Journal of Machine Learning Research 13 (2012) 723-773 Submitted 4/08; Revised 11/11; Published 3/12

A Kernel Two-Sample Test

Arthur Gretton[*] ARTHUR.GRETTON@GMAIL.COM
MPI for Intelligent Systems
Spemannstrasse 38
72076 Tübingen, Germany

Karsten M. Borgwardt[†] KARSTEN.BORGWARDT@TUEBINGEN.MPG.DE
Machine Learning and Computational Biology Research Group
Max Planck Institutes Tübingen
Spemannstrasse 38
72076 Tübingen, Germany

Malte J. Rasch[‡] MALTE@MAIL.BNU.EDU.CN
19 XinJieKouWai St.
State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University,
Beijing, 100875, P.R. China

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
MPI for Intelligent Systems
Spemannstrasse 38
72076, Tübingen, Germany

Alexander Smola[§] ALEX@SMOLA.ORG
Yahoo! Research
2821 Mission College Blvd
Santa Clara, CA 95054, USA

Editor: Nicolas Vayatis

Abstract

We propose a framework for analyzing and comparing distributions, which we use to construct statistical tests to determine if two samples are drawn from different distributions. Our test statistic is the largest difference in expectations over functions in the unit ball of a reproducing kernel Hilbert space (RKHS), and is called the *maximum mean discrepancy* (MMD). We present two distribution-free tests based on large deviation bounds for the MMD, and a third test based on the asymptotic distribution of this statistic. The MMD can be computed in quadratic time, although efficient linear time approximations are available. Our statistic is an instance of an integral probability metric, and various classical metrics on distributions are obtained when alternative function classes are used in place of an RKHS. We apply our two-sample tests to a variety of problems, including attribute matching for databases using the Hungarian marriage method, where they perform strongly. Excellent performance is also obtained when comparing distributions over graphs, for which these are the first such tests.

∗. Also at Gatsby Computational Neuroscience Unit, CSML, 17 Queen Square, London WC1N 3AR, UK.
†. This work was carried out while K.M.B. was with the Ludwig-Maximilians-Universität München.
‡. This work was carried out while M.J.R. was with the Graz University of Technology.
§. Also at The Australian National University, Canberra, ACT 0200, Australia.

Keywords: kernel methods, two-sample test, uniform convergence bounds, schema matching, integral probability metric, hypothesis testing

1. Introduction

We address the problem of comparing samples from two probability distributions, by proposing statistical tests of the null hypothesis that these distributions are equal against the alternative hypothesis that these distributions are different (this is called the two-sample problem). Such tests have application in a variety of areas. In bioinformatics, it is of interest to compare microarray data from identical tissue types as measured by different laboratories, to detect whether the data may be analysed jointly, or whether differences in experimental procedure have caused systematic differences in the data distributions. Equally of interest are comparisons between microarray data from different tissue types, either to determine whether two subtypes of cancer may be treated as statistically indistinguishable from a diagnosis perspective, or to detect differences in healthy and cancerous tissue. In database attribute matching, it is desirable to merge databases containing multiple fields, where it is not known in advance which fields correspond: the fields are matched by maximising the similarity in the distributions of their entries.

We test whether distributions p and q are different on the basis of samples drawn from each of them, by finding a well behaved (e.g., smooth) function which is large on the points drawn from p, and small (as negative as possible) on the points from q. We use as our test statistic the difference between the mean function values on the two samples; when this is large, the samples are likely from different distributions. We call this test statistic the Maximum Mean Discrepancy (MMD).

Clearly the quality of the MMD as a statistic depends on the class $\mathcal{F}$ of smooth functions that define it. On one hand, $\mathcal{F}$ must be "rich enough" so that the population MMD vanishes if and only if $p = q$. On the other hand, for the test to be consistent in power, $\mathcal{F}$ needs to be "restrictive" enough for the empirical estimate of the MMD to converge quickly to its expectation as the sample size increases. We will use the unit balls in characteristic reproducing kernel Hilbert spaces (Fukumizu et al., 2008; Sriperumbudur et al., 2010b) as our function classes, since these will be shown to satisfy both of the foregoing properties. We also review classical metrics on distributions, namely the Kolmogorov-Smirnov and Earth-Mover's distances, which are based on different function classes; collectively these are known as integral probability metrics (Müller, 1997). On a more practical note, the MMD has a reasonable computational cost, when compared with other two-sample tests: given m points sampled from p and n from q, the cost is $O(m+n)^2$ time. We also propose a test statistic with a computational cost of $O(m+n)$: the associated test can achieve a given Type II error at a lower overall computational cost than the quadratic-cost test, by looking at a larger volume of data.

We define three nonparametric statistical tests based on the MMD. The first two tests are distribution-free, meaning they make no assumptions regarding p and q, albeit at the expense of being conservative in detecting differences between the distributions. The third test is based on the asymptotic distribution of the MMD, and is in practice more sensitive to differences in distribution at small sample sizes. The present work synthesizes and expands on results of Gretton et al. (2007a,b) and Smola et al. (2007),[1] who in turn build on the earlier work of Borgwardt et al. (2006). Note that

1. In particular, most of the proofs here were not provided by Gretton et al. (2007a), but in an accompanying technical report (Gretton et al., 2008a), which this document replaces.

the latter addresses only the third kind of test, and that the approach of Gretton et al. (2007a,b) is rigorous in its treatment of the asymptotic distribution of the test statistic under the null hypothesis.

We begin our presentation in Section 2 with a formal definition of the MMD. We review the notion of a characteristic RKHS, and establish that when $\mathcal{F}$ is a unit ball in a characteristic RKHS, then the population MMD is zero if and only if $p = q$. We further show that universal RKHSs in the sense of Steinwart (2001) are characteristic. In Section 3, we give an overview of hypothesis testing as it applies to the two-sample problem, and review alternative test statistics, including the L_2 distance between kernel density estimates (Anderson et al., 1994), which is the prior approach closest to our work. We present our first two hypothesis tests in Section 4, based on two different bounds on the deviation between the population and empirical MMD. We take a different approach in Section 5, where we use the asymptotic distribution of the empirical MMD estimate as the basis for a third test. When large volumes of data are available, the cost of computing the MMD (quadratic in the sample size) may be excessive: we therefore propose in Section 6 a modified version of the MMD statistic that has a linear cost in the number of samples, and an associated asymptotic test. In Section 7, we provide an overview of methods related to the MMD in the statistics and machine learning literature. We also review alternative function classes for which the MMD defines a metric on probability distributions. Finally, in Section 8, we demonstrate the performance of MMD-based two-sample tests on problems from neuroscience, bioinformatics, and attribute matching using the Hungarian marriage method. Our approach performs well on high dimensional data with low sample size; in addition, we are able to successfully distinguish distributions on graph data, for which ours is the first proposed test.

A Matlab implementation of the tests is at `www.gatsby.ucl.ac.uk/~gretton/mmd/mmd.htm`.

2. The Maximum Mean Discrepancy

In this section, we present the maximum mean discrepancy (MMD), and describe conditions under which it is a metric on the space of probability distributions. The MMD is defined in terms of particular function spaces that witness the difference in distributions: we therefore begin in Section 2.1 by introducing the MMD for an arbitrary function space. In Section 2.2, we compute both the population MMD and two empirical estimates when the associated function space is a reproducing kernel Hilbert space, and in Section 2.3 we derive the RKHS function that witnesses the MMD for a given pair of distributions.

2.1 Definition of the Maximum Mean Discrepancy

Our goal is to formulate a statistical test that answers the following question:

Problem 1 *Let x and y be random variables defined on a topological space $\mathcal{X}$, with respective Borel probability measures p and q. Given observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$, independently and identically distributed (i.i.d.) from p and q, respectively, can we decide whether $p \neq q$?*

Where there is no ambiguity, we use the shorthand notation $\mathbf{E}_x[f(x)] := \mathbf{E}_{x\sim p}[f(x)]$ and $\mathbf{E}_y[f(y)] := \mathbf{E}_{y\sim q}[f(y)]$ to denote expectations with respect to p and q, respectively, where $x \sim p$ indicates x has distribution p. To start with, we wish to determine a criterion that, in the population setting, takes on a unique and distinctive value only when $p = q$. It will be defined based on Lemma 9.3.2 of Dudley (2002).

Lemma 1 *Let $(\mathcal{X},d)$ be a metric space, and let p,q be two Borel probability measures defined on $\mathcal{X}$. Then $p=q$ if and only if $\mathbf{E}_x(f(x)) = \mathbf{E}_y(f(y))$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.*

Although $C(\mathcal{X})$ in principle allows us to identify $p=q$ uniquely, it is not practical to work with such a rich function class in the finite sample setting. We thus define a more general class of statistic, for as yet unspecified function classes $\mathcal{F}$, to measure the disparity between p and q (Fortet and Mourier, 1953; Müller, 1997).

Definition 2 *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let p,q,x,y,X,Y be defined as above. We define the maximum mean discrepancy (MMD) as*

$$\mathrm{MMD}\left[\mathcal{F},p,q\right] := \sup_{f\in\mathcal{F}} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]\right). \tag{1}$$

In the statistics literature, this is known as an integral probability metric (Müller, 1997). A biased[2] *empirical estimate of the MMD is obtained by replacing the population expectations with empirical expectations computed on the samples X and Y,*

$$\mathrm{MMD}_b\left[\mathcal{F},X,Y\right] := \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i)\right). \tag{2}$$

We must therefore identify a function class that is rich enough to uniquely identify whether $p=q$, yet restrictive enough to provide useful finite sample estimates (the latter property will be established in subsequent sections).

2.2 The MMD in Reproducing Kernel Hilbert Spaces

In the present section, we propose as our MMD function class $\mathcal{F}$ the unit ball in a reproducing kernel Hilbert space $\mathcal{H}$. We will provide finite sample estimates of this quantity (both biased and unbiased), and establish conditions under which the MMD can be used to distinguish between probability measures. Other possible function classes $\mathcal{F}$ are discussed in Sections 7.1 and 7.2.

We first review some properties of $\mathcal{H}$ (Schölkopf and Smola, 2002). Since $\mathcal{H}$ is an RKHS, the operator of evaluation δ_x mapping $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$ is continuous. Thus, by the Riesz representation theorem (Reed and Simon, 1980, Theorem II.4), there is a feature mapping $\phi(x)$ from $\mathcal{X}$ to $\mathbb{R}$ such that $f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$. This feature mapping takes the canonical form $\phi(x) = k(x,\cdot)$ (Steinwart and Christmann, 2008, Lemma 4.19), where $k(x_1,x_2) : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is positive definite, and the notation $k(x,\cdot)$ indicates the kernel has one argument fixed at x, and the second free. Note in particular that $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x,y)$. We will generally use the more concise notation $\phi(x)$ for the feature mapping, although in some cases it will be clearer to write $k(x,\cdot)$.

We next extend the notion of feature map to the embedding of a probability distribution: we will define an element $\mu_p \in \mathcal{H}$ such that $\mathbf{E}_x f = \langle f, \mu_p\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, which we call the *mean embedding* of p. Embeddings of probability measures into reproducing kernel Hilbert spaces are well established in the statistics literature: see Berlinet and Thomas-Agnan (2004, Chapter 4) for further detail and references. We begin by establishing conditions under which the mean embedding μ_p exists (Fukumizu et al., 2004, p. 93), (Sriperumbudur et al., 2010b, Theorem 1).

2. The empirical MMD defined below has an upward bias—we will define an unbiased statistic in the following section.

Lemma 3 *If $k(\cdot,\cdot)$ is measurable and $\mathbf{E}_x\sqrt{k(x,x)} < \infty$ then $\mu_p \in \mathcal{H}$.*

Proof The linear operator $T_p f := \mathbf{E}_x f$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_p f| = |\mathbf{E}_x f| \leq \mathbf{E}_x |f| = \mathbf{E}_x |\langle f, \phi(x)\rangle_{\mathcal{H}}| \leq \mathbf{E}_x \left(\sqrt{k(x,x)}\, \|f\|_{\mathcal{H}}\right).$$

Hence by the Riesz representer theorem, there exists a $\mu_p \in \mathcal{H}$ such that $T_p f = \langle f, \mu_p\rangle_{\mathcal{H}}$. If we set $f = \phi(t) = k(t,\cdot)$, we obtain $\mu_p(t) = \langle \mu_p, k(t,\cdot)\rangle_{\mathcal{H}} = \mathbf{E}_x k(t,x)$: in other words, the mean embedding of the distribution p is the expectation under p of the canonical feature map. ■

We next show that the MMD may be expressed as the distance in $\mathcal{H}$ between mean embeddings (Borgwardt et al., 2006).

Lemma 4 *Assume the condition in Lemma 3 for the existence of the mean embeddings μ_p, μ_q is satisfied. Then*

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \left\|\mu_p - \mu_q\right\|^2_{\mathcal{H}}.$$

Proof

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{F}, p, q] &= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1} \left(\mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)]\right)\right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1} \left\langle \mu_p - \mu_q, f\right\rangle_{\mathcal{H}}\right]^2 \\
&= \left\|\mu_p - \mu_q\right\|^2_{\mathcal{H}}.
\end{aligned}$$

■

We now establish a condition on the RKHS $\mathcal{H}$ under which the mean embedding μ_p is injective, which indicates that $\mathrm{MMD}[\mathcal{F}, p, q]$ is a metric[3] on the Borel probability measures on $\mathcal{X}$. Evidently, this property will not hold for all $\mathcal{H}$: for instance, a polynomial RKHS of degree two cannot distinguish between distributions with the same mean and variance, but different kurtosis (Sriperumbudur et al., 2010b, Example 3). The MMD is a metric, however, when $\mathcal{H}$ is a *universal* RKHS, defined on a compact metric space $\mathcal{X}$. Universality requires that $k(\cdot,\cdot)$ be continuous, and $\mathcal{H}$ be dense in $C(\mathcal{X})$ with respect to the L_∞ norm. Steinwart (2001) proves that the Gaussian and Laplace RKHSs are universal.

Theorem 5 *Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact metric space $\mathcal{X}$, with associated continuous kernel $k(\cdot,\cdot)$. Then $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ if and only if $p = q$.*

Proof The proof follows Cortes et al. (2008, Supplementary Appendix), whose approach is clearer than the original proof of Gretton et al. (2008a, p. 4).[4] First, it is clear that $p = q$ implies

3. According to Dudley (2002, p. 26) a metric $d(x,y)$ satisfies the following four properties: symmetry, triangle inequality, $d(x,x) = 0$, and $d(x,y) = 0 \implies x = y$. A pseudo-metric only satisfies the first three properties.
4. Note that the proof of Cortes et al. (2008) requires an application the of dominated convergence theorem, rather than using the Riesz representation theorem to show the existence of the mean embeddings μ_p and μ_q as we did in Lemma 3.

$\mathrm{MMD}[\mathcal{F},p,q]$ is zero. We now prove the converse. By the universality of $\mathcal{H}$, for any given $\varepsilon > 0$ and $f \in C(\mathcal{X})$ there exists a $g \in \mathcal{H}$ such that

$$\|f-g\|_\infty \leq \varepsilon.$$

We next make the expansion

$$|\mathbf{E}_x f(x) - \mathbf{E}_y(f(y))| \leq |\mathbf{E}_x f(x) - \mathbf{E}_x g(x)| + |\mathbf{E}_x g(x) - \mathbf{E}_y g(y)| + |\mathbf{E}_y g(y) - \mathbf{E}_y f(y)|.$$

The first and third terms satisfy

$$|\mathbf{E}_x f(x) - \mathbf{E}_x g(x)| \leq \mathbf{E}_x |f(x) - g(x)| \leq \varepsilon.$$

Next, write

$$\mathbf{E}_x g(x) - \mathbf{E}_y g(y) = \langle g, \mu_p - \mu_q \rangle_{\mathcal{H}} = 0,$$

since $\mathrm{MMD}[\mathcal{F},p,q] = 0$ implies $\mu_p = \mu_q$. Hence

$$|\mathbf{E}_x f(x) - \mathbf{E}_y(f(y))| \leq 2\varepsilon$$

for all $f \in C(\mathcal{X})$ and $\varepsilon > 0$, which implies $p = q$ by Lemma 1. ■

While our result establishes the mapping μ_p is injective for universal kernels on compact domains, this result can also be shown in more general cases. Fukumizu et al. (2008) introduce the notion of *characteristic kernels*, these being kernels for which the mean map is injective. Fukumizu et al. establish that Gaussian and Laplace kernels are characteristic on $\mathbb{R}^d$, and thus that the associated MMD is a metric on distributions for this domain. Sriperumbudur et al. (2008, 2010b) and Sriperumbudur et al. (2011a) further explore the properties of characteristic kernels, providing a simple condition to determine whether translation invariant kernels are characteristic, and investigating the relation between universal and characteristic kernels on non-compact domains.

Given we are in an RKHS, we may easily obtain the squared MMD, $\left\|\mu_p - \mu_q\right\|^2_{\mathcal{H}}$, in terms of kernel functions, and a corresponding unbiased finite sample estimate.

Lemma 6 *Given x and x' independent random variables with distribution p, and y and y' independent random variables with distribution q, the squared population* MMD *is*

$$\mathrm{MMD}^2[\mathcal{F},p,q] = \mathbf{E}_{x,x'}\left[k(x,x')\right] - 2\mathbf{E}_{x,y}\left[k(x,y)\right] + \mathbf{E}_{y,y'}\left[k(y,y')\right],$$

where x' is an independent copy of x with the same distribution, and y' is an independent copy of y. An unbiased *empirical estimate is a sum of two U-statistics and a sample average,*

$$\begin{aligned}\mathrm{MMD}_u^2[\mathcal{F},X,Y] &= \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n k(y_i,y_j) \\ &\quad - \frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n k(x_i,y_j). \end{aligned} \tag{3}$$

When $m = n$, a slightly simpler empirical estimate may be used. Let $Z := (z_1,\ldots,z_m)$ be m i.i.d. random variables, where $z := (x,y) \sim p \times q$ (i.e., x and y are independent). An unbiased estimate of MMD^2 *is*

$$\mathrm{MMD}_u^2[\mathcal{F},X,Y] = \frac{1}{(m)(m-1)}\sum_{i\neq j}^m h(z_i,z_j), \tag{4}$$

which is a one-sample U-statistic with

$$h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i).$$

Proof Starting from the expression for $\mathrm{MMD}^2[\mathcal{F}, p, q]$ in Lemma 4,

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{F}, p, q] &= \left\|\mu_p - \mu_q\right\|_{\mathcal{H}}^2 \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\ &= \mathbf{E}_{x,x'} \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \mathbf{E}_{y,y'} \langle \phi(y), \phi(y') \rangle_{\mathcal{H}} - 2\mathbf{E}_{x,y} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}},\end{aligned}$$

The proof is completed by applying $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x')$; the empirical estimates follow straightforwardly, by replacing the population expectations with their corresponding U-statistics and sample averages. This statistic is unbiased following Serfling (1980, Chapter 5). ■

Note that MMD_u^2 may be negative, since it is an unbiased estimator of $(\mathrm{MMD}[\mathcal{F}, p, q])^2$. The only terms missing to ensure nonnegativity, however, are $h(z_i, z_i)$, which were removed to remove spurious correlations between observations. Consequently we have the bound

$$\mathrm{MMD}_u^2 + \frac{1}{m(m-1)} \sum_{i=1}^{m} k(x_i, x_i) + k(y_i, y_i) - 2k(x_i, y_i) \geq 0.$$

Moreover, while the empirical statistic for $m = n$ is an unbiased estimate of MMD^2, it does not have minimum variance, since we ignore the cross-terms $k(x_i, y_i)$, of which there are $O(n)$. From (3), however, we see the minimum variance estimate is almost identical (Serfling, 1980, Section 5.1.4).

