Department Logothetis, email: nikos.logothetis@tuebingen.mpg.de

This report is available in PDF–format via anonymous ftp at ftp://ftp.kyb.tuebingen.mpg.de/pub/mpi-memos/pdf/techrepTitle.pdf. The complete series of Technical Reports is documented at: http://www.kyb.tuebingen.mpg.de/techreports.html

Behaviour and Convergence of the Constrained Covariance

Arthur Gretton, Alexander Smola, Olivier Bousquet, Ralf Herbrich, Bernhard Schölkopf, and Nikos Logothetis

Abstract. We discuss reproducing kernel Hilbert space (RKHS)-based measures of statistical dependence, with emphasis on constrained covariance (COCO), a novel criterion to test dependence of random variables. We show that COCO is a test for independence if and only if the associated RKHSs are universal. That said, no independence test exists that can distinguish dependent and independent random variables in all circumstances. Dependent random variables can result in a COCO which is arbitrarily close to zero when the source densities are highly non-smooth, which can make dependence hard to detect empirically. All current kernel-based independence tests share this behaviour. Finally, we demonstrate exponential convergence between the population and empirical COCO, which implies that COCO does not suffer from slow learning rates when used as a dependence test.

---

1 Introduction

Tests to determine the dependence or independence of random variables are well established in statistical analysis. Some approaches require density estimation as an intermediate step ([9] is a classic study); while others assume a parametric model of how the variables were obtained from independent random variables, as in blind source separation [8]. In this paper we propose a non-parametric independence criterion, which relies on the fact that the random variables[1] x, y are independent if and only if

$$\mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})] = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})]. \tag{1.1}$$

for bounded, continuous functions f, g (see for instance [10, 14]). The proposed criterion works by maximising the discrepancy between the empirical estimates of the LHS and RHS of (1.1) over pre-specified function classes $f \in \mathcal{F}$ and $g \in \mathcal{G}$, and comparing the discrepancy to the amount of deviation that can be expected from the fact that we are dealing with empirical estimates rather than expectations. We call our criterion the constrained covariance (COCO).[2]

The results presented here build on recent work published on the subject of kernel based dependence measures. In particular, the canonical correlation between functions in a reproducing kernel Hilbert space (KCC), defined in [1] for a variety of kernels and in [11] for splines, can be used as a test of independence. Indeed, in the Gaussian case, Bach and Jordan show the KCC to be zero if and only if its two arguments are statistically independent. In Section 3, we characterise *all* reproducing kernel Hilbert spaces (RKHSs) for which this property holds (both for COCO and KCC): these are required to be *universal* (the RKHS must be dense in the space of continuous functions [17]). Specifically, the Gaussian and Laplace kernels are universal, as are many exponential-based kernels; polynomial kernels, however, are not universal.

We next demonstrate in Section 4 that for a fixed-size, finite sample of *dependent* random variables, there exists no test that can reliably detect that the random variables are dependent. To clarify how this might affect our criterion, we prove that the population COCO can be made arbitrarily small when certain smoothness assumptions on the density are violated, which makes it difficult to detect dependence on the basis of a finite sample. This is also true of other related kernel dependence measures, including the kernel mutual information (KMI) in [5], and kernel generalised variance (KGV) in [1], both of which were shown in [5] to be upper bounds near independence on the Parzen window estimate of the mutual information. Thus, as in all dependence tests, any inference made is subject to certain assumptions about the underlying generative process - the present work describes these assumptions explicitly for the first time, in the case of kernel-based tests.

Next, we give two bounds, based on Rademacher averages, which describe *exponential* convergence between the population and empirical COCO. The first assures us that the population COCO is small when the empirical COCO is small; the second shows that the population COCO is large when the empirical COCO is large (both statements apply with high probability). These results are very interesting, in that they illustrate a broader phenomenon: slow learning rates do not occur in dependence testing, even though they are unavoidable in regression and classification [4, Ch. 7]. This might appear surprising in the specific case of COCO, since this criterion is optimised in the course of kernelised PLS regression (assuming a kernelised output space: see the discussion in [2]). Another important consequence of the bounds is that *any* dependence between the random variables will be detected rapidly *as the sample size increases*, even though perfect dependence detection is impossible for fixed sample size.[3]

2 Definitions and Background

Before presenting our main results, we begin our discussion with some relevant definitions and background theory, covering both classical independence criteria and RKHSs.[4] Let $(\Omega, \mathcal{A}, \mathbf{P}_{\mathsf{x},\mathsf{y}})$ be a probability space.

[1] We write random variables *sans serif*.

[2] In [5], this was called the kernel covariance (KC).

[3] This technical report is intended as a technical supplement to [6], which was submitted to AISTATS 2005. Proofs in this report, which were given only in sketch form in [6], include Sections 4.2, 5.1, and 5.2.

[4] See [10] and [16] for more detail on these topics.

Consider random variables $\mathsf{x} : (\Omega, \mathcal{A}) \to (U, \mathcal{U})$ and $\mathsf{y} : (\Omega, \mathcal{A}) \to (V, \mathcal{V})$, where U and V are complete metric spaces, and $\mathcal{U}$ and $\mathcal{V}$ their respective Borel σ-algebras. The covariance between x and y is defined as follows.

Definition 1 (Covariance). The covariance of two random variables x, y is given as

$$\mathrm{cov}(\mathsf{x}, \mathsf{y}) := \mathbf{E}_{\mathsf{x},\mathsf{y}}[\mathsf{xy}] - \mathbf{E}_{\mathsf{x}}[\mathsf{x}]\mathbf{E}_{\mathsf{y}}[\mathsf{y}]. \tag{2.1}$$

For our purposes, the notion of independence of random variables is best expressed using the following characterisation:

Theorem 2 (Independence). *The random variables* x *and* y *are independent if and only if* $\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) = 0$ *for each pair* (f, g) *of bounded, continuous functions.*

This theorem suggests the following definition as an independence test.

Definition 3 (Constrained Covariance (COCO)). Given function classes $\mathcal{F}, \mathcal{G}$ we define the *constrained covariance* as

$$\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{F}, \mathcal{G}) := \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left[\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y}))\right]. \tag{2.2}$$

(when $\mathcal{F}$ and $\mathcal{G}$ are unit balls in their respective vector spaces, then this is just the norm of the covariance operator mapping $\mathcal{G}$ to $\mathcal{F}$: see [13]). Given n independent observations $\boldsymbol{z} := ((x_1, y_1), \ldots, (x_n, y_n)) \subset (\mathcal{X} \times \mathcal{Y})^n$, its empirical estimate is defined as

$$\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; \mathcal{F}, \mathcal{G}) := \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left[\frac{1}{n}\sum_{i=1}^{n} f(x_i)g(y_i) - \frac{1}{n^2}\sum_{i=1}^{n} f(x_i)\sum_{j=1}^{n} g(y_j)\right]$$

It follows from Theorem 2 that if $\mathcal{F}, \mathcal{G}$ are the sets of continuous functions bounded by 1 we have $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{F}, \mathcal{G}) = 0$ if and only if x and y are independent. In other words, COCO and $\mathrm{COCO}_{\mathrm{emp}}$ are criteria which can be tested *directly* without the need for an intermediate density estimator (in general, the distributions may not even have densities). It is also clear, however, that unless $\mathcal{F}, \mathcal{G}$ are restricted in further ways, $\mathrm{COCO}_{\mathrm{emp}}$ will always be large, due to the rich choice of functions available. A *non-trivial dependence measure* is thus obtained using function classes that do not give an everywhere-zero empirical average, yet which still guarantee that COCO is zero if and only if its arguments are independent. A tradeoff between the restrictiveness of these function classes and the convergence of $\mathrm{COCO}_{\mathrm{emp}}$ to COCO can be accomplished using standard tools from uniform convergence theory (see Section 5). It turns out (Section 3) that unit-radius balls in universal reproducing kernel Hilbert spaces constitute function classes that yield non-trivial dependence estimates. To demonstrate this, we will use certain properties of these spaces [15]. A reproducing kernel Hilbert space is a Hilbert space $\mathcal{F}$ for which at each $x \in \mathcal{X}$, the point evaluation functional, $\delta_x : \mathcal{F} \to \mathbb{R}$, which maps $f \in \mathcal{F}$ to $f(x) \in \mathbb{R}$, is continuous. To each reproducing kernel Hilbert space, there corresponds a unique positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (the reproducing kernel), which constitutes the inner product on this space; this is guaranteed by the Moore-Aronszajn theorem.

In RKHSs the representer theorem [16] holds, stating that the solution of an optimisation problem, dependent only on the function evaluations on a set of observations and on RKHS norms, lies in the span of the kernel functions evaluated on the observations. This property is next used to specify an easily computed expression for $\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G)$ where F and G are respectively unit balls in the reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$. The proof may be found in [5], although we give a condensed version here.

