A Kernel Two-Sample Test for Functional Data

WORK IN PROGRESS PUT ONLINE FOR MLSS TÜBINGEN 2020

George Wynne[1] and Andrew B. Duncan[1,2]

[1]Imperial College London, Department of Mathematics
[2]The Alan Turing Institute

Abstract

Kernel two-sample tests have shown good empirical performance and facilitate statistical tests for non-Euclidean data. This paper investigates the situation where functional data is observed, meaning each data point corresponds to a discretised function. We highlight issues of the naive application of the kernel based test resulting from a Euclidean view of the data, which motivates a functional view point. This motivates a broad new family of kernels over function spaces. Theoretical properties of this new family are investigated and numerical experiments performed on a range of synthetic and real world data.

1 Introduction

Non-parametric two-sample tests for equality of distributions are widely studied in statistics, driven by applications in goodness-of-fit tests, anomaly and change-point detection and clustering. Classical examples of such tests include the Kolmogorov-Smirnov test [33, 58, 51] and Wald-Wolfowitz runs test [73] with subsequent multivariate extensions, such as [18]. This paper deals with the scenario where each data sample is the discretisation of a function, for example a time series or random surface. This presents challenges for the application of classical tests since the discretisation often results in high dimensional data and since the data is function it has properties not present in standard data in $\mathbb{R}^d$ such as temporal differences, periodicity and smoothness. It is therefore necessary to expand upon the classical tests when dealing with such data.

Functional data analysis [27, 29] is the study of data which arises as a discretisation of a random function. Techniques are employed which view the random function as a random variable in a function space as opposed to viewing the discretisation as a high dimensional vector [29]. To handle the computation of objects in infinite dimensional spaces, functional principal components [27] are commonly computed to represent the functions as finite dimensional objects. Existing two-sample tests for functional data commonly apply a classical finite dimensional test to this reduced dimension representation or apply a distance based test statistic and commonly focus on only a

difference of mean or covariance structure, as opposed to an arbitrary difference in the two distributions. A non-exhaustive list of existing works include [43] which used the Anderson-Darling test to test for arbitrary difference in distributions, [28] used an L^2 based distance statistic to detect difference in mean, [40] used a Hilbert-Schmidt norm based statistic to test for different covariance structure, [24] used a functional Cramer Von-Mises statistic too test for arbitrary difference in distributions and an approach based on the Schilling test and the Wilcoxon test was proposed in [7].

In the past two decades kernels have seen a surge of use in statistical applications [37, 22, 67, 6]. In particular, kernel based two-sample testing [22, 21] has become a popular method for two-sample testing. A kernel is employed to facilitate a mapping, called a mean-embedding, of the two distributions being tested for equality into an infinite dimensional reproducing kernel Hilbert space and then distance is measured in that space, called the maximum mean discrepancy (MMD). The potency of this technique is that the kernel trick can be used to represent this distance in terms of integrals of the kernel, which may be easily and efficiently be approximated by a Monte Carlo estimate using samples from the two distributions. In fact, the formulation of this feature expansion based approach is equivalent to an integral probability metric representation. The advantage of kernel based tests is that there is no assumption made on the distributions and kernels can be constructed on arbitrary input spaces, not just Euclidean spaces, such as graphs and strings [55, 19] meaning two-sample tests over these non-standard domains can be performed. An important point is that the performance of the test greatly depends on the kernel and the domain. In the case of Euclidean space some work has been completed to investigate the performance of kernel based tests in high dimensions [46, 48] but it still remains a largely open question. In particular, the theoretical properties of the kernel based tests when performed using functional data has received very little attention, despite large activity on the problem in the FDA literature [40, 43, 24, 28].

The first step in such analysis would be the proposal of an appropriate kernel over a function space. In [10] a kernel which uses the signature feature map was introduced but did not build upon existing theory of kernel based tests. An alternative approach would be to adapt an existing kernel on $\mathbb{R}^d$ but with a function space distance replacing the Euclidean distance. This was done for a Gauss kernel in [11] which employed an L^2 distance but in the context of kernel ridge regression, not kernel two-sample testing. Indeed, it is not obvious that an L^2 distance is appropiate for statistical testing. Related to kernel based tests are energy distance tests [68, 69], the relationship was made clear in [54]. Investigation into the impact of the choice of distance in distanced based tests for functional data is studied in [9, 8, 77] and a distance other than L^2 for the functional data was advocated. This motivates the investigation into kernels which involve distances other than L^2 in their formulation.

The aim of this paper is to build upon the finite dimensional theory of high-dimensional kernel based two-sample tests to provide evidence that when dealing with discretised functional data, kernels over function spaces are the appropiate tool. We propose a new class of kernels over function spaces and then investigate their statistical properties and practical implementation. Our specific contributions are the following

1. Section 4 identifies the scaling effects caused by functional data when observations are increasingly dense and the two scaling regimes caused by pointwise independent and dependent randomness in the functions.

2. Section 5 defines a broad class of kernels over Hilbert spaces as natural generalisations of kernels over $\mathbb{R}^d$ and prove they lead to valid tests that can identify arbitrary difference in distributions. A criteria for convergence of MMD implying weak convergence is stated.
3. Section 6 provides a statistical description of the effect of fitting curves to the discretised functional data before the test is performed, highlights the relationship between weak convergence and MMD and gives closed form expressions for MMD and mean-embeddings when the two distributions are Gaussian processes.
4. Section 7 outlines the impact on testpower of our kernels and multiple examples of valid kernels.
5. Section 8 contains multiple numerical simulations using real and synthetic data to reinforce the theoretical arguments made and compare to existing methods.

The remainder of the paper is as follows. Section 2 covers preliminaries of modelling random functional data such as the Karhunen-Loève expansion and Gaussian measures. Section 3 recalls some important properties of kernels and their associated reproducing kernel Hilbert spaces, defines maximum mean discrepancy and the kernel two-sample test. Section 4 outlines the scaling of test power that occurs when an increasingly finer observation mesh is used for functional data. Two distinct scalings are highlighted, when the function randomness is independent identically distributed pointwise random noise versus dependent pointwise noise. Section 5 defines a broad class of kernels and offers an integral feature map interpretation of them as well as outlining when the kernels are characteristic, meaning the two-sample test is valid. Section 6 highlights the statistical impact of fitting curves to discretised functions before performing the test. A relationship between MMD and weak convergence is highlighted and closed form expressions for the MMD and mean-embeddings when the distributions are Gaussian processes are given. Section 7 discusses the impact of test power of certain hyper parameters, here the hyperparameter being a choice of operator that is used in the kernel, and gives multiple examples of choices for the kernel hyper parameters and a computational speed up using random Fourier features. Section 8 contains multiple numerical experiements validating the theory in the paper, a simulation is performed to validate the scaling arguments of Section 4 and synthetic and real data sets are used to compare the performance of the kernel based test against existing functional two-sample tests.

2 Hilbert Space Modelling of Functional Data

In this paper we shall follow the Hilbert space approach to functional data analysis and use this section to outline the required preliminaries, see [12, 29] for a comprehensive review. Before discussing random functions we establish notation for families of operators that will be used extensively. Let $\mathcal{X}$ be a real, separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ then $L(\mathcal{X})$ denotes the set of bounded linear maps from $\mathcal{X}$ to itself, $L^+(\mathcal{X})$ denotes the subset of $L(\mathcal{X})$ of operators that are self-adjoint (also known as symmetric) and non-negative, meaning $\langle Tx, y \rangle \geq 0 \, \forall x, y \in \mathcal{X}$. The subset of $L^+(\mathcal{X})$ of trace class operators is denoted $L_1^+(\mathcal{X})$ and by the spectral theorem [63, Theorem A.5.13] such operators can be diagonalised. This means for every

$T \in L_1^+(\mathcal{X})$ there exists an orthonormal basis of eigenfunctions $\{e_n\}_{n=1}^{\infty}$ in $\mathcal{X}$ such that $Tx = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n$, where $\{\lambda_n\}_{n=1}^{\infty}$ are non-negative eigenvalues and the trace satisfies $\mathrm{Tr}(T) = \sum_{n=1}^{\infty} \lambda_n < \infty$.

We now outline the Karhunen-Loève expansion of stochastic processes. Let $x(\cdot)$ be a stochastic process in $\mathcal{X} = L^2([0,1])$, note the following will hold for a stochastic process taking values in any real, separable Hilbert space but we focus on $L^2([0,1])$ since it is the most common setting for functional data. Suppose that the pointwise covariance function $\mathbb{E}[x(s)x(t)] = k(s,t)$ is continuous, then the mean function $m(t) = \mathbb{E}[X(t)]$ is also in $\mathcal{X}$. Define the covariance operator associated with X by $C_k \colon \mathcal{X} \to \mathcal{X}$, $C_k y(t) = \int_0^1 k(s,t)y(s)ds$. Then $C_k \in L_1^+(\mathcal{X})$ and denote the spectral decomposition $C_k y = \sum_{n=1}^{\infty} \lambda_n \langle y, e_n \rangle e_n$. The Karhunen-Loève (KL) expansion [65, Theorem 11.4] provides a characterisation of the law of the process $x(\cdot)$ in terms of an infinite-series expansion. More specifically, we can write

$$x(\cdot) \sim m + \sum_{n=1}^{\infty} \lambda_n^{\frac{1}{2}} \eta_n e_n(\cdot)$$

where $\{\eta_n\}_{n=1}^{\infty}$ are unit-variance uncorrelated random variables. Additionally, Mercer's theorem [64] provides an expansion of the covariance as $k(s,t) = \sum_{n=1}^{\infty} \lambda_n e_n(s) e_n(t)$ where the convergence is uniform.

An important case of random functions are Gaussian processes [49]. Given a positive definite kernel k and a function m we say x is a Gaussian process with mean function m and covariance function k if for every finite collection of points $\{s_n\}_{n=1}^{N}$ the random vector $(x(s_1), \ldots, x(s_N))$ is a multivariate Gaussian random variable with mean vector $(m(s_1), \ldots, m(s_N))$ and covariance matrix $k(s_n, s_m)_{n,m=1}^{N}$. The mean function and covariance function completely determines the Gaussian process. We write $x \sim \mathcal{GP}(m,k)$ to denote the Gaussian process with mean function m and covariance function k. If $x \sim \mathcal{GP}(0,k)$ then in the Karhunen-Loève representation $\eta_n \sim \mathcal{N}(0,1)$ and the η_n are all independent.

Gaussian processes that take values in $\mathcal{X}$ can be associated with Gaussian measures on $\mathcal{X}$. Gaussian measures are natural generalisations of Gaussian distributions on $\mathbb{R}^d$ to infinite dimensional spaces, which are defined by a mean element and covariance operator rather than a mean vector and covariance matrix, for an introduction see [44, Chapter 1]. Specifically $x \sim \mathcal{GP}(m,k)$ can be associated with the Gaussian measure N_{m,C_k} with mean m and covariance operator C_k, the covariance operator associated with k as outlined above. Similarly given any $m \in \mathcal{X}$ and $C \in L_1^+(\mathcal{X})$ then there exists a Gaussian measure $N_{m,C}$ with mean m and covariance operator C [44, Theorem 1.12]. In fact, the Gaussian measure $N_{m,C}$ is characterised as the unique probability measure on $\mathcal{X}$ with Fourier transform $\widehat{N}_{m,C}(y) = \exp(i\langle m, y \rangle - \frac{1}{2}\langle Cy, y \rangle)$. Finally, if C is injective then a Gaussian measure with covariance operator C is called non-degenerate and has full support on $\mathcal{X}$ [44, Proposition 1.25].

3 Reproducing Kernel Hilbert Spaces and Maximum Mean Discrepancy

This section will outline what a kernel and a reproducing kernel Hilbert space is with examples and associated references. The backbone of the kernel two-sample test, max-

imum mean discrepency, is outlined and the associated estimators and testing procedure is discussed.