The biased statistic in (2) may also be easily computed following the above reasoning. Substituting the empirical estimates $\mu_X := \frac{1}{m}\sum_{i=1}^{m} \phi(x_i)$ and $\mu_Y := \frac{1}{n}\sum_{i=1}^{n} \phi(y_i)$ of the feature space means based on respective samples X and Y, we obtain

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) \right]^{\frac{1}{2}}. \tag{5}$$

Note that the U-statistics of (3) have been replaced by V-statistics. Intuitively we expect the empirical test statistic $\mathrm{MMD}[\mathcal{F}, X, Y]$, whether biased or unbiased, to be small if $p = q$, and large if the distributions are far apart. It costs $O((m+n)^2)$ time to compute both statistics.

2.3 Witness Function of the MMD for RKHSs

We define the witness function f^* to be the RKHS function attaining the supremum in (1), and its empirical estimate $\hat{f}^*$ to be the function attaining the supremum in (2). From the reasoning in Lemma 4, it is clear that

$$\begin{aligned} f^*(t) &\propto \langle \phi(t), \mu_p - \mu_q \rangle_{\mathcal{H}} &&= \mathbf{E}_x[k(x,t)] - \mathbf{E}_y[k(y,t)], \\ \hat{f}^*(t) &\propto \langle \phi(t), \mu_X - \mu_Y \rangle_{\mathcal{H}} &&= \tfrac{1}{m}\textstyle\sum_{i=1}^{m} k(x_i, t) - \tfrac{1}{n}\sum_{i=1}^{n} k(y_i, t). \end{aligned}$$

where we have defined $\mu_X = m^{-1}\sum_{i=1}^{m} \phi(x_i)$, and μ_Y by analogy. The result follows since the unit vector v maximizing $\langle v, x \rangle_{\mathcal{H}}$ in a Hilbert space is $v = x/\|x\|_{\mathcal{H}}$.

We illustrate the behavior of MMD in Figure 1 using a one-dimensional example. The data X and Y were generated from distributions p and q with equal means and variances, with p Gaussian

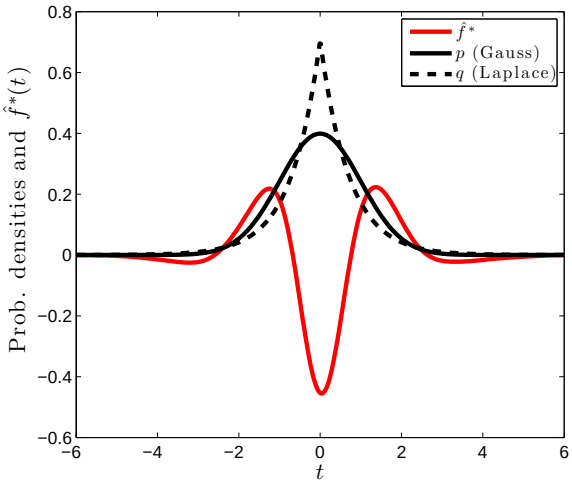

Figure 1: Illustration of the function maximizing the mean discrepancy in the case where a Gaussian is being compared with a Laplace distribution. Both distributions have zero mean and unit variance. The function $\hat{f}^*$ that witnesses the MMD has been scaled for plotting purposes, and was computed empirically on the basis of 2×10^4 samples, using a Gaussian kernel with $\sigma = 0.5$.

and q Laplacian. We chose $\mathcal{F}$ to be the unit ball in a Gaussian RKHS. The empirical estimate $\hat{f}^*$ of the function f^* that witnesses the MMD—in other words, the function maximizing the mean discrepancy in (1)—is smooth, negative where the Laplace density exceeds the Gaussian density (at the center and tails), and positive where the Gaussian density is larger. The magnitude of $\hat{f}^*$ is a direct reflection of the amount by which one density exceeds the other, insofar as the smoothness constraint permits it.

3. Background Material

We now present three background results. First, we introduce the terminology used in statistical hypothesis testing. Second, we demonstrate via an example that even for tests which have asymptotically no error, we cannot guarantee performance at any fixed sample size without making assumptions about the distributions. Third, we review some alternative statistics used in comparing distributions, and the associated two-sample tests (see also Section 7 for an overview of additional integral probability metrics).

3.1 Statistical Hypothesis Testing

Having described a metric on probability distributions (the MMD) based on distances between their Hilbert space embeddings, and empirical estimates (biased and unbiased) of this metric, we address the problem of determining whether the empirical MMD shows a *statistically significant* difference between distributions. To this end, we briefly describe the framework of statistical hypothesis testing as it applies in the present context, following Casella and Berger (2002, Chapter 8). Given i.i.d.

samples $X \sim p$ of size m and $Y \sim q$ of size n, the statistical test, $T(X,Y) : \mathcal{X}^m \times \mathcal{X}^n \mapsto \{0,1\}$ is used to distinguish between the null hypothesis $\mathcal{H}_0 : p = q$ and the alternative hypothesis $\mathcal{H}_A : p \neq q$. This is achieved by comparing the test statistic[5] $\mathrm{MMD}[\mathcal{F},X,Y]$ with a particular threshold: if the threshold is exceeded, then the test rejects the null hypothesis (bearing in mind that a zero population MMD indicates $p = q$). The acceptance region of the test is thus defined as the set of real numbers below the threshold. Since the test is based on finite samples, it is possible that an incorrect answer will be returned. A Type I error is made when $p = q$ is rejected based on the observed samples, despite the null hypothesis having generated the data. Conversely, a Type II error occurs when $p = q$ is accepted despite the underlying distributions being different. The *level* α of a test is an upper bound on the probability of a Type I error: this is a design parameter of the test which must be set in advance, and is used to determine the threshold to which we compare the test statistic (finding the test threshold for a given α is the topic of Sections 4 and 5). The *power* of a test against a particular member of the alternative class $\mathcal{H}_A$ (i.e., a specific (p,q) such that $p \neq q$) is the probability of wrongly accepting $p = q$ in this instance. A consistent test achieves a level α, and a Type II error of zero, in the large sample limit. We will see that the tests proposed in this paper are consistent.

3.2 A Negative Result

Even if a test is consistent, it is not possible to distinguish distributions with high probability at a given, *fixed* sample size (i.e., to provide guarantees on the Type II error), without prior assumptions as to the nature of the difference between p and q. This is true regardless of the two-sample test used. There are several ways to illustrate this, which each give insight into the kinds of differences that might be undetectable for a given number of samples. The following example[6] is one such illustration.

Example 1 *Assume we have a distribution p from which we have drawn m i.i.d. observations. We construct a distribution q by drawing m^2 i.i.d. observations from p, and defining a discrete distribution over these m^2 instances with probability m^{-2} each. It is easy to check that if we now draw m observations from q, there is at least a $\binom{m^2}{m}\frac{m!}{m^{2m}} > 1 - e^{-1} > 0.63$ probability that we thereby obtain an m sample from p. Hence no test will be able to distinguish samples from p and q in this case. We could make the probability of detection arbitrarily small by increasing the size of the sample from which we construct q.*

3.3 Previous Work

We next give a brief overview of some earlier approaches to the two sample problem for multivariate data. Since our later experimental comparison is with respect to certain of these methods, we give abbreviated algorithm names in italics where appropriate: these should be used as a key to the tables in Section 8.

5. This may be biased or unbiased.
6. This is a variation of a construction for independence tests, which was suggested in a private communication by John Langford.

3.3.1 L_2 DISTANCE BETWEEN PARZEN WINDOW ESTIMATES

The prior work closest to the current approach is the Parzen window-based statistic of Anderson et al. (1994). We begin with a short overview of the Parzen window estimate and its properties (Silverman, 1986), before proceeding to a comparison with the RKHS approach. We assume a distribution p on $\mathbb{R}^d$, which has an associated density function f_p. The Parzen window estimate of this density from an i.i.d. sample X of size m is

$$\hat{f}_p(x) = \frac{1}{m}\sum_{i=1}^{m} \kappa(x_i - x), \text{ where } \kappa \text{ satisfies } \int_{\mathcal{X}} \kappa(x)\,dx = 1 \text{ and } \kappa(x) \geq 0.$$

We may rescale κ according to $\frac{1}{h_m^d}\kappa\left(\frac{x}{h_m}\right)$ for a bandwidth parameter h_m. To simplify the discussion, we use a single bandwidth h_{m+n} for both $\hat{f}_p$ and $\hat{f}_q$. Assuming m/n is bounded away from zero and infinity, consistency of the Parzen window estimates for f_p and f_q requires

$$\lim_{m,n\to\infty} h_{m+n}^d = 0 \quad \text{and} \quad \lim_{m,n\to\infty} (m+n)h_{m+n}^d = \infty. \tag{6}$$

We now show the L_2 distance between Parzen windows density estimates is a special case of the biased MMD in Equation (5). Denote by $D_r(p,q) := \left\| f_p - f_q \right\|_r$ the L_r distance between the densities f_p and f_q corresponding to the distributions p and q, respectively. For $r = 1$ the distance $D_r(p,q)$ is known as the Lévy distance (Feller, 1971), and for $r = 2$ we encounter a distance measure derived from the Renyi entropy (Gokcay and Principe, 2002). Assume that $\hat{f}_p$ and $\hat{f}_q$ are given as kernel density estimates with kernel $\kappa(x - x')$, that is, $\hat{f}_p(x) = m^{-1}\sum_{i=1}^{m}\kappa(x_i - x)$ and $\hat{f}_q(y)$ is defined by analogy. In this case

$$\begin{aligned} D_2(\hat{f}_p, \hat{f}_q)^2 &= \int \left[\frac{1}{m}\sum_{i=1}^{m}\kappa(x_i - z) - \frac{1}{n}\sum_{i=1}^{n}\kappa(y_i - z)\right]^2 dz \\ &= \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i - x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i - y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i - y_j), \end{aligned}$$

where $k(x-y) = \int \kappa(x-z)\kappa(y-z)dz$. By its definition $k(x-y)$ is an RKHS kernel, as it is an inner product between $\kappa(x-z)$ and $\kappa(y-z)$ on the domain $\mathcal{X}$.

We now describe the asymptotic performance of a two-sample test using the statistic $D_2(\hat{f}_p, \hat{f}_q)^2$. We consider the power of the test under local departures from the null hypothesis. Anderson et al. (1994) define these to take the form

$$f_q = f_p + \delta g, \tag{7}$$

where $\delta \in \mathbb{R}$, and g is a fixed, bounded, integrable function chosen to ensure that f_q is a valid density for sufficiently small $|\delta|$. Anderson et al. consider two cases: the kernel bandwidth converging to zero with increasing sample size, ensuring consistency of the Parzen window estimates of f_p and f_q; and the case of a fixed bandwidth. In the former case, the minimum distance with which the test can discriminate f_p from f_q is[7] $\delta = (m+n)^{-1/2}h_{m+n}^{-d/2}$. In the latter case, this minimum distance is $\delta = (m+n)^{-1/2}$, under the assumption that the Fourier transform of the kernel κ does not vanish

7. Formally, define s_α as a threshold for the statistic $D_2\left(\hat{f}_p, \hat{f}_q\right)^2$, chosen to ensure the test has level α, and let $\delta = (m+n)^{-1/2}h_{m+n}^{-d/2}c$ for some fixed $c \neq 0$. When $m,n \to \infty$ such that m/n is bounded away from 0 and ∞, and

on an interval (Anderson et al., 1994, Section 2.4), which implies the kernel k is characteristic (Sriperumbudur et al., 2010b). The power of the L_2 test against local alternatives is greater when the kernel is held fixed, since for *any* rate of decrease of h_{m+n} with increasing sample size, δ will decrease more slowly than for a fixed kernel.

An RKHS-based approach generalizes the L_2 statistic in a number of important respects. First, we may employ a much larger class of characteristic kernels that cannot be written as inner products between Parzen windows: several examples are given by Steinwart (2001, Section 3) and Micchelli et al. (2006, Section 3) (these kernels are universal, hence characteristic). We may further generalize to kernels on structured objects such as strings and graphs (Schölkopf et al., 2004), as done in our experiments (Section 8). Second, even when the kernel may be written as an inner product of Parzen windows on $\mathbb{R}^d$, the D_2^2 statistic with fixed bandwidth no longer converges to an L_2 distance between probability density functions, hence it is more natural to define the statistic as an integral probability metric for a particular RKHS, as in Definition 2. Indeed, in our experiments, we obtain good performance in experimental settings where the dimensionality greatly exceeds the sample size, and density estimates would perform very poorly[8] (for instance the Gaussian toy example in Figure 5B, for which performance actually improves when the dimensionality increases; and the microarray data sets in Table 1). This suggests it is not necessary to solve the more difficult problem of density estimation in high dimensions to do two-sample testing.

Finally, the kernel approach leads us to establish consistency against a larger class of local alternatives to the null hypothesis than that considered by Anderson et al. In Theorem 13, we prove consistency against a class of alternatives encoded in terms of the mean embeddings of p and q, which applies to any domain on which RKHS kernels may be defined, and not only densities on $\mathbb{R}^d$. This more general approach also has interesting consequences for distributions on $\mathbb{R}^d$: for instance, a local departure from $\mathcal{H}_0$ occurs when p and q differ at increasing frequencies in their respective characteristic functions. This class of local alternatives cannot be expressed in the form δg for fixed g, as in (7). We discuss this issue further in Section 5.

3.3.2 MMD FOR MULTINOMIALS

Assume a finite domain $\mathcal{X} := \{1, \ldots, d\}$, and define the random variables x and y on $\mathcal{X}$ such that $p_i := P(x = i)$ and $q_j := P(y = j)$. We embed x into an RKHS $\mathcal{H}$ via the feature mapping $\phi(x) := e_x$, where e_s is the unit vector in $\mathbb{R}^d$ taking value 1 in dimension s, and zero in the remaining entries. The kernel is the usual inner product on $\mathbb{R}^d$. In this case,

$$\mathrm{MMD}^2[\mathcal{F}, p, q] = \|p - q\|_{\mathbb{R}^d}^2 = \sum_{i=1}^{d} (p_i - q_i)^2 . \tag{8}$$

Harchaoui et al. (2008, Section 1, long version) note that this L_2 statistic may not be the best choice for finite domains, citing a result of Lehmann and Romano (2005, Theorem 14.3.2) that Pearson's

assuming conditions (6), the limit

$$\pi(c) := \lim_{(m+n)\to\infty} \Pr_{\mathcal{H}_A}\left(D_2\left(\hat{f}_p, \hat{f}_q\right)^2 > s_\alpha\right)$$

is well-defined, and satisfies $\alpha < \pi(c) < 1$ for $0 < |c| < \infty$, and $\pi(c) \to 1$ as $c \to \infty$.

8. The L_2 error of a kernel density estimate converges as $O(n^{-4/(4+d)})$ when the optimal bandwidth is used (Wasserman, 2006, Section 6.5).

Chi-squared statistic is optimal for the problem of goodness of fit testing for multinomials.[9] It would be of interest to establish whether an analogous result holds for two-sample testing in a wider class of RKHS feature spaces.

3.3.3 FURTHER MULTIVARIATE TWO-SAMPLE TESTS

Biau and Gyorfi (2005) *(Biau)* use as their test statistic the L_1 distance between discretized estimates of the probabilities, where the partitioning is refined as the sample size increases. This space partitioning approach becomes difficult or impossible for high dimensional problems, since there are too few points per bin. For this reason, we use this test only for low-dimensional problems in our experiments.

A generalisation of the Wald-Wolfowitz runs test to the multivariate domain was proposed and analysed by Friedman and Rafsky (1979) and Henze and Penrose (1999) *(FR Wolf)*, and involves counting the number of edges in the minimum spanning tree over the aggregated data that connect points in X to points in Y. The resulting test relies on the asymptotic normality of the test statistic, and is not distribution-free under the null hypothesis for finite samples (the test threshold depends on p, as with our asymptotic test in Section 5; by contrast, our tests in Section 4 are distribution-free). The computational cost of this method using Kruskal's algorithm is $O((m+n)^2 \log(m+n))$, although more modern methods improve on the $\log(m+n)$ term: see Chazelle (2000) for details. Friedman and Rafsky (1979) claim that calculating the matrix of distances, which costs $O((m+n)^2)$, dominates their computing time; we return to this point in our experiments (Section 8). Two possible generalisations of the Kolmogorov-Smirnov test to the multivariate case were studied by Bickel (1969) and Friedman and Rafsky (1979). The approach of Friedman and Rafsky *(FR Smirnov)* in this case again requires a minimal spanning tree, and has a similar cost to their multivariate runs test.

A more recent multivariate test was introduced by Rosenbaum (2005). This entails computing the minimum distance non-bipartite matching over the aggregate data, and using the number of pairs containing a sample from both X and Y as a test statistic. The resulting statistic is distribution-free under the null hypothesis at finite sample sizes, in which respect it is superior to the Friedman-Rafsky test; on the other hand, it costs $O((m+n)^3)$ to compute. Another distribution-free test *(Hall)* was proposed by Hall and Tajvidi (2002): for each point from p, it requires computing the closest points in the aggregated data, and counting how many of these are from q (the procedure is repeated for each point from q with respect to points from p). As we shall see in our experimental comparisons, the test statistic is costly to compute; Hall and Tajvidi consider only tens of points in their experiments.

4. Tests Based on Uniform Convergence Bounds

In this section, we introduce two tests for the two-sample problem that have exact performance guarantees at finite sample sizes, based on uniform convergence bounds. The first, in Section 4.1, uses the McDiarmid (1989) bound on the biased MMD statistic, and the second, in Section 4.2, uses a Hoeffding (1963) bound for the unbiased statistic.

9. A goodness of fit test determines whether a sample from p is drawn from a *known* target multinomial q. Pearson's Chi-squared statistic weights each term in the sum (8) by its corresponding q_i^{-1}.

4.1 Bound on the Biased Statistic and Test

We establish two properties of the MMD, from which we derive a hypothesis test. First, we show that regardless of whether or not $p = q$, the empirical MMD converges in probability at rate $O((m+n)^{-\frac{1}{2}})$ to its population value. This shows the consistency of statistical tests based on the MMD. Second, we give probabilistic bounds for large deviations of the empirical MMD in the case $p = q$. These bounds lead directly to a threshold for our first hypothesis test. We begin by establishing the convergence of $\mathrm{MMD}_b[\mathcal{F}, X, Y]$ to $\mathrm{MMD}[\mathcal{F}, p, q]$. The following theorem is proved in A.2.