Lemma 4 (Value of $\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G)$). *Denote by* $\mathcal{F}, \mathcal{G}$ *RKHSs on the domains* $\mathcal{X}$ *and* $\mathcal{Y}$ *respectively and let* F, G *be the unit balls in the corresponding RKHS. Then*

$$\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G) = \frac{1}{n}\sqrt{\|\bar{K}^f \bar{K}^g\|_2} \tag{2.3}$$

where $\bar{K}^f$ is the matrix obtained by the projection $\bar{K}^f = PK^fP$ with projection operator $P_{ij} = \delta_{ij} - \frac{1}{n}$ and Gram matrix $K^f_{ij} = k_f(x_i, x_j)$. $\bar{K}^g$ is defined by analogy using the kernel of $\mathcal{G}$ (which might be different from that of $\mathcal{F}$).

Proof. By the representer theorem, the solution of the maximisation problem arising from $\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G)$ is given by $f = \sum_{i=1}^n \alpha_i k_f(x_i, x)$ and $g = \sum_{j=1}^n \beta_j k_g(y_j, y)$. Hence

$$\begin{aligned}
\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G) &= \sup_{\alpha^\top K^f \alpha \leq 1, \beta^\top K^g \beta \leq 1} \frac{1}{n}\alpha^\top K^f K^g \beta - \frac{1}{n^2}\alpha^\top K^f \vec{1}\vec{1}^\top K^g \beta \\
&= \sup_{\|\alpha\|, \|\beta\| \leq 1} \frac{1}{n}\alpha^\top (K^f)^{\frac{1}{2}} P (K^g)^{\frac{1}{2}} \beta \\
&= \frac{1}{n}\|(K^f)^{\frac{1}{2}} P (K^g)^{\frac{1}{2}}\|
\end{aligned}$$

Squaring the argument in the norm, rearranging, and using the fact that $P = PP$ proves the theorem. Here we defined $\vec{1} \in \mathbb{R}^n$ to be the vector of ones. □

A second theorem which will be crucial in our proofs is Mercer's theorem, which provides a decomposition of the kernel into eigenfunctions and eigenvalues.

Theorem 5 (Mercer's Theorem). *Let $k(\cdot,\cdot) \in L_\infty(\mathcal{X}^2)$ be a symmetric real valued function with an associated positive definite integral operator with normalised orthogonal eigenfunctions $\varphi_p \in L_2(\mathcal{X})$, sorted such that the associated eigenvalues $\tilde{k}_p$ do not increase. Then for almost all $x_i \in \mathcal{X}$ and $x_j \in \mathcal{X}$, the series*

$$k(x_i, x_j) := \sum_{p=1}^{\infty} \tilde{k}_p \varphi_p(x_i)\varphi_p(x_j)$$

converges absolutely and uniformly. In addition, the series $\sum_{i=1}^\infty |\tilde{k}_i|$.

Finally, we give kernel-dependent decay rates for the coefficients used to expand functions in $\mathcal{F}$ in terms of the set of basis functions $\{\varphi_i(\cdot)\}$ from Mercer's theorem.

Lemma 6 (Rate of decay of expansion coefficients). *Let $f \in \mathcal{F}$, where $f(x) := \sum_{i=1}^\infty \tilde{f}_i \varphi_i(x)$. Then as long as $(\tilde{k}_i)^{-1}$ increases super-linearly with i, $(|\tilde{f}_i|) \in \ell_1$ and there exists an $l_0 \in \mathbb{N}$ such that for all $\epsilon > 0$ and all $l > l_0$, $|\tilde{f}_l| < \epsilon$.*

Proof. This holds since for any $f \in \mathcal{F}$, $\|f\|^2_{\mathcal{F}} := \sum_{i=1}^\infty \tilde{f}_i^2 \left(\tilde{k}_i\right)^{-1} < \infty$. □

The super-linearity requirement in Lemma 6 is satisfied by many kernels, including the Gaussian (for which the $((\tilde{k}_m)^{-1})$ increase as $\exp(m^2)$); see [16]. We assume hereafter that our kernel satisfies the requirements of Lemma 6.

3 A Test for Independence

We now characterise the class of kernels for which COCO is a non-trivial test of dependence. The main result is given in Theorem 8, in which we demonstrate that COCO constitutes such a test when $\mathcal{F}$ and $\mathcal{G}$ are RKHSs with a universal kernel [17].

Definition 7 (Universal kernel). A *continuous kernel* $k(\cdot,\cdot)$ on a compact metric space $(\mathcal{X}, d)$ is called universal if and only if the RKHS $\mathcal{F}$ induced by the kernel is dense in $C(\mathcal{X})$ with respect to the topology induced by the infinity norm $\|f - g\|_\infty$.

For instance, [17] shows the following two kernels are universal on compact subsets of $\mathbb{R}^d$:

$$k(x,x') = \exp\left(-\lambda\|x-x'\|^2\right) \text{ and}$$
$$k(x,x') = \exp\left(-\lambda\|x-x'\|\right) \text{ for } \lambda > 0.$$

We now state our main result for this section.

Theorem 8 (COCO is only zero at independence for universal kernels). *Denote by $\mathcal{F}, \mathcal{G}$ RKHSs with universal kernels k_f, k_g on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively and let F, G be the unit balls in the corresponding RKHSs. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$ if and only if x, y are independent.*

Proof. It is clear that $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ is zero if x and y are independent. We prove the converse by contradiction, using the starting assumptions $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; B(\mathcal{X}), B(\mathcal{Y})) = c$ for some $c > 0$ (here $B(\mathcal{X})$ denotes the subset of $C(\mathcal{X})$ of continuous functions bounded by 1 in the $L_\infty(\mathcal{X})$, and $B(\mathcal{Y})$ is defined in an analogous manner) and $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$. There exist two sequences of functions $f_n \in C(\mathcal{X})$ and $g_n \in C(\mathcal{Y})$, satisfying $\|f_n\|_\infty \leq 1, \|g_n\|_\infty \leq 1$, for which

$$\lim_{n\to\infty} \mathrm{cov}(f_n(\mathsf{x}), g_n(\mathsf{y})) = c.$$

More to the point, there exists an n^* for which $\mathrm{cov}(f_{n^*}, (\mathsf{x}) g_{n^*}(\mathsf{y})) \geq c/2$. We know that $\mathcal{F}$ and $\mathcal{G}$ are respectively dense in $C(\mathcal{X})$ and $C(\mathcal{Y})$: this means that for all $1/3 > \epsilon > 0$, we can find some $f^* \in \mathcal{F}$ (and an analogous $g^* \in \mathcal{G}$) satisfying $\|f^* - f_{n^*}\|_\infty < \epsilon = \frac{c}{24}$. Writing as $\tilde{f}(\mathsf{x}) := f^*(\mathsf{x}) - f_{n^*}(\mathsf{x}) + f_{n^*}(\mathsf{x})$ (with an analogous $\tilde{g}(\mathsf{y})$ definition), we obtain

$$\begin{aligned}
&\mathrm{cov}(f^*(\mathsf{x}), g^*(\mathsf{y})) \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}}\left[\tilde{f}(\mathsf{x})\tilde{g}(\mathsf{y})\right] - \mathbf{E}_{\mathsf{x}}(\tilde{f}(\mathsf{x}))\mathbf{E}_{\mathsf{y}}(\tilde{g}(\mathsf{y})) \\
&\geq \mathrm{cov}(f_{n^*}(\mathsf{x}), g_{n^*}(\mathsf{y})) - 2\epsilon\left|\mathbf{E}_{\mathsf{x}}\left(f_{n^*}(\mathsf{x})\right)\right| \\
&\qquad - 2\epsilon\left|\mathbf{E}_{\mathsf{y}}\left(g_{n^*}(\mathsf{y})\right)\right| - 2\epsilon^2 \\
&\geq \frac{c}{2} - 6\frac{c}{24} = \frac{c}{4} > 0.
\end{aligned}$$

This contradicts the assumption that $\mathrm{cov}(f^*(\mathsf{x}), g^*(\mathsf{y})) = 0$, and completes the proof. □

The kernel dependence tests (COCO, KMI, KGV, and KCC) are generalised in [5, 1] to a greater number of random variables, providing tests of pairwise independence.

4 Limitations of Independence Tests

4.1 General independence tests

In this section, we illustrate with a simple example that for a finite sample, there exists no test of independence which can reliably (i.e. with high probability) distinguish dependence from independence. This discussion is intended as a complement to the next section, where we explicitly construct dependent random variables which are difficult for the empirical COCO to distinguish from independence. We illustrate the case where $\mathcal{X}$ is countable, but our reasoning applies equally to continuous spaces.