3.1 Kernels and Reproducing Kernel Hilbert Spaces

Given a nonempty set $\mathcal{X}$ a kernel is a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is symmetric, meaning $k(x, y) = k(y, x)$, for all $x, y \in \mathcal{X}$, and positive definite, that is, the matrix $\{k(x_n, x_m);\, n, m \in \{1, \ldots, N\}\}$ is positive semi-definite, for all $\{x_n\}_{n=1}^N \subset \mathcal{X}$ and for $N \in \mathbb{N}$. For each kernel k there is an associated Hilbert space of functions over $\mathcal{X}$ known as the reproducing kernel Hilbert space (RKHS) denoted $\mathcal{H}_k(\mathcal{X})$ [3, 63, 16]. RKHSs have found numerous applications in function approximation and inference for decades since their original application to spline interpolation [72]. For a detailed survey, refer to [50, 42]. The RKHS associated with k satisfies the following two properties i). $k(\cdot, x) \in \mathcal{H}(\mathcal{X})$ for all $x \in \mathcal{X}$ ii). $\langle f, k(\cdot, x)\rangle_{\mathcal{H}(\mathcal{X})} = f(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{H}(\mathcal{X})$. The latter is known as the reproducing property. The RKHS is constructed from the kernel in a natural way. The linear span of a kernel k with one input fixed, i.e,

$$\mathcal{H}_0(\mathcal{X}) = \left\{ \sum_{n=1}^{N} a_n k(\cdot, x_n) \colon N \in \mathbb{N}, \{a_n\}_{n=1}^N \subset \mathbb{R}, \{x_n\}_{n=1}^N \subset \mathcal{X} \right\}$$

is a pre-Hilbert space equipped with the following inner product

$$\langle f, g\rangle_{\mathcal{H}_0(\mathcal{X})} = \sum_{n=1}^{N} \sum_{m=1}^{M} a_n b_m k(x_n, y_m)$$

where $f = \sum_{n=1}^N a_n k(\cdot, x_n)$ and $g = \sum_{m=1}^M b_m k(\cdot, y_m)$. The RKHS $\mathcal{H}_k(\mathcal{X})$ of k is then obtained from $\mathcal{H}_0(\mathcal{X})$ through completion. More specifically $\mathcal{H}_k(\mathcal{X})$ is the set of functions which are pointwise limits of Cauchy sequences in $\mathcal{H}_0(\mathcal{X})$ [3, Theorem 3]. The relationship between kernels and RKHS is one-to one, for every kernel the RKHS is unique and for every Hilbert space of functions such that there exists a function k satisfying the two properties above it may be concluded that the k is unique and a kernel. This result is known as the Aronszajn theorem [3, Theorem 3].

A kernel k on $\mathcal{X} \subseteq \mathbb{R}^d$ is said to be *translation-invariant* if it can be written as $k(x, y) = \phi(x - y)$ for some ϕ. Bochner's theorem, Theorem 8 in the Appendix, tells us that if k is continuous and translation invariant then there exists a Borel meaure on $\mathcal{X}$ such that $\hat{\mu}_k(x - y) = k(x, y)$ and we call μ_k the spectral measure of k. The spectral measure is an important tool in the analysis of kernel methods and shall become important later when discussing the two-sample problem.

We now give multiple examples of kernels over $\mathbb{R}^d$, function spaces and different structured domains. The most commonly used kernel over $\mathbb{R}^d$ is the Gauss kernel $k(x, y) = e^{-\|x-y\|^2/2\gamma^2}$ which is widely used in regression and classification and has accumulated a large amount of theoretical interest, see for example [63, Chapter 4.4]. Another kernel on $\mathbb{R}^d$ is the Matérn kernel which is a popular choice in the spatial modelling literature. Not many kernels over function spaces have been theoretically investigated, a Gauss type kernel whose input is a compact subset of $L^2([0, 1])$ was studied in [11, Example 3] and a kernel based on the signature feature map was proposed in [10]. Kernels over probability distributions were studied in [26] and over

graphs in [53] and references for more examples of kernels over structured domains are [55, 19].

3.2 Maximum Mean Discrepancy

Given a kernel k and associated RKHS $\mathcal{H}_k(\mathcal{X})$ let $\mathcal{P}$ be the set of Borel probability measures on $\mathcal{X}$ and assuming k is measurable define $\mathcal{P}_k \subset \mathcal{P}$ such that $\int k(x,x)^{\frac{1}{2}} dP(x) < \infty$ for all $P \in \mathcal{P}_k$. Note that $\mathcal{P}_k = \mathcal{P}$ if and only if k is bounded [62, Proposition 2] which is very common in practise and shall be the case for all kernels considered in this paper. For $P, Q \in \mathcal{P}_k$ we define the *Maximum Mean Discrepancy* denoted $\mathrm{MMD}_k(P,Q)$ as follows

$$\mathrm{MMD}_k(P,Q) = \sup_{\|f\|_{\mathcal{H}_k(\mathcal{X})} \leq 1} \left| \int f dP - \int f dQ \right|$$

This is an *integral probability metric* [38, 62] and without further assumptions defines a pseudo-metric on $\mathcal{P}_k$, which permits the possibility that $\mathrm{MMD}_k(P,Q) = 0$ but $P \neq Q$.

We introduce the *mean embedding* $\Phi_k P$ of $P \in \mathcal{P}_k$ into $\mathcal{H}_k(\mathcal{X})$ defined by $\Phi_k P = \int k(\cdot, x) dP(x)$. This can be viewed as the mean in $\mathcal{H}_k(\mathcal{X})$ of the function $k(x,\cdot)$ with respect to P in the sense of a Bochner integral [29, Section 2.6]. Following [62, Section 2] this allows us to write

$$\begin{aligned}
\mathrm{MMD}_k(P,Q)^2 &= \left(\sup_{\|f\|_{\mathcal{H}_k(\mathcal{X})} \leq 1} \left| \int f dP - \int f dQ \right| \right)^2 \\
&= \left(\sup_{\|f\|_{\mathcal{H}_k(\mathcal{X})} \leq 1} |\langle \Phi_k P - \Phi_k Q, f \rangle| \right)^2 \\
&= \|\Phi_k P - \Phi_k Q\|^2_{\mathcal{H}_k(\mathcal{X})} \qquad (1)
\end{aligned}$$

The crucial observation which motivates the use of MMD as an effective measure of discrepancy is that the supremum can be eliminated using the reproducing property of the inner product [62, Section 2]. This yields the following closed form representation

$$\begin{aligned}
\mathrm{MMD}_k(P,Q)^2 = &\int \int k(x,x') dP(x) dP(x') + \int \int k(y,y') dQ(y) dQ(y') \\
&- 2 \int \int k(x,y) dP(x) dQ(y) \qquad (2)
\end{aligned}$$

It is clear that MMD_k is a metric over $\mathcal{P}_k$ if and only if the map $\Phi_k \colon \mathcal{P}_k \to \mathcal{H}_k(\mathcal{X})$ is injective. Given a subset $\mathfrak{P} \subseteq \mathcal{P}_k$, a kernel is *characteristic to* $\mathfrak{P}$ if the map Φ_k is injective over $\mathfrak{P}$. In the case that $\mathfrak{P} = \mathcal{P}$ we just say that k is characteristic. Various works have provided sufficient conditions for a kernel over finite dimensional spaces to be characteristic, see for example [62, 61, 57]. One important advantage of the kernel based approach is that properties of the estimator are agnostic of the input space, so even in the main scenario of this paper where $\mathcal{X}$ is infinite dimensional we may continue to use many helpful results to understand the properties of the estimator.

Given independent samples $X_n = \{x_i\}_{i=1}^n$ from P and $Y_m = \{y_i\}_{i=1}^m$ from Q we wish to estimate $\mathrm{MMD}_k(P,Q)^2$. A number of estimators have been proposed. For

clarity of presentation we shall assume that $m = n$, but stress that all of the following can be generalised to situations where the two data-sets are unbalanced. Given samples X_n and Y_n, the following U-statistic is an unbiased estimator of $\text{MMD}_k^2(P, Q)^2$

$$\widehat{\text{MMD}}_k(X_n, Y_n)^2 \coloneqq \frac{1}{n(n-1)} \sum_{i \neq j}^{n} h(z_i, z_j) \tag{3}$$

where $z_i = (x_i, y_i)$ and $h(z_i, z_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$. This estimator can be evaluated in $O(n^2)$ time. A unbiased linear time estimator proposed in [30] is given by

$$\widehat{\text{MMD}}_{k,\text{lin}}(X_n, Y_n)^2 \coloneqq \frac{2}{n} \sum_{i=1}^{n/2} h(z_{2i-1}, z_{2i}), \tag{4}$$

where we assume that n is even. While the cost for computing $\widehat{\text{MMD}}_{k,\text{lin}}(X_n, Y_n)^2$ is only $O(n)$ this comes at the cost of reduced efficiency, i.e. $\text{Var}(\widehat{\text{MMD}}_k(X_n, Y_n)^2) < \text{Var}(\widehat{\text{MMD}}_{k,\text{lin}}(X_n, Y_n)^2)$, see for example [66]. Various probabilistic bounds have been derived on the error between the estimator and $\text{MMD}_k(P, Q)^2$, including [22, Theorem 10, Theorem 15]. In particular, both estimators are consistent and asymptotically normal.

3.3 The Kernel Two-Sample Test

Given independent samples $X_n = \{x_i\}_{i=1}^n$ from P and $Y_n = \{y_i\}_{i=1}^n$ from Q we seek to test the hypothesis $H_0 \colon P = Q$ against the alternative hypothesis $H_1 \colon P \neq Q$ without making any distributional assumptions. The *kernel two-sample test* of [22] employs an estimator of MMD as the test statistic. More specifically, fixing a characteristic kernel k, we reject H_0 if

$$\widehat{\text{MMD}}_k(X_n, Y_n)^2 > c_\alpha,$$

where c_α is a threshold selected to ensure a false-positive rate of α. While we do not have a closed-form expression for c_α, it can be estimated using a permutation bootstrap. More specifically, we repeatedly shuffle $X_n \cup Y_n$, each time splitting into two data sets X_n' and Y_n', from which $\widehat{\text{MMD}}_k(X_n', Y_n')^2$ is calculated. An estimator of the threshold $\hat{c}_\alpha$ is then obtained as the $(1-\alpha)$-th quantile of the resulting empirical distribution. The same test procedure may be performed using the linear time MMD estimator as the test statistic.

4 Power of Kernel Two-Sample Testing

A test is characterised by its false-positive rate α and its its false-negative rate β. Decreasing one rate typical results in the other rate increasing. The power of a test is a measure of its ability to correctly reject the null hypothesis. More specifically, fixing α, one defines the power of the test at α to be $\phi = 1 - \beta$. A test is *consistent* if $\phi \to 1$ as $n \to \infty$. It is evident from previous works that the properties of the kernel will have a very significant impact on the power of the test. The MMD estimator is

asymptotically normal, under the alternative hypothesis, for large n, the power of the test satisfies

$$\mathbb{P}\left(n\widehat{\text{MMD}}_k(X_n, Y_n)^2 > \widehat{c}_\alpha\right) \approx \Phi\left(\sqrt{n}\frac{\text{MMD}_k(P,Q)^2}{2\sqrt{\xi}} - \frac{c_\alpha}{2\sqrt{n\xi}}\right) \tag{5}$$

where Φ is the CDF for a standard Gaussian distribution and $\xi = \xi_1 = \text{Var}_z\left[\mathbb{E}_{z'}[h(z,z')]\right]$ for the quadratic-time estimator, and $\xi = \xi_2 = \text{Var}_{z,z'}\left[h(z,z')\right]$ if the linear time estimator is used instead [48, 34]. It follows that test power can be maximised by maximising $\text{MMD}_k(P,Q)^2/2\sqrt{\xi}$ which can be seen as a signal-to-noise-ratio [34]. In previous works, methods have been proposed for increasing test power by optimising the kernel parameters using the signal-to-noise-ratio as an objective [67, 48, 34].

4.1 Testing in High Dimensions

To further develop an understanding of the power of the two-sample test in high dimensions, as well as the of the influence of the kernel, we shall focus on the Gaussian kernel. We focus on the signal-to-noise of the linear time estimator since the expressions are simpler and it is widely used in practise, the same asymptotic behaivour occurs with the quadratic time estimator. Following [48] we seek to characterise the power of a test under a shift-of-mean alternative. More specifically, let $P = \mathcal{N}(0, \Sigma)$ and $Q = \mathcal{N}(\mu, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite and $\mu \in \mathbb{R}^d$. Given independent samples X_n and Y_n from P and Q respectively, we seek to test whether $P = Q$ against the alternative hypothesis that $\mu \neq 0$. We focus on approximating quantities associated with the linear time test for simplicity, similar results hold for the quadratic time test. The following result approximates the signal-to-noise ratio in the $d \to \infty$ limit.