Theorem 7 *Let p, q, X, Y be defined as in Problem 1, and assume $0 \leq k(x,y) \leq K$. Then*

$$\Pr_{X,Y}\left\{|\mathrm{MMD}_b[\mathcal{F}, X, Y] - \mathrm{MMD}[\mathcal{F}, p, q]| > 2\left((K/m)^{\frac{1}{2}} + (K/n)^{\frac{1}{2}}\right) + \varepsilon\right\} \leq 2\exp\left(\frac{-\varepsilon^2 mn}{2K(m+n)}\right),$$

where $\Pr_{X,Y}$ denotes the probability over the m-sample X and n-sample Y.

Our next goal is to refine this result in a way that allows us to define a test threshold under the null hypothesis $p = q$. Under this circumstance, the constants in the exponent are slightly improved. The following theorem is proved in Appendix A.3.

Theorem 8 *Under the conditions of Theorem 7 where additionally $p = q$ and $m = n$,*

$$\mathrm{MMD}_b[\mathcal{F}, X, Y] \leq \underbrace{m^{-\frac{1}{2}}\sqrt{2\mathbf{E}_{x,x'}\left[k(x,x) - k(x,x')\right]}}_{B_1(\mathcal{F},p)} + \varepsilon \leq \underbrace{(2K/m)^{1/2}}_{B_2(\mathcal{F},p)} + \varepsilon,$$

both with probability at least $1 - \exp\left(-\frac{\varepsilon^2 m}{4K}\right)$.

In this theorem, we illustrate two possible bounds $B_1(\mathcal{F}, p)$ and $B_2(\mathcal{F}, p)$ on the bias in the empirical estimate (5). The first inequality is interesting inasmuch as it provides a link between the bias bound $B_1(\mathcal{F}, p)$ and kernel size (for instance, if we were to use a Gaussian kernel with large σ, then $k(x,x)$ and $k(x,x')$ would likely be close, and the bias small). In the context of testing, however, we would need to provide an additional bound to show convergence of an empirical estimate of $B_1(\mathcal{F}, p)$ to its population equivalent. Thus, in the following test for $p = q$ based on Theorem 8, we use $B_2(\mathcal{F}, p)$ to bound the bias.[10]

Corollary 9 *A hypothesis test of level α for the null hypothesis $p = q$, that is, for $\mathrm{MMD}[\mathcal{F}, p, q] = 0$, has the acceptance region $\mathrm{MMD}_b[\mathcal{F}, X, Y] < \sqrt{2K/m}\left(1 + \sqrt{2\log\alpha^{-1}}\right)$.*

We emphasize that this test is distribution-free: the test threshold does not depend on the particular distribution that generated the sample. Theorem 7 guarantees the consistency of the test against fixed alternatives, and that the Type II error probability decreases to zero at rate $O\left(m^{-1/2}\right)$, assuming $m = n$. To put this convergence rate in perspective, consider a test of whether two normal distributions have equal means, given they have unknown but equal variance (Casella and Berger, 2002, Exercise 8.41). In this case, the test statistic has a Student-t distribution with $n + m - 2$ degrees of freedom, and its Type II error probability converges at the same rate as our test.

It is worth noting that bounds may be obtained for the deviation between population mean embeddings μ_p and the empirical embeddings μ_X in a completely analogous fashion. The proof

10. Note that we use a tighter bias bound than Gretton et al. (2007a).

requires symmetrization by means of a *ghost sample*, that is, a second set of observations drawn from the same distribution. While not the focus of the present paper, such bounds can be used to perform inference based on moment matching (Altun and Smola, 2006; Dudík and Schapire, 2006; Dudík et al., 2004).

4.2 Bound on the Unbiased Statistic and Test

The previous bounds are of interest since the proof strategy can be used for general function classes with well behaved Rademacher averages (see Sriperumbudur et al., 2010a). When $\mathcal{F}$ is the unit ball in an RKHS, however, we may very easily define a test via a convergence bound on the unbiased statistic MMD_u^2 in Lemma 4. We base our test on the following theorem, which is a straightforward application of the large deviation bound on U-statistics of Hoeffding (1963, p. 25).

Theorem 10 *Assume* $0 \leq k(x_i, x_j) \leq K$, *from which it follows* $-2K \leq h(z_i, z_j) \leq 2K$. *Then*

$$\Pr_{X,Y}\left\{\mathrm{MMD}_u^2(\mathcal{F}, X, Y) - \mathrm{MMD}^2(\mathcal{F}, p, q) > t\right\} \leq \exp\left(\frac{-t^2 m_2}{8K^2}\right)$$

where $m_2 := \lfloor m/2 \rfloor$ *(the same bound applies for deviations of* $-t$ *and below).*

A consistent statistical test for $p = q$ using MMD_u^2 is then obtained.

Corollary 11 *A hypothesis test of level* α *for the null hypothesis* $p = q$ *has the acceptance region* $\mathrm{MMD}_u^2 < (4K/\sqrt{m})\sqrt{\log(\alpha^{-1})}$.

This test is distribution-free. We now compare the thresholds of the above test with that in Corollary 9. We note first that the threshold for the biased statistic applies to an estimate of MMD, whereas that for the unbiased statistic is for an estimate of MMD^2. Squaring the former threshold to make the two quantities comparable, the squared threshold in Corollary 9 decreases as m^{-1}, whereas the threshold in Corollary 11 decreases as $m^{-1/2}$. Thus for sufficiently large[11] m, the McDiarmid-based threshold will be lower (and the associated test statistic is in any case biased upwards), and its Type II error will be better for a given Type I bound. This is confirmed in our Section 8 experiments. Note, however, that the rate of convergence of the squared, biased MMD estimate to its population value remains at $1/\sqrt{m}$ (bearing in mind we take the square of a biased estimate, where the bias term decays as $1/\sqrt{m}$).

Finally, we note that the bounds we obtained in this section and the last are rather conservative for a number of reasons: first, they do not take the actual distributions into account. In fact, they are finite sample size, distribution-free bounds that hold even in the worst case scenario. The bounds could be tightened using localization, moments of the distribution, etc.: see, for example, Bousquet et al. (2005) and de la Peña and Giné (1999). Any such improvements could be plugged straight into Theorem 19. Second, in computing bounds rather than trying to characterize the distribution of $\mathrm{MMD}[\mathcal{F}, X, Y]$ explicitly, we force our test to be conservative by design. In the following we aim for an exact characterization of the asymptotic distribution of $\mathrm{MMD}[\mathcal{F}, X, Y]$ instead of a bound. While this will not satisfy the uniform convergence requirements, it leads to superior tests in practice.

11. In the case of $\alpha = 0.05$, this is $m \geq 12$.

5. Test Based on the Asymptotic Distribution of the Unbiased Statistic

We propose a third test, which is based on the asymptotic distribution of the unbiased estimate of MMD^2 in Lemma 6. This test uses the asymptotic distribution of MMD_u^2 under $\mathcal{H}_0$, which follows from results of Anderson et al. (1994, Appendix) and Serfling (1980, Section 5.5.2): see Appendix B.1 for the proof.

Theorem 12 *Let $\tilde{k}(x_i, x_j)$ be the kernel between feature space mappings from which the mean embedding of p has been subtracted,*

$$\begin{aligned}\tilde{k}(x_i, x_j) &:= \left\langle \phi(x_i) - \mu_p, \phi(x_j) - \mu_p \right\rangle_{\mathcal{H}} \\ &= k(x_i, x_j) - \mathbf{E}_x k(x_i, x) - \mathbf{E}_x k(x, x_j) + \mathbf{E}_{x,x'} k(x, x'),\end{aligned} \tag{9}$$

where x' is an independent copy of x drawn from p. Assume $\tilde{k} \in L_2(\mathcal{X} \times \mathcal{X}, p \times p)$ (i.e., the centred kernel is square integrable, which is true for all p when the kernel is bounded), and that for $t = m + n$, $\lim_{m,n\to\infty} m/t \to \rho_x$ and $\lim_{m,n\to\infty} n/t \to \rho_y := (1 - \rho_x)$ for fixed $0 < \rho_x < 1$. Then under $\mathcal{H}_0$, MMD_u^2 converges in distribution according to

$$t\mathrm{MMD}_u^2[\mathcal{F}, X, Y] \underset{D}{\to} \sum_{l=1}^{\infty} \lambda_l \left[(\rho_x^{-1/2} a_l - \rho_y^{-1/2} b_l)^2 - (\rho_x \rho_y)^{-1} \right], \tag{10}$$

where $a_l \sim \mathcal{N}(0,1)$ and $b_l \sim \mathcal{N}(0,1)$ are infinite sequences of independent Gaussian random variables, and the λ_i are eigenvalues of

$$\int_{\mathcal{X}} \tilde{k}(x, x') \psi_i(x) dp(x) = \lambda_i \psi_i(x').$$

We illustrate the MMD density under both the null and alternative hypotheses by approximating it empirically for $p = q$ and $p \neq q$. Results are plotted in Figure 2.

Our goal is to determine whether the empirical test statistic MMD_u^2 is so large as to be outside the $1 - \alpha$ quantile of the null distribution in (10), which gives a level α test. Consistency of this test against local departures from the null hypothesis is provided by the following theorem, proved in Appendix B.2.

Theorem 13 *Define ρ_x, ρ_y, and t as in Theorem 12, and write $\mu_q = \mu_p + g_t$, where $g_t \in \mathcal{H}$ is chosen such that $\mu_p + g_t$ remains a valid mean embedding, and $\|g_t\|_{\mathcal{H}}$ is made to approach zero as $t \to \infty$ to describe local departures from the null hypothesis. Then $\|g_t\|_{\mathcal{H}} = ct^{-1/2}$ is the minimum distance between μ_p and μ_q distinguishable by the test.*

An example of a local departure from the null hypothesis is described earlier in the discussion of the L_2 distance between Parzen window estimates (Section 3.3.1). The class of local alternatives considered in Theorem 13 is more general, however: for instance, Sriperumbudur et al. (2010b, Section 4) and Harchaoui et al. (2008, Section 5, long version) give examples of classes of perturbations g_t with decreasing RKHS norm. These perturbations have the property that p differs from q at increasing frequencies, rather than simply with decreasing amplitude.

One way to estimate the $1 - \alpha$ quantile of the null distribution is using the bootstrap on the aggregated data, following Arcones and Giné (1992). Alternatively, we may approximate the null

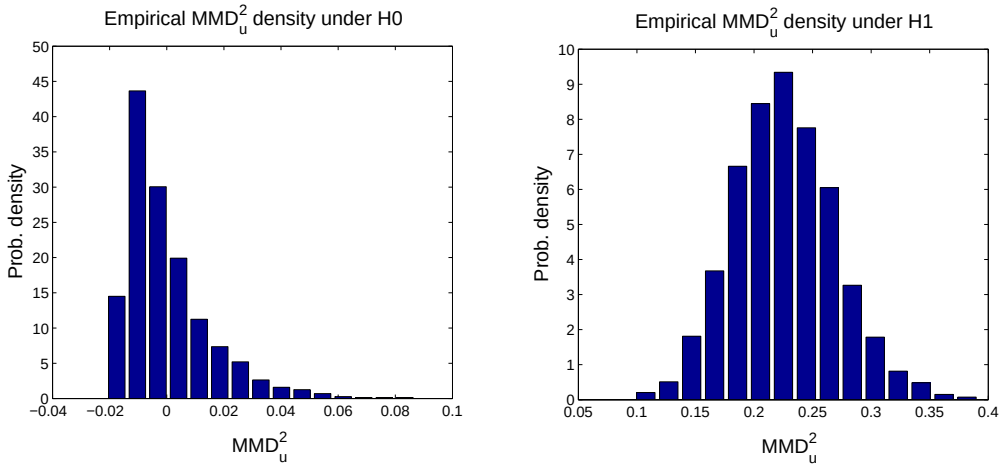

Figure 2: **Left:** Empirical distribution of the MMD under $\mathcal{H}_0$, with p and q both Gaussians with unit standard deviation, using 50 samples from each. **Right:** Empirical distribution of the MMD under $\mathcal{H}_A$, with p a Laplace distribution with unit standard deviation, and q a Laplace distribution with standard deviation $3\sqrt{2}$, using 100 samples from each. In both cases, the histograms were obtained by computing 2000 independent instances of the MMD.

distribution by fitting Pearson curves to its first four moments (Johnson et al., 1994, Section 18.8). Taking advantage of the degeneracy of the U-statistic, we obtain for $m = n$

$$\begin{aligned}
\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^2\right) &= \frac{2}{m(m-1)}\mathbf{E}_{z,z'}\left[h^2(z,z')\right] \text{ and} \\
\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^3\right) &= \frac{8(m-2)}{m^2(m-1)^2}\mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}\left(h(z,z'')h(z',z'')\right)\right] + O(m^{-4})
\end{aligned} \tag{11}$$

(see Appendix B.3), where $h(z,z')$ is defined in Lemma 6, $z = (x,y) \sim p \times q$ where x and y are independent, and z', z'' are independent copies of z. The fourth moment $\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^4\right)$ is not computed, since it is both very small, $O(m^{-4})$, and expensive to calculate, $O(m^4)$. Instead, we replace the kurtosis[12] with a lower bound due to Wilkins (1944), $\mathrm{kurt}\left(\mathrm{MMD}_u^2\right) \geq \left(\mathrm{skew}\left(\mathrm{MMD}_u^2\right)\right)^2 + 1$. In Figure 3, we illustrate the Pearson curve fit to the null distribution: the fit is good in the upper quantiles of the distribution, where the test threshold is computed. Finally, we note that two alternative empirical estimates of the null distribution have more recently been proposed by Gretton et al. (2009): a consistent estimate, based on an empirical computation of the eigenvalues λ_l in (10); and an alternative Gamma approximation to the null distribution, which has a smaller computational cost but is generally less accurate. Further detail and experimental comparisons are given by Gretton et al.

12. The kurtosis is defined in terms of the fourth and second moments as $\mathrm{kurt}\left(\mathrm{MMD}_u^2\right) = \frac{\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^4\right)}{\left[\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^2\right)\right]^2} - 3$.

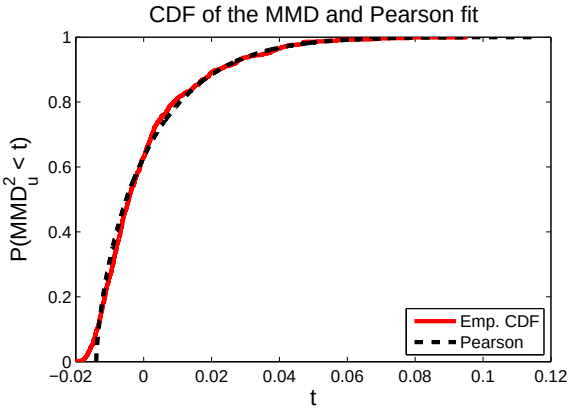

Figure 3: Illustration of the empirical CDF of the MMD and a Pearson curve fit. Both p and q were Gaussian with zero mean and unit variance, and 50 samples were drawn from each. The empirical CDF was computed on the basis of 1000 randomly generated MMD values. To ensure the quality of fit was determined only by the accuracy of the Pearson approximation, the moments used for the Pearson curves were also computed on the basis of these 1000 samples. The MMD used a Gaussian kernel with $\sigma = 0.5$.

6. A Linear Time Statistic and Test

The MMD-based tests are already more efficient than the $O(m^2 \log m)$ and $O(m^3)$ tests described in Section 3.3.3 (assuming $m = n$ for conciseness). It is still desirable, however, to obtain $O(m)$ tests which do not sacrifice too much statistical power. Moreover, we would like to obtain tests which have $O(1)$ storage requirements for computing the test statistic, in order to apply the test to data streams. We now describe how to achieve this by computing the test statistic using a subsampling of the terms in the sum. The empirical estimate in this case is obtained by drawing pairs from X and Y respectively *without* replacement.

Lemma 14 *Define $m_2 := \lfloor m/2 \rfloor$, assume $m = n$, and define $h(z_1, z_2)$ as in Lemma 6. The estimator*

$$\mathrm{MMD}_l^2[\mathcal{F}, X, Y] := \frac{1}{m_2} \sum_{i=1}^{m_2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i}))$$

can be computed in linear time, and is an unbiased estimate of $\mathrm{MMD}^2[\mathcal{F}, p, q]$.

While it is expected that MMD_l^2 has higher variance than MMD_u^2 (as we will see explicitly later), it is computationally much more appealing. In particular, the statistic can be used in stream computations with need for only $O(1)$ memory, whereas MMD_u^2 requires $O(m)$ storage and $O(m^2)$ time to compute the kernel h on all interacting pairs.

Since MMD_l^2 is just the average over a set of random variables, Hoeffding's bound and the central limit theorem readily allow us to provide both uniform convergence and asymptotic statements with little effort. The first follows directly from Hoeffding (1963, Theorem 2).

Theorem 15 *Assume* $0 \leq k(x_i, x_j) \leq K$. *Then*

$$\Pr_{X,Y}\left\{\mathrm{MMD}_l^2(\mathcal{F}, X, Y) - \mathrm{MMD}^2(\mathcal{F}, p, q) > t\right\} \leq \exp\left(\frac{-t^2 m_2}{8K^2}\right)$$

where $m_2 := \lfloor m/2 \rfloor$ *(the same bound applies for deviations of* $-t$ *and below).*

Note that the bound of Theorem 10 is identical to that of Theorem 15, which shows the former is rather loose. Next we invoke the central limit theorem (e.g., Serfling, 1980, Section 1.9).

Corollary 16 *Assume* $0 < \mathbf{E}\left(h^2\right) < \infty$. *Then* MMD_l^2 *converges in distribution to a Gaussian according to*

$$m^{\frac{1}{2}}\left(\mathrm{MMD}_l^2 - \mathrm{MMD}^2\left[\mathcal{F}, p, q\right]\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_l^2\right),$$

where $\sigma_l^2 = 2\left[\mathbf{E}_{z,z'} h^2(z,z') - \left[\mathbf{E}_{z,z'} h(z,z')\right]^2\right]$, *where we use the shorthand* $\mathbf{E}_{z,z'} := \mathbf{E}_{z,z' \sim p \times q}$.

The factor of 2 arises since we are averaging over only $\lfloor m/2 \rfloor$ observations. It is instructive to compare this asymptotic distribution with that of the quadratic time statistic MMD_u^2 under $\mathcal{H}_A$, when $m = n$. In this case, MMD_u^2 converges in distribution to a Gaussian according to

$$m^{\frac{1}{2}}\left(\mathrm{MMD}_u^2 - \mathrm{MMD}^2\left[\mathcal{F}, p, q\right]\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_u^2\right),$$

where $\sigma_u^2 = 4\left(\mathbf{E}_z\left[\left(\mathbf{E}_{z'} h(z,z')\right)^2\right] - \left[\mathbf{E}_{z,z'}\left(h(z,z')\right)\right]^2\right)$ (Serfling, 1980, Section 5.5). Thus for MMD_u^2, the asymptotic variance is (up to scaling) the variance of $\mathbf{E}_{z'}[h(z,z')]$, whereas for MMD_l^2 it is $\mathrm{Var}_{z,z'}[h(z,z')]$.