We begin with some notation. Consider a set $\mathcal{P}$ of probability distributions $\mathbf{P}_{\mathsf{x}}$, where x contains m entries. The set $\mathcal{P}$ is split into two subsets: $\mathcal{P}_i$ contains distributions $\mathbf{P}_{\mathsf{x}}^{(i)}$ of mutually independent random variables $\mathbf{P}_{\mathsf{x}}^{(i)} = \prod_{j=1}^m \mathbf{P}_{\mathsf{x}_j}$, and $\mathcal{P}_d$ contains distributions $\mathbf{P}_{\mathsf{x}}^{(d)}$ of dependent random variables. We next introduce a test $\Delta(\boldsymbol{x})$, which takes a data set[5] $\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}$, and returns

$$\begin{aligned}
\Delta(\boldsymbol{x}) = 1 \quad &: \quad \boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(d)}, \\
\Delta(\boldsymbol{x}) = 0 \quad &: \quad \boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}
\end{aligned}$$

[5] We denote by $\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}$ the drawing of n i.i.d. samples $\boldsymbol{x} := (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ from $\mathbf{P}_{\mathsf{x}}$.

Given that the test sees only a finite sample, it cannot determine with complete certainty whether the data are drawn from $\mathbf{P}_{\mathsf{x}^n}^{(d)}$ or $\mathbf{P}_{\mathsf{x}^n}^{(i)}$. We call Δ an α-test when

$$\sup_{\mathbf{P}_{\mathsf{x}}^{(i)} \in \mathcal{P}_i} \mathbf{E}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}} (\Delta(\boldsymbol{x}) = 1) \leq \alpha;$$

in other words α upper bounds the probability of a Type I error. Our theorem is as follows:

Theorem 9 (Universal limit on dependence tests). *For any α-test, some fixed $n \in \mathbb{N}$, and any $1 - \alpha > \epsilon > 0$, there exists $\mathbf{P}_{\mathsf{x}} \notin \mathcal{P}_i$ such that*

$$\mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}} (\Delta(\boldsymbol{x}) = 0) \geq 1 - \alpha - \epsilon;$$

in other words, a dependence test with a low Type I error can have a severe Type II error.

Proof. We introduce a distribution $\mathbf{P}_{\mathsf{x}}^{(\gamma)} := \gamma \mathbf{P}_{\mathsf{x}}^{(i)} + (1 - \gamma)\mathbf{P}_{\mathsf{x}}^{(d)}$, where $0 \leq \gamma < 1$. Clearly, random variables drawn from $\mathbf{P}_{\mathsf{x}}^{(\gamma)}$ are dependent. The probability of a Type II error for this mixture is then

$$\begin{aligned}
\mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(\gamma)}} (\Delta(\boldsymbol{x}) = 0) &\overset{(a)}{=} \sum_{\boldsymbol{x}} \mathbf{P}_{\mathsf{x}^n}^{(\gamma)}(\boldsymbol{x}) \mathbb{I}_{\Delta(\boldsymbol{x})=0} \\
&= \sum_{\boldsymbol{x}} \prod_{k=1}^{n} \mathbf{P}_{\mathsf{x}}^{(\gamma)}(\mathbf{x}_k) \mathbb{I}_{\Delta(\boldsymbol{x})=0} \\
&> \sum_{\boldsymbol{x}} \prod_{k=1}^{n} \gamma \mathbf{P}_{\mathsf{x}}^{(i)}(\mathbf{x}_k) \mathbb{I}_{\Delta(\boldsymbol{x})=0} \\
&= \gamma^n \mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}} (\Delta(\boldsymbol{x}) = 0) \\
&= \gamma^n (1 - \alpha)
\end{aligned}$$

where the sum following (a) is over all possible draws of $\boldsymbol{x}$ from $\mathbf{P}_{\mathsf{x}^n}^{(\gamma)}$, and $\mathbb{I}_A$ is the indicator function for event A. Taking γ very close to 1 (i.e. making the dependent distribution very unlikely in the mixture) proves the theorem. □

4.2 Kernel independence tests

We prove the existence of a dependent probability distribution for which COCO is small, but with a large covariance between certain functions in $\mathcal{F}$ and $\mathcal{G}$; we then demonstrate that this also holds for the KCC, KMI, and KGV. Although the population COCO is *not* zero for this density, its small size will make this dependence hard to detect unless a large data sample is available. We illustrate this phenomenon by specifying a particular joint density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ chosen such that $\mathrm{cov}(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y}))$ is large for some large l (meaning x, y have a non-trivial dependence), but $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ is small (making it hard to detect a non-zero value of the *population* COCO on the basis of a *finite sample*, as in the previous section). The intuition behind our argument is made clear by re-writing COCO for RKHSs as

$$\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) = \sup_{f \in \mathcal{F},\, g \in \mathcal{G}} \frac{\mathrm{cov}\,(f(\mathsf{x}), g(\mathsf{y}))}{\|f\|_{\mathcal{F}} \|g\|_{\mathcal{G}}}. \tag{4.1}$$

This will obviously be small when the RKHS norms in the denominator are much larger than the covariance in the numerator: we will see that this motivates our choice of density. More specifically, high order eigenfunctions of the kernel[6] ($\varphi_l(x)$ and $\varphi_l(y)$ for large l) have large RKHS norms, a fact widely exploited in regression as a roughness penalty [16]. Thus, if the high order eigenfunctions are prominent

[6] See Theorem 5 for a definition of the eigenfunctions. Note that the kernels in $\mathcal{F}$ and $\mathcal{G}$ may not be identical, and the eigenfunctions $\varphi_i(x)$ and $\varphi_j(y)$ might therefore be different. We use the *arguments* of the eigenfunctions to distinguish between them, since this is unambiguous and avoids messy notation.

in $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ (i.e., for highly non-smooth densities), we expect COCO to be small even when there exists an l for which $\mathrm{cov}(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y}))$ is large.[7]

Theorem 10 (Dependent random variables can have small COCO). *Assume $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert spaces with $(\tilde{k}^x_m)^{-1}$ and $(\tilde{k}^y_m)^{-1}$ increasing superlinearly with m (see Lemma 6), and with respective eigenfunctions $\varphi_i(x)$ and $\varphi_j(y)$ absolutely bounded.*[8] *Then there exists a density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ for which* $\mathrm{cov}\,(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y})) \geq \beta - \epsilon$ *for non-trivial β and arbitrarily small $\epsilon > 0$, yet for which* $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) < \gamma$ *for an arbitrarily small $\gamma > 0$.*

Proof. We begin by constructing a density for which $\mathrm{cov}\,(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y})) \geq \beta - \epsilon$, where ϵ is small for large enough l. This is written

$$\mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y) = \alpha_l + \beta\varphi_l(x)\varphi_l(y) \tag{4.2}$$

where $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y) \geq 0$ and $\int \mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y)dxdy = 1$. The first constraint requires $\alpha_l - \beta \min_{x,y}(\varphi_l(x)\varphi_l(y)) \geq 0$, which can be satisfied as long as the $\varphi_l(x)$ and $\varphi_l(y)$ are absolutely bounded. The second constraint affects the covariance between kernel eigenfunctions,

$$\begin{aligned} \widetilde{C}_{i,j} &= \mathrm{cov}(\varphi_i(x), \varphi_j(y)) \\ &:= \mathbf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \mathbf{E}_{\mathsf{x}}(\varphi_i(\mathsf{x}))\mathbf{E}_{\mathsf{y}}(\varphi_j(\mathsf{y})). \end{aligned} \tag{4.3}$$

Indeed, this constraint causes $\widetilde{\mathbf{C}}$ to have i, jth entries

$$\widetilde{C}_{i\neq l, j\neq l} := \epsilon_{ij}, \qquad \widetilde{C}_{l,l} := \beta + \epsilon_{ll}, \tag{4.4}$$

where ϵ_{ij} denotes a quantity with absolute value arbitrarily small for large enough l (the proof is in Appendix A.1).