Lemma 1. *Let P and Q be as above and suppose that $k(x,y) = \exp(-\frac{1}{2\gamma^2}\|x-y\|^2)$ then for d large, if $\gamma^2 \asymp d$*

$$\frac{\text{MMD}_k(P,Q)^2}{\sqrt{2\xi_2}} \approx \frac{\|\mu\|^2}{\sqrt{8\text{Tr}(\Sigma^2) + 8\mu^T\Sigma\mu}} \tag{6}$$

The proof is deferred to Section 9.2 in the Appendix. Combining this with Equation 5 means we may view Equation 6 as a quantity to optimise if we wish to increase the power of the test when dealing with such P, Q. Note that in the case $\Sigma = \sigma^2 I_{d\times d}$ we recover the result detailed in [48, Theorem 1]. Indeed, in this situation as $d \to \infty$,

$$\frac{\text{MMD}_k(P,Q)^2}{\sqrt{2\xi_2}} \approx \frac{\|\mu\|^2}{\sqrt{8d\sigma^4 + 8\sigma^2\|\mu\|^2}}.$$

As noted informally in [48, 47], the scaling $\gamma^2 \asymp d$ coincides with that arising from the *median heuristic* where γ^2 is chosen to be the median of the pairwise square distances between all pairs of samples from P and Q. More explicitly, given samples of random vectors in $\mathbb{R}^d$ denoted $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^m$ from P, Q respectively then the median heuristic amounts to setting

$$\gamma^2 = \text{Median}\big\{\|a-b\|^2 \colon a, b \in \{x_i\}_{i=1}^n \cup \{y_i\}_{i=1}^m, a \neq b\big\} \tag{7}$$

When P and Q are Gaussian we can prove the following bounds on the scaling of the median heuristic with respect to dimension.

Lemma 2. *Let $P = \mathcal{N}(\mu_1, \Sigma_1)$ and $Q = \mathcal{N}(\mu_2, \Sigma_2)$ be independent Gaussian distributions on $\mathbb{R}^d$ such that $\mathrm{Tr}(\Sigma_i) \asymp d$ and $\|\mu_i\|^2 \lesssim d$ for $i = 1, 2$ and d large then*

$$\mathrm{Median}\{\|x - y\|^2, x \sim P, y \sim Q\} \asymp d$$

In line with the conclusions of [46] Lemmas 1 and 2 indicate that, at least for the case of testing equality of means for multivariate Gaussians, when using a Gaussian kernel the median heuristic provides a good choice for the lengthscale parameter γ^2.

4.2 Testing for Discretised Functional Data

We now consider the case where the data arises as discretisations of random functions. We assume the random functions lie in $L^2([0,1]^r)$ with probability one. Assume that for each realisation we observe N input output pairs for each function realisation as an N dimensional vector $x_N = (x(s_1), \ldots, x(s_N))$ and assume that the points cover $[0,1]^r$ regularly enough so that the Riemann sum property $\frac{1}{N}\sum_{i=1}^N x(s_i)^2 \to \int_{[0,1]^r} x(s)^2\, ds$ holds as $N \to \infty$. For example the observations locations may be a uniform grid on $[0,1]^r$ with M divisions in each dimension, meaning $N = M^r$ and $N \to \infty$ as $M \to \infty$ meaning the discretisation in each dimension gets finer. To relate this change of notation to the previous subsection we see that the dimension of the random vector here is the number of observations of each random function, therefore increasing the dimension corresponds to increasing the sampling resolution of the functions.

To understand the influence of the sampling resolution on the test power, let us consider the simple problem of detecting a non-zero mean given Gaussian process observations. That is, we consider $P = \mathcal{GP}(0, k_0)$ and $Q = \mathcal{GP}(m, k_0)$ where $0 \neq m \in L^2([0,1]^r)$ and k_0 is continuous. The probability measures associated to the discretised data are $P_N = \mathcal{N}(0, K_N)$ and $Q_N = \mathcal{N}(m_N, K_N)$, respectively, where $K_N = (k_0(s_i, s_j))_{i,j=1}^N$ and $m_N = (m(s_1), \ldots, m(s_N))$. By Lemma 2 therefore $\gamma^2 \asymp N$, the conditions of Lemma 2 are met since for N large $\|m_N\| \approx N\|m\|_{L^2([0,1]^r)}$ by Riemann sum scaling and the sum of the matrix eigenvalues is approximately equal to N multiplied by the sum of the eigenvalues of the kernel obtained from Mercer's theorem [56, Section 1].

Similar scaling arguments yield

$$\mathrm{Tr}(K_N^2) \approx N^2 \int_{[0,1]^r}\int_{[0,1]^r} k_0(s,t)dsdt = N^2\|C_{k_0}\|_{HS}^2$$
$$m_N^\top K_N m_N \approx N^2 \int_{[0,1]^r}\int_{[0,1]^r} m(s)k_0(s,t)m(t)dsdt = N^2\|C_{k_0}^{1/2}m\|_{L^2([0,1]^r)}^2$$

where C_{k_0} is the covariance operator associated with k_0 and $\|\cdot\|_{HS}$ denotes the Hilbert Schmidt norm. Plugging these terms into (6) we obtain, for large N,

$$\frac{\mathrm{MMD}_{k_N}(P_N, Q_N)^2}{\sqrt{2\xi_2}} \approx \frac{\|m\|_{L^2([0,1]^r)}^2}{\sqrt{8\|C_{k_0}\|_{HS}^2 + 8\|C_{k_0}^{1/2}m\|_{L^2([0,1]^r)}^2}} \tag{8}$$

Letting $\{\lambda_n\}_{n=1}^{\infty}$ be the eigenvalues associated with the eigenbasis $\{e_n\}_{n=1}^{\infty}$ for C_{k_0} then we can rewrite Equation (8) as

$$\frac{\text{MMD}_{k_N}(P_N, Q_N)^2}{\sqrt{2\xi_2}} \approx \frac{\sum_{n=1}^{\infty} m_n^2}{\sqrt{8\sum_{n=1}^{\infty}\lambda_n^2 + 8\sum_{n=1}^{\infty}\lambda_n m_n^2}}, \tag{9}$$

where $m_n = \langle m, e_n\rangle_{L^2([0,1]^r)}$. Equation 9 show that, for this scaling of γ, as $N \to \infty$ the signal to noise ratio converges to a quantity that does not depend on N nor on the actual locations of the points, as long as the Riemann sum property holds. Since Equation 9 relates to the power of the test through Equation 5 we see that asymptotically the statistical power is independent of the sampling resolution.

These calculations yield some important observations about two-sample kernel tests for functional data. Firstly, the signal-to-noise ratio increases as the noise becomes decorrelated. Indeed, consider the extreme case where the noise is white, so that the kernel is $k_0(s,t) = \sigma^2\delta_{st}$ meaning $K_N = \sigma^2 I_{N\times N}$. This kernel is not continuous so we cannot use the same calculations performed to obtain Equation 8 but instead we may directly substitute into Equation 6 the quantities related to K_N in this case to see

$$\frac{\text{MMD}_{k_N}(P_N, Q_N)^2}{\sqrt{2\xi_2}} \approx \frac{\sqrt{N}\|m\|^2_{L^2([0,1]^r)}}{\sqrt{8\sigma^4 + 8\sigma^2\|m\|^2_{L^2([0,1]^r)}}} \tag{10}$$

meaning the power increases as the mesh-size increases. This coincides with the intuition that a for white noise processes, refining the discretisation mesh will always yield more information about a particular realisation since each observation reveals new information as the noise is independent. Numerical simulation of this phenomenon is performed in Section 8.

Secondly, choosing the scaling $\gamma^2 = \gamma_0^2 N$ for some $\gamma_0 > 0$ results in the following functional form for the kernel k_N as N becomes large, assuming the inputs are x_N, y_N the discretisations of some $x, y \in L^2([0,1]^r)$

$$k_N(x_N, y_N) = \exp\left(-\frac{1}{\gamma_0^2 N}\sum_{i=1}^{N}(x(s_i) - y(s_i))^2\right) \approx \exp\left(-\frac{1}{\gamma_0^2}\|x - y\|^2_{L^2([0,1]^r)}\right)$$

We note that this is the only scaling of γ with respect to N for which such a functional form holds for the kernel.

To summarise we see that if one observes a discretised version of random functions and employs the finite dimensional kernel test with $\gamma^2 \asymp N$ then, as long as the observation points satisfy the Riemann sum property, as the number of observations increases the test power will become independent of the sampling resolution and can be expressed in terms of function space norms. Additionally, this choice of scaling for γ^2, which occurs when employing the commonly used median heuristic, causes the kernel to converge to a quantity that depends on function space norms. This motivates defining kernels directly on functions lying in the Hilbert space $L^2([0,1]^r)$, or indeed other Hilbert spaces, and studying the associated kernel two-sample test for distributions over function spaces.

5 Kernels and RKHS on Function Spaces

For the rest of the paper, unless specified otherwise, for example in Theorem 3, the spaces $\mathcal{X}, \mathcal{Y}$ will be real, separable Hilbert spaces with inner products and norms $\langle \cdot, \cdot \rangle_{\mathcal{X}}, \langle \cdot, \cdot \rangle_{\mathcal{Y}}, \|\cdot\|_{\mathcal{X}}, \|\cdot\|_{\mathcal{Y}}$ and when $\mathcal{X} = \mathcal{Y}$, for example in Theorem 1, we will denote the inner product and norm by $\langle \cdot, \cdot \rangle, \|\cdot\|$. We adopt the notation in Section 2 for various families of operators.

5.1 SE-T kernel

Motivated by the scaling discussions in Section 4 we define a kernel that acts directly on a Hilbert space.

Definition 1. *For $T\colon \mathcal{X} \to \mathcal{Y}$ the squared-exponential T kernel (SE-T) is defined as*

$$k_T(x, y) = e^{-\frac{1}{2}\|T(x)-T(y)\|_{\mathcal{Y}}^2}$$

We use the name squared-exponential instead of Gauss because the SE-T kernel is not always the Fourier transform of a Gaussian distribution whereas the Gauss kernel on $\mathbb{R}^d$ is, which is a key distinction and is relevant for our proofs. Lemma 3 assures us this function is a kernel. This definition allows us to adapt results about the Gauss kernel on $\mathbb{R}^d$ to the SE-T kernel since it is the natural infinite dimensional generalisation. For example the following theorem characterises the RKHS of the SE-T kernel for a certain choice of T, as was done in the finite dimensional case in [36]. Before we state the result we introduce the infinite dimensional generalisation of a multi-index, define Γ to be the set of summable sequences indexed by $\mathbb{N}$ taking values in $\mathbb{N} \cup \{0\}$ and for $\gamma \in \Gamma$ set $|\gamma| = \sum_{n=1}^{\infty} \gamma_n$, so $\gamma \in \Gamma$ if and only if $\gamma_n = 0$ for all but finitely many $n \in \mathbb{N}$ meaning Γ is a countable set. We set $\Gamma_n = \{\gamma \in \Gamma\colon |\gamma| = n\}$ and the notation $\sum_{|\gamma|\geq 0}$ shall mean $\sum_{n=0}^{\infty} \sum_{\gamma \in \Gamma_n}$ which is a countable sum.

Theorem 1. *Let $T \in L^+(\mathcal{X})$ be of the form $Tx = \sum_{n=1}^{\infty} \lambda_n^{1/2} \langle x, e_n \rangle e_n$ with convergence in $\mathcal{X}$ for some orthonormal basis $\{e_n\}_{n=1}^{\infty}$ and bounded positive coefficients $\{\lambda_n\}_{n=1}^{\infty}$ then the RKHS of the SE-T kernel is*

$$\mathcal{H}_{k_T}(\mathcal{X}) = \left\{ F(x) = e^{-\frac{1}{2}\|Tx\|^2} \sum_{|\gamma|\geq 0} w_\gamma x^\gamma \colon \sum_{|\gamma|\geq 0} \frac{\gamma!}{\lambda^\gamma} w_\gamma^2 < \infty \right\}$$

where $x^\gamma = \prod_{n=1}^{\infty} x_n^{\gamma_n}$, $\lambda^\gamma = \prod_{n=1}^{\infty} \lambda_n^{\gamma_n}$ and $\gamma! = \prod_{n=1}^{\infty} \gamma_n!$ and $\mathcal{H}_{k_T}(\mathcal{X})$ is equipped with the inner product

$$\langle F, G \rangle_{\mathcal{H}_{k_T}(\mathcal{X})} = \sum_{|\gamma|\geq 0} \frac{\gamma!}{\lambda^\gamma} w_\gamma v_\gamma$$

where $F(x) = e^{-\frac{1}{2}\|Tx\|^2} \sum_{|\gamma|\geq 0} w_\gamma x^\gamma$, $G(x) = e^{-\frac{1}{2}\|Tx\|^2} \sum_{|\gamma|\geq 0} v_\gamma x^\gamma$

Remark 1. *In the proof of Theorem 1 an orthonormal basis of $\mathcal{H}_{k_T}(\mathcal{X})$ is given which resembles the infinite dimensional Hermite polynomials which are used throughout infinite dimensional analysis and probability theory, for example see [13, Chapter 10] and [39, Chapter 2]. In particular they are used to define Sobolev spaces for functions*

over a real, separable Hilbert space [13, Theorem 9.2.12] which raises the interesting and, as far as we are aware, open question of how $\mathcal{H}_{k_T}(\mathcal{X})$ relates to such Sobolev spaces for different choices of T.