We end by noting another potential approach to reducing the cost of computing an empirical MMD estimate, by using a low rank approximation to the Gram matrix (Fine and Scheinberg, 2001; Williams and Seeger, 2001; Smola and Schölkopf, 2000). An incremental computation of the MMD based on such a low rank approximation would require $O(md)$ storage and $O(md)$ computation (where d is the rank of the approximate Gram matrix which is used to factorize *both* matrices) rather than $O(m)$ storage and $O(m^2)$ operations. That said, it remains to be determined what effect this approximation would have on the distribution of the test statistic under $\mathcal{H}_0$, and hence on the test threshold.

7. Related Metrics and Learning Problems

The present section discusses a number of topics related to the maximum mean discrepancy, including metrics on probability distributions using non-RKHS function classes (Sections 7.1 and 7.2), the relation with set kernels and kernels on probability measures (Section 7.3), an extension to kernel measures of independence (Section 7.4), a two-sample statistic using a distribution over witness functions (Section 7.5), and a connection to outlier detection (Section 7.6).

7.1 The MMD in Other Function Classes

The definition of the maximum mean discrepancy is by no means limited to RKHS. In fact, any function class $\mathcal{F}$ that comes with uniform convergence guarantees and is sufficiently rich will enjoy the above properties. Below, we consider the case where the scaled functions in $\mathcal{F}$ are dense in $C(\mathcal{X})$ (which is useful for instance when the functions in $\mathcal{F}$ are norm constrained).

Definition 17 *Let $\mathcal{F}$ be a subset of some vector space. The star $S[\mathcal{F}]$ of a set $\mathcal{F}$ is*

$$S[\mathcal{F}] := \{\alpha f | f \in \mathcal{F} \text{ and } \alpha \in [0,\infty)\}$$

Theorem 18 *Denote by $\mathcal{F}$ the subset of some vector space of functions from $\mathcal{X}$ to $\mathbb{R}$ for which $S[\mathcal{F}] \cap C(\mathcal{X})$ is dense in $C(\mathcal{X})$ with respect to the $L_\infty(\mathcal{X})$ norm. Then $\mathrm{MMD}[\mathcal{F},p,q] = 0$ if and only if $p = q$, and $\mathrm{MMD}[\mathcal{F},p,q]$ is a metric on the space of probability distributions. Whenever the star of $\mathcal{F}$ is* not *dense, the* MMD *defines a pseudo-metric space.*

Proof It is clear that $p = q$ implies $\mathrm{MMD}[\mathcal{F},p,q] = 0$. The proof of the converse is very similar to that of Theorem 5. Define $\mathcal{H} := S(\mathcal{F}) \cap C(\mathcal{X})$. Since by assumption $\mathcal{H}$ is dense in $C(\mathcal{X})$, there exists an $h^* \in \mathcal{H}$ satisfying $\|h^* - f\|_\infty < \varepsilon$ for all $f \in C(\mathcal{X})$. Write $h^* := \alpha^* g^*$, where $g^* \in \mathcal{F}$. By assumption, $\mathbf{E}_x g^* - \mathbf{E}_y g^* = 0$. Thus we have the bound

$$\begin{aligned} |\mathbf{E}_x f(x) - \mathbf{E}_y(f(y))| &\leq |\mathbf{E}_x f(x) - \mathbf{E}_x h^*(x)| + \alpha^* |\mathbf{E}_x g^*(x) - \mathbf{E}_y g^*(y)| + |\mathbf{E}_y h^*(y) - \mathbf{E}_y f(y)| \\ &\leq 2\varepsilon \end{aligned}$$

for all $f \in C(\mathcal{X})$ and $\varepsilon > 0$, which implies $p = q$ by Lemma 1.

To show $\mathrm{MMD}[\mathcal{F},p,q]$ is a metric, it remains to prove the triangle inequality. We have

$$\begin{aligned} \sup_{f\in\mathcal{F}} \left|E_p f - E_q f\right| + \sup_{g\in\mathcal{F}} \left|E_q g - E_r g\right| &\geq \sup_{f\in\mathcal{F}} \left[\left|E_p f - E_q f\right| + \left|E_q f - E_r\right|\right] \\ &\geq \sup_{f\in\mathcal{F}} \left|E_p f - E_r f\right|. \end{aligned}$$

■

Note that any uniform convergence statements in terms of $\mathcal{F}$ allow us immediately to characterize an estimator of $\mathrm{MMD}(\mathcal{F},p,q)$ explicitly. The following result shows how (this reasoning is also the basis for the proofs in Section 4, although here we do not restrict ourselves to an RKHS).

Theorem 19 *Let $\delta \in (0,1)$ be a confidence level and assume that for some $\varepsilon(\delta,m,\mathcal{F})$ the following holds for samples $\{x_1,\ldots,x_m\}$ drawn from p:*

$$\Pr_X \left\{ \sup_{f\in\mathcal{F}} \left| \mathbf{E}_x[f] - \frac{1}{m}\sum_{i=1}^m f(x_i) \right| > \varepsilon(\delta,m,\mathcal{F}) \right\} \leq \delta.$$

In this case we have that,

$$\Pr_{X,Y}\left\{|\mathrm{MMD}[\mathcal{F},p,q] - \mathrm{MMD}_b[\mathcal{F},X,Y]| > 2\varepsilon(\delta/2,m,\mathcal{F})\right\} \leq \delta,$$

where $\mathrm{MMD}_b[\mathcal{F},X,Y]$ is taken from Definition 2.

Proof The proof works simply by using convexity and suprema as follows:

$$\begin{aligned} &|\mathrm{MMD}[\mathcal{F},p,q] - \mathrm{MMD}_b[\mathcal{F},X,Y]| \\ =& \left| \sup_{f\in\mathcal{F}} |\mathbf{E}_x[f] - \mathbf{E}_y[f]| - \sup_{f\in\mathcal{F}} \left| \frac{1}{m}\sum_{i=1}^m f(x_i) - \frac{1}{n}\sum_{i=1}^n f(y_i) \right| \right| \\ \leq& \sup_{f\in\mathcal{F}} \left| \mathbf{E}_x[f] - \mathbf{E}_y[f] - \frac{1}{m}\sum_{i=1}^m f(x_i) + \frac{1}{n}\sum_{i=1}^n f(y_i) \right| \\ \leq& \sup_{f\in\mathcal{F}} \left| \mathbf{E}_x[f] - \frac{1}{m}\sum_{i=1}^m f(x_i) \right| + \sup_{f\in\mathcal{F}} \left| \mathbf{E}_y[f] - \frac{1}{n}\sum_{i=1}^n f(y_i) \right|. \end{aligned}$$

Bounding each of the two terms via a uniform convergence bound proves the claim. ■

This shows that $\mathrm{MMD}_b[\mathcal{F},X,Y]$ can be used to estimate $\mathrm{MMD}[\mathcal{F},p,q]$, and that the quantity is asymptotically unbiased.

Remark 20 (Reduction to Binary Classification) *As noted by Friedman (2003), any classifier which maps a set of observations $\{z_i,l_i\}$ with $z_i \in \mathcal{X}$ on some domain $\mathcal{X}$ and labels $l_i \in \{\pm 1\}$, for which uniform convergence bounds exist on the convergence of the empirical loss to the expected loss, can be used to obtain a similarity measure on distributions—simply assign $l_i = 1$ if $z_i \in X$ and $l_i = -1$ for $z_i \in Y$ and find a classifier which is able to separate the two sets. In this case maximization of $\mathbf{E}_x[f] - \mathbf{E}_y[f]$ is achieved by ensuring that as many $z \sim p(z)$ as possible correspond to $f(z) = 1$, whereas for as many $z \sim q(z)$ as possible we have $f(z) = -1$. Consequently neural networks, decision trees, boosted classifiers and other objects for which uniform convergence bounds can be obtained can be used for the purpose of distribution comparison. Metrics and divergences on distributions can also be defined explicitly starting from classifiers. For instance, Sriperumbudur et al. (2009, Section 2) show the* MMD *minimizes the expected risk of a classifier with linear loss on the samples X and Y, and Ben-David et al. (2007, Section 4) use the error of a hyperplane classifier to approximate the $\mathcal{A}$-distance between distributions (Kifer et al., 2004). Reid and Williamson (2011) provide further discussion and examples.*

7.2 Examples of Non-RKHS Function Classes

Other function spaces $\mathcal{F}$ inspired by the statistics literature can also be considered in defining the MMD. Indeed, Lemma 1 defines an MMD with $\mathcal{F}$ the space of bounded continuous real-valued functions, which is a Banach space with the supremum norm (Dudley, 2002, p. 158). We now describe two further metrics on the space of probability distributions, namely the Kolmogorov-Smirnov and Earth Mover's distances, and their associated function classes.

7.2.1 KOLMOGOROV-SMIRNOV STATISTIC

The Kolmogorov-Smirnov (K-S) test is probably one of the most famous two-sample tests in statistics. It works for random variables $x \in \mathbb{R}$ (or any other set for which we can establish a total order). Denote by $F_p(x)$ the cumulative distribution function of p and let $F_X(x)$ be its empirical counterpart,

$$F_p(z) := \Pr\{x \leq z \text{ for } x \sim p\} \text{ and } F_X(z) := \frac{1}{|X|}\sum_{i=1}^{m} 1_{z \leq x_i}.$$

It is clear that F_p captures the properties of p. The Kolmogorov metric is simply the L_∞ distance $\|F_X - F_Y\|_\infty$ for two sets of observations X and Y. Smirnov (1939) showed that for $p = q$ the limiting distribution of the empirical cumulative distribution functions satisfies

$$\lim_{m,n\to\infty} \Pr_{X,Y}\left\{\left[\tfrac{mn}{m+n}\right]^{\frac{1}{2}} \|F_X - F_Y\|_\infty > x\right\} = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \text{ for } x \geq 0, \tag{12}$$

which is distribution independent. This allows for an efficient characterization of the distribution under the null hypothesis $\mathcal{H}_0$. Efficient numerical approximations to (12) can be found in numerical analysis handbooks (Press et al., 1994). The distribution under the alternative $p \neq q$, however, is unknown.

The Kolmogorov metric is, in fact, a special instance of $\mathrm{MMD}[\mathcal{F}, p, q]$ for a certain Banach space (Müller, 1997, Theorem 5.2).

Proposition 21 *Let $\mathcal{F}$ be the class of functions $\mathcal{X} \to \mathbb{R}$ of bounded variation*[13] *1. Then* $\mathrm{MMD}[\mathcal{F}, p, q] = \left\|F_p - F_q\right\|_\infty$.

7.2.2 EARTH-MOVER DISTANCES

Another class of distance measures on distributions that may be written as maximum mean discrepancies are the Earth-Mover distances. We assume $(\mathcal{X}, \rho)$ is a separable metric space, and define $\mathcal{P}_1(\mathcal{X})$ to be the space of probability measures on $\mathcal{X}$ for which $\int \rho(x,z)dp(z) < \infty$ for all $p \in \mathcal{P}_1(\mathcal{X})$ and $x \in \mathcal{X}$ (these are the probability measures for which $\mathbf{E}_x|x| < \infty$ when $\mathcal{X} = \mathbb{R}$). We then have the following definition (Dudley, 2002, p. 420).

Definition 22 (Monge-Wasserstein metric) *Let $p \in \mathcal{P}_1(\mathcal{X})$ and $q \in \mathcal{P}_1(\mathcal{X})$. The Monge-Wasserstein distance is defined as*

$$W(p,q) := \inf_{\mu \in M(p,q)} \int \rho(x,y)d\mu(x,y),$$

where $M(p,q)$ is the set of joint distributions on $\mathcal{X} \times \mathcal{X}$ with marginals p and q.

We may interpret this as the cost (as represented by the metric $\rho(x,y)$) of transferring mass distributed according to p to a distribution in accordance with q, where μ is the movement schedule. In general, a large variety of costs of moving mass from x to y can be used, such as psycho-optical similarity measures in image retrieval (Rubner et al., 2000). The following theorem provides the link with the MMD (Dudley, 2002, Theorem 11.8.2).

Theorem 23 (Kantorovich-Rubinstein) *Let $p \in \mathcal{P}_1(\mathcal{X})$ and $q \in \mathcal{P}_1(\mathcal{X})$, where $\mathcal{X}$ is separable. Then a metric on $\mathcal{P}_1(S)$ is defined as*

$$W(p,q) = \|p-q\|_L^* = \sup_{\|f\|_L \le 1} \left| \int f\, d(p-q) \right|,$$

where

$$\|f\|_L := \sup_{x \neq y \in \mathcal{X}} \frac{|f(x) - f(y)|}{\rho(x,y)}$$

is the Lipschitz seminorm[14] *for real valued f on $\mathcal{X}$.*

A simple example of this theorem is as follows (Dudley, 2002, Exercise 1, p. 425).

Example 2 *Let $\mathcal{X} = \mathbb{R}$ with associated $\rho(x,y) = |x-y|$. Then given f such that $\|f\|_L \le 1$, we use integration by parts to obtain*

$$\left| \int f\, d(p-q) \right| = \left| \int (F_p - F_q)(x) f'(x) dx \right| \le \int \left| (F_p - F_q) \right| (x) dx,$$

13. A function f defined on $[a,b]$ is of bounded variation C if the total variation is bounded by C, that is, the supremum over all sums

$$\sum_{1 \le i \le n} |f(x_i) - f(x_{i-1})|,$$

where $a \le x_0 \le \ldots \le x_n \le b$ (Dudley, 2002, p. 184).

14. A seminorm satisfies the requirements of a norm besides $\|x\| = 0$ only for $x = 0$ (Dudley, 2002, p. 156).

where the maximum is attained for the function g with derivative $g' = 2\mathbf{1}_{F_p > F_q} - 1$ (and for which $\|g\|_L = 1$). We recover the L_1 distance between distribution functions,

$$W(P,Q) = \int \left|(F_p - F_q)\right|(x)dx.$$

One may further generalize Theorem 23 to the set of all laws $\mathcal{P}(\mathcal{X})$ on arbitrary metric spaces $\mathcal{X}$ (Dudley, 2002, Proposition 11.3.2).

Definition 24 (Bounded Lipschitz metric) *Let p and q be laws on a metric space $\mathcal{X}$. Then*

$$\beta(p,q) := \sup_{\|f\|_{BL} \leq 1} \left| \int f \, d(p-q) \right|$$

is a metric on $\mathcal{P}(\mathcal{X})$, where f belongs to the space of bounded Lipschitz functions with norm

$$\|f\|_{BL} := \|f\|_L + \|f\|_\infty .$$

Empirical estimates of the Monge-Wasserstein and Bounded Lipschitz metrics on $\mathbb{R}^d$ are provided by Sriperumbudur et al. (2010a).

7.3 Set Kernels and Kernels Between Probability Measures

Gärtner et al. (2002) propose kernels for Multi-Instance Classification (MIC) which deal with sets of observations. The purpose of MIC is to find estimators which are able to infer that if some elements in a set satisfy a certain property, then the set of observations also has this property. For instance, a dish of mushrooms is poisonous if it contains any poisonous mushrooms. Likewise a keyring will open a door if it contains a suitable key. One is only given the ensemble, however, rather than information about which instance of the set satisfies the property.

The solution proposed by Gärtner et al. (2002) is to map the ensembles $X_i := \{x_{i1}, \ldots, x_{im_i}\}$, where i is the ensemble index and m_i the number of elements in the ith ensemble, jointly into feature space via

$$\phi(X_i) := \frac{1}{m_i} \sum_{j=1}^{m_i} \phi(x_{ij}),$$

and to use the latter as the basis for a kernel method. This simple approach affords rather good performance. With the benefit of hindsight, it is now understandable why the kernel

$$k(X_i, X_j) = \frac{1}{m_i m_j} \sum_{u,v}^{m_i, m_j} k(x_{iu}, x_{jv})$$

produces useful results: it is simply the kernel between the empirical means in feature space $\left\langle \mu(X_i), \mu(X_j) \right\rangle$ (Hein et al., 2004, Equation 4). Jebara and Kondor (2003) later extended this setting by smoothing the empirical densities before computing inner products.

Note, however, that the empirical mean embedding μ_X may not be the best statistic to use for MIC: we are only interested in determining whether *some* instances in the domain have the desired property, rather than making a statement regarding the distribution over all instances. Taking this into account leads to an improved algorithm (Andrews et al., 2003).

7.4 Kernel Measures of Independence

We next demonstrate the application of MMD in determining whether two random variables x and y are independent. In other words, assume that pairs of random variables (x_i, y_i) are jointly drawn from some distribution $p := p_{xy}$. We wish to determine whether this distribution factorizes; that is, whether $q := p_x \times p_y$ is the same as p. One application of such an independence measure is in independent component analysis (Comon, 1994), where the goal is to find a linear mapping of the observations x_i to obtain mutually independent outputs. Kernel methods were employed to solve this problem by Bach and Jordan (2002), Gretton et al. (2005a,b), and Shen et al. (2009). In the following we re-derive one of the above kernel independence measures as a distance between mean embeddings (see also Smola et al., 2007).

We begin by defining

$$\begin{aligned} \mu[p_{xy}] &:= \mathbf{E}_{x,y}\left[v((x,y),\cdot)\right] \\ \text{and } \mu[p_x \times p_y] &:= \mathbf{E}_x\mathbf{E}_y\left[v((x,y),\cdot)\right]. \end{aligned}$$

Here we assume $\mathcal{V}$ is an RKHS over $\mathcal{X}\times\mathcal{Y}$ with kernel $v((x,y),(x',y'))$. If x and y are dependent, then $\mu[p_{xy}] \neq \mu[p_x \times p_y]$. Hence we may use $\Delta(\mathcal{V}, p_{xy}, p_x \times p_y) := \|\mu[p_{xy}] - \mu[p_x \times p_y]\|_{\mathcal{V}}$ as a measure of dependence.

Now assume that $v((x,y),(x',y')) = k(x,x')l(y,y')$, that is, the RKHS $\mathcal{V}$ is a direct product $\mathcal{H}\otimes\mathcal{G}$ of RKHSs on $\mathcal{X}$ and $\mathcal{Y}$. In this case it is easy to see that

$$\begin{aligned} \Delta^2(\mathcal{V}, p_{xy}, p_x \times p_y) &= \|\mathbf{E}_{xy}\left[k(x,\cdot)l(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\right]\mathbf{E}_y\left[l(y,\cdot)\right]\|_{\mathcal{V}}^2 \\ &= \mathbf{E}_{xy}\mathbf{E}_{x'y'}\left[k(x,x')l(y,y')\right] - 2\mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'y'}\left[k(x,x')l(y,y')\right] \\ &\quad + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}\left[k(x,x')l(y,y')\right]. \end{aligned}$$

The latter is also the squared Hilbert-Schmidt norm of the cross-covariance operator between RKHSs (Gretton et al., 2005a): for characteristic kernels, this is zero if and only if x and y are independent.