We next expand the functions f and g which define COCO (i.e. elements of the respective RKHSs at which the supremum is attained) as $f(x) = \sum_{i=1}^{\infty} \tilde{f}_i\varphi_i(x)$ and $g(y) = \sum_{j=1}^{\infty} \tilde{g}_j\varphi_j(y)$ (the expansion coefficients are written as vectors $\tilde{\mathbf{f}}, \tilde{\mathbf{g}}$). Using these expansions, the numerator of (4.1) becomes $\mathrm{cov}(f(x), g(y)) = \tilde{\mathbf{f}}^\top \widetilde{\mathbf{C}}\tilde{\mathbf{g}}$, and

$$\begin{aligned} \tilde{\mathbf{f}}^\top \widetilde{\mathbf{C}}\tilde{\mathbf{g}} &\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |\tilde{f}_i|\,|\tilde{g}_j|\,\epsilon + |\tilde{f}_l|\,|\tilde{g}_l|\,\beta \\ &= \|\tilde{\mathbf{f}}\|_1\,\|\tilde{\mathbf{g}}\|_1\,\epsilon + |\tilde{f}_l|\,|\tilde{g}_l|\,\beta, \end{aligned}$$

where we replace all entries in $\widetilde{\mathbf{C}}$ with their expressions in (4.4), and $\epsilon = \max_{i,j} |\epsilon_{ij}|$ is small. Lemma 6 ensures that $\|\tilde{\mathbf{f}}\|_1$ and $\|\tilde{\mathbf{g}}\|_1$ both converge. In the case of the remaining term $|\tilde{f}_l|\,|\tilde{g}_l|\,\beta$, we divide through by the norms in the denominator of COCO to get

$$\frac{|\tilde{f}_l|\,|\tilde{g}_l|\,\beta}{\sqrt{\sum_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i^f\right)^{-1}}\sqrt{\sum_{j=1}^{\infty} \tilde{g}_j^2 \left(\tilde{k}_j^g\right)^{-1}}} \leq \beta\sqrt{\tilde{k}_l^f \tilde{k}_l^g},$$

and the right hand side approaches zero as $l \to \infty$ thanks to Theorem 5. □

We now prove the KCC [1] has the same limitation, being upper bounded by a constant multiple of

[7] This reasoning can be extended to motivate kernel choice for the detection of particular dependencies, although this is beyond the scope of the present study. Note also that an alternative Parzen-window based interpretation of kernel choice is given in [5].

[8] This condition is not satisfied for all Mercer kernels: see [16, Exercise 2.24]. The assumption holds in most everyday cases we encounter (e.g. the Fourier basis), however, so it is reasonable in this context.

COCO. The KCC is defined as

$$\begin{aligned}
&\mathcal{J}_\kappa\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{F}, \mathcal{G}\right) \\
&:= \sup_{f\in\mathcal{F},\, g\in\mathcal{G}} \frac{\operatorname{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)}{\sqrt{\operatorname{var}\left(f\left(\mathsf{x}\right)\right)+\kappa\|f\|_{\mathcal{F}}^2}\sqrt{\operatorname{var}\left(g\left(\mathsf{y}\right)\right)+\kappa\|g\|_{\mathcal{G}}^2}} \\
&\leq \kappa^{-1}\|f^*\|_{\mathcal{F}}^{-1}\|g^*\|_{\mathcal{G}}^{-1}\operatorname{cov}\left(f^*(\mathsf{x}), g^*(\mathsf{y})\right) \\
&\leq \kappa^{-1}\mathrm{COCO}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G\right),
\end{aligned}$$

where f^*, g^* attain the supremum in the first line, and we assume f and g to be bounded.

Finally, we demonstrate that the KMI [5] and KGV [1], which are respectively extensions to COCO and the KCC, have the same property. This follows since the KMI can be written as $-\frac{1}{2}\log(\prod_{i=1}^n(1-\rho_i^2))$, where $|\rho_i|$ are upper bounded by COCO, and the KGV as $-\frac{1}{2}\log(\prod_{i=1}^n(1-\gamma_i^2))$, where the $|\gamma_i|$ are upper bounded by the KCC. Small COCO will therefore cause small KMI, and small KCC will cause small KGV.

5 Bounds

We give two convergence bounds in this section. The first (and simplest) guarantees small population COCO when the empirical COCO is small; the second, which has a more involved derivation, guarantees that if the empirical COCO is large, then the population COCO is also large. A consequence of these bounds is that the empirical COCO converges to the population COCO at speed $1/\sqrt{n}$. This means that if we define the independence test $\Delta(\boldsymbol{z})$ (Section 4.1) as the indicator that COCO is larger than a term of the form $C\sqrt{\log(1/\alpha)/n}$ with C a constant, then $\Delta(\boldsymbol{z})$ is an α-test with type II error upper bounded by a term approaching zero as $1/\sqrt{n}$.

5.1 Upper bound

We require two bounds for the proof in this section. The first is the standard Hoeffding bound [7].

Theorem 11 (Hoeffding's inequality). *Consider a collection of n i.i.d. random variables $(\mathsf{z}_1, \ldots, \mathsf{z}_n)$ such that $a_i \leq \mathsf{z}_i \leq b_i$ for $1 \leq i \leq n$. Then for $0 < t < 1 - \mathbf{E}_{\mathsf{z}}(\mathsf{z})$,*

$$\mathbf{P}_{\mathsf{z}^n}\left(\frac{1}{n}\sum_{i=1}^n \mathsf{z}_i - \mathbf{E}_{\mathsf{z}}(\mathsf{z}) \geq t\right) \leq e^{-2nt^2/\sum_{i=1}^n (b_i-a_i)^2}.$$

The second theorem, from [7, p. 25], applies to U-statistics of the kind we encounter in calculating covariances.

Theorem 12 (Positive deviation bound for one sample U-statistics). *Consider a collection of n i.i.d. random variables $(\mathsf{z}_1, \ldots, \mathsf{z}_n)$. We define the U-statistic*

$$\mathsf{u} := \frac{1}{n(n-1)\ldots(n-r+1)}\sum_{\mathbf{i}_r^n} h_{i_1,\ldots,i_r}(\mathsf{z}_{i_1}, \ldots, \mathsf{z}_{i_r}),$$

where the index set $\mathbf{i}_r^n$ is the set of all r-tuples drawn without replacement from $\{1, \ldots, n\}$, and the function h is called the kernel *of the U-statistic. If $a \leq h \leq b$,*

$$\mathbf{P}_{\mathsf{u}}(\mathsf{u} - \mathbf{E}_{\mathsf{u}}(\mathsf{u}) \geq t) \leq \exp\left(\frac{-2t^2\lceil n/r\rceil}{(b-a)^2}\right).$$

We now state the bound.

Theorem 13 (Upper bound on population COCO). *Assume that functions in F and G are bounded a.s. by 1. Then for $n > 1$ and all $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{f \in F, g \in G} \mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) \leq \sup_{f \in F, g \in G} \widehat{\mathrm{cov}}(f(x), g(y)) + \sqrt{\frac{2\log(2/\delta)}{n(\sqrt{2}-1)^2}}.$$

where we denote the empirical covariance based on the sample $\boldsymbol{z}$ as

$$\widehat{\mathrm{cov}}(f(x), g(y)) := \frac{1}{n(n-1)} \sum_{i \neq j} f(x_i) g(y_j) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(y_i).$$

Proof. First, we note that

$$\sup_{f \in F, g \in G} \mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) - \sup_{f \in F, g \in G} \widehat{\mathrm{cov}}(f(x), g(y)) \leq \sup_{f \in F, g \in G} \left(\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) - \widehat{\mathrm{cov}}(f(x), g(y)) \right).$$

We can therefore ignore the suprema, and treat only the random variables $\mathsf{f} := f(\mathsf{x})$ and $\mathsf{g} := g(\mathsf{y})$ (thus, when we are considering a sample of size n, we write $\mathsf{f}_i = f(\mathsf{x}_i)$ and $\mathsf{g}_j = g(\mathsf{y}_j)$). To find the bound, we make the split

$$\begin{aligned}
\mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\mathrm{cov}(\mathsf{f}, \mathsf{g}) - \widehat{\mathrm{cov}}(f, g) \geq t \right) \quad \leq \quad & \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(-\frac{1}{n} \sum_{i=1}^{n} \mathsf{f}_i \mathsf{g}_i + \mathbf{E}_{\mathsf{f},\mathsf{g}}(\mathsf{fg}) \geq (1-\alpha) t \right) \\
& + \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \mathsf{f}_i \mathsf{g}_j - \mathbf{E}_{\mathsf{f}}(\mathsf{f}) \mathbf{E}_{\mathsf{g}}(\mathsf{g}) \geq \alpha t \right)
\end{aligned}$$