For the two-sample test to be valid we need the SE-T and IMQ-T kernels to be characteristic meaning the mean-embedding is injective over $\mathcal{P}$ so the test can tell the difference between any two probability measures. To understand the problem better we again leverage results regarding the Gauss kernel on $\mathbb{R}^d$, in particular the proof in [62, Theorem 9] that the Gauss kernel on $\mathbb{R}^d$ is characteristic. This uses the fact that the Gauss kernel on $\mathbb{R}^d$ is the Fourier transform of a Gaussian distribution on $\mathbb{R}^d$ whose full support implies the kernel is characteristic. By choosing T such that the SE-T kernel is the Fourier transform of a Gaussian measure on $\mathcal{X}$ that has full support we can use the same argument.

Theorem 2. *Let $T \in L_1^+(\mathcal{X})$ then the SE-T kernel is characteristic if and only if T is non-degenerate.*

This is dissatisfyingly limiting since $T \in L_1^+(\mathcal{X})$ is a restrictive assumption, for example it does not include $T = I$ the identity operator. We shall employ a limit argument to reduce the requirements on T. To this end we define admissible maps.

Definition 2. *A map $T\colon \mathcal{X} \to \mathcal{Y}$ is called admissible if it is Borel measurable, continuous and injective.*

The next result provides a broad family of kernels which are characteristic. It applies for $\mathcal{X}$ being more general than a real, separable Hilbert space. A Polish space is a separable, completely metrizable topological space. Multiple examples of admissible T are given in Section 7 and are examined numerically in Section 8.

Theorem 3. *Let $\mathcal{X}$ be a Polish space, $\mathcal{Y}$ a real, separable Hilbert space and T an admissible map then the SE-T kernel is characteristic.*

A critical result used in the proof is the Minlos-Sazanov theorem, detailed as Theorem 9 in the Appendix, which is an infinite dimensional version of Bochner's theorem. The result allows us to identify spectral properties of the SE-T kernel which are used to deduce characteristicness.

5.2 Integral Kernel Formulation

For certain choices of T the SE-T kernel falls into a family of *integral kernels*. Let $k_0\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a bounded function, $C \in L_1^+(\mathcal{X})$ and N_C the corresponding mean zero Gaussian measure on $\mathcal{X}$ and define $k_{C,k_0}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows

$$k_{C,k_0}(x, y) \coloneqq \int_{\mathcal{X}} k_0\left(\langle x, h\rangle, \langle y, h\rangle\right) dN_C(h)$$

Proposition 1. *If k_0 is a kernel over $\mathbb{R} \times \mathbb{R}$ then k_{C,k_0} is a kernel over $\mathcal{X} \times \mathcal{X}$. If k_0 is also continuous and translation invariant with spectral measure μ such that there exists an interval $(a, b) \subset \mathbb{R}$ with $\mu(U) > 0$ for every open subset $U \subset (a, b)$ then k_{C,k_0} is characteristic.*

We provide two examples of when k_{C,k_0} yields SE-T kernels. If $k_0(x,y) = \cos(x-y)$ then k_{C,k_0} is the SE-$C^{\frac{1}{2}}$ kernel

$$k_{C,k_0}(x,y) = \widehat{N}_C(x-y) = e^{-\frac{1}{2}\|x-y\|_C^2} = e^{-\frac{1}{2}\sum_{n=1}^{\infty}\lambda_n(x_n-y_n)^2}$$

where $\|x-y\|_C^2 = \langle C(x-y), x-y\rangle$, $\{\lambda_n\}_{n=1}^{\infty}$ are the eigenvalues of C and $x_n = \langle x, e_n\rangle$ are the coefficients with respect to the eigenfunction basis $\{e_n\}_{n=1}^{\infty}$ of C.

Secondly, let $\alpha \in \mathbb{R}$ and assume C is non-degenerate and set k_0 to be the complex exponential of α multiplied the by white noise mapping associated with C, see [44, Section 1.2.4], then k_{C,k_0} is the SE-αI kernel.

$$k_{C,k_0}(x,y) = k_{\alpha I}(x,y) = e^{-\frac{\alpha}{2}\|x-y\|^2} \tag{11}$$

Note that $k_{\alpha I}$ is not the Fourier transform of any Gaussian measure on $\mathcal{X}$ [35, Proposition 1.2.11] which shows how the integral kernel framework is more general than using only Fourier transform of Gaussian measures to obtain kernels as was done in Theorem 2.

The integral framework can yield non-SE type kernels. Let N_1 be the measure associated with the Gaussian distribution $\mathcal{N}(0,1)$ on $\mathbb{R}$, C be non-degenerate and $k_0(x,y) = \widehat{N}_1(x-y)$ then we have

$$\begin{aligned}
k_{C,k_0}(x,y) &= \int_{\mathcal{X}}\int_{\mathbb{R}} e^{iz\langle h, x-y\rangle} dN_1(z) dN_C(h) \\
&= \frac{1}{\sqrt{2\pi\|x-y\|_C^2}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{izw} e^{\frac{-w^2}{2\|x-y\|_C^2}} dw dN_1(z) &(12)\\
&= \int_{\mathbb{R}} e^{-\frac{z^2\|x-y\|_C^2}{2}} dN_1(z) &(13)\\
&= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{z^2}{2}(\|x-y\|_C^2+1)} dz \\
&= \frac{1}{\sqrt{\|x-y\|_C^2+1}} &(14)
\end{aligned}$$

where Equation 12 is obtained by noting $\langle h, x-y\rangle$ has distribution $\mathcal{N}(0, \|x-y\|_C^2)$ [44, Corollary 1.19] and then using standard exponential integral identities. From the first equality it is clear to see $k_C(x,y) = \hat{\nu}(x-y)$ where ν is the law of the random variable in $\mathcal{X}$ given by aX where $a \sim \mathcal{N}(0,1)$ and $X \sim N_C$ are independent. As was done for the SE-T kernel we may extend the definition past $C \in L_1^+(\mathcal{X})$.

Definition 3. *For $T\colon \mathcal{X} \to \mathcal{Y}$ the inverse multi-quadric T kernel (IMQ-T) is defined as*

$$k_T(x,y) = \frac{1}{\sqrt{\|T(x)-T(y)\|_{\mathcal{Y}}^2+1}}$$

By using Proposition 1 we immediately obtain that if $T \in L_1^+(\mathcal{X})$ and T is non-degenerate then the IMQ-T kernel is characteristic. But by the same limiting argument as Theorem 3 and the integral kernel formulation of IMQ-T we obtain a more general result.

Corollary 1. *Under the same conditions as Theorem 3 the IMQ-T kernel is characteristic.*

6 MMD and Testing for Measures on Function Spaces

In the previous section we provided a large class of characteristic kernels. By definition, a characteristic kernel k has an injective mean-embedding Φ_k so by Equation 1 we may conclude that for such kernels MMD is a metric on $\mathcal{P}$. The testing procedure for functional data is exactly the same as for finite dimensional data, described in Subsection 3.3, and MMD being a metric ensures the test is valid. This highlights the strength and generality of kernel based statistical methods, once the kernel satisfies the required properties related to the input space the statistical procedures can be carried out the same way regardless of the input space.

However, when dealing with functional data in practise one does not observe the entire function, rather a discretised set of input-output pairs. If observation points are sparse enough to render numerical computation of required quantities ineffective then approximations of the true underlying functions would be taken and then the testing procedure would be performed on the approximate curves.

To this end let $P \in \mathcal{P}$ and $x \sim P$ be a random element in $\mathcal{X}$ then we denote the discrete, possibly noise corrupted, observation of x by $\tilde{x}$. Any reconstruction method that maps to $\mathcal{X}$ used to obtain x from $\tilde{x}$ shall be denoted $\mathcal{R}$, for example $\mathcal{R}$ could be a spline interpolation or a kernel ridge regression function [16]. The two-sample test would then be performed on the reconstructed data $\mathcal{R}\tilde{x}$.

Proposition 2. *Let k_T be the SE-T or IMQ-T kernel $P, Q \in \mathcal{P}$ with $X_n = \{x_i\}_{i=1}^n$ i.i.d. samples from P and $Y_n = \{y_i\}_{i=1}^n$ i.i.d. samples from Q and observed the discrete data $\tilde{X}_n = \{\tilde{x}_i\}_{i=1}^n$ and $\tilde{Y}_n = \{\tilde{y}_i\}_{i=1}^n$. Let $\mathcal{R}$ be any reconstruction method mapping into $\mathcal{X}$ then*

$$\left|\widehat{\mathrm{MMD}}_{k_T}(X_n, Y_n)^2 - \widehat{\mathrm{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2\right| \leq \frac{4L}{n}\sum_{i=1}^{n}\|T(\mathcal{R}\tilde{x}_i) - T(x_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_i) - T(y_i)\|_{\mathcal{Y}}$$

where $\widehat{\mathrm{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2$ is the estimate of $\mathrm{MMD}^2_{k_T}(P, Q)^2$ based on the data reconstructed by $\mathcal{R}$ and $L = \frac{1}{\sqrt{e}}$ if k_T is the SE-T kernel and $L = \frac{2}{3\sqrt{3}}$ if k_T is the IMQ-T kernel.

An analogous result can be derived for the linear time estimator with the same proof technique. While Proposition 2 provides a statement on the approximation of $\widehat{\mathrm{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2$ we are primarily concerned with its statistical properties. It is known that a scaled version of the exact estimator estimator converges in distribution to a Gaussian [22, Theorem 16] and we will next show a similar result for the reconstructed data.

First we introduce notation regarding the input observation distributions. Given $P \in \mathcal{P}$ and $x \sim P$ denote by $\tilde{x}_{u(n)}$ the discretisation of x over a set of observation points drawn from the distribution U with cardinality $u(n)$. The same notation but with V, v instead of U, u is used for disctretised samples from Q. For example if the input space was $[0, 1]$ then one could have $u(n) = n^2$ and U as the law of n^2 i.i.d. samples from Uniform$(0, 1)$. Given a collection of samples from P, Q assume that the observation point sampling is i.i.d. across the samples. The point of $u(n), v(n)$ is to

facilitate statements regarding the number of observation points required per function sample compared to the number of function samples.

Theorem 4. *Let k_T be the SE-T or IMQ-T kernel $P, Q \in \mathcal{P}$ with $X_n = \{x_i\}_{i=1}^n$ i.i.d. samples from P and $Y_n = \{y_i\}_{i=1}^n$ i.i.d. samples from Q and observed the discrete data $\tilde{X}_n = \{\tilde{x}_{i,u(n)}\}_{i=1}^n$ and $\tilde{Y}_n = \{\tilde{y}_{i,v(n)}\}_{i=1}^n$ with observation point distributions U, V and cardinalities $u(n), v(n)$ respectively. Assume $U, V, u(n), v(n)$ are such that $n^{\frac{1}{2}}\mathbb{E}_{x\sim P,U}[\|T(\mathcal{R}\tilde{x}_{u(n)}) - T(x)\|_{\mathcal{Y}}] \to 0$ and $n^{\frac{1}{2}}\mathbb{E}_{y\sim Q,V}[\|T(\mathcal{R}\tilde{y}_{v(n)}) - T(y)\|_{\mathcal{Y}}] \to 0$ then*

$$n^{\frac{1}{2}}\left(\widehat{\text{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2 - \text{MMD}_{k_T}(P,Q)^2\right) \xrightarrow{d} \mathcal{N}(0, \xi)$$

where $\xi = 4\text{Var}_z\left[\mathbb{E}_{z'}[h(z,z')]\right]$.

A similar result can be derived for the linear time estimator by using the linear time estimator version of Proposition 2. The assumptions of the theorem ensure that the reconstruction method $\mathcal{R}$, observation point distributions U, V and observation point cardinalities $u(n), v(n)$ allow the estimator based on reconstructed data to converge to the estimator based on exact data. We now discuss two examples of when these assumptions hold. The first has $\mathcal{R}$ as a kernel interpolant and the second as the posterior mean of a Gaussian process, for more details on each of these reconstruction methods see [49, 31]. For both examples assume T is Lipschitz continuous with Lipschitz constant L_T, $\mathcal{X} = L^2(\Omega)$ for some compact $\Omega \subset \mathbb{R}^r$ and with probability one samples from P, Q lie in $W_2^\tau(\Omega)$ for some $\tau > \frac{r}{2}$, for a discussion on this assumption in the case P, Q are Gaussian processes see [31, Section 4].