Theorem 25 *Denote by C_{xy} the covariance operator between random variables x and y, drawn jointly from p_{xy}, where the functions on $\mathcal{X}$ and $\mathcal{Y}$ are the reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ respectively. Then the Hilbert-Schmidt norm $\|C_{xy}\|_{\mathrm{HS}}$ equals $\Delta(\mathcal{V}, p_{xy}, p_x \times p_y)$.*

Empirical estimates of this quantity are as follows:

Theorem 26 *Denote by K and L the kernel matrices on X and Y respectively, and by $H = I - \mathbf{1}/m$ the projection matrix onto the subspace orthogonal to the vector with all entries set to 1 (where $\mathbf{1}$ is an $m \times m$ matrix of ones). Then $m^{-2}\operatorname{tr} HKHL$ is an estimate of Δ^2 with bias $O(m^{-1})$. The deviation from Δ^2 is $O_P(m^{-1/2})$.*

Gretton et al. (2005a) provide explicit constants. In certain circumstances, including in the case of RKHSs with Gaussian kernels, the empirical Δ^2 may also be interpreted in terms of a smoothed difference between the joint empirical characteristic function (ECF) and the product of the marginal ECFs (Feuerverger, 1993; Kankainen, 1995). This interpretation does not hold in all cases, however, for example, for kernels on strings, graphs, and other structured spaces. An illustration of the witness function $f^* \in \mathcal{V}$ from Section 2.3 is provided in Figure 4, for the case of dependence detection. This is a smooth function which has large magnitude where the joint density is most different from the product of the marginals.

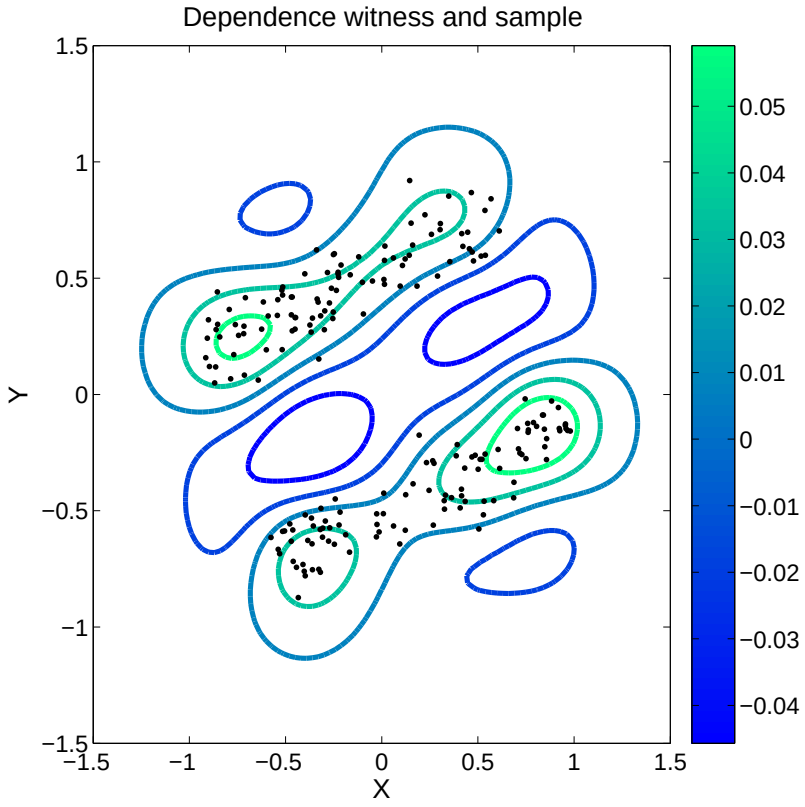

Figure 4: Illustration of the function maximizing the mean discrepancy when MMD is used as a measure of dependence. A sample from dependent random variables x and y is shown in black, and the associated function $\hat{f}^*$ that witnesses the MMD is plotted as a contour. The latter was computed empirically on the basis of 200 samples, using a Gaussian kernel with $\sigma = 0.2$.

We remark that a hypothesis test based on the above kernel statistic is more complicated than for the two-sample problem, since the product of the marginal distributions is in effect simulated by permuting the variables of the original sample. Further details are provided by Gretton et al. (2008b).

7.5 Kernel Statistics Using a Distribution over Witness Functions

Shawe-Taylor and Dolia (2007) define a distance between distributions as follows: let $\mathcal{H}$ be a set of functions on $\mathcal{X}$ and r be a probability distribution over $\mathcal{H}$. Then the distance between two distributions p and q is given by

$$D(p,q) := \mathbf{E}_{f\sim r(f)} \left| \mathbf{E}_x[f(x)] - \mathbf{E}_y[f(y)] \right|. \tag{13}$$

That is, we compute the average distance between p and q with respect to a distribution over test functions. The following result shows the relation with the MMD, and is due to Song et al. (2008, Section 6).

Lemma 27 *Let $\mathcal{H}$ be a reproducing kernel Hilbert space, $f \in \mathcal{H}$, and assume $r(f) = r(\|f\|_{\mathcal{H}})$ with finite $\mathbf{E}_{f\sim r}[\|f\|_{\mathcal{H}}]$. Then $D(p,q) = C\left\|\mu_p - \mu_q\right\|_{\mathcal{H}}$ for some constant C which depends only on $\mathcal{H}$ and r.*

Proof By definition $\mathbf{E}_x[f(x)] = \langle \mu_p, f\rangle_{\mathcal{H}}$. Using linearity of the inner product, Equation (13) equals

$$\begin{aligned}&\int \left|\langle \mu_p - \mu_q, f\rangle_{\mathcal{H}}\right| \mathrm{d}r(f)\\ &= \left\|\mu_p - \mu_q\right\|_{\mathcal{H}} \int \left|\left\langle \frac{\mu_p - \mu_q}{\left\|\mu_p - \mu_q\right\|_{\mathcal{H}}}, f\right\rangle_{\mathcal{H}}\right| \mathrm{d}r(f),\end{aligned}$$

where the integral is independent of p, q. To see this, note that for any p, q, $\frac{\mu_p - \mu_q}{\left\|\mu_p - \mu_q\right\|_{\mathcal{H}}}$ is a unit vector which can be transformed into the first canonical basis vector (for instance) by a rotation which leaves the integral invariant, bearing in mind that r is rotation invariant. ■

7.6 Outlier Detection

An application related to the two sample problem is that of outlier detection: this is the question of whether a novel point is generated from the same distribution as a particular i.i.d. sample. In a way, this is a special case of a two sample test, where the second sample contains only one observation. Several methods essentially rely on the distance between a novel point to the sample mean in feature space to detect outliers.

For instance, Davy et al. (2002) use a related method to deal with nonstationary time series. Likewise Shawe-Taylor and Cristianini (2004, p. 117) discuss how to detect novel observations by using the following reasoning: the probability of being an outlier is bounded both as a function of the spread of the points in feature space and the uncertainty in the empirical feature space mean (as bounded using symmetrisation and McDiarmid's tail bound).

Instead of using the sample mean and variance, Tax and Duin (1999) estimate the center and radius of a minimal enclosing sphere for the data, the advantage being that such bounds can potentially lead to more reliable tests for single observations. Schölkopf et al. (2001) show that the minimal enclosing sphere problem is equivalent to novelty detection by means of finding a hyperplane separating the data from the origin, at least in the case of radial basis function kernels.

8. Experiments

We conducted distribution comparisons using our MMD-based tests on data sets from three real-world domains: database applications, bioinformatics, and neurobiology. We investigated both uniform convergence approaches (MMD_b with the Corollary 9 threshold, and MMD_u^2 H with the Corollary 11 threshold); the asymptotic approaches with bootstrap (MMD_u^2 B) and moment matching to Pearson curves (MMD_u^2 M), both described in Section 5; and the asymptotic approach using the linear time statistic (MMD_l^2) from Section 6. We also compared against several alternatives from

the literature (where applicable): the multivariate t-test, the Friedman-Rafsky Kolmogorov-Smirnov generalisation *(Smir)*, the Friedman-Rafsky Wald-Wolfowitz generalisation *(Wolf)*, the Biau-Györfi test *(Biau)* with a uniform space partitioning, and the Hall-Tajvidi test *(Hall)*. See Section 3.3 for details regarding these tests. Note that we do not apply the Biau-Györfi test to high-dimensional problems (since the required space partitioning is no longer possible), and that MMD is the only method applicable to structured data such as graphs.

An important issue in the practical application of the MMD-based tests is the selection of the kernel parameters. We illustrate this with a Gaussian RBF kernel, where we must choose the kernel width σ (we use this kernel for univariate and multivariate data, but not for graphs). The empirical MMD is zero both for kernel size $\sigma = 0$ (where the aggregate Gram matrix over X and Y is a unit matrix), and also approaches zero as $\sigma \to \infty$ (where the aggregate Gram matrix becomes uniformly constant). We set σ to be the median distance between points in the aggregate sample, as a compromise between these two extremes: this remains a heuristic, similar to those described in Takeuchi et al. (2006) and Schölkopf (1997), and the optimum choice of kernel size is an ongoing area of research. We further note that setting the kernel using the sample being tested may cause changes to the asymptotic distribution: in particular, the analysis in Sections 4 and 5 assumes the kernel not to be a function of the sample. An analysis of the convergence of MMD when the kernel is adapted on the basis of the sample is provided by Sriperumbudur et al. (2009), although the asymptotic distribution in this case remains a topic of research. As a practical matter, however, the median heuristic has not been observed to have much effect on the asymptotic distribution, and in experiments is indistinguishable from results obtained by computing the kernel on a small subset of the sample set aside for this purpose. See Appendix C for more detail.

8.1 Toy Example: Two Gaussians

In our first experiment, we investigated the scaling performance of the various tests as a function of the dimensionality d of the space $\mathcal{X} \subset \mathbb{R}^d$, when both p and q were Gaussian. We considered values of d up to 2500: the performance of the MMD-based tests cannot therefore be explained in the context of density estimation (as in Section 3.3.1), since the associated density estimates are necessarily meaningless here. The levels for all tests were set at $\alpha = 0.05$, $m = n = 250$ samples were used, and results were averaged over 100 repetitions. In the first case, the distributions had different means and unit variance. The percentage of times the null hypothesis was correctly rejected over a set of Euclidean distances between the distribution means (20 values logarithmically spaced from 0.05 to 50), was computed as a function of the dimensionality of the normal distributions. In case of the t-test, a ridge was added to the covariance estimate, to avoid singularity (the ratio of largest to smallest eigenvalue was ensured to be at most 2). In the second case, samples were drawn from distributions $\mathcal{N}(0, \mathbf{I})$ and $\mathcal{N}(0, \sigma^2 \mathbf{I})$ with different variance. The percentage of null rejections was averaged over 20 σ values logarithmically spaced from $10^{0.01}$ to 10. The t-test was not compared in this case, since its output would have been irrelevant. Results are plotted in Figure 5.

In the case of Gaussians with differing means, we observe the t-test performs best in low dimensions, however its performance is severely weakened when the number of samples exceeds the number of dimensions. The performance of MMD_u^2 M is comparable to the t-test in low dimensions, and outperforms all other methods in high dimensions. The worst performance is obtained for MMD_u^2 H, though MMD_b also does relatively poorly: this is unsurprising given that these tests

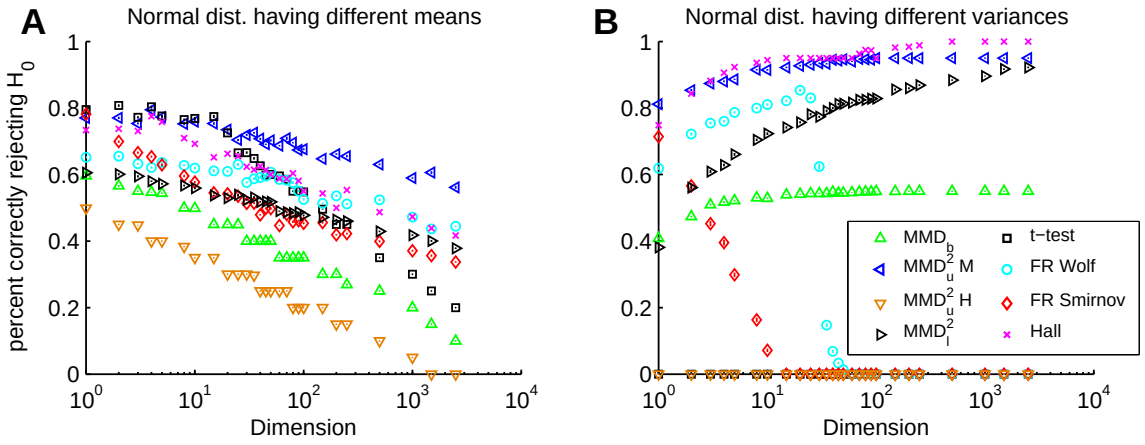

Figure 5: Type II performance of the various tests when separating two Gaussians, with test level $\alpha = 0.05$. **A** Gaussians having same variance and different means. **B** Gaussians having same mean and different variances.

derive from distribution-free large deviation bounds, and the sample size is relatively small. Remarkably, MMD_l^2 performs quite well compared with the Section 3.3.3 tests in high dimensions.

In the case of Gaussians of differing variance, the *Hall* test performs best, followed closely by MMD_u^2 M. *FR Wolf* and (to a much greater extent) *FR Smirnov* both have difficulties in high dimensions, failing completely once the dimensionality becomes too great. The linear-cost test MMD_l^2 again performs surprisingly well, almost matching the MMD_u^2 M performance at the highest dimensionality. Both MMD_u^2 H and MMD_b perform poorly, the former failing completely: this is one of several illustrations we will encounter of the much greater tightness of the Corollary 9 threshold over that in Corollary 11.

8.2 Data Integration

In our next application of MMD, we performed distribution testing for data integration: the objective being to aggregate two data sets into a single sample, with the understanding that both original samples were generated from the same distribution. Clearly, it is important to check this last condition before proceeding, or an analysis could detect patterns in the new data set that are caused by combining the two different source distributions. We chose several real-world settings for this task: we compared microarray data from normal and tumor tissues (Health status), microarray data from different subtypes of cancer (Subtype), and local field potential (LFP) electrode recordings from the Macaque primary visual cortex (V1) with and without spike events (Neural Data I and II, as described in more detail by Rasch et al., 2008). In all cases, the two data sets have different statistical properties, but the detection of these differences is made difficult by the high data dimensionality (indeed, for the microarray data, density estimation is impossible given the sample size and data dimensionality, and no successful test can rely on accurate density estimates as an intermediate step).

Data Set	Attr.	MMD_b	MMD_u^2 H	MMD_u^2 B	MMD_u^2 M	t-test	Wolf	Smir	Hall
Neural Data I	Same	100.0	100.0	96.5	96.5	100.0	97.0	95.0	96.0
	Different	38.0	100.0	**0.0**	**0.0**	42.0	**0.0**	10.0	49.0
Neural Data II	Same	100.0	100.0	94.6	95.2	100.0	95.0	94.5	96.0
	Different	99.7	100.0	3.3	3.4	100.0	**0.8**	31.8	5.9
Health status	Same	100.0	100.0	95.5	94.4	100.0	94.7	96.1	95.6
	Different	100.0	100.0	1.0	**0.8**	100.0	2.8	44.0	35.7
Subtype	Same	100.0	100.0	99.1	96.4	100.0	94.6	97.3	96.5
	Different	100.0	100.0	**0.0**	**0.0**	100.0	**0.0**	28.4	0.2

Table 1: Distribution testing for data integration on multivariate data. Numbers indicate the percentage of repetitions for which the null hypothesis (p=q) was accepted, given $\alpha = 0.05$. Sample size (dimension; repetitions of experiment): Neural I 4000 (63; 100) ; Neural II 1000 (100; 1200); Health Status 25 (12,600; 1000); Subtype 25 (2,118; 1000).

Data Set	Attr.	MMD_b	MMD_u^2 H	MMD_u^2 B	MMD_u^2 M	t-test	Wolf	Smir	Hall	Biau
BIO	Same	100.0	100.0	93.8	94.8	95.2	90.3	95.8	95.3	99.3
	Different	20.0	52.6	**17.2**	17.6	36.2	**17.2**	18.6	17.9	42.1
FOREST	Same	100.0	100.0	96.4	96.0	97.4	94.6	99.8	95.5	100.0
	Different	3.9	11.0	**0.0**	**0.0**	0.2	3.8	**0.0**	50.1	**0.0**
CNUM	Same	100.0	100.0	94.5	93.8	94.0	98.4	97.5	91.2	98.5
	Different	14.9	52.7	2.7	**2.5**	19.17	22.5	11.6	79.1	50.5
FOREST10D	Same	100.0	100.0	94.0	94.0	100.0	93.5	96.5	97.0	100.0
	Different	86.6	100.0	**0.0**	**0.0**	**0.0**	**0.0**	1.0	72.0	100.0

Table 2: Naive attribute matching on univariate (BIO, FOREST, CNUM) and multivariate (FOREST10D) data. Numbers indicate the percentage of times the null hypothesis $p = q$ was accepted with $\alpha = 0.05$, pooled over attributes. Sample size (dimension; attributes; repetitions of experiment): BIO 377 (1; 6; 100); FOREST 538 (1; 10; 100); CNUM 386 (1; 13; 100); FOREST10D 1000 (10; 2; 100).

We applied our tests to these data sets in the following fashion. Given two data sets A and B, we either chose one sample from A and the other from B *(attributes = different)*; or both samples from either A or B *(attributes = same)*. We then repeated this process up to 1200 times. Results are reported in Table 1. Our asymptotic tests perform better than all competitors besides *Wolf*: in the latter case, we have greater Type II error for one neural data set, lower Type II error on the Health Status data (which has very high dimension and low sample size), and identical (error-free) performance on the remaining examples. We note that the Type I error of the bootstrap test on the Subtype data set is far from its design value of 0.05, indicating that the Pearson curves provide a better threshold estimate for these low sample sizes. For the remaining data sets, the Type I errors of the Pearson and Bootstrap approximations are close. Thus, for larger data sets, the bootstrap is to be preferred, since it costs $O(m^2)$, compared with a cost of $O(m^3)$ for the Pearson curves (due to the cost of computing (11)). Finally, the uniform convergence-based tests are too conservative, with MMD_b finding differences in distribution only for the data with largest sample size, and MMD_u^2 H never finding differences.

8.3 Computational Cost

We next investigate the tradeoff between computational cost and performance of the various tests, with a particular focus on how the quadratic-cost MMD tests from Sections 4 and 5 compare with the linear time MMD-based asymptotic test from Section 6. We consider two 1-D data sets (CNUM and FOREST) and two higher-dimensional data sets (FOREST10D and NEUROII). Results are plotted in Figure 6. If cost is not a factor, then the MMD_u^2 B shows best overall performance as a function of sample size, with a Type II error dropping to zero as fast or faster than competing approaches in three of four cases, and narrowly trailing *FR Wolf* in the remaining case (FOREST10D). That said, for data sets CNUM, FOREST, and FOREST10D, the linear time MMD achieves a given Type II error at a far smaller computational cost than MMD_u^2 B, albeit by looking at a great deal more data. In the CNUM case, however, the linear test is not able to achieve zero error even for the largest data set size. For the NEUROII data, attaining zero Type II error has about the same cost for both approaches. The difference in cost of MMD_u^2 B and MMD_b is due to the bootstrapping required for the former, which produces a constant offset in cost between the two (here 150 resamplings were used).