Next, we apply the theorems of Hoeffding to do the bound. For any $0 \leq \alpha \leq 1$,

$$\begin{aligned}
& \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\mathrm{cov}_{\mathsf{f},\mathsf{g}}(\mathsf{f}, \mathsf{g}) - \widehat{\mathrm{cov}}(f, g) \geq t \right) \\
\leq \quad & \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\mathbf{E}_{\mathsf{f},\mathsf{g}}(\mathsf{fg}) - \mathbf{E}_{\mathsf{f}}(\mathsf{f}) \mathbf{E}_{\mathsf{g}}(\mathsf{g}) - \frac{1}{n} \sum_{i=1}^{n} \mathsf{f}_i \mathsf{g}_i + \frac{1}{n(n-1)} \sum_{i \neq j} \mathsf{f}_i \mathsf{g}_j \geq t \right) \\
\leq \quad & \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(-\frac{1}{n} \sum_{i=1}^{n} \mathsf{f}_i \mathsf{g}_i + \mathbf{E}_{\mathsf{f},\mathsf{g}}(\mathsf{fg}) \geq (1-\alpha) t \right) + \mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \mathsf{f}_i \mathsf{g}_j - \mathbf{E}_{\mathsf{f}}(\mathsf{f}) \mathbf{E}_{\mathsf{g}}(\mathsf{g}) \geq \alpha t \right)
\end{aligned}$$

We replace $\mathsf{z}_i = \mathsf{f}_i \mathsf{g}_i$in Theorem 11 and use $|\mathsf{z}_i| \leq 1$ to obtain

$$\mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(-\frac{1}{n} \sum_{i=1}^{n} \mathsf{f}_i \mathsf{g}_i + \mathbf{E}_{\mathsf{f},\mathsf{g}}(\mathsf{fg}) \geq (1-\alpha) t \right) \leq \exp\left(-(1-\alpha)^2 n t^2 / 2 \right).$$

We likewise repace $\mathsf{z}_i = (\mathsf{f}_i, \mathsf{g}_i)$ in Theorem 12, and use the kernel $h((\mathsf{f}_i, \mathsf{g}_i), (\mathsf{f}_j \mathsf{g}_j)) = \mathsf{f}_i \mathsf{g}_j$, giving

$$\mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\frac{1}{n(n-1)} \sum_{i \neq j} \mathsf{f}_i \mathsf{g}_j - \mathbf{E}_{\mathsf{f}}(\mathsf{f}) \mathbf{E}_{\mathsf{g}}(\mathsf{g}) \geq \alpha t \right) \leq \exp\left(-\alpha^2 \lceil n/2 \rceil t^2 / 2 \right)$$

Combining these two results, we obtain

$$\begin{aligned}
\mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\mathrm{cov}_{\mathsf{f},\mathsf{g}}(\mathsf{f}, \mathsf{g}) - \widehat{\mathrm{cov}}(f, g) \geq t \right) \quad & \leq \quad \exp\left(-(1-\alpha)^2 n t^2 / 2 \right) + \exp\left(-\alpha^2 \lceil n/2 \rceil t^2 / 2 \right) \\
& \leq \quad \exp\left(-\alpha^2 n t^2 / 4 \right) + \exp\left(-(1-\alpha)^2 n t^2 / 2 \right)
\end{aligned}$$

We complete the proof by setting $\alpha = 2 - \sqrt{2}$, which gives both exponents the same argument. Thus

$$\mathbf{P}_{\mathsf{f}^n, \mathsf{g}^n} \left(\mathrm{cov}_{\mathsf{f},\mathsf{g}}(\mathsf{f}, \mathsf{g}) - \widehat{\mathrm{cov}}(f, g) \geq t \right) \leq 2 \exp\left(-(\sqrt{2}-1)^2 n t^2 / 2 \right).$$

This is easily rearranged to give the bound in the form of Theorem 13. ∎

5.2 Lower bound

A lower bound on the population COCO is harder to compute, since we have to deal with the suprema. We begin with McDiarmid's inequality [12].

Theorem 14 (McDiarmid's inequality). *Let $h : \mathcal{Z}^n \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, n\}$, there exist $c_i < \infty$ for which*

$$\sup_{\boldsymbol{z} \in \mathcal{Z}^n, \tilde{z} \in \mathcal{Z}} |h(z_1, \ldots z_n) - h(z_1, \ldots z_{i-1}, \tilde{z}, z_{i+1}, \ldots, z_n)| \leq c_i.$$

Then for all measures $\mathbf{P}_\mathsf{z}$ and every $t > 0$,

$$\mathbf{P}_{\mathsf{z}^n} \left(h(\mathsf{x}) - \mathbf{E}_{\mathsf{z}^n}(h(\mathsf{x})) > t \right) < \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

We now introduce the main theorem.

Theorem 15 (Lower bound on population COCO). *Assume functions in F and G are bounded a.s. by 1, and that the functions $k_f(x, x) \leq 1$ and $k_g(y, y) \leq 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then for $n > 1$ and all $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{f \in F, g \in G} \widehat{\mathrm{cov}}(f, g) \leq \sup_{f \in F, g \in G} \mathrm{cov}(f, g) + \frac{134}{\sqrt{n}} + \sqrt{\frac{18 \log 2/\delta}{n}}.$$

Proof. We begin with a rearrangement of the suprema;

$$\begin{aligned}
&\sup_{f \in F, g \in G} \widehat{\mathrm{cov}}(f(x), g(y)) - \sup_{f \in F, g \in G} \mathrm{cov}_{\mathsf{x},\mathsf{y}}(f(\mathsf{x}), g(\mathsf{y})) \\
\leq\ &\sup_{f \in F, g \in G} \left(\widehat{\mathrm{cov}}(f(x), g(y)) - \mathrm{cov}_{\mathsf{x},\mathsf{y}}(f(\mathsf{x}), g(\mathsf{y})) \right) \\
\leq\ &\sup_{f \in F, g \in G} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x}) g(\mathsf{y}) - \frac{1}{n} \sum_{i=1}^n f(x_i) g(y_i) \right) + \qquad (5.1) \\
&\sup_{f \in F, g \in G} \left(\frac{1}{n(n-1)} \sum_{i \neq j} f(x_i) g(y_j) - \mathbf{E}_\mathsf{x} f(\mathsf{x}) \mathbf{E}_\mathsf{y} g(\mathsf{y}) \right). \qquad (5.2)
\end{aligned}$$

When upper bounding the above, we treat the deviations of (5.1) and (5.2) separately (as we did in the last section), thus splitting the total deviation as $t = \alpha t + (1 - \alpha)t$. The first term is bounded using McDiarmid and symmetrisation in the usual way (see Appendix A.2), giving

$$\begin{aligned}
&\mathbf{P}_{\mathsf{x}^n, \mathsf{y}^n} \left[\sup_{f \in F, g \in G} \left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x}) g(\mathsf{y}) - \frac{1}{n} \sum_{i=1}^n f(\mathsf{x}_i) g(\mathsf{y}_i) \right) \right. \\
\geq\ &\left. \mathbf{E}_{\mathsf{x}^n, \mathsf{y}^n, \boldsymbol{\sigma}} \left(\sup_{f \in F, g \in G} \frac{2}{n} \sum_{i=1}^n \sigma_i f(\mathsf{x}_i) g(\mathsf{y}_i) \right) + (1 - \alpha)t \right] \leq e^{-n(1-\alpha)^2 t^2/2} \qquad (5.3)
\end{aligned}$$

In the case of the second term, we begin with McDiarmid to get

$$\begin{aligned}
&\mathbf{P}_{\mathsf{x}^n, \mathsf{y}^n} \left(\sup_{f \in F, g \in G} \left[\frac{1}{n(n-1)} \sum_{i \neq j} f_i g_j - \mathbf{E}_\mathsf{x} f \mathbf{E}_\mathsf{y} g \right] \geq \right. \\
&\underbrace{\left. \mathbf{E}_{\mathsf{x}^n, \mathsf{y}^n} \sup_{f \in F, g \in G} \left[\frac{1}{n(n-1)} \sum_{i \neq j} f_i g_j - \mathbf{E}_\mathsf{x} f \mathbf{E}_\mathsf{y} g \right] + \alpha t \right)}_{(a)} \qquad (5.4) \\
&\leq e^{\frac{-n\alpha^2 t^2}{8}}. \qquad (5.5)
\end{aligned}$$

(the proof is in Appendix A.3). Our next step is to replace (a) with an upper bound based on Rademacher averages. We cannot symmetrise (a) directly: instead, we first apply the Hoeffding decomposition and then decouple, following [3]. The Hoeffding decomposition yields

$$\begin{aligned}
&\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sup_{f\in F, g\in G} \left[\frac{1}{n(n-1)} \sum_{i\neq j} f_i g_j - \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g \right] \\
\leq\ & \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n (f(\mathsf{x}_i)g(\mathsf{y}_j) - f(\mathsf{x}_i)\mathbf{E}_{\mathsf{y}} g - \mathbf{E}_{\mathsf{x}} f\, g(\mathsf{y}_i) + \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g) \right) \\
&+ \mathbf{E}_{\mathsf{x}^n} \sup_{f,g} \left(\frac{2}{n} \sum_{i=1}^n (f(\mathsf{x}_i) - \mathbf{E}_{\mathsf{x}} f)\mathbf{E}_{\mathsf{y}} g \right).
\end{aligned}$$