Example 1. *Assume there is no noise corruption in the observations so that results from the scattered data approximation [74] literature are applicable. Set $\mathcal{R}$ as the kernel interpolant using a kernel with RKHS norm equivalent to $W_2^\nu(\Omega)$ with $\nu > \frac{r}{2}$ and U, V deterministic point selection procedures that place point quasi-uniformly, for example regularly placed grid points, see [74, Chapter 4], then*

$$\begin{aligned}\mathbb{E}_{x\sim P,U}[\|T(\mathcal{R}\tilde{x}_{u(n)} - x)\|_{\mathcal{Y}}] &\leq L_T\mathbb{E}_{x\sim P,U}[\|\mathcal{R}\tilde{x}_{u(n)} - x\|_{L^2(\Omega)}] \\ &\leq Cu(n)^{-(\tau\wedge\nu)/r} = o(n^{-\frac{1}{2}})\end{aligned}$$

for some constant $C > 0$ with an analogous bound holding for the reconstruction of y with $v(n)$ instead of $u(n)$. If $u(n) = v(n) = n$ then the error is $o(n^{-\frac{1}{2}})$ and the conditions of Theorem 4 are satisfied. For a proof of the error bound and examples of quasi-uniform point sets see [76].

Example 2. *Assume that the observations are corrupted with i.i.d. $\mathcal{N}(0, \sigma^2)$ noise and U, V are the uniform distribution on Ω we employ Bayesian non-parametric error bounds [20]. Starting with a mean zero Gaussian process prior, with covariance kernel having RKHS norm equivalent to $W_2^\nu(\Omega)$ for some $\nu > r$, set $\mathcal{R}$ as the posterior mean obtained by conditioning on the noisy observations. Then [71, Theorem 5] states*

$$\begin{aligned}\mathbb{E}_{x\sim P,U}[\|T(\mathcal{R}\tilde{x}_{u(n)} - x)\|_{\mathcal{Y}}] &\leq L_T\mathbb{E}_{x\sim P,U}[\|\mathcal{R}\tilde{x}_{u(n)} - x\|_{L^2(\Omega)}] \\ &\leq Cu(n)^{-\frac{((\nu-r/2)\wedge\tau)}{2\nu}}\end{aligned}$$

for some constant $C > 0$ with an analogous bound holding for the reconstruction of y with $v(n)$ instead of $u(n)$. Therefore if $u(n) = v(n) = n^\gamma$ for any $\gamma > \nu/((\nu -$

$r/2) \wedge \tau)$ *then the error is* $o(n^{-\frac{1}{2}})$ *so the conditions of Theorem 4 are satisfied. This highlights how the required cardinality of the discretisation depends on the differing smoothnesses of the approximating function and true function, this in turn impacts the strength of the approximation. In this case the data is assumed to be noisy, therefore more observations of each function are needed as opposed to the previous scenario.*

6.1 Weak Convergence and Mean Embeddings of GPs

We know that when k is characteristic that MMD is a metric on $\mathcal{P}$ so it is natural to identify the topology it generates and in particular how it relates to the standard topology for elements of $\mathcal{P}$, the weak topology.

Theorem 5. *Let* $\mathcal{X}$ *be a Polish space,* k *a bounded, continuous, characteristic kernel on* $\mathcal{X} \times \mathcal{X}$ *and* $P \in \mathcal{P}$ *then* $P_n \xrightarrow{w} P$ *implies* $\text{MMD}_k(P_n, P) \to 0$ *and if* $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}$ *is tight then* $\text{MMD}_k(P_n, P) \to 0$ *implies* $P_n \xrightarrow{w} P$ *where* $\xrightarrow{w}$ *denotes weak convergence.*

For a discussion on weak convergence and tightness see [5]. The tightness is used to compensate for the lack of standard compactness assumptions on $\mathcal{X}$ which is often required in finite dimensions. For example the unit ball in a Hilbert space is compact if and only if the Hilbert space is finite dimensional which is often not the case for function spaces. In particular, in [10] an example where $\text{MMD}_k(P_n, P) \to 0$ but P_n but does converge to P was given without the assumption of tightness. A precise characterisation of the relationship between MMD and weak convergence over a Polish space is an open problem.

A key property of the SE-T kernel is that the mean-embedding $\Phi_{k_T} P$ and $\text{MMD}_{k_T}(P, Q)^2$ have closed form solutions when P, Q are Gaussian measures. Using the natural correspondence between Gaussian measures and Gaussian processes from Section 2 we may get closed form expressions for Gaussian processes. This addresses the open question regarding the link between Bayesian non-parametrics methods and kernel mean-embeddings that was discussed in [37, Section 6.2].

Before stating the next result we need to introduce the concept of determinant for an operator, for $S \in L_1(\mathcal{X})$ define $\|I + S\| = \prod_{n=1}^{\infty}(1 + \lambda_n)$ where $\{\lambda_n\}_{n=1}^{\infty}$ are the eigenvalues of S. The equality $\|(I + S)(I + R)\| = \|I + S\|\|I + R\|$ holds and is frequently used.

Theorem 6. *Let* k_T *be the SE-*T *kernel for some* $T \in L^+(\mathcal{X})$ *and* $P = N_{a,S}$ *be a non-degenerate Gaussian measure on* $\mathcal{X}$ *then*

$$\Phi_{k_T}(N_{a,S})(x) = \|I + TST\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I+TST)^{-1}T(x-a), T(x-a)\rangle}$$

Theorem 7. *Let* k_T *be the SE-*T *kernel for some* $T \in L^+(\mathcal{X})$ *and* $P = N_{a,S}, Q = N_{b,R}$ *be non-degenerate Gaussian measures on* $\mathcal{X}$ *then*

$$\begin{aligned}\text{MMD}_{k_T}(P, Q)^2 = &\|I + 2TST\|^{-\frac{1}{2}} + \|I + 2TRT\|^{-\frac{1}{2}} \\ &- 2\|(I + TST)(I + (TRT)^{\frac{1}{2}}(I + TST)^{-1}(TRT)^{\frac{1}{2}})\|^{-\frac{1}{2}} \\ &\times e^{-\frac{1}{2}\langle (I+T(S+R)T)^{-1}T(a-b), T(a-b)\rangle}\end{aligned}$$

These results outline the geometry of Gaussian measures with respect to the distance induced by the SE-T kernel. We see that the means only occur in the formula

through their difference and if both mean elements are zero then the distance is measured purely in terms of the spectrum of the covariance operators.

Corollary 2. *Under the Assumptions of Theorem 7 and that T, S, R commute then*

$$\begin{aligned}\mathrm{MMD}_{k_T}(P,Q)^2 = &\|I+2TST\|^{-\frac{1}{2}} + \|I+2TRT\|^{-\frac{1}{2}} \\ &- 2\|I+T(S+R)T\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I+T(S+R)T)^{-1}T(a-b), T(a-b)\rangle}\end{aligned}$$

7 Impact and Choice of T

This section outlines the impact on power of the choice of T and gives multiple examples of how to construct admissible T for different testing scenarios. Subsection 7.1 investigates the mean shift problem for two Gaussian processes and using closed form expressions for push-forward measures highlights how Equation 8 behaves for different choices of T. Subsection 7.2 provides strategies for constructing admissible maps and outlines three broad classes.

7.1 Impact of T on Test Power

Using Equation 8 and the Riemann sum scaling we can conclude that for $\mathcal{X} = L^2([0,1]^r)$ and the SE-I kernel, for two GPs that differ only in mean $P = \mathcal{GP}(0, k_0), Q = \mathcal{GP}(m, k_0)$ with k_0 continuous the surrogate for test power is

$$\frac{\|m\|^2_{L^2([0,1]^r)}}{\sqrt{8\|C_{k_0}\|^2_{HS} + 8\|C_{k_0}^{1/2}m\|^2_{L^2([0,1]^r)}}} \tag{15}$$

If one uses the SE-T kernel with $T \in L^+(\mathcal{X})$ then we can use a push-forward measure argument to rewrite this surrogate for power. Using the correspondance between GPs and GMs we may write $P = N_{C_{k_0}}, Q = N_{m,C_{k_0}}$ and note the the push-forward measures through T are $T_\# P = N_{TC_{k_0}T}, T_\# Q = N_{Tm,TC_{k_0}T}$ [44, Proposition 1.1.8] therefore performing the two-sample test on P, Q with the SE-T kernel is the same as performing the two-sample test on $T_\# P, T_\# Q$ with the SE-I kernel. Substituting this into Equation 15 gives the new power surrogate

$$\frac{\|Tm\|^2_{L^2([0,1]^r)}}{\sqrt{8\|TC_{k_0}T\|^2_{HS} + 8\|T^{\frac{1}{2}}C_{k_0}^{\frac{1}{2}}T^{\frac{3}{2}}m\|^2_{L^2([0,1]^r)}}} \tag{16}$$

We numerically investigate how Equation 16 behaves in an idealised scenario. Assume T shares an eigenbasis $\{e_n\}_{n=1}^\infty$ with C_{k_0} and T has eigenvalues $\lambda_n = n^{-\alpha_T}$, C_{k_0} has eigenvalues $\kappa_n = n^{-\alpha_{k_0}}$ and $m_n = \langle m, e_n\rangle_{L^2([0,1]^r)} = n^{-\alpha_m}$. The intuition is that if m is smoother than the eigenvalues of C_{k_0}, meaning $\alpha_m > \alpha_{k_0}$, then smoothing of the signal, which corresponds to a larger α_T, will help detect the signal since it reduces the variance and hence increase the power surrogate. However if α_T is too large then the signal will be smoothed too much to detect anything and the surrogate will decrease.

Figure 1 confirms this intuition. The plot shows, for $\alpha_m = 2.5$, how for different values of α_{k_0}, α_T impact the value of the surrogate. When $\alpha_{k_0} = 1.2$, meaning

the eigenvalues of C_{k_0} decay slower than the mean coefficients and thus dominate the signal, smoothing increases the surrogate until $\alpha_T \approx 1$ after which the surrogate decreases. The benefit of smoothing becomes smaller as α_{k_0} gets larger, meaning the strength of the variance gets smaller. Indeed at $\alpha_{k_0} = 2.1$ the variance eigenvalues decay is comparable to the mean coefficient decay $\alpha_m = 2.5$ therefore smoothing only deteriorates the surrogate. Although the surrogate values vary slightly in the plot the impact of smoothing in performance of the test can be large, as seen in Section 8.

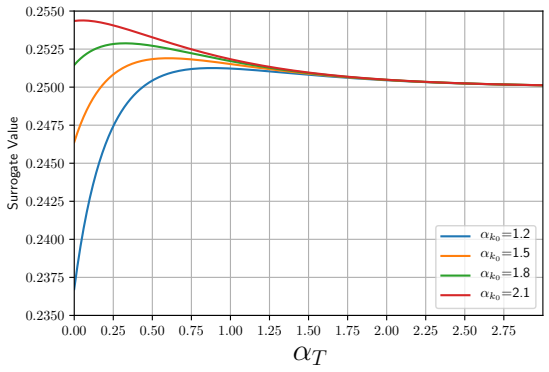

Figure 1: Value of the power surrogate (Equation 16) with $\alpha_m = 2.5$ for different values of α_{k_0}, α_T

This is an idealised scenario as we have assumed that T and C_{k_0} share an eigenbasis with their eigenvalues parameterised in a convenient manner. However it still demonstrates the smoothing trade off that occurs given certain choices of T. Further analytically investigating the impact of different choices of T is beyond the scope of this paper but we believe it is an important question.

7.2 Examples of Admissible T

We now present examples and techniques to construct maps T that are admissible. Three categories will be discussed, integral operators induced by kernels, non-linear maps induced by kernels, and situation specific maps.

For the first catergory assume $\mathcal{X} = \mathcal{Y} = L^2(\Omega)$ for some compact $\Omega \subset \mathbb{R}^r$ and let k_0 be a measurable kernel over $\Omega \times \Omega$ and set $T_{k_0} = C_{k_0}$ where C_{k_0} is the covariance operator associated with k_0, see Section 2. We call k_0 an *admissible kernel* if T_{k_0} is admissible. If k_0 is continuous then by Mercer's theorem $T_{k_0}x = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n$ for some positive sequence $\{\lambda_n\}_{n=1}^{\infty}$ and orthonormal set $\{e_n\}_{n=1}^{\infty}$ [63, Chapter 4.5]. To be admissible T_{k_0} needs to be injective which is equivalent to $\{e_n\}_{n=1}^{\infty}$ forming a basis [64, Proof of Theorem 3.1]. Call k_0 *integrally strictly positive definite* (ISPD) if $\int_\Omega \int_\Omega x(s)k_0(s,t)x(t)dsdt > 0$ for all non-zero $x \in \mathcal{X}$. Recall that if k_0 is translation invariant then by Theorem 8 there exists a measure μ_{k_0} such that $\hat{\mu}_{k_0}(s-t) = k(s,t)$.