The *t*-test also performs well in three of the four problems, and in fact represents the best cost-performance tradeoff in these three data sets (i.e., while it requires much more data than MMD_u^2 B for a given Type II error rate, it costs far less to compute). The *t*-test assumes that only the difference in means is important in distinguishing the distributions, and it requires an accurate estimate of the within-sample covariance; the test fails completely on the NEUROII data. We emphasise that the Kolmogorov-Smirnov results in 1-D were obtained using the classical statistic, and not the Friedman-Rafsky statistic, hence the low computational cost. The cost of both Friedman-Rafsky statistics is therefore given by the *FR Wolf* cost in this case. The latter scales similarly with sample size to the quadratic time MMD tests, confirming Friedman and Rafsky's observation that obtaining the pairwise distances between sample points is the dominant cost of their tests. We also remark on the unusual behaviour of the Type II error of the *FR Wolf* test in the FOREST data set, which worsens for increasing sample size.

We conclude that the approach to be recommended for two-sample testing will depend on the data available: for small amounts of data, the best results are obtained using every observation to maximum effect, and employing the quadratic time MMD_u^2 B test. When large volumes of data are available, a better option is to look at each point only once, which can yield lower Type II error for a given computational cost. It may also be worth doing a t-test first in this case, and only running more sophisticated nonparametric tests if the t-test accepts the null hypothesis, to verify the distributions are identical in more than just mean.

8.4 Attribute Matching

Our final series of experiments addresses automatic attribute matching. Given two databases, we want to detect corresponding attributes in the schemas of these databases, based on their data-content (as a simple example, two databases might have respective fields Wage and Salary, which are assumed to be observed via a subsampling of a particular population, and we wish to automatically determine that both Wage and Salary denote to the same underlying attribute). We use a two-sample test on pairs of attributes from two databases to find corresponding pairs.[15] This procedure

15. Note that corresponding attributes may have different distributions in real-world databases. Hence, schema matching cannot solely rely on distribution testing.

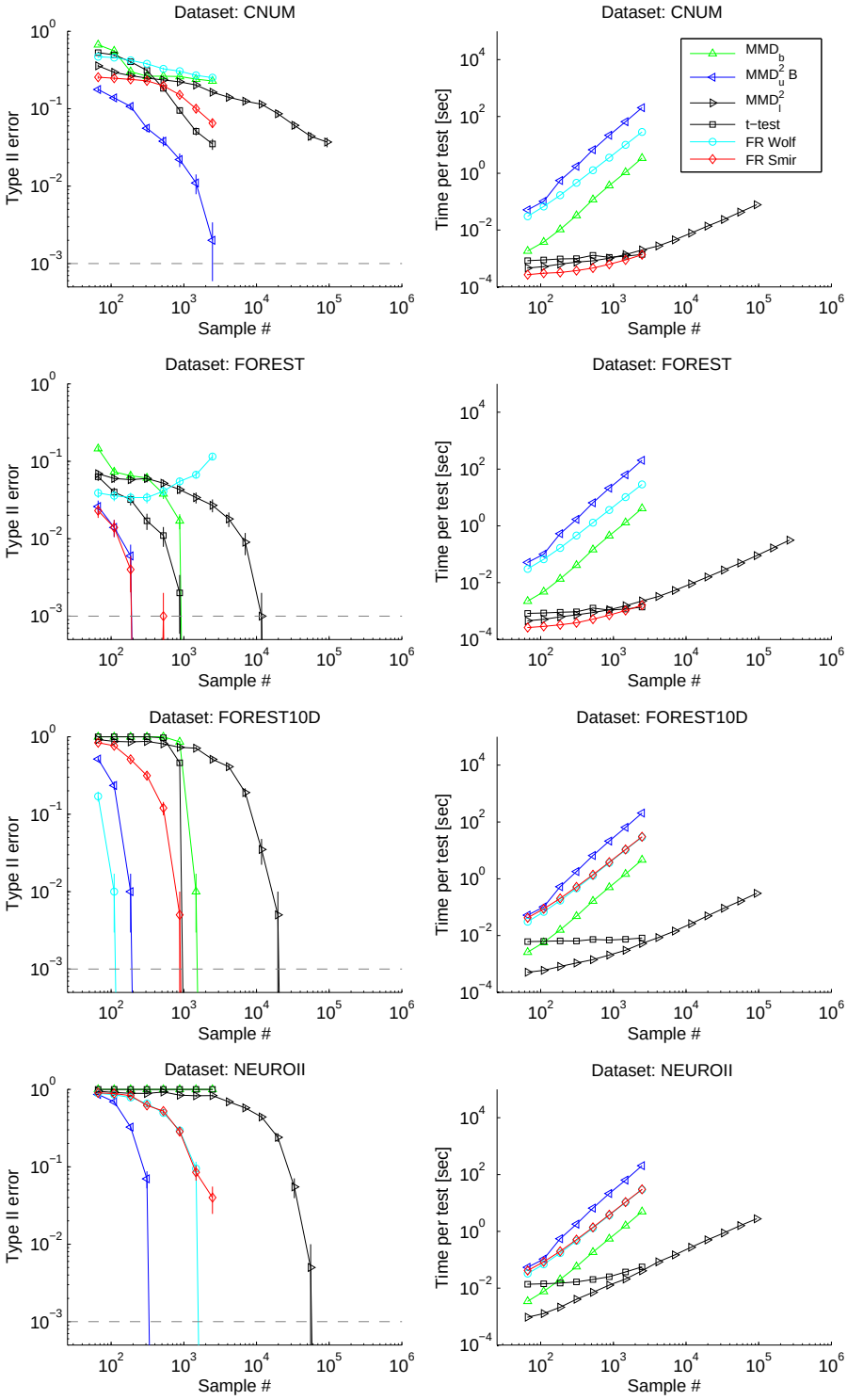

Figure 6: Linear-cost vs quadratic-cost MMD. The first column shows Type II performance, and the second shows runtime. The dashed grey horizontal line indicates zero Type II error (required due to log y-axis).

is also called *table matching* for tables from different databases. We performed attribute matching as follows: first, the data set D was split into two halves A and B. Each of the n attributes in A (and B, resp.) was then represented by its instances in A (resp. B). We then tested all pairs of attributes from A and from B against each other, to find the optimal assignment of attributes $A_1, \ldots, A_n$ from A to attributes $B_1, \ldots, B_n$ from B. We assumed that A and B contain the same number of attributes.

As a naive approach, we could assume that any possible pair of attributes might correspond, and thus that every attribute of A needs to be tested against all the attributes of B to find the optimal match. We report results for this naive approach, aggregated over all pairs of possible attribute matches, in Table 2. We used three data sets: the census income data set from the UCI KDD archive (CNUM), the protein homology data set from the 2004 KDD Cup (BIO) (Caruana and Joachims, 2004), and the forest data set from the UCI ML archive (Blake and Merz, 1998). For the final data set, we performed univariate matching of attributes (FOREST) and multivariate matching of tables (FOREST10D) from two different databases, where each table represents one type of forest. Both our asymptotic MMD_u^2-based tests perform as well as or better than the alternatives, notably for CNUM, where the advantage of MMD_u^2 is large. Unlike in Table 1, the next best alternatives are not consistently the same across all data: for example, in BIO they are *Wolf* or *Hall*, whereas in FOREST they are *Smir*, *Biau*, or the t-test. Thus, MMD_u^2 appears to perform more consistently across the multiple data sets. The Friedman-Rafsky tests do not always return a Type I error close to the design parameter: for instance, *Wolf* has a Type I error of 9.7% on the BIO data set (on these data, MMD_u^2 has the joint best Type II error without compromising the designed Type I performance). Finally, MMD_b performs much better than in Table 1, although surprisingly it fails to reliably detect differences in FOREST10D. The results of MMD_u^2 H are also improved, although it remains among the worst performing methods.

A more principled approach to attribute matching is also possible. Assume that $\phi(A) = (\phi_1(A_1), \phi_2(A_2), ..., \phi_n(A_n))$: in other words, the kernel decomposes into kernels on the individual attributes of A (and also decomposes this way on the attributes of B). In this case, MMD^2 can be written $\sum_{i=1}^n \|\mu_i(A_i) - \mu_i(B_i)\|^2$, where we sum over the MMD terms on each of the attributes. Our goal of optimally assigning attributes from B to attributes of A via MMD is equivalent to finding the optimal permutation π of attributes of B that minimizes $\sum_{i=1}^n \|\mu_i(A_i) - \mu_i(B_{\pi(i)})\|^2$. If we define $C_{ij} = \|\mu_i(A_i) - \mu_i(B_j)\|^2$, then this is the same as minimizing the sum over $C_{i,\pi(i)}$. This is the linear assignment problem, which costs $O(n^3)$ time using the Hungarian method (Kuhn, 1955).

While this may appear to be a crude heuristic, it nonetheless defines a semi-metric on the sample spaces X and Y and the corresponding distributions p and q. This follows from the fact that matching distances are proper metrics if the matching cost functions are metrics. We formalize this as follows:

Theorem 28 *Let p, q be distributions on $\mathbb{R}^d$ and denote by p_i, q_i the marginal distributions on the i-th variable. Moreover, denote by Π the symmetric group on $\{1, \ldots, d\}$. The following distance, obtained by optimal coordinate matching, is a semi-metric.*

$$\Delta[\mathcal{F}, p, q] := \min_{\pi \in \Pi} \sum_{i=1}^{d} \text{MMD}[\mathcal{F}, p_i, q_{\pi(i)}].$$

Proof Clearly $\Delta[\mathcal{F}, p, q]$ is nonnegative, since it is a sum of nonnegative quantities. Next we show the triangle inequality. Denote by r a third distribution on $\mathbb{R}^d$ and let $\pi_{p,q}, \pi_{q,r}$ and $\pi_{p,r}$ be the

distance minimizing permutations over the associated pairs from $\{p,q,r\}$. It follows that

$$\begin{aligned}\Delta[\mathcal{F},p,q]+\Delta[\mathcal{F},q,r] &= \sum_{i=1}^{d}\mathrm{MMD}[\mathcal{F},p_i,q_{\pi_{p,q}(i)}]+\sum_{i=1}^{d}\mathrm{MMD}[\mathcal{F},q_i,r_{\pi_{q,r}(i)}]\\ &\geq \sum_{i=1}^{d}\mathrm{MMD}[\mathcal{F},p_i,r_{[\pi_{p,q}\circ\pi_{q,r}](i)}]\geq\Delta[\mathcal{F},p,r].\end{aligned}$$

The first inequality follows from the triangle inequality on MMD,

$$\mathrm{MMD}[\mathcal{F},p_i,q_{\pi_{p,q}(i)}]+\mathrm{MMD}[\mathcal{F},q_{\pi_{p,q}(i)},r_{[\pi_{p,q}\circ\pi_{q,r}](i)}]\geq\mathrm{MMD}[\mathcal{F},p_i,r_{[\pi_{p,q}\circ\pi_{q,r}](i)}].$$

The second inequality is a result of minimization over π. ■

We tested this 'Hungarian approach' to attribute matching via MMD_u^2 B on three univariate data sets (BIO, CNUM, FOREST) and for table matching on a fourth (FOREST10D). To study MMD_u^2 B on structured data, we used two data sets of protein graphs (PROTEINS and ENZYMES) and used the graph kernel for proteins from Borgwardt et al. (2005) for table matching via the Hungarian method (the other tests were not applicable to these graph data). The challenge here is to match tables representing one functional class of proteins (or enzymes) from data set A to the corresponding tables (functional classes) in B. Results are shown in Table 3. Besides on the BIO and CNUM data sets, MMD_u^2 B made no errors.

Data Set	Data type	No. attributes	Sample size	Repetitions	% correct
BIO	univariate	6	377	100	90.0
CNUM	univariate	13	386	100	99.8
FOREST	univariate	10	538	100	100.0
FOREST10D	multivariate	2	1000	100	100.0
ENZYME	structured	6	50	50	100.0
PROTEINS	structured	2	200	50	100.0

Table 3: Hungarian Method for attribute matching via MMD_u^2 B on univariate (BIO, CNUM, FOREST), multivariate (FOREST10D), and structured (ENZYMES, PROTEINS) data ($\alpha = 0.05$; "% correct" is the percentage of correct attribute matches over all repetitions).

9. Conclusion

We have established three simple multivariate tests for comparing two distributions p and q, based on samples of size m and n from these respective distributions. Our test statistic is the maximum mean discrepancy (MMD), defined as the maximum deviation in the expectation of a function evaluated on each of the random variables, taken over a sufficiently rich function class: in our case, a reproducing kernel Hilbert space (RKHS). Equivalently, the statistic can be written as the norm of the difference between distribution feature means in the RKHS. We do not require density estimates as an intermediate step. Two of our tests provide Type I error bounds that are exact and distribution-free for finite sample sizes. We also give a third test based on quantiles of the asymptotic distribution

of the associated test statistic. All three tests can be computed in $O((m+n)^2)$ time, however when sufficient data are available, a linear time statistic can be used, which in our experiments was able to achieve a given Type II error at smaller computational cost, by looking at many more samples than the quadratic-cost tests.

We have seen in Section 7 that several classical metrics on probability distributions can be written as integral probability metrics with function classes that are not Hilbert spaces, but rather Banach or seminormed spaces (for instance the Kolmogorov-Smirnov and Earth Mover's distances). It is therefore of interest to establish under what conditions one could write these discrepancies in terms of norms of differences of mean embeddings. Sriperumbudur et al. (2011b) provide expressions for the maximum mean discrepancy in terms of mean embeddings in reproducing kernel Banach spaces. When the Banach space is not an RKBS, the question of establishing a mean embedding interpretation for the MMD remains open.

We also note (following Section 7.3) that the MMD for RKHSs is associated with a particular kernel between probability distributions. Hein et al. (2004) describe several further such kernels, which induce corresponding distances between feature space distribution mappings: these may in turn lead to new and powerful two-sample tests.

Two recent studies have shown that additional divergence measures between distributions can be obtained empirically through optimization in a reproducing kernel Hilbert space. Harchaoui et al. (2008) define a two-sample test statistic arising from the kernel Fisher discriminant, rather than the difference of RKHS means; and Nguyen et al. (2008) obtain a KL divergence estimate by approximating the ratio of densities (or its log) with a function in an RKHS. By design, both these kernel-based statistics prioritise different features of p and q when measuring the divergence between distributions, and the resulting effects on distinguishability of distributions are therefore of interest.

Acknowledgments

We would like to thank the anonymous referees, whose suggestions greatly improved the paper; Bharath Sriperumbudur, for thoroughly proofreading the final draft; Sivaraman Balakrishnan, Philipp Berens, Olivier Bousquet, Corinna Cortes, Omri Guttman, Peter Hall, Matthias Hein, John Langford, Mehryar Mohri, Novi Quadrianto, Le Song, and Vishy Vishwanathan, for constructive discussions; Patrick Warnat (DKFZ, Heidelberg), for providing the microarray data sets; and Nikos Logothetis, for providing the neural data sets. National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and by the Austrian Science Fund (FWF), project # S9102-N04.

Appendix A. Large Deviation Bounds for Tests with Finite Sample Guarantees

This section contains proofs of the theorems of Section 4.1. We begin in Section A.1 with a review of McDiarmid's inequality and the Rademacher average of a function class. We prove Theorem 7 in Section A.2, and Theorem 8 in Section A.3.

A.1 Preliminary Definitions and Theorems

We need the following theorem, due to McDiarmid (1989).

Theorem 29 (McDiarmid's inequality) *Let $f : \mathcal{X}^m \to \mathbb{R}$ be a function such that for all $i \in \{1,\dots,m\}$, there exist $c_i < \infty$ for which*

$$\sup_{X\in\mathcal{X}^m,\tilde{x}\in\mathcal{X}} |f(x_1,\dots x_m) - f(x_1,\dots x_{i-1},\tilde{x},x_{i+1},\dots,x_m)| \leq c_i.$$

Then for all probability measures p and every $\varepsilon > 0$,

$$\Pr_X \left(f(X) - \mathbf{E}_X(f(X)) > t\right) < \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right),$$

where $\mathbf{E}_X$ denotes the expectation over the m random variables $x_i \sim p$, and $\Pr_X$ denotes the probability over these m variables.

We also define the Rademacher average of the function class $\mathcal{F}$ with respect to the m-sample X.

Definition 30 (Rademacher average of $\mathcal{F}$ on X) *Let $\mathcal{F}$ be the unit ball in an RKHS on the domain $\mathcal{X}$, with kernel bounded according to $0 \leq k(x,y) \leq K$. Let X be an i.i.d. sample of size m drawn according to a probability measure p on $\mathcal{X}$, and let σ_i be i.i.d and take values in $\{-1,1\}$ with equal probability. We define the Rademacher average*

$$\begin{aligned} R_m(\mathcal{F},X) &:= \mathbf{E}_\sigma \sup_{f\in\mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right| \\ &\leq (K/m)^{1/2}, \end{aligned}$$

where the upper bound is due to Bartlett and Mendelson (2002, Lemma 22), and $\mathbf{E}_\sigma$ denotes the expectation over all the σ_i. Similarly, we define

$$R_m(\mathcal{F},p) := \mathbf{E}_{x,\sigma} \sup_{f\in\mathcal{F}} \left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right|.$$

A.2 Bound when p and q May Differ

We want to show that the absolute difference between $\mathrm{MMD}(\mathcal{F},p,q)$ and $\mathrm{MMD}_b(\mathcal{F},X,Y)$ is close to its expected value, independent of the distributions p and q. To this end, we prove three intermediate results, which we then combine. The first result we need is an upper bound on the absolute difference between $\mathrm{MMD}(\mathcal{F},p,q)$ and $\mathrm{MMD}_b(\mathcal{F},X,Y)$. We have

$$\begin{aligned} & |\mathrm{MMD}(\mathcal{F},p,q) - \mathrm{MMD}_b(\mathcal{F},X,Y)| \\ = & \left|\sup_{f\in\mathcal{F}} (\mathbf{E}_x(f) - \mathbf{E}_y(f)) - \sup_{f\in\mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^m f(x_i) - \frac{1}{n}\sum_{i=1}^n f(y_i)\right)\right| \\ \leq & \underbrace{\sup_{f\in\mathcal{F}} \left|\mathbf{E}_x(f) - \mathbf{E}_y(f) - \frac{1}{m}\sum_{i=1}^m f(x_i) + \frac{1}{n}\sum_{i=1}^n f(y_i)\right|}_{\Delta(p,q,X,Y)}. \end{aligned} \tag{14}$$

Second, we provide an upper bound on the difference between $\Delta(p,q,X,Y)$ and its expectation. Changing either of x_i or y_i in $\Delta(p,q,X,Y)$ results in changes in magnitude of at most $2K^{1/2}/m$ or $2K^{1/2}/n$, respectively. We can then apply McDiarmid's theorem, given a denominator in the exponent of

$$m\left(2K^{1/2}/m\right)^2 + n\left(2K^{1/2}/n\right)^2 = 4K\left(\frac{1}{m}+\frac{1}{n}\right) = 4K\frac{m+n}{mn},$$

to obtain

$$\Pr_{X,Y}\left(\Delta(p,q,X,Y) - \mathbf{E}_{X,Y}\left[\Delta(p,q,X,Y)\right] > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2 mn}{2K(m+n)}\right). \tag{15}$$

For our final result, we exploit symmetrisation, following, for example, van der Vaart and Wellner (1996, p. 108), to upper bound the expectation of $\Delta(p,q,X,Y)$. Denoting by X' an i.i.d sample of size m drawn independently of X (and likewise for Y'), we have

$$\begin{aligned}
& \mathbf{E}_{X,Y}\left[\Delta(p,q,X,Y)\right] \\
= \;& \mathbf{E}_{X,Y}\sup_{f\in\mathcal{F}}\left|\mathbf{E}_x(f) - \frac{1}{m}\sum_{i=1}^m f(x_i) - \mathbf{E}_y(f) + \frac{1}{n}\sum_{i=1}^n f(y_j)\right| \\
= \;& \mathbf{E}_{X,Y}\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X'}\left(\frac{1}{m}\sum_{i=1}^m f(x_i')\right) - \frac{1}{m}\sum_{i=1}^m f(x_i) - \mathbf{E}_{Y'}\left(\frac{1}{n}\sum_{i=1}^n f(y_j')\right) + \frac{1}{n}\sum_{i=1}^n f(y_j)\right| \\
\underset{(a)}{\leq} \;& \mathbf{E}_{X,Y,X',Y'}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(x_i') - \frac{1}{m}\sum_{i=1}^m f(x_i) - \frac{1}{n}\sum_{i=1}^n f(y_j') + \frac{1}{n}\sum_{i=1}^n f(y_j)\right| \\
= \;& \mathbf{E}_{X,Y,X',Y',\sigma,\sigma'}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i\left(f(x_i') - f(x_i)\right) + \frac{1}{n}\sum_{i=1}^n \sigma_i'\left(f(y_j') - f(y_j)\right)\right| \\
\underset{(b)}{\leq} \;& \mathbf{E}_{X,X',\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i\left(f(x_i') - f(x_i)\right)\right| + \mathbf{E}_{Y,Y',\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(y_j') - f(y_j)\right)\right| \\
\underset{(c)}{\leq} \;& 2\left[R_m(\mathcal{F},p) + R_n(\mathcal{F},q)\right]. \\
\underset{(d)}{\leq} \;& 2\left[(K/m)^{1/2} + (K/n)^{1/2}\right],
\end{aligned} \tag{16}$$

where (a) uses Jensen's inequality, (b) uses the triangle inequality, (c) substitutes Definition 30 (the Rademacher average), and (d) bounds the Rademacher averages, also via Definition 30.