The second term is easily symmetrised using the argument in Appendix A.2, giving

$$\begin{aligned}
\mathbf{E}_{\mathsf{x}^n} \sup_{f,g} \left(\frac{2}{n} \sum_{i=1}^n (f(\mathsf{x}_i) - \mathbf{E}_{\mathsf{x}} f)\mathbf{E}_{\mathsf{y}} g \right) &\leq \frac{4}{n} \mathbf{E}_{\mathsf{x}^n,\boldsymbol{\sigma}} \sup_{f,g} \sum_{i=1}^n (\sigma_i f(\mathsf{x}_i)\mathbf{E}_{\mathsf{y}} g) \\
&\leq \frac{4}{n} \mathbf{E}_{\mathsf{x}^n,\mathsf{y}'^n\boldsymbol{\sigma}} \sup_{f,g} \sum_{i=1}^n (\sigma_i f(\mathsf{x}_i) g(\mathsf{y}'_i)),
\end{aligned}$$

where the final line uses Jensen's inequality to bring the expectation outside the supremum, σ_i are Rademacher random variables that take values in $\{-1, 1\}$ with equal probability, and y'_i are independent copies of y_i. The first term cannot be symmetrised, as the sum is not over i.i.d. terms. Thus we apply Theorem 3.1.1. to replace the y_i in this term with independent copies y'_i, giving the upper bound

$$\begin{aligned}
&\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n f(\mathsf{x}_i)g(\mathsf{y}_j) - f(\mathsf{x}_i)\mathbf{E}_{\mathsf{y}} g - \mathbf{E}_{\mathsf{x}} f\, g(\mathsf{y}_i) + \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g \right) \leq \\
&8\mathbf{E}_{\mathsf{x}^n,\mathsf{y}'^n} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n f(\mathsf{x}_i)g(\mathsf{y}'_j) - f(\mathsf{x}_i)\mathbf{E}_{\mathsf{y}} g - \mathbf{E}_{\mathsf{x}} f\, g(\mathsf{y}'_i) + \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g \right) \qquad (5.6)
\end{aligned}$$

where we use the constant in [3, p. 102] (which is better than the general constant in [3, Theorem 3.1.1]). We can then apply the exact proof strategy of [3, Theorem 3.5.3] (the proof used is on the bottom of p. 140) to this decoupled quantity, which gives

$$\begin{aligned}
(5.6) \leq\ & 2 \times 2 \times 8\, \mathbf{E} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n \sigma_i \sigma'_j \left(f(\mathsf{x}_i)g(\mathsf{y}'_j) - f(\mathsf{x}_i)\mathbf{E}_{\mathsf{y}} g - \mathbf{E}_{\mathsf{x}} f\, g(\mathsf{y}'_i) + \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g \right) \right) \\
\leq\ & 32\, \mathbf{E} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n \sigma_i \sigma'_j \left(f(\mathsf{x}_i)g(\mathsf{y}'_j) - f(\mathsf{x}_i)g(\mathsf{y}''_j) - f(\mathsf{x}''_i)g(\mathsf{y}'_j) + f(\mathsf{x}'_i)g(\mathsf{y}''_j) \right) \right) \\
\leq\ & 4 \times 32\, \mathbf{E} \sup_{f,g} \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{i\neq j}^n \sigma_i \sigma'_j f(\mathsf{x}_i) g(\mathsf{y}'_j) \right),
\end{aligned}$$

where x'_i, x''_i are independent copies of x_i, and y'_i, y''_i are independent copies of y_i, and we use Jensen in the second line.[9]

[9] We no longer provide the random variables in the subscript of the expectation, since this would be unwieldy.

To conclude the proof, it turns out that we do not need to explicitly deal with these additional copies: instead, we apply a simple additional bound (see Appendix A.4) to get

$$\begin{aligned} & \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sup_{f\in F, g\in G} \left[\frac{1}{n(n-1)} \sum_{i\neq j} f_i g_j - \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g \right] \\ \leq \quad & \frac{128}{n(n-1)} \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sqrt{\sum_{i\neq j} k_f(\mathsf{x}_i,\mathsf{x}_i) k_g(\mathsf{y}_i,\mathsf{y}_i)} + \frac{4}{n} \mathbf{E}_{\mathsf{x}^n,\mathsf{y}'^n} \sqrt{\sum_{i=1}^{n} k_f(\mathsf{x}_i,\mathsf{x}_i) k_g(\mathsf{y}'_i,\mathsf{y}'_i)} \end{aligned}$$

and then substitute $k_f(x,x) \leq 1$ and $k_g(y,y) \leq 1$. Setting $\alpha = 2/3$ in (5.3) and (5.5) (to give the same exponent for both deviations), and bearing in mind that $\frac{1}{\sqrt{n(n-1)}} < \frac{1}{\sqrt{n}}$ when $n > 1$, completes the proof. □

A Proofs

A.1 Covariance between eigenfunctions

We prove that there exists a density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ (specifically, that in 4.2) for which $\mathrm{cov}(\varphi_l(x), \varphi_l(y))$ is large for big enough l, yet with a small covariance $\mathrm{cov}(\varphi_i(x), \varphi_j(y))$ between *any* other pair of eigenfunctions ($i \neq l$, $j \neq l$). We begin by defining the expansions of the constant functions $e(x) = 1$ on $\mathcal{X}$ and $e(y) = 1$ on $\mathcal{Y}$. Using the notation $\mathbb{I}_A$ to denote a function which is one when A holds, and zero otherwise, we have

$$\begin{array}{rcccccc} e(x) & := & \mathbb{I}_{x\in\mathcal{X}} & = & \sum_{p=1}^{\infty} \varphi_p(x)\langle 1, \varphi_p(x)\rangle_{L_2(\mathcal{X})} & =: & \sum_{p=1}^{\infty} \tilde{e}_p^x \varphi_p(x) \\ e(y) & := & \mathbb{I}_{y\in\mathcal{Y}} & = & \sum_{q=1}^{\infty} \varphi_q(y)\langle 1, \varphi_q(y)\rangle_{L_2(\mathcal{X})} & =: & \sum_{q=1}^{\infty} \tilde{e}_q^y \varphi_q(y) \end{array}$$

where we use the notation $\tilde{e}_p^x := \langle 1, \varphi_p(x)\rangle_{L_2(\mathcal{X})}$ and $\tilde{e}_q^y := \langle 1, \varphi_q(y)\rangle_{L_2(\mathcal{Y})}$, consistent with Lemma 6.

We now begin the proof. We first write a matrix $\widetilde{\mathbf{C}}$ (having infinite size) of covariances $\mathrm{cov}(\varphi_i(x), \varphi_j(y))$, with (i,j)th entry

$$\begin{aligned} \widetilde{C}_{i,j} &:= \mathrm{cov}(\varphi_i(x), \varphi_j(y)) \\ &= \mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \mathsf{E}_{\mathsf{x}}(\varphi_i(\mathsf{x}))\mathsf{E}_{\mathsf{y}}(\varphi_j(\mathsf{y})) \\ &= \mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\mathbb{I}_{\mathsf{y}\in\mathcal{Y}})\mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_j(\mathsf{y})\mathbb{I}_{\mathsf{x}\in\mathcal{X}}) \\ &= \mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\varphi_i(\mathsf{x})\sum_{q=1}^{\infty}\tilde{e}_q^{(y)}\varphi_q(\mathsf{y})\right)\mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\varphi_j(\mathsf{y})\sum_{p=1}^{\infty}\tilde{e}_p^{(x)}\varphi_p(\mathsf{x})\right) \\ &= \mathsf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \sum_{q=1}^{\infty}\tilde{e}_q^{(y)}\mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\varphi_i(\mathsf{x})\varphi_q(\mathsf{y})\right)\sum_{p=1}^{\infty}\tilde{e}_p^{(x)}\mathsf{E}_{\mathsf{x},\mathsf{y}}\left(\varphi_p(\mathsf{x})\varphi_j(\mathsf{y})\right). \end{aligned}$$

Our next step is to use the density from (4.2) in the expectations above, which allows us to prove that only $\mathrm{cov}(\varphi_l(x), \varphi_l(y))$ is large. We remind the reader that this density takes the form

$$\begin{aligned} \mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y) &= \alpha_l e(x)e(y) + \beta\varphi_l(x)\varphi_l(y) \\ &= \alpha_l\left(\sum_{p=1}^{\infty}\tilde{e}_p^{(x)}\varphi_p(x)\right)\left(\sum_{q=1}^{\infty}\tilde{e}_q^{(y)}\varphi_q(y)\right) + \beta\varphi_l(x)\varphi_l(y). \end{aligned}$$