Proposition 3. *Let $\Omega \subset \mathbb{R}^r$ be compact and k_0 a continuous kernel on Ω, if k_0 is ISPD then k_0 is admissible. In particular, if k_0 is continuous and translation invariant and μ_{k_0} has full support on Ω then k_0 is admissible.*

For multiple examples of ISPD kernels see [61] and of μ_{k_0} see [62]. Using the product to convolution property of the Fourier transform one can construct k_0 such that μ_{k_0} has full support relatively easily or modify standard integral operators which aren't admissible. For example, for some $F \in \mathbb{N}$ consider the kernel

$$k_{\cos(F)}(s,t) = \sum_{n=1}^{F} \cos(2\pi n(s-t))$$

on $[0,1]^2$ whose spectral measure if a sum of Dirac measures so does not have full support. If the Dirac measures are convolved with a Gaussian then they would be smoothed out and would result in a measure with full support. Since convolution in the kernel frequency domain corresponds to a product of kernels the new kernel $k_{\text{c-exp}(F,l)}(s,t) = e^{-\frac{1}{2l^2}(s-t)^2} k_{\cos(F)}(s,t)$ satisfies the conditions of Proposition 3. This technique of frequency modification has found success in modelling audio signals, see [75, Section 3.4]. In fact any operator of the form $Tx = \sum_{n=1}^{\infty} \lambda_n \langle x, e_n \rangle e_n$ for positive, bounded $\{\lambda_n\}_{n=1}^{\infty}$ and an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ is admissible even if it is not induced by a kernel, for example the functional Mahalanobis distance [4].

The second category of T is non-linear maps induced by kernels. We will dicuss a broad class which lift the samples into a higher dimensional output space. Let $\Omega \subset \mathbb{R}^{r_1}$ and k_0 be a continuous kernel on $\mathbb{R}^{r_2} \times \mathbb{R}^{r_2}$. Define the lifting map $L_{k_0} \colon \mathcal{X} = L^2(\Omega, \mathbb{R}^{r_2}) \to L^2(\Omega, \mathcal{H}_{k_0}(\mathbb{R}^{r_2})) = \mathcal{Y}$ as $L_{k_0}(x)(s) = k_0(x(s), \cdot)$ where $L^2(\Omega, \mathcal{H}_{k_0}(\mathbb{R}^{r_2}))$ is the space of equivalence classes with norm

$$\|x\|^2_{L^2(\Omega, \mathcal{H}_{k_0}(\mathbb{R}^{r_2}))} = \int_{\Omega} \|x(s)\|^2_{\mathcal{H}_{k_0}(\mathbb{R}^{r_2})} ds$$

This is a separable Hilbert space with basis $\{h_n \otimes e_n\}_{n=1}^{\infty}$ where $\{e_n\}_{n=1}^{n}$ is a basis of $L^2(\Omega)$ and $\{h_n\}_{n=1}^{\infty}$ is a basis of $\mathcal{H}_{k_0}(\mathbb{R}^{r_2})$ which exists since k_0 is continuous and $\mathbb{R}^{r_2}$ is separable [63, Lemma 4.33]. The map L_{k_0} is admissible if the map $s \to k(s, \cdot)$ is injective over Ω and a sufficient condition for this is k_0 is characteristic over the set of Borel measures over $\mathbb{R}^{r_2}$ with numerous examples of such kernels known [62]. This class of non-linear maps was discussed in [10] and acts as a a feature expansion of the samples. Due to its non-linear nature it is harder to interpret its impact on power as opposed to linear maps.

The third category is scenario specific choices. By this we mean maps T whose structure is specified to the testing problem at hand. For example, while the kernel two-sample test may be applied for distributions with arbitrary difference one may tailor it for a specific testing problem, such a difference of covariance operator. In this scenario higher order moments should be emphasised and we present two examples of T which do this. First, for an injective polynomial $p(\xi)$, one could set $T(x) = p(x)$ as long as integrability conditions are fulfilled. For instance let $\Omega \subset \mathbb{R}^r$ be bounded $\mathcal{Y} = L^2(\Omega)$ and $p(x) = x + x^2 + x^3$ then one would need $\mathcal{X} \subset L^6(\Omega)$. Another example of map to emphasise higher order moments is to let $\mathcal{X} \subset L^4(\Omega)$ and $\mathcal{Y}$ the direct sum of $L^2(\Omega)$ with itself equipped with the norm $\|(x, x')\|^2_{\mathcal{Y}} = \|x\|^2_{L^2} + \|x'\|^2_{L^2}$ and $T(x) = (x, x^2)$. This map captures second order differences and first order differences individually, as opposed to the polynomial map which combines them.

7.3 Random Fourier Features

If $\mathcal{X} = \mathcal{Y} = L^2(\Omega)$ for some compact $\Omega \subset \mathbb{R}^r$ and $T = C_{k_0}^{\frac{1}{2}}$ for come continuous kernel k_0 then $\|T(x) - T(y)\|^2 = \|x - y\|^2_{C_{k_0}}$ is a double integral. If the discretisation of the samples consists of N points in each dimension then approximating this double integral will have cost $O(N^{2r})$. Since the assumptions on k_0 imply that $C_{k_0} \in L_1^+(\mathcal{X})$ we may use the representation of the SE-T as the Fourier transform of the Gaussian measure $N_{C_{k_0}}$ corresponding to $\mathcal{GP}(0, k_0)$ and employ Random Fourier Features [45] to obtain the Monte Carlo approximation

$$e^{-\frac{1}{2}\|x-y\|^2_{C_{k_0}}} = \int_{\mathcal{X}} e^{i\langle x-y,h\rangle} dN_{C_{k_0}}(y) \approx \frac{1}{N_F} \sum_{n=1}^{N_{\text{F}}} \cos(\langle g_n, x - y\rangle)$$

where $N_{\text{F}} \in \mathbb{N}$ and $\{g_n\}_{n=1}^{N_{\text{F}}}$ are i.i.d. samples from $\mathcal{GP}(0, k_0)$. When k_0 has a closed form Mercer expansion then samples from $\mathcal{GP}(0, k_0)$ can be approximated by truncating the Karhunen-Loève expansion $g_n \approx \sum_{m=1}^{N_{\text{KL}}} \lambda_m^{\frac{1}{2}} \eta_{n,m} e_m$ where $N_{\text{KL}} \in \mathbb{N}$ and $\eta_{n,m} \sim \mathcal{N}(0, 1)$ are i.i.d. Therefore the computational cost of this approximate procedure is $O(N_{\text{F}} N_{\text{KL}} N^r)$ which could provide significant speed ups when N is large.

7.4 Tuning T

All of the operators outlined above have associated hyperparameters. It is outside the scope of this paper to investigate new methods to choose these parameters but we do believe it is important future work. Multiple methods have been proposed using the surrogrates for test power outlined in Section 4 [67, 23, 34]. All of these considered finite dimensional data but due to the kernel mapping the same methods apply to infinite dimensional data. It is beyond the scope of this paper to investigate these methods.

8 Numerical Simulations

In this section we perform numerical simulations on real and synthetic data to reinforce the theoretical results.

8.1 Power Scaling of Functional Data

Verification of the power scaling when performing the mean shift two-sample test using functional data, discussed in Section 4, is performed. Specifically we perform the two-sample test using the SE-I kernel with $X \sim \mathcal{GP}(0, k_l)$ and $Y \sim \mathcal{GP}(m, k_l)$ where $m(t) = 0.05$ for $t \in [0, 1]$ and $k_l(s, t) = e^{-\frac{1}{2l^2}(s-t)^2}$ with 50 samples from each distribution. This is repeated 500 times to calculate power with 1000 permutations used in the bootstrap to simulate the null. The observation points are a uniform grid on $[0, 1]$ with N points, meaning N will be the dimension of the observed discretised function vectors. The parameter l dictates the dependency of the fluctations. Small l means less dependency between the random function values so the covariance matrix is closer to the identity. When the random functions are m with $\mathcal{N}(0, 1)$ i.i.d. corruption the corresponding value of l is zero which essentially means $k_l(x, y) = \delta_{xy}$. In this case the

scaling of power is expected to follow Equation 10 and grow asymptotically as $\sqrt{N}$. On the other hand if $l > 0$ the fluctuations within each random function are dependent and we expect scaling as Equation 8 which does not grow asymptotically with N.

Figure 2 confirms this theory showing that power increases with a finer observation mesh, meaning dimensionality of the observation increases, only when there is no dependence in the random functions values. We see some increase of power as dimension increases for the case of small dependency however the rate of increase is much smaller than the i.i.d. setting.

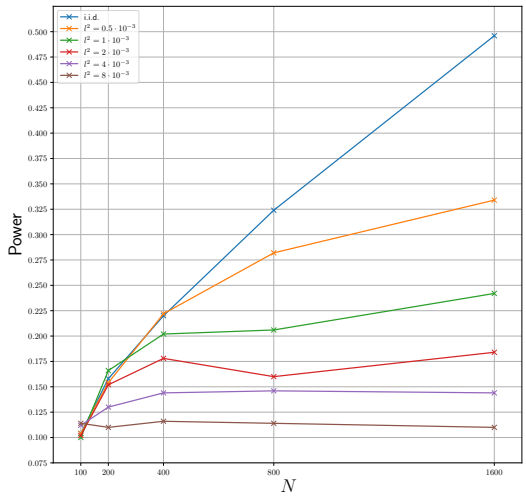

Figure 2: Test power as dimension increases given different point dependency strengths

8.2 Synthetic Data

The tests are all performed using the SE-T kernel for five different choices of T and unless stated otherwise $\mathcal{Y} = L^2([0,1])$ and we use the short hand L^2 for $L^2([0,1])$ and n_x, n_y will denote the sample sizes of the two samples. To calculate power each test is repeated 500 times and 1000 permutations are used in the bootstrap to simulate the null distribution.

ID will denote $T = I$. CEXP will denote $T = T_{k_0}$ with $k_0 = k_{\text{c-exp}(20,\sqrt{10})}$ the cosine exponential kernel. POLY will denote $T(x) = x + x^2 + x^3$. SQR will denote $T(x) = (x, x^2)$ with $\mathcal{Y}$ the direct sum of $L^2([0,1])$ with itself as detailed in Subsection 7.2. FPCA will denote $Tx = \sum_{n=1}^{F} \lambda_n \langle x, e_n \rangle e_n$ where λ_n, e_n are empirical functional principal components and principal values computed from the union of the two collections of samples with F choosen such that 95% is explained, see [27] for a discussion on functional principal components.

For each of these five scenarios power will be calculated using the kernel $\exp\left(-\frac{1}{2\gamma^2}\|T(x) - T(y)\|_{\mathcal{Y}}^2\right)$ where, for all but SQR scenario, we use the median

heuristic $\gamma^2 = \text{Median}\big\{\|T(a) - T(b)\|_{\mathcal{Y}}^2 \colon a, b \in \{x_i\}_{i=1}^{n_X} \cup \{y_i\}_{i=1}^{n_Y}, a \neq b\big\}$. As the SQR scenario involves two norms in the exponent two calculations of median heuristic are needed so that the kernel used is $\exp(-\frac{1}{2\gamma_1^2}\|x - y\|_{L^2}^2 - \frac{1}{2\gamma_2^2}\|x^2 - y^2\|_{L^2}^2)$ with $\gamma_i^2 = \text{Median}\big\{\|a^j - b^j\|_{L^2}^2 \colon a, b \in \{x_i\}_{i=1}^{n_X} \cup \{y_i\}_{i=1}^{n_Y}, a \neq b\big\}$ for $j = 1, 2$.

Difference of Mean

We compare to the Functional Anderson-Darling (FAD) test in [43] which involves computing functional principal components and then doing multiple Anderson-Darling tests. Independent realisations $\{x_i\}_{i=1}^{n_x}$ and $\{y_j\}_{j=1}^{n_y}$ of the random functions x, y over $[0, 1]$ are observed on a grid of 100 uniform points with $n_X = n_Y = 100$ and observation noise $\mathcal{N}(0, 0.25)$. The two distributions are

$$x(t) \sim t + \xi_{10}\sqrt{2}\sin(2\pi t) + \xi_5\sqrt{2}\cos(2\pi t)$$
$$y(t) \sim t + \delta t^3 + \gamma_{10}\sqrt{2}\sin(2\pi t) + \gamma_5\sqrt{2}\cos(2\pi t)$$

with $\xi_5, \gamma_5 \overset{i.i.d}{\sim} \mathcal{N}(0, 5)$ and $\xi_{10}, \gamma_{10} \overset{i.i.d}{\sim} \mathcal{N}(0, 10)$. The δ parameter measures the deviation from the null hypothesis that X, Y have the same distribution The range of the parameter is $\delta \in \{0, 0.5, 1, 1.5, 2\}$.

Figure 3 shows CEXP performing best among all the choices which makes sense since this choice explicitly smooths the signal to make the mean more identifiable compared to the noise. We see that FPCA performs poorly because the principal components are deduced entirely from the covariance structure and do not represent the mean difference well. Except from FPCA all choices of T out perform the FAD method. This is most likley because the FAD method involves computing multiple principle components, an estimation which is inherently random, and computes mulitple FAD tests with a Bonferroni correction which can cause too harsh a requirement for significance. There is a slight inflation of test size, meaning rejection is larger than 5% when the null hypothesis is true.