Having established our preliminary results, we proceed to the proof of Theorem 7.

Proof (Theorem 7) Combining Equations (15) and (16), gives

$$\Pr_{X,Y}\left(\Delta(p,q,X,Y) - 2\left[(K/m)^{1/2} + (K/n)^{1/2}\right] > \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2 mn}{2K(m+n)}\right).$$

Substituting Equation (14) yields the result.

■

A.3 Bound when $p = q$ and $m = n$

In this section, we derive the Theorem 8 result, namely the large deviation bound on the MMD when $p = q$ and $m = n$. Note also that we consider only positive deviations of $\mathrm{MMD}_b(\mathcal{F}, X, Y)$ from $\mathrm{MMD}(\mathcal{F}, p, q)$, since negative deviations are irrelevant to our hypothesis test. The proof follows the same three steps as in the previous section. The first step in (14) becomes

$$\begin{aligned}\mathrm{MMD}_b(\mathcal{F}, X, Y) - \mathrm{MMD}(\mathcal{F}, p, q) &= \mathrm{MMD}_b(\mathcal{F}, X, X') - 0 \\ &= \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \left(f(x_i) - f(x_i') \right) \right). \end{aligned} \tag{17}$$

The McDiarmid bound on the difference between (17) and its expectation is now a function of $2m$ observations in (17), and has a denominator in the exponent of $2m\left(2K^{1/2}/m\right)^2 = 8K/m$. We use a different strategy in obtaining an upper bound on the expected (17), however: this is now

$$\begin{aligned} & \mathbf{E}_{X,X'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \left(f(x_i) - f(x_i') \right) \right] \\ = & \frac{1}{m} \mathbf{E}_{X,X'} \left\| \sum_{i=1}^{m} \left(\phi(x_i) - \phi(x_i') \right) \right\| \\ = & \frac{1}{m} \mathbf{E}_{X,X'} \left[\sum_{i=1}^{m} \sum_{j=1}^{m} \left(k(x_i, x_j) + k(x_i', x_j') - k(x_i, x_j') - k(x_i', x_j) \right) \right]^{\frac{1}{2}} \\ \leq & \frac{1}{m} \left[2m \mathbf{E}_x k(x,x) + 2m(m-1) \mathbf{E}_{x,x'} k(x,x') - 2m^2 \mathbf{E}_{x,x'} k(x,x') \right]^{\frac{1}{2}} \\ = & \left[\frac{2}{m} \mathbf{E}_{x,x'} \left(k(x,x) - k(x,x') \right) \right]^{\frac{1}{2}} \qquad (18) \\ \leq & \left(2K/m \right)^{1/2}. \qquad (19) \end{aligned}$$

We remark that both (18) and (19) bound the amount by which our biased estimate of the population MMD exceeds zero under $\mathcal{H}_0$. Combining the three results, we find that under $\mathcal{H}_0$,

$$\begin{aligned} \Pr_{X,X'} \left(\mathrm{MMD}_b(\mathcal{F}, X, X') - \left[\frac{2}{m} \mathbf{E}_{x,x'} \left(k(x,x) - k(x,x') \right) \right]^{\frac{1}{2}} > \varepsilon \right) &< \exp\left(\frac{-\varepsilon^2 m}{4K} \right) \quad \text{and} \\ \Pr_{X,X'} \left(\mathrm{MMD}_b(\mathcal{F}, X, X') - (2K/m)^{1/2} > \varepsilon \right) &< \exp\left(\frac{-\varepsilon^2 m}{4K} \right). \end{aligned}$$

Appendix B. Proofs for Asymptotic Tests

We derive results needed in the asymptotic test of Section 5. Appendix B.1 describes the distribution of the empirical MMD under $\mathcal{H}_0$ (i.e., $p = q$). Appendix B.2 establishes consistency of the test under local departures from $\mathcal{H}_0$. Appendix B.3 contains derivations of the second and third moments of the empirical MMD, also under $\mathcal{H}_0$.

B.1 Convergence of the Empirical MMD under $\mathcal{H}_0$

In this appendix, we prove Theorem 12, which describes the distribution of the unbiased estimator $\mathrm{MMD}_u^2[\mathcal{F},X,Y]$ under the null hypothesis. Thus, throughout this section, the reader should bear in mind that y now has the same distribution as x, that is, $y \sim p$. We first recall from Lemma 6 in Section 2.2 the population expression,

$$\mathrm{MMD}^2[\mathcal{F},p,q] := \mathbf{E}_{x,x'}k(x,x') + \mathbf{E}_{y,y'}k(y,y') - 2\mathbf{E}_{x,y}k(x,y),$$

and its empirical counterpart,

$$\begin{aligned}\mathrm{MMD}_u^2[\mathcal{F},X,Y] &= \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k(y_i,y_j) \\ &\quad -\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i,y_j). \end{aligned} \tag{20}$$

We begin with the asymptotic analysis of $\mathrm{MMD}_u^2[\mathcal{F},X,Y]$ under the null hypothesis. This is based on the reasoning of Anderson et al. (1994, Appendix), bearing in mind the following changes:

- we do not need to deal with the bias terms S_{1j} in Anderson et al. (1994, Appendix) that vanish for large sample sizes, since our statistic is unbiased;
- we require greater generality, since our kernels are not necessarily inner products in L_2 between probability density functions (although this is a special case: see Section 3.3.1).

We first transform each term in the sum (20) by centering. Under $\mathcal{H}_0$, both x and y have the same mean embedding μ_p. Thus we replace each instance of $k(x_i,x_j)$ in the sum with a kernel $\tilde{k}(x_i,x_j)$ between feature space mappings from which the mean has been subtracted,

$$\begin{aligned}\tilde{k}(x_i,x_j) &:= \left\langle \phi(x_i)-\mu_p, \phi(x_j)-\mu_p \right\rangle_{\mathcal{H}} \\ &= k(x_i,x_j) - \mathbf{E}_x k(x_i,x) - \mathbf{E}_x k(x,x_j) + \mathbf{E}_{x,x'}k(x,x').\end{aligned}$$

The centering terms cancel across the three terms (the distance between the two points is unaffected by an identical global shift in both the points). This gives the equivalent form of the empirical MMD,

$$\begin{aligned}\mathrm{MMD}_u^2[\mathcal{F},X,Y] &= \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\tilde{k}(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tilde{k}(y_i,y_j) \\ &\quad -\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{k}(x_i,y_j), \end{aligned} \tag{21}$$

where each of the three sums has expected value zero. Note in particular that the U-statistics in $\tilde{k}(x_i,x_j)$ are degenerate, meaning

$$\mathbf{E}_x\tilde{k}(x,v) = \mathbf{E}_x k(x,v) - \mathbf{E}_{x,x'}k(x,x') - \mathbf{E}_x k(x,v) + \mathbf{E}_{x,x'}k(x,x') = 0. \tag{22}$$

We define the operator $S_{\tilde{k}} : L_2(p) \to \mathcal{F}$ satisfying

$$S_{\tilde{k}}g(x) := \int_{\mathcal{X}} \tilde{k}(x,x')g(x')dp(x').$$

According to Reed and Simon (1980, Theorem VI.23), this operator is Hilbert-Schmidt, and hence compact, if and only if the kernel $\tilde{k}$ is square integrable under p,

$$\tilde{k} \in L_2(\mathcal{X} \times \mathcal{X}, p \times p). \tag{23}$$

We may write the kernel $\tilde{k}(x_i, x_j)$ in terms of eigenfunctions $\psi_l(x)$ with respect to the probability measure p,

$$\tilde{k}(x, x') = \sum_{l=1}^{\infty} \lambda_l \psi_l(x) \psi_l(x'), \tag{24}$$

where

$$\begin{aligned} \int_{\mathcal{X}} \tilde{k}(x, x') \psi_i(x) dp(x) &= \lambda_i \psi_i(x'), \\ \int_{\mathcal{X}} \psi_i(x) \psi_j(x) dp(x) &= \delta_{ij}, \end{aligned} \tag{25}$$

and the convergence is in $L_2(\mathcal{X} \times \mathcal{X}, p \times p)$. Since the operator is Hilbert-Schmidt, we have by Reed and Simon (1980, Theorem VI.22) that $\sum \lambda_i^2 < \infty$.

Using the degeneracy of the U-statistic in (22), then when $\lambda_i \neq 0$,

$$\begin{aligned} \lambda_i \mathbf{E}_{x'} \psi_i(x') &= \int_{\mathcal{X}} \mathbf{E}_{x'} \tilde{k}(x, x') \psi_i(x) dp(x) \\ &= 0, \end{aligned}$$

and hence

$$\mathbf{E}_x \psi_i(x) = 0. \tag{26}$$

In other words, the eigenfunctions $\psi_i(x)$ are zero mean and uncorrelated.

We now use these results to find the asymptotic distribution of (21). First,

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \tilde{k}(x_i, x_j) &= \frac{1}{m} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \sum_{l=1}^{\infty} \lambda_l \psi_l(x_i) \psi_l(x_j) \\ &= \frac{1}{m} \sum_{l=1}^{\infty} \lambda_l \left(\left(\sum_i \psi_l(x_i) \right)^2 - \sum_i \psi_l^2(x_i) \right) \\ &\underset{D}{\to} \sum_{l=1}^{\infty} \lambda_l (a_l^2 - 1), \end{aligned} \tag{27}$$

where $a_l \sim \mathcal{N}(0,1)$ are i.i.d., and the final relation denotes convergence in distribution, which is proved by Serfling (1980, Section 5.5.2) using (25) and (26).[16] Given that the random variables $a_l^2 - 1$ are zero mean with finite variance, it can be shown either via Kolmogorov's inequality or by the Martingale convergence theorem that the above sum converges almost surely if $\sum_{l=1}^{\infty} \lambda_l^2 < \infty$ (Grimmet and Stirzaker, 2001, Chapter 7.11 Exercise 30). As we have seen, this is guaranteed under the assumption (23).

Likewise

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{k}(y_i, y_j) \underset{D}{\to} \sum_{l=1}^{\infty} \lambda_l (b_l^2 - 1),$$

16. Simply replace $\tilde{h}_2(x_i, x_j)$ with $\tilde{k}(x_i, x_j)$ in Serfling (1980, top of p. 196).

where $b_l \sim \mathcal{N}(0,1)$ independent of the a_l, and

$$\frac{1}{\sqrt{mn}} \sum_{i=1}^{m} \sum_{j=1}^{n} \tilde{k}(x_i, y_j) \underset{D}{\to} \sum_{l=1}^{\infty} \lambda_l a_l b_l, \tag{28}$$

both jointly in distribution with (27), where (28) is proved at the end of the section. We now combine these results. Define $t = m+n$, and assume $\lim_{m,n\to\infty} m/t \to \rho_x$ and $\lim_{m,n\to\infty} n/t \to \rho_y := (1-\rho_x)$ for fixed $0 < \rho_x < 1$. Then

$$\begin{aligned} t\mathrm{MMD}_u^2[\mathcal{F}, X, Y] &\underset{D}{\to} \rho_x^{-1} \sum_{l=1}^{\infty} \lambda_l (a_l^2 - 1) + \rho_y^{-1} \sum_{l=1}^{\infty} \lambda_l (b_l^2 - 1) - \frac{2}{\sqrt{\rho_x \rho_y}} \sum_{l=1}^{\infty} \lambda_l a_l b_l \\ &= \sum_{l=1}^{\infty} \lambda_l \left[(\rho_x^{-1/2} a_l - \rho_y^{-1/2} b_l)^2 - (\rho_x \rho_y)^{-1} \right]. \end{aligned}$$

Proof (Equation 28) The proof is a modification of the result for convergence of degenerate U-statistics of Serfling (1980, Section 5.5.2). We only provide those details that differ from the proof of Serfling, and otherwise refer to the steps in the original proof as needed. First, using (24) to expand out the centred kernel, we may write

$$T_{mn} := \frac{1}{\sqrt{mn}} \sum_{i=1}^{m} \sum_{j=1}^{n} \tilde{k}(x_i, y_j) = \frac{1}{\sqrt{mn}} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{\infty} \lambda_l \psi_l(x_i) \psi_l(y_j).$$

We define a truncation of this sum,

$$T_{mnL} := \frac{1}{\sqrt{mn}} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{l=1}^{L} \lambda_l \psi_l(x_i) \psi_l(y_j).$$

The target distribution is written

$$V = \sum_{l=1}^{\infty} \lambda_l a_l b_l,$$

and its truncation is

$$V_L := \sum_{l=1}^{L} \lambda_l a_l b_l.$$

Our goal is to show

$$\left| \mathbf{E}_{X,Y} \left(e^{\imath s T_{mn}} \right) - \mathbf{E}_{a,b} \left(e^{\imath s V} \right) \right|$$

vanishes for all s as m and n increase, where the expectation $\mathbf{E}_{X,Y}$ is over all sample points, which implies $T_{mn} \underset{D}{\to} V$ (Dudley, 2002, Theorem 9.8.2). We achieve this via the upper bound

$$\begin{aligned} \left| \mathbf{E}_{X,Y} \left(e^{\imath s T_{mn}} \right) - \mathbf{E}_{a,b} \left(e^{\imath s V} \right) \right| &\leq \left| \mathbf{E}_{X,Y} \left(e^{\imath s T_{mn}} \right) - \mathbf{E}_{XY} \left(e^{\imath s T_{mnL}} \right) \right| + \left| \mathbf{E}_{XY} \left(e^{\imath s T_{mnL}} \right) - \mathbf{E}_{a,b} \left(e^{\imath s V_L} \right) \right| \\ &\quad + \left| \mathbf{E}_{a,b} \left(e^{\imath s V_L} \right) - \mathbf{E}_{a,b} \left(e^{\imath s V} \right) \right|, \end{aligned}$$

where we need to show that for large enough L, each of the three terms vanish.

First term: We first show that for large enough L, T_{mn} and T_{mnL} are close in distribution. From Serfling (1980, p. 197),

$$\left| \mathbf{E}_{X,Y} \left(e^{\imath s T_{mn}} \right) - \mathbf{E}_{X,Y} \left(e^{\imath s T_{mnL}} \right) \right| \leq |s| \left[\mathbf{E}_{X,Y} \left(T_{mn} - T_{mnL} \right)^2 \right]^{1/2},$$

and we may write the difference between the full sum and its truncation as

$$T_{mn} - T_{mnL} = \frac{1}{\sqrt{mn}} \sum_{i=1}^{m} \sum_{j=1}^{n} \underbrace{\left(\tilde{k}(x_i, y_j) - \sum_{l=1}^{L} \lambda_l \psi_l(x_i) \psi_l(y_j) \right)}_{g_K(x_i, y_j)}.$$

Each of the properties (Serfling, 1980, Equations (6a)-(6c) p. 197) still holds for g_K, namely

$$\begin{aligned} \mathbf{E}_{x,x'} \left(g_K(x,x') \right) &= 0, \\ \mathbf{E}_{x,x'} \left(g_K^2(x,x') \right) &= \sum_{l=L+1}^{\infty} \lambda_l^2, \\ \mathbf{E}_{x} \left(g_K(x,x') \right) &= 0. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{E}_{X,Y} \left(T_{mn} - T_{mnL} \right)^2 &= \frac{1}{mn} \sum_{i=1}^{m} \sum_{q=1}^{m} \sum_{j=1}^{n} \sum_{r=1}^{n} \mathbf{E}_{x_i, x_q, y_j, y_r} \left[g_K(x_i, y_j) g_K(x_q, y_r) \right] \\ &= \begin{cases} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{E}_{x,x'} \left(g_K^2(x,x') \right) & i = q \text{ and } j = r, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

where we have used that $p = q$ under $\mathcal{H}_0$, which allows us to replace $\mathbf{E}_{x,y}$ with $\mathbf{E}_{x,x'}$ in the final line. It follows that for large enough L,

$$\begin{aligned} |s| \left[\mathbf{E}_{X,Y} \left(T_{mn} - T_{mnL} \right)^2 \right]^{1/2} &= |s| \left[\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{E}_{x,x'} \left(g_K^2(x,x') \right) \right]^{1/2} \\ &= |s| \left[\sum_{l=L+1}^{\infty} \lambda_l^2 \right]^{1/2} \\ &< \varepsilon. \end{aligned}$$

Second term: We show that

$$T_{mnL} \underset{D}{\to} V_L \tag{29}$$

as $m \to \infty$ and $n \to \infty$. We rewrite T_{mnL} as

$$T_{mnL} = \sum_{l=1}^{L} \lambda_l \left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \psi_l(x_i) \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_l(y_j) \right).$$

Define the length L vectors W_m and W'_n having lth entries

$$W_{ml} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \psi_l(x_i), \qquad W'_{nl} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_l(y_j),$$

respectively. These have mean and covariance

$$\mathbf{E}_X(W_{ml}) = 0, \qquad \mathrm{Cov}_{X,Y}(W_{ml}, W_{ml'}) = \begin{cases} 1 & l = l', \\ 0 & l \neq l'. \end{cases}$$

Moreover, the vectors W_m and W'_n are independent. The result (29) then holds by the Lindberg-Lévy CLT (Serfling, 1980, Theorem 1.9.1A).