This expression contains two as yet unknown constants α_l and β, which are constrained by the requirement that $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)$ be a density.[10] Enforcing $\mathbf{f}_{\mathsf{x},\mathsf{y}} \geq 0$ requires $\alpha_l - \beta\min_{x,y}(\varphi_l(x)\varphi_l(y)) \geq 0$, which can be satisfied as long as the $\varphi_l(x)$ and $\varphi_l(y)$ are absolutely bounded; the density having unit integral implies

$$\begin{aligned} 1 &= \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy \\ &= \int_{\mathcal{X}\times\mathcal{Y}} e(x)e(y)\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy \\ &= \alpha_l\int_{\mathcal{X}\times\mathcal{Y}}\left(\sum_{p=1}^{\infty}\tilde{e}_p^x\varphi_p(x)\right)^2\left(\sum_{q=1}^{\infty}\tilde{e}_q^y\varphi_q(y)\right)^2 dxdy \\ &\qquad +\beta\int_{\mathcal{X}\times\mathcal{Y}}\left(\sum_{p=1}^{\infty}\tilde{e}_p^x\varphi_p(x)\right)\left(\sum_{q=1}^{\infty}\tilde{e}_q^y\varphi_q(y)\right)\varphi_l(x)\varphi_l(y)dxdy \\ &= \alpha_l M_x M_y + \beta\tilde{e}_l^x\tilde{e}_l^y, \end{aligned} \tag{A.1}$$

where for ease of notation we define $M_x := \sum_{p=1}^{\infty}\left(\tilde{e}_p^x\right)^2$ and $M_y = \sum_{q=1}^{\infty}\left(\tilde{e}_q^y\right)^2$. From Lemma 6, the series $\tilde{e}_i^x$ and $\tilde{e}_j^y$ are absolutely convergent: thus for a sufficiently large l, $|\tilde{e}_l^x|$ and $|\tilde{e}_l^y|$ are small. We rearrange (A.1) to get

$$(1 - \beta|\tilde{e}_l^x\tilde{e}_l^y|)(M_xM_y)^{-1} \leq \alpha_l \leq (1 + \beta|\tilde{e}_l^x\tilde{e}_l^y|)(M_xM_y)^{-1}. \tag{A.2}$$

[10] This implies two constraints: $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y) \geq 0$ on $\mathcal{X}\times\mathcal{Y}$ and $\int_{\mathcal{X}\times\mathcal{Y}}\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)dxdy = 1$.

We now substitute $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x,y)$ into (4.3). After simplifying, we find $\widetilde{\mathbf{C}}$ has i,jth entries

$$\widetilde{C}_{i,j} = \begin{cases} \alpha_l \tilde{e}_i^x \tilde{e}_j^y (1 - \alpha_l M_x M_y) & := & \epsilon_{ij} & i,j \neq l, \\ \alpha_l \tilde{e}_l^x \tilde{e}_j^y - (\beta \tilde{e}_l^y + \alpha_l \tilde{e}_l^x M_y)\left(\alpha_l \tilde{e}_j^y M_x\right) & := & \epsilon_{lj} & i = l, j \neq l, \\ \alpha_l \tilde{e}_i^x \tilde{e}_l^y - (\alpha_l \tilde{e}_i^x M_y)\left(\beta \tilde{e}_l^x + \alpha_l \tilde{e}_l^y M_x\right) & := & \epsilon_{jl} & i \neq l, j = l, \\ \beta + \alpha_l \tilde{e}_l^x \tilde{e}_l^y - (\beta \tilde{e}_l^y + \alpha_l \tilde{e}_l^x M_y)\left(\beta \tilde{e}_l^x + \alpha_l \tilde{e}_l^y M_x\right) & := & \beta + \epsilon_{ll} & i = j = l, \end{cases}$$

where ϵ_{ij} denotes a quantity with small absolute value for large enough l, and we use the bounds on α_l from (A.2) to determine which $\widetilde{C}_{i,j}$ will be small. The density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ thus satisfies $\mathrm{cov}(\varphi_i(x), \varphi_j(y)) \geq \beta - \epsilon$ when $i = j = l$, and $\mathrm{cov}(\varphi_i(x), \varphi_j(y)) < \epsilon$ elsewhere, where $\epsilon = \max_{i.j} |\epsilon_{ij}|$.

A.2 Symmetrisation and McDiarmid

In this section, we prove the bound

$$\begin{aligned} &\mathbf{P}_{\mathsf{x}^n,\mathsf{y}^n}\left[\sup_{f\in F, g\in G}\left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x})g(\mathsf{y}) - \frac{1}{n}\sum_{i=1}^n f(\mathsf{x}_i)g(\mathsf{y}_i)\right)\right. \\ &\geq \left.\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n,\boldsymbol{\sigma}}\left(\sup_{f\in F, g\in G}\frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathsf{x}_i)g(\mathsf{y}_i)\right) + t\right] \leq e^{-nt^2/2} \end{aligned}$$

where the σ_i are Rademacher variables (i.e. σ_i and σ_j are i.i.d. for $i \neq j$ and $\mathbf{P}_{\sigma_i}(\sigma_i = 1) = \mathbf{P}_{\sigma_i}(\sigma_i = -1) = 1/2$). First, we simplify our notation. We define $\mathsf{z}_i = (\mathsf{x}_i, \mathsf{y}_i)$, and $h(\mathsf{z}_1 \ldots, \mathsf{z}_n) = \sup_{f\in F, g\in G}\left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x})g(\mathsf{y}) - \frac{1}{n}\sum_{i=1}^n f(\mathsf{x}_i)g(\mathsf{y}_i)\right)$ in Theorem 14. Given that functions f and g are assumed absolutely bounded by 1, it is clear that $|c_i| \leq 1/n$ for all c_i in Theorem 14. Thus.

$$\begin{aligned} &\mathbf{P}_{\mathsf{x}^n,\mathsf{y}^n}\left[\sup_{f\in F, g\in G}\left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x})g(\mathsf{y}) - \frac{1}{n}\sum_{i=1}^n f(\mathsf{x}_i)g(\mathsf{y}_i)\right)\right. \\ &\geq \left.\underbrace{\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n}\left(\sup_{f\in F, g\in G}\left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x})g(\mathsf{y}) - \frac{1}{n}\sum_{i=1}^n f(\mathsf{x}_i)g(\mathsf{y}_i)\right)\right)}_{(a)} + t\right] \leq e^{-nt^2/2} \end{aligned}$$

Next we symmetrise, which allows us to replace (a) with an upper bound. To simplify notation, we introduce a new function class $L := F \otimes G$ of functions taking the form $l(x,y) = f(x)g(y)$. We denote the original sample as $\boldsymbol{z}$, and the ghost sample (the i.i.d. copy of $\boldsymbol{z}$) as $\widetilde{\boldsymbol{z}}$.

$$
\begin{aligned}
& \mathbf{E}_{\mathsf{z}^n}\left(\sup_{l\in L}\left(\mathbf{E}_{\mathsf{z}}(l(\mathsf{z}))-\frac{1}{n}\sum_{i=1}^{n} l(\mathsf{z}_i)\right)\right) \\
= \;& \mathbf{E}_{\mathsf{z}^n}\left(\sup_{l\in L}\left(\mathbf{E}_{\tilde{\mathsf{z}}^n}\left(\frac{1}{n}\sum_{i=1}^{n} l(\tilde{\mathsf{z}}_i)\right)-\frac{1}{n}\sum_{i=1}^{n} l(\mathsf{z}_i)\right)\right) \\
\overset{(a)}{\leq} \;& \mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\tilde{\mathsf{z}}^n}\sup_{l\in L}\left(\frac{1}{n}\sum_{i=1}^{n} l(\tilde{\mathsf{z}}_i)-l(\mathsf{z}_i)\right) \\
\overset{(b)}{=} \;& \frac{1}{n}\mathbf{E}_{\mathsf{z}^{m-1},\tilde{\mathsf{z}}_k}\mathbf{E}_{\tilde{\mathsf{z}}^{m-1},\mathsf{z}_k}\sup_{l\in L}\left(\sum_{i\neq k} l(\tilde{\mathsf{z}}_i)-l(\mathsf{z}_i)+l(\mathsf{z}_k)-l(\tilde{\mathsf{z}}_k)\right) \\
\overset{(c)}{=} \;& \mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\tilde{\mathsf{z}}^n}\sup_{l\in L}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\tilde{\mathsf{z}}_i)+\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[-l(\mathsf{z}_i)\right]\right) \\
\overset{(d)}{\leq} \;& \mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\tilde{\mathsf{z}}^n}\mathbf{E}_{\boldsymbol{\sigma}}\left(\sup_{l\in L}\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\tilde{\mathsf{z}}_i)+\sup_{l\in L}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left[-l(\mathsf{z}_i)\right]\right) \\
\overset{(e)}{=} \;& \mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\boldsymbol{\sigma}}\left(\sup_{l\in L}\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\mathsf{z}_i)\right)+\mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\boldsymbol{\sigma}}\left(\sup_{l\in L}\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\mathsf{z}_i)\right) \\
= \;& 2\mathbf{E}_{\mathsf{z}^n}\mathbf{E}_{\boldsymbol{\sigma}}\left(\sup_{l\in L}\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\mathsf{z}_i)\right)
\end{aligned}
$$