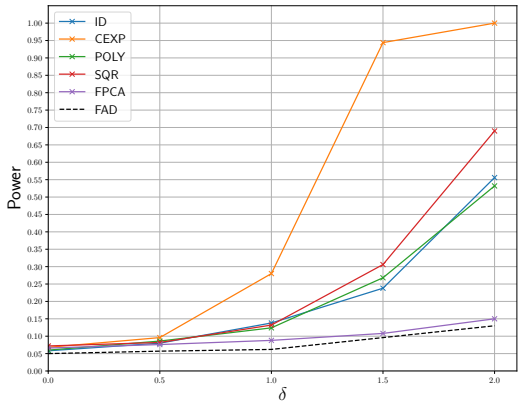

Figure 3: Test power under mean difference for different choices of T.

Difference of Variance

We investigate two synthetic data sets, the first from [43] and the second from [41]. The first represents a difference in covariance in a specific frequency and the second a difference across all frequencies.

In the first data set $n_x = n_y = 100$, observations are made on the same uniform grid of 100 points and the observation noise is $\mathcal{N}(0, 0.25)$. The two distributions are

$$x(t) \sim \xi_{10}\sqrt{2}\sin(2\pi t) + \xi_5\sqrt{2}\cos(2\pi t)$$
$$y(t) \sim \gamma_{10+\delta}\sqrt{2}\sin(2\pi t) + \gamma_5\sqrt{2}\cos(2\pi t)$$

with $\xi_5, \gamma_5 \overset{i.i.d}{\sim} \mathcal{N}(0, 5)$ and $\xi_{10} \sim \mathcal{N}(0, 10)$ and $\gamma_{10+\delta} \sim \mathcal{N}(0, 10 + \delta)$. Therefore the difference in covariance structure is manifested in the first frequency. The range of the parameter is $\delta \in \{0, 5, 10, 15, 20\}$ and we again compare against the FAD test.

Figure 4 shows that POLY and SQR perform the best which is too be expected since they capture higher order behaviour of the random functions. FPCA performs well since the distributions vary explcitly in the distributions of the coefficients of the functional principal components. CEXP performs almost identically to ID since it designed to improve performance on mean shift tests.

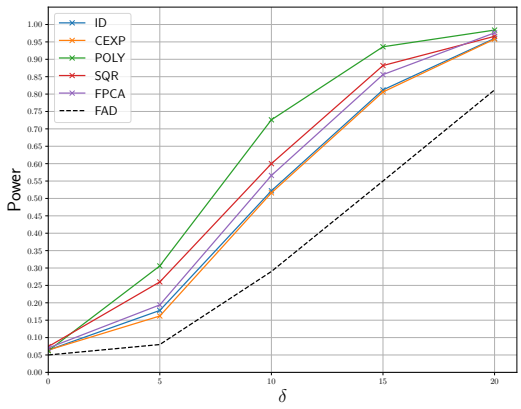

Figure 4: Test power under variance difference in one frequency for different choices of T.

The second dataset is from [41] and we compare against the data reported there of a bootstrap Hilbert-Schmidt norm (BOOT-HS) test [41, Section 2.2] and a functional principal component chi-squared (FPCA-χ) test from [17], which is similar to the test in [40]. The number of function samples is $n_x = n_y = 25$ and each sample is observed on a uniform grid over $[0, 1]$ consisting of 500 points. The first distribution is defined as

$$x(t) \sim \sum_{n=1}^{10} \xi_n n^{-\frac{1}{2}}\sqrt{2}\sin(\pi n t) + \eta_n n^{-\frac{1}{2}}\sqrt{2}\cos(\pi n t)$$

where ξ_n, η_n are i.i.d. Student's t-distribution random variables with 5 degrees of freedom. For $\delta \in \mathbb{R}$ the other function distribution is $y \sim \delta x'$ where X' is an i.i.d. copy of x. When $\delta = 1$ the two distributions are the same. The entire covariance structure of Y is different from that of X when $\delta \neq 1$ which is in contrast the previous numerical example where the covariance structure differed at only one frequency. The range of the deviation parameter is $\delta \in \{1, 1.2, 1.4, 1.6, 1.8, 2\}$.

Figure 5 shows again that POLY, SQR performs the best. The BOOT-HS and FPCA-χ tests are both conservative, providing rejection rates below 5% when the null is true as opposed the the kernel based tests which all lie at or very close to the 5% level.

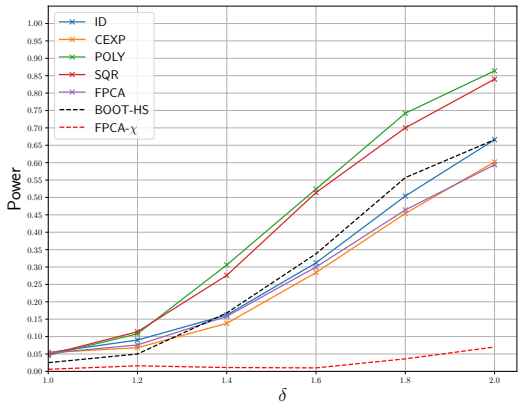

Figure 5: Test power under variance difference across all frequencies for different choices of T.

Difference of Higher Orders

Data from [24] is used when performing the test. The random functions X, Y are distributed as

$$x(t) \sim \sum_{n=1}^{15} e^{-\frac{n}{2}} \xi_n^x \psi_n(t)$$
$$y(t) \sim \sum_{n=1}^{15} e^{-\frac{n}{2}} \xi_{n,1}^y \psi_n(t) + \delta \sum_{n=1}^{15} n^{-2} \xi_{n,2}^Y \psi_n^*(t)$$

with $\xi_n^x, \xi_{n,1}^y, \xi_{n,2}^y \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, $\psi_1(t) = 1$, $\psi_n(t) = \sqrt{2}\sin((k-1)\pi t)$ for $n > 1$ and $\psi_1^*(t) = 1$, $\psi_n^*(t) = \sqrt{2}\cos((k-1)\pi(2t-1))$ if $n > 1$ is even, $\psi_n^*(t) = \sqrt{2}\sin((k-1)\pi(2t-1))$ if $n > 1$ is odd. The observation noise for x is $\mathcal{N}(0, 0.01)$ and for y is $\mathcal{N}(0, 0.09)$. The range of the parameter is $\delta \in \{0, 1, 2, 3, 4\}$ and we compare against the FAD test and the Carmer Von-Mises test in [24]. The number of samples is $n_x = n_y = 15$ and for each random function 20 observation locations are

sampled randomly according to p_x or p_y with p_x being the uniform distribution on $[0, 1]$ and p_y the distribution with density function $0.8 + 0.4t$ on $[0, 1]$.

Since the data is noisy and irregularly sampled, curves were fit to the data before the test was performed. The posterior mean of a Gaussian process with noise parameter $\sigma^2 = 0.01$ was fit to each data sample using a Matérn-1.5 kernel $k_{\text{Mat}}(s,t) = (1 + \sqrt{3}(s-t))e^{-\sqrt{3}(s-t)}$. Given a random function realisation x, denote the set of observation locations by $U = \{u_i\}_{i=1}^{20}$ and vector of observed noisy function values by $\tilde{x}$ then the reconstructed function is $\mathcal{R}(\tilde{x})(s) = k_{sU}(k_{UU} + \sigma^2 I_{N\times N})^{-1}\tilde{x}$ where $k_{sU} = (k_{\text{Mat}}(s,u_1),\ldots,k_{\text{Mat}}(s,u_n))^T$, $k_{UU} = (k_{\text{Mat}}(u_i,u_j))_{1\leq i,j\leq n}$. For the derivation of this formula see [49]. This curve fitting technique is common and the parameter σ^2 was choosen so that the curves had visually good fits.

Figure 6 shows that the performance across the different choices of T does not vary much and that POLY, SQR attain the best performance.

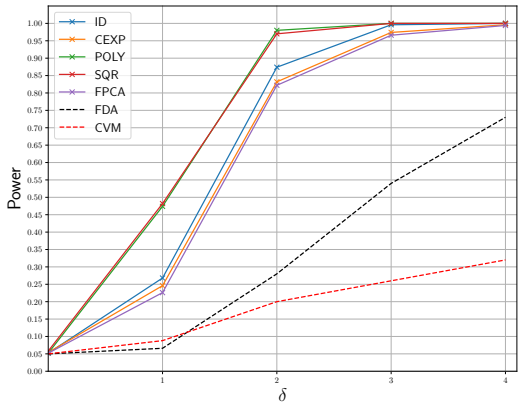

Figure 6: Test power under difference of higher orders for different choices of T.

8.3 Real Data

Berkeley Growth Data

We now perform tests on the Berkeley growth dataset which contains the height of 39 male and 54 female children from age 1 to 18 and 31 locations. The data can be found in the R package fda. We perform the two sample test on this data for the five different choices of T with $\gamma = 1$ or choosen via the median heuristic outlined in the previous subsection. To identify the effect on test performance of sample size we perform random subsampling of the datasets and repeat the test to calculate test power. For each sample size $N \in \{5, 15, 25, 35\}$ we sample N functions from each data set and perform the test, this is repeated 500 times to calculate test power. The results are plotted in Figure 7. Similarly, to investigate the size of the test we sample two disjoint subsets of size $N \in \{5, 15, 25\}$ from the female data set and perform the test and record whether the null was incorrectly rejected, this is repeated 500 times to obtain

N	ID	CEXP	POLY	SQR	FPCA
5	5	4.8	4.6	5.6	5
15	4.4	4.6	4.2	5.2	4.6
25	5	4.6	5	5.8	5.6

Table 1: Empirical size, meaning the rate of rejection of the null when the null is true, of the two-sample test performed on the male Berkeley growth data for different sample sizes across different choices of T. The values are written as percentages, a value above 5 shows too frequent rejection and below shows too conservative a test.

a rate of incorrect rejection of the null. This investigation of the size of the test was performed using the median heuristic and the results are reported in Table 1.

Figure 7 shows POLY, SQR performing the best closely followed by ID. Table 1 shows nearly all the tests have the correct empirical size.

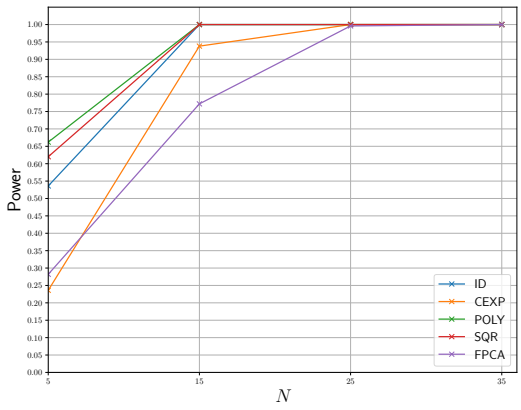

Figure 7: Test power under subsamples of size N using Berkeley growth data.

NEU Steel Data

We perform the two-sample test on two classes from the North Eastern University steel defect dataset [60, 25, 14]. The dataset consists of 200×200 pixel grey scale images of defects of steel surfaces with 6 different classes of defects and 300 images in each class. We perform the test on the two classes which are most visually similar called rolled-in scale and crazing. See the URL [59] for further description of the dataset. For each sample size $N \in \{10, 20, 30, 40\}$ we sample N images from each class and perform the test, this is repeated 500 times to calculate test power. Again we assess the empirical size by sampling two distinct subsets from one class, the rolled-in class, for sample sizes $N \in \{10, 20, 30, 40\}$ and repeat this 500 times and report the rate of incorrect null rejection.

N	ID	CEXP	POLY	SQR	FPCA
10	4.2	4.2	5.4	5.8	5.8
20	4.8	4.8	6.8	5.8	6.2
30	5.2	6.6	7.2	6.2	6.6
40	6.4	7	4.4	4.4	4.8

Table 2: Empirical size, meaning the rate of rejection of the null when the null is true, of the two-sample test performed on the crazing class from the NEU steel defect data, for different sample sizes across different choices of T. The values are written as percentages, a value above 5 shows too frequent rejection and below shows too conservative a test.

Figure 8 shows POLY, SQR having the best performance, PCA performs well and so does CEXP. Table 2 shows that the empirical size is inflated under some choices of T especially CEXP. Once the test is performed with 40 samples the sizes return to an acceptable level for POLY, SQR, FPCA. This inflation of empirical size should be taken into account when viewing the powers of the tests.