Third term: From Serfling (1980, p. 199), we have

$$\left|\mathbf{E}_{a,b}\left(e^{\imath sV_L}\right) - \mathbf{E}_{a,b}\left(e^{\imath sV}\right)\right| \leq |s|\left[\mathbf{E}_{a,b}\left(V - V_L\right)^2\right]^{1/2}.$$

We can bound the right hand term by

$$\begin{aligned}
\mathbf{E}_{a,b}\left(V - V_L\right)^2 &= \mathbf{E}_{a,b}\left(\sum_{l=L+1}^{\infty} \lambda_l a_l b_l\right)^2 \\
&= \sum_{l=L+1}^{\infty} \lambda_l^2 \mathbf{E}_y\left(a_l^2\right)\mathbf{E}_z\left(b_l^2\right) \\
&= \sum_{l=L+1}^{\infty} \lambda_l^2 \\
&\leq \varepsilon
\end{aligned}$$

for L sufficiently large. ■

B.2 Alternative Distribution: Consistency Against Local Alternatives

We prove Theorem 13, which gives the power against a local alternative hypothesis of a two-sample test based on MMD_u^2. The proof modifies a result of Anderson et al. (1994, Section 2.4), where we consider a more general class of local departures from the null hypothesis (rather than the class of perturbed densities described in Section 3.3.1).

First, we recall our test statistic,

$$\begin{aligned}
\mathrm{MMD}_u^2[\mathcal{F}, X, Y] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i, x_j) \\
+ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j).
\end{aligned}$$

We begin by transforming this statistic by centering the samples X and Y in feature space by μ_p and μ_q, respectively; unlike the $\mathcal{H}_0$ case, however, $\mu_p \neq \mu_q$, and the new statistic MMD_c^2 is *not* the same as MMD_u^2. The first term is centered as in (9). The second and third terms are respectively replaced by

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\langle \phi(y_i) - \mu_q, \phi(y_j) - \mu_q\right\rangle_{\mathcal{H}}$$

and

$$\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\langle \phi(x_i) - \mu_p, \phi(y_j) - \mu_q\right\rangle_{\mathcal{H}}.$$

The resulting centred statistic is

$$\mathrm{MMD}_c^2[\mathcal{F},X,Y] = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\left\langle \phi(x_i)-\mu_p, \phi(x_j)-\mu_p\right\rangle_{\mathcal{H}} + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left\langle \phi(y_i)-\mu_q, \phi(y_j)-\mu_q\right\rangle_{\mathcal{H}} - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\langle \phi(x_i)-\mu_p, \phi(y_j)-\mu_q\right\rangle_{\mathcal{H}}.$$

We write $\mu_q = \mu_p + g_t$, where $g_t \in \mathcal{H}$ is chosen such that $\mu_p + g_t$ remains a valid distribution embedding, and $\|g_t\|_{\mathcal{H}}$ can be made to approach zero to describe local departures from the null hypothesis. The difference between the original statistic and the centred statistic is then

$$\begin{aligned}
&\mathrm{MMD}_u^2[\mathcal{F},X,Y] - \mathrm{MMD}_c^2[\mathcal{F},X,Y]\\
=&\frac{2}{m}\sum_{i=1}^{m}\langle \mu_p, \phi(x_i)\rangle_{\mathcal{H}} - \langle \mu_p,\mu_p\rangle_{\mathcal{H}} + \frac{2}{n}\sum_{i=1}^{n}\left\langle \mu_q, \phi(y_i)\right\rangle_{\mathcal{H}} - \left\langle \mu_q,\mu_q\right\rangle_{\mathcal{H}}\\
&\quad - \frac{2}{m}\sum_{i=1}^{m}\left\langle \mu_q, \phi(x_i)\right\rangle_{\mathcal{H}} - \frac{2}{n}\sum_{i=1}^{n}\langle \mu_p, \phi(y_i)\rangle_{\mathcal{H}} + 2\left\langle \mu_p,\mu_q\right\rangle_{\mathcal{H}}\\
=&\frac{2}{n}\sum_{i=1}^{n}\left\langle g_t, \phi(y_i)-\mu_q\right\rangle_{\mathcal{H}} - \frac{2}{m}\sum_{i=1}^{m}\langle g_t, \phi(x_i)-\mu_p\rangle_{\mathcal{H}} + \langle g_t, g_t\rangle_{\mathcal{H}}.
\end{aligned}$$

We next show g_t can be used to encode a local departure from the null hypothesis. Define $t = m+n$, and assume $\lim_{m,n\to\infty} m/t \to \rho_x$ and $\lim_{m,n\to\infty} n/t \to \rho_y := (1-\rho_x)$ where $0 < \rho_x < 1$. Consider the case where the departure from the null hypothesis satisfies $\|g_t\|_{\mathcal{H}} = ct^{-1/2}$. Then, as $t\to\infty$,

$$t\mathrm{MMD}_c^2[\mathcal{F},X,Y] \underset{D}{\to} \sum_{l=1}^{\infty}\lambda_l\left[(\rho_x^{-1/2}a_l + \rho_y^{-1/2}b_l)^2 - (\rho_x\rho_y)^{-1}\right] =: S$$

as before, since the distance between μ_p and μ_q vanishes for large t (as $\|g_t\|_{\mathcal{H}} \to 0$). Next, the terms

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle \frac{g_t}{\|g_t\|_{\mathcal{H}}}, \phi(y_i)-\mu_q\right\rangle_{\mathcal{H}} \quad \text{and} \quad \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left\langle \frac{g_t}{\|g_t\|_{\mathcal{H}}}, \phi(x_i)-\mu_p\right\rangle_{\mathcal{H}}$$

in the difference between MMD_u^2 and MMD_c^2 are straightforward sums of independent zero mean random variables, and have Gaussian asymptotic distribution. Defining u_y to be the zero mean Gaussian random variable associated with the first term,

$$\begin{aligned}
\frac{t}{n}\sum_{i=1}^{n}\left\langle g_t, \phi(y_i)-\mu_q\right\rangle_{\mathcal{H}} &= \frac{t}{n}\left(ct^{-1/2}\right)\sum_{i=1}^{n}\left\langle \frac{g_t}{\|g_t\|_{\mathcal{H}}}, \phi(y_i)-\mu_q\right\rangle_{\mathcal{H}}\\
&\underset{D}{\to} c\rho_y^{-1/2}u_y.
\end{aligned}$$

Likewise,

$$\frac{t}{m}\sum_{i=1}^{m}\langle g_t, \phi(x_i)-\mu_p\rangle_{\mathcal{H}} \underset{D}{\to} c\rho_x^{-1/2}u_x,$$

where u_x is a zero mean Gaussian random variable independent of u_y (note, however, that u_x and u_y are correlated with terms in S, and are defined on the same probability space as a_l and b_l in this sum). Finally,

$$t\left\langle g_t, g_t\right\rangle_{\mathcal{H}} = c^2.$$

This leads to our main result: given the threshold s_α, then

$$\Pr_{\mathcal{H}_A}\left(tMMD_u^2 > s_\alpha\right) \to \Pr\left(S + 2c\left(\rho_x^{-1/2}u_x - \rho_y^{-1/2}u_y\right) + c^2 > s_\alpha\right),$$

which is constant in t, and increases as $c \to \infty$. Thus, $\|g_t\|_{\mathcal{H}} = ct^{-1/2}$ is the minimum distance between μ_p and μ_q distinguishable by the asymptotic MMD-based test.

B.3 Moments of the Empirical MMD Under $\mathcal{H}_0$

In this section, we compute the moments of the U-statistic in Section 5 for $m = n$, under the null hypothesis conditions

$$\mathbf{E}_{z,z'}h(z,z') = 0, \tag{30}$$

and, importantly,

$$\mathbf{E}_{z'}h(z,z') = 0. \tag{31}$$

Note that the latter implies the former.

Variance/2nd moment: This was derived by Hoeffding (1948, p. 299), and is also described by Serfling (1980, Lemma A p. 183). Applying these results,

$$\begin{aligned}
&\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^2\right) \\
&= \left(\frac{2}{n(n-1)}\right)^2\left[\frac{n(n-1)}{2}(n-2)(2)\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] + \frac{n(n-1)}{2}\mathbf{E}_{z,z'}\left[h^2(z,z')\right]\right] \\
&= \frac{2(n-2)}{n(n-1)}\mathbf{E}_z\left[(\mathbf{E}_{z'}h(z,z'))^2\right] + \frac{2}{n(n-1)}\mathbf{E}_{z,z'}\left[h^2(z,z')\right] \\
&= \frac{2}{n(n-1)}\mathbf{E}_{z,z'}\left[h^2(z,z')\right],
\end{aligned}$$

where the first term in the penultimate line is zero due to (31). Note that variance and 2nd moment are the same under the zero mean assumption.

3rd moment: We consider the terms that appear in the expansion of $\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^3\right)$. These are all of the form

$$\left(\frac{2}{n(n-1)}\right)^3\mathbf{E}(h_{ab}h_{cd}h_{ef}),$$

where we shorten $h_{ab} = h(z_a, z_b)$, and we know z_a and z_b are always independent. Most of the terms vanish due to (30) and (31). The first terms that remain take the form

$$\left(\frac{2}{n(n-1)}\right)^3\mathbf{E}(h_{ab}h_{bc}h_{ca}),$$

and there are

$$\frac{n(n-1)}{2}(n-2)(2)$$

of them, which gives us the expression

$$\left(\frac{2}{n(n-1)}\right)^3 \frac{n(n-1)}{2}(n-2)(2)\mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}\left(h(z,z'')h(z',z'')\right)\right]$$
$$= \frac{8(n-2)}{n^2(n-1)^2}\mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}\left(h(z,z'')h(z',z'')\right)\right]. \quad (32)$$

Note the scaling $\frac{8(n-2)}{n^2(n-1)^2} \sim \frac{1}{n^3}$. The remaining non-zero terms, for which $a = c = e$ and $b = d = f$, take the form

$$\left(\frac{2}{n(n-1)}\right)^3 \mathbf{E}_{z,z'}\left[h^3(z,z')\right],$$

and there are $\frac{n(n-1)}{2}$ of them, which gives

$$\left(\frac{2}{n(n-1)}\right)^2 \mathbf{E}_{z,z'}\left[h^3(z,z')\right].$$

However $\left(\frac{2}{n(n-1)}\right)^2 \sim n^{-4}$ so this term is negligible compared with (32). Thus, a reasonable approximation to the third moment is

$$\mathbf{E}\left(\left[\mathrm{MMD}_u^2\right]^3\right) \approx \frac{8(n-2)}{n^2(n-1)^2}\mathbf{E}_{z,z'}\left[h(z,z')\mathbf{E}_{z''}\left(h(z,z'')h(z',z'')\right)\right].$$

Appendix C. Empirical Evaluation of the Median Heuristic for Kernel Choice

In this appendix, we provide an empirical evaluation of the median heuristic for kernel choice, described at the start of Section 8: according to this heuristic, the kernel bandwidth is set at the median distance between points in the aggregate sample over p and q (in the case of a Gaussian kernel on $\mathbb{R}^d$). We investigated three kernel choice strategies: kernel selection on the entire sample from p and q; kernel selection on a hold-out set (10% of data), and testing on the remaining 90%; and kernel selection *and* testing on 90% of the available data. These strategies were evaluated on the Neural Data I data set described in Section 8.2, using a Gaussian kernel, and both the bootstrap and Pearson curve methods for selecting the test threshold. Results are plotted in Figure 7. We note that the Type II error of each approach follows the same trend. The Type II errors of the second and third approaches are indistinguishable, and the first approach has a slightly lower Type II error (as it is computed on slightly more data). In this instance, the null distribution with the kernel bandwidth set using the tested data is not substantially different to that obtained when a held-out set is used.

References

Y. Altun and A.J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proc. Annual Conf. Computational Learning Theory*, LNCS, pages 139–153. Springer, 2006.

N. Anderson, P. Hall, and D. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50:41–54, 1994.

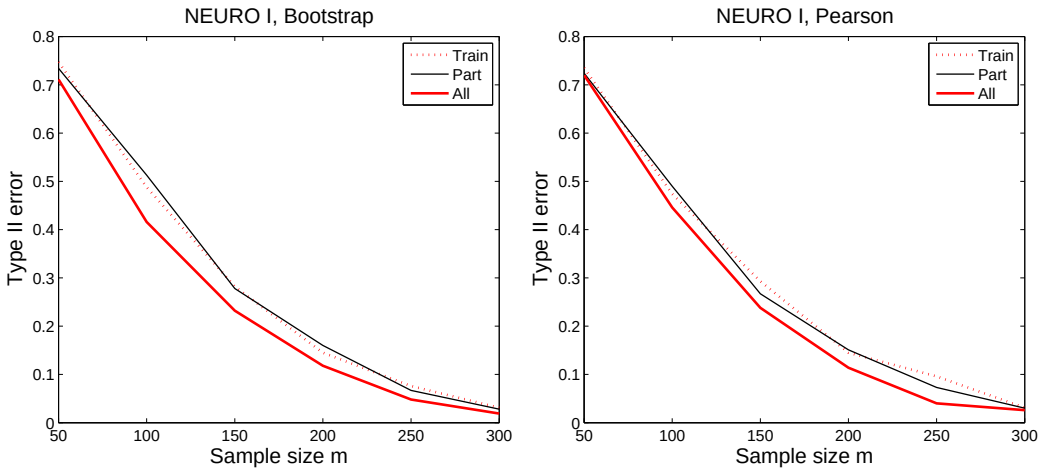

Figure 7: Type II error on the Neural Data I set, for kernel computed via the median heuristic on the full data set ("All"), kernel computed via the median heuristic on a 10% hold-out set ("Train"), and kernel computed via the median heuristic on 90% of the data ("Part"). Results are plotted over 1000 repetitions. **Left:** Bootstrap results. **Right:** Pearson curve results.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.

M. Arcones and E. Giné. On the bootstrap of *u* and *v* statistics. *The Annals of Statistics*, 20(2): 655–674, 1992.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

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

G. Biau and L. Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

P. Bickel. A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL http://www.ics.uci.edu/∼mlearn/MLRepository.html.

K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics (ISMB)*, 21(Suppl 1):i47–i56, Jun 2005.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics (ISMB)*, 22(14): e49–e57, 2006.

O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323– 375, 2005.

R. Caruana and T. Joachims. KDD cup. 2004. URL http://kodiak.cs.cornell.edu/kddcup/index.html.

G. Casella and R. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

B. Chazelle. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *Journal of the ACM*, 47:1028–1047, 2000.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 5254 of *Lecture Notes in Computer Science*, pages 38–53. Springer, 2008.

M. Davy, A. Gretton, A. Doucet, and P. J. W. Rayner. Optimized support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12):442–445, 2002.

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

M. Dudík and R. E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 123–138. Springer Verlag, 2006.

M. Dudík, S. Phillips, and R.E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 472–486. Springer Verlag, 2004.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, UK, 2002.

W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley and Sons, New York, 2nd edition, 1971.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3): 419–433, 1993.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

R. Fortet and E. Mourier. Convergence de la réparation empirique vers la réparation théorique. *Ann. Scient. École Norm. Sup.*, 70:266–285, 1953.

J. Friedman. On multivariate goodness-of-fit and two-sample testing. Technical Report SLAC-PUB-10325, University of Stanford Statistics Department, 2003.

J. Friedman and L. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):697–717, 1979.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann Publishers Inc., 2002.

E. Gokcay and J.C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–171, 2002.

A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 63–77. Springer-Verlag, 2005a.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 15*, pages 513–520, Cambridge, MA, 2007a. MIT Press.

A. Gretton, K. Borgwardt, M. Rasch, B. Schlkopf, and A. Smola. A kernel approach to comparing distributions. *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-07)*, pages 1637–1641, 2007b.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two sample problem. Technical Report 157, MPI for Biological Cybernetics, 2008a.

A. Gretton, K. Fukumizu, C.-H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, Cambridge, MA, 2008b. MIT Press.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, Red Hook, NY, 2009. Curran Associates Inc.

G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, third edition, 2001.

P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems 20*, pages 609–616. MIT Press, Cambridge, MA, 2008.

M. Hein, T.N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in SVMs. In *Proceedings of the 26th DAGM Symposium*, pages 270–277, Berlin, 2004. Springer.

N. Henze and M. Penrose. On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298, 1999.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

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

T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Proceedings of the Annual Conference on Computational Learning Theory*, volume 2777 of *LNCS*, pages 57–71, Heidelberg, Germany, 2003. Springer-Verlag.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1*. John Wiley and Sons, 2nd edition, 1994.

A. Kankainen. *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*. PhD thesis, University of Jyväskylä, 1995.

D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the International Conference on Very Large Data Bases*, pages 180–191. VLDB Endowment, 2004.

H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7: 2651–2667, 2006.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

X.L. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems 20*, pages 1089–1096. MIT Press, Cambridge, MA, 2008.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C. The Art of Scientific Computation*. Cambridge University Press, Cambridge, UK, 1994.

M. Rasch, A. Gretton, Y. Murayama, W. Maass, and N. K. Logothetis. Predicting spiking activity from local field potentials. *Journal of Neurophysiology*, 99:1461–1476, 2008.

M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis*. Academic Press, San Diego, 1980.

M. Reid and R. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

P. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society B*, 67(4):515–530, 2005.

Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

B. Schölkopf. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997. Download: http://www.kernel-machines.org.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

B. Schölkopf, K. Tsuda, and J.-P. Vert. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004.

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

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

J. Shawe-Taylor and A. Dolia. A framework for probability density estimation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 468–475, 2007.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel-based independent component analysis. *IEEE Transactions on Signal Processing*, 57:3498 – 3511, 2009.

B. W. Silverman. *Density Estimation for Statistical and Data Analysis*. Monographs on statistics and applied probability. Chapman and Hall, London, 1986.

N.V. Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow University Mathematics Bulletin*, 2:3–26, 1939. University of Moscow.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the International Conference on Machine Learning*, pages 911–918, San Francisco, 2000. Morgan Kaufmann Publishers.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 4754, pages 13–31. Springer, 2007.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the International Conference on Machine Learning*, pages 992–999. ACM, 2008.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 111–122, 2008.

B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schoelkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, Red Hook, NY, 2009. Curran Associates Inc.

B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. Lanckriet. Non-parametric estimation of integral probability metrics. In *International Symposium on Information Theory*, pages 1428 – 1432, 2010a.

B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010b.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011a.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint. In *Advances in Neural Information Processing Systems 24*. Curran Associates Inc., Red Hook, NY, 2011b.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

I. Takeuchi, Q. V. Le, T. Sears, and A. J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7, 2006.

D. M. J. Tax and R. P. W. Duin. Data domain description by support vectors. In *Proceedings ESANN*, pages 251–256, Brussels, 1999. D Facto.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

J. E. Wilkins. A note on skewness and kurtosis. *The Annals of Mathematical Statistics*, 15(3): 333–335, 1944.

C. K. I. Williams and M. Seeger. Using the Nystrom method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA, 2001. MIT Press.