(a) The supremum is convex: we can permute expectation and supremum using Jensen's inequality.
(b) $\tilde{\mathsf{z}}_k$ and z_k have the same distribution, so we can swap them.
(c) This is true for any $\boldsymbol{\sigma} \in \{-1,1\}^n$.
(d) The supremum is convex. $\mathbf{E}_{\boldsymbol{\sigma}}$ is taken assuming $\mathbf{P}_\sigma(\sigma=1)=\mathbf{P}_\sigma(\sigma=-1)=1/2$, which is used in the next step.
(e) Here $\mathbf{E}_{\mathsf{z}^n}$ and $\mathbf{E}_{\tilde{\mathsf{z}}^n}$ effectively mean the same thing. The 2nd term loses its minus due to the symmetry of $\mathbf{P}_\sigma$.

The final term is a rademacher average for class $\mathcal{F}$.

A.3 McDiarmid for the U-statistic

In this section, we prove

$$
\begin{aligned}
& \mathbf{P}_{\mathsf{x}^n,\mathsf{y}^n}\left(\sup_{f\in F,g\in G}\left(\frac{1}{n(n-1)}\sum_{i\neq j} f(\mathsf{x}_i)g(\mathsf{x}_j)-\mathbf{E}_{\mathsf{x}}f\mathbf{E}_{\mathsf{y}}g\right)\geq\right. \\
& \left.\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n}\sup_{f\in F,g\in G}\left[\frac{1}{n(n-1)}\sum_{i\neq j} f(\mathsf{x}_i)g(\mathsf{x}_j)-\mathbf{E}_{\mathsf{x}}f\mathbf{E}_{\mathsf{y}}g\right]+t\right) \\
& \leq e^{\frac{-nt^2}{8}}.
\end{aligned}
$$

We define the new variable $z_i := (x_i, y_i)$, and set $h(z_1,\ldots,z_n) := \sup_{f\in F,g\in G}(l(z_1,\ldots,z_n))$ where $l(z_1,\ldots,z_n)=\frac{1}{n(n-1)}\sum_{i\neq j} f_i g_j - \mathbf{E}_{\mathsf{x}}f\mathbf{E}_{\mathsf{y}}g$. Given we assume $|f|\leq 1$ and $|g|\leq 1$,

$$
\begin{aligned}
c_k \leq & |h(z_1,\ldots,z_k,\ldots,z_n) - h(z_1,\ldots,z'_k,\ldots,z_n)| \\
\leq & \sup_{f\in F, g\in G} |l(z_1,\ldots,z_k,\ldots,z_n) - l(z_1,\ldots,z'_k,\ldots,z_n)| \\
= & \frac{1}{n(n-1)} \sup_{f\in F,g\in G} \left| \sum_{i\neq j\neq k} f_i g_j - \sum_{i\neq j\neq k} f_i g_j + f(x_k)\sum_{j\neq k} g_j + g(y_k)\sum_{i\neq k} f_i - f(x'_k)\sum_{j\neq k} g_j - g(y'_k)\sum_{i\neq k} f_i \right| \\
= & \frac{1}{n(n-1)} \sup_{f\in F,g\in G} \left| (f(x_k) - f(x'_k))\sum_{j\neq k} g_j + (g(y_k) - g(y'_k))\sum_{i\neq k} f_i \right| \\
\leq & \frac{1}{n(n-1)} \sup_{f\in F,g\in G} \left(\left| (f(x_k) - f(x'_k))\sum_{j\neq k} g_j \right| + \left| (g(y_k) - g(y'_k))\sum_{i\neq k} f_i \right| \right) \\
\leq & \frac{2}{n(n-1)} \sup_{f\in F,g\in G} \left(\left| \sum_{j\neq k} g_j \right| + \left| \sum_{i\neq k} f_i \right| \right) \\
\leq & \frac{4}{n}
\end{aligned}
$$

The exponent in McDiarmid's theorem (Theorem 14) then becomes

$$
\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \leq \exp\left(-\frac{2t^2}{\frac{16n}{n^2}}\right) = \exp\left(-\frac{nt^2}{8}\right)
$$

as required.

A.4 Bounds on the decoupled Rademacher average and chaos

In this section, we show how to bound the decoupled Rademacher average and chaos in such a manner as to avoid computing them empirically.

Lemma 16 (Bounds on the Rademacher average and chaos). *We have*

$$
\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n,\boldsymbol{\sigma}} \left(\sup_{f,g} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathsf{x}_i) g(\mathsf{y}_i) \right) \leq \frac{1}{n} \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sqrt{\sum_{i=1}^n k(\mathsf{x}_i,\mathsf{x}_i) l(\mathsf{y}_i,\mathsf{y}_i)},
$$

and

$$
\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n,\boldsymbol{\sigma}} \left(\sup_{f,g} \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j\neq i}^n \sigma_i\sigma_j f(\mathsf{x}_i) g(\mathsf{y}_j) \right) \leq \frac{1}{n(n-1)} \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sqrt{\sum_{i=1}^n \sum_{j\neq i}^n k(\mathsf{x}_i,\mathsf{x}_i) l(\mathsf{y}_i,\mathsf{y}_i)}.
$$

Proof. We start with

$$
\sup_{f,g} \sum_{i=1}^n \sigma_i f(\mathsf{x}_i) g(\mathsf{y}_i) = \left\| \sum_{i=1}^n \sigma_i Q_{\mathsf{x}_i,\mathsf{y}_i} \right\|,
$$

where $Q_{\mathsf{x}_i,\mathsf{y}_i}$ is the rank one operator mapping g to $k(\mathsf{x}_i,.)g(\mathsf{y}_i)$. We can upper bound this norm by the Hilbert-Schmidt norm to get

$$
\sup_{f,g} \sum_{i=1}^n \sigma_i f(\mathsf{x}_i) g(\mathsf{y}_i) \leq \sqrt{\sum_{i,j} \sigma_i\sigma_j \left\langle Q_{\mathsf{x}_i,\mathsf{y}_i}, Q_{\mathsf{x}_j,\mathsf{y}_j} \right\rangle_2}.
$$

Taking expectations and using Jensen's inequality, we get

$$\begin{aligned}\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n,\boldsymbol{\sigma}} \sup_{f,g} \sum_{i=1}^n \sigma_i f(\mathsf{x}_i) g(\mathsf{y}_i) &\leq \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sqrt{\sum_{i=1}^n \|Q_{\mathsf{x}_i,\mathsf{y}_i}\|_2^2} \\ &= \mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n} \sqrt{\sum_{i=1}^n k(\mathsf{x}_i,\mathsf{x}_i) l(\mathsf{y}_i,\mathsf{y}_i)}.\end{aligned}$$

This gives the first result. A similar reasoning can then be applied to the Rademacher chaos. □

References

[1] F. Bach and M. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

[2] G.H. Bakır, A. Gretton, M. Franz, and B. Schölkopf. Multivariate regression with stiefel constraints. Technical Report 101, Max Planck Institute for Biological Cybernetics, 2004.

[3] V. de la Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, New York, 1999.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of mathematics*. Springer, New York, 1996.

[5] A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. Technical report, MPI for Biological Cybernetics, 2003.

[6] A. Gretton, A. Smola, O. Bousquet, R. Herbrich, A. Belitski, M. Augath, Y. Murayama, J. Pauls, B. Schölkopf, and N. Logothetis. Kernel constrained covariance for dependence measurement. In *AIStats (submitted)*, 2005.

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[8] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2001.

[9] Yu. I. Ingster. An asymptotically minimax test of the hypothesis of independence. *J. Soviet Math.*, 44:466–476, 1989.

[10] J. Jacod and P. Protter. *Probability Essentials*. Springer, New York, 2000.

[11] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *J. R. Stat. Soc. B*, 55(3):725–740, 1993.

[12] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989. Cambridge University Press.

[13] E. Mourier. Éléments aléatoires dans un éspace de Banach. *Ann. Inst. H. Poincaré Sect B.*, 161:161–244, 1953.

[14] A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

[15] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific and Technical, Harlow, UK, 1988.

[16] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT press, Cambridge, MA, 2002.

[17] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2, 2001.