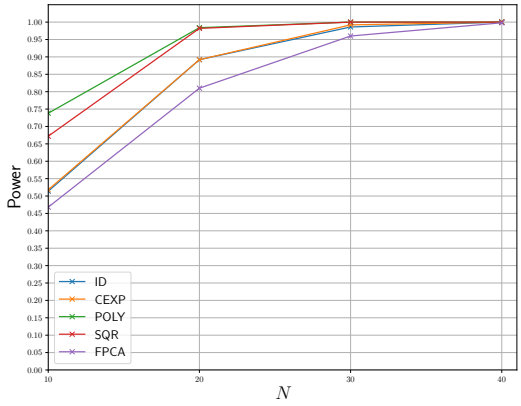

Figure 8: Test power under subsamples of size N using NEU steel data.

Acknowledgments

GW was supported by an EPSRC Industrial CASE award [18000171] in partnership with Shell.

References

[1] S. Albeverio and S. Mazzucchi. An introduction to infinite-dimensional oscillatory and probabilistic integrals. In *Stochastic Analysis: A Series of Lectures*, pages 1–54. Springer Basel, 2015.

[2] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer New York, 1984.

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

[4] J. R. Berrendero, B. Bueno-Larraz, and A. Cuevas. On mahalanobis distance in functional settings. *Journal of Machine Learning Research*, 21(9):1–33, 2020.

[5] P. Billingsley. *Weak Convergence of Measures*. Society for Industrial and Applied Mathematics, 1971.

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

[7] A. Cabana, A. M. Estrada, J. Pena, and A. J. Quiroz. Permutation tests in the two-sample problem for functional data. In *Functional Statistics and Related Fields*, pages 77–85. Springer, 2017.

[8] S. Chakraborty and X. Zhang. A new framework for distance and kernel-based metrics in high dimensions. *arXiv preprint arXiv:1909.13469*, 2019.

[9] H. Chen, P. T. Reiss, and T. Tarpey. Optimally weighted l2 distance for functional data. *Biometrics*, 70(3):516–525, 2014.

[10] I. Chevyrev and H. Oberhauser. Signature moments to characterize laws of stochastic processes, 2018.

[11] A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 406–414. 2010.

[12] A. Cuevas. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23, 2014.

[13] G. Da Prato and J. Zabczyk. *Second Order Partial Differential Equations in Hilbert Spaces*. Cambridge University Press, 2002.

[14] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng. PGA-net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Transactions on Industrial Informatics*, pages 1–1, 2019.

[15] S. N. Ethier and T. G. Kurtz, editors. *Markov Processes*. John Wiley & Sons, Inc., 1986.

[16] G. Fasshauer and M. McCourt. *Kernel-based Approximation Methods using MATLAB*. WORLD SCIENTIFIC, June 2014.

[17] S. FREMDT, J. G. STEINEBACH, L. HORVÁTH, and P. KOKOSZKA. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152, 2012.

[18] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.

[19] T. Gärtner. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 5(1):49, 2003.

[20] S. Ghosal and A. W. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

[21] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, Mar. 2012.

[23] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1205–1213. Curran Associates, Inc., 2012.

[24] P. Hall and I. V. Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 17(4):1511–1531, 2007.

[25] Y. He, K. Song, Q. Meng, and Y. Yan. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1493–1504, 2020.

[26] M. Hein, T. N. Lal, and O. Bousquet. Hilbertian metrics on probability measures and their application in svm's. In *Joint Pattern Recognition Symposium*, pages 270–277. Springer, 2004.

[27] L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

[28] L. Horváth, P. Kokoszka, and R. Reeder. Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):103–122, 2012.

[29] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Ltd, 2015.

[30] W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 262–271. Curran Associates, Inc., 2017.

[31] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

[32] A. S. Kechris. *Classical Descriptive Set Theory*. Springer New York, 1995.

[33] A. Kolmogorov-Smirnov, A. Kolmogorov, and M. Kolmogorov. Sulla determinazione empírica di uma legge di distribuzione. 1933.

[34] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests, 2020.

[35] S. Maniglia and A. Rhandi. Gaussian measures on separable hilbert spaces and applications. 2004, 01 2004.

[36] H. Q. Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2009.

[37] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

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

[39] I. Nourdin and G. Peccati. *Normal Approximations with Malliavin Calculus*. Cambridge University Press, 2009.

[40] V. M. Panaretos, D. Kraus, and J. H. Maddocks. Second-order comparison of gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[41] E. Paparoditis and T. Sapatinas. Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, 103(3):727–733, 2016.

[42] V. I. Paulsen and M. Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2016.

[43] G.-M. Pomann, A.-M. Staicu, and S. Ghosh. A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(3):395–414, Jan. 2016.

[44] G. D. Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Berlin Heidelberg, 2006.

[45] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

[46] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[47] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3571–3577. AAAI Press, 2015.

[48] A. Ramdas, S. J. Reddi, B. Póczos, A. R. Singh, and L. A. Wasserman. On the high-dimensional power of linear-time kernel two-sample testing under mean-difference alternatives. *ArXiv*, abs/1411.6314, 2014.

[49] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[50] S. Saitoh and Y. Sawano. *Theory of reproducing kernels and applications*. Springer, 2016.

[51] P. Schmid. On the kolmogorov and smirnov limit theorems for discontinuous distribution functions. *The Annals of Mathematical Statistics*, 29(4):1011–1027, 1958.

[52] I. J. Schoenberg. Metric spaces and completely monotone functions. *The Annals of Mathematics*, 39(4):811, 1938.

[53] B. Schölkopf, K. Tsuda, and J.-P. Vert. *Support vector machine applications in computational biology*. MIT press, 2004.

[54] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

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

[56] J. Shawe-Taylor, C. K. I. Williams, and N. Cristianiniand J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005.

[57] C.-J. Simon-Gabriel and B. Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *JMLR*, 19(44):1–29, 2018.

[58] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

[59] K. Song and Y. Yan. Neu steel dataset description. `http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html`.

[60] K. Song and Y. Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013.

[61] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *JMLR*, 12:2389–2410, 2011.

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

[63] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008.

[64] I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[65] T. Sullivan. *Introduction to Uncertainty Quantification*. Springer International Publishing, 2015.

[66] D. J. Sutherland. Unbiased estimators for the variance of mmd estimators, 2019.

[67] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy, 2016.

[68] G. J. Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003.

[69] G. J. Székely, M. L. Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

[70] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces*. Springer Netherlands, 1987.

[71] A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

[72] G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

[73] A. Wald and J. Wolfowitz. On a test whether two samples are from the same distribution. *Ann. Math. Stat*, 11:147–162, 1940.

[74] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.

[75] W. J. Wilkinson. *Gaussian Process Modelling for Audio Signals*. PhD thesis, Queen Mary Univeristy London, 2019.

[76] G. Wynne, F.-X. Briol, and M. Girolami. Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness. *arXiv:2001.10818*, 2020.

[77] C. Zhu, S. Yao, X. Zhang, and X. Shao. Distance-based and rkhs-based dependence metrics in high dimension. *Annals of Statistics*, 2019. To appear.

9 Appendix

9.1 Bochner and Minlos-Sazanov Theorem

Bochner's theorem provides an exact relationship between continuous, translation invariant kernels on $\mathbb{R}^d$, meaning $k(x, y) = \phi(x - y)$ for some continuous ϕ, and the Fourier transforms of finite Borel measures on $\mathbb{R}^d$. For a proof see [74, Theorem 6.6].

Theorem 8 (Bochner). *A continuous function $\phi\colon \mathbb{R}^d \to \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite Borel measure μ_ϕ on $\mathbb{R}^d$*

$$\hat{\mu}_\phi(x) \coloneqq \int_{\mathbb{R}^d} e^{ix^T y} d\mu_\phi(y) = \phi(x)$$

Bochner's theorem does not continue to hold in infinite dimensions, for example the kernel $k(x, y) = e^{-\frac{1}{2}\|x-y\|^2}$ on $\mathcal{X}$ is not the Fourier transform of a finite Borel measure on $\mathcal{X}$ [35, Proposition 1.2.11]. Instead, a stronger continuity property is required, this is the content of the Minlos-Sazanov theorem. For a proof see [35, Theorem 1.1.5] or [70, Theorem VI.1.1].

Theorem 9 (Minlos-Sazanov). *Let $\mathcal{X}$ be a real, seperable Hilbert space and $\phi\colon \mathcal{X} \to \mathbb{C}$ a positive definite function on $\mathcal{X}$ then the following are equivalent*

1. *ϕ is the Fourier transform of a finite Borel measure on $\mathcal{X}$*
2. *There exists $C \in L_1^+(\mathcal{X})$ such that ϕ is continuous with respect to the norm induced by C given by $\|x\|_C^2 = \langle Cx, x\rangle$*

The existence of such an operator is a much stronger continuity property than standard continuity on $\mathcal{X}$ and will be crucial in the proof of Theorem 10, one of our main results. To see that continuity with respect to such a C is stronger than usual continuity consider the following example. Fix any $\varepsilon > 0$ and assume we only know that $\phi\colon \mathcal{X} \to \mathbb{R}$ is continuous and for simplicity assume that $\phi(0) = 1$, then we know there exists some $\delta > 0$ such that $\|x\| < \delta$ implies $|\phi(x) - 1| < \varepsilon$ meaning we have control of $\phi(x)$ over the bounded set $\|x\| < \delta$. If ϕ is continuous with respect to $\|\cdot\|_C$ for some $C \in L_1^+(\mathcal{X})$ then we know there exists some $\delta' > 0$ such that $\|x\|_C < \delta'$ implies $|\phi(x) - 1| < \varepsilon$ so we have control of $\phi(x)$ over the *unbounded* set $\|x\|_C < \delta'$. To see this set is unbounded let $\{\lambda_n, e_n\}_{n=1}^\infty$ be the orthonormal eigensystem of C and for $n \in \mathbb{N}$ let $y_n = \frac{\delta' e_n}{2\lambda_n}$ if $\lambda_n > 0$ otherwise $y_n = ne_n$, then since $C \in L_1^+(\mathcal{X})$ we know $\lambda_n \to 0$ so $\|y_n\| \to \infty$. Since we used elements from the eigensystem it is clear that $\|y_n\|_C \leq \delta'/2$. Therefore we have constructed a subset of the ball with respect to $\|\cdot\|_C$ of radius δ' that has unbounded norm.

A way to see that this additional continuity assumption is needed is to note that, informally speaking, there is an intimate relationship between the decay of the density of a measure and the continuity of its Fourier transform. In the case of measures on $\mathcal{X}$ we know that the random variables must satisfy $\|x\|^2 = \sum_{n=1}^\infty \langle x, e_n\rangle^2 < \infty$, for any orthonormal basis $\{e_n\}_{m=1}^\infty$ which imposes a tail condition on the high frequencies of the distribution. This is a decay condition on the probability measure and so one should expect a corresponding continuity condition on the Fourier transform, which is manifested in Theorem 9 through C being trace class.

9.2 Proofs for Section 4

Proof of Lemma 1. The proof follows closely the approximations made in [48] in particular the use of the truncated Taylor series for the squared-exponential kernel. The derivation is a series of tedious Gaussian integrals but we shall highlight the key steps and relation to the referenced work which takes great care to make clear the derivation. We are using the kernel $k(x,y) = \exp(-\frac{1}{2\gamma^2}\|x-y\|^2)$ which has $2\gamma^2$ in the denominator instead of γ^2 as used in [48]. Random variables distributed according to P are denoted x, x' and distributed according to Q denoted y, y'. We now state multiple classical multivariate Gaussian integral identities to be used throughout the calculations

$$\begin{aligned}
&\int \|y\|^2 dQ(y) = \|\mu\|^2 + \text{Tr}(\Sigma) \\
&\int \|y\|^4 dQ(y) = 2\text{Tr}(\Sigma^2) + \text{Tr}(\Sigma)^2 + 4\mu^T\Sigma\mu + \|\mu\|^4 + 2\text{Tr}(\Sigma)\|\mu\|^2 \\
&\int\int \langle x,y\rangle^2 dP(x)dQ(y) = \text{Tr}(\Sigma^2) + \mu^T\Sigma\mu \\
&\int\int\int \|x-x'\|^2\|x-y\|^2 dP(x)dP(x')d\mathcal{N}(\mu,\Sigma)(y) \\
&\quad = 2\text{Tr}(\Sigma^2) + 4\text{Tr}(\Sigma)^2 + 2\text{Tr}(\Sigma)\|\mu\|^2
\end{aligned}$$

The approximation of $\text{MMD}_k(P,Q)^2$ is no different from [48, Lemma 1] and given by $\|\mu\|^2/\gamma^2$ so we focus on approximating $\sqrt{2\xi_2}$. By [48, Lemma 2]

$$\begin{aligned}
\xi_2 &= \mathbb{E}_{x,x'}[k(x,x')^2] + \mathbb{E}_{y,y'}[k(y,y')^2] + 2\mathbb{E}_{x,y}[k(x,y)^2] \\
&\quad + 2\mathbb{E}_{x,x'}[k(x,x')]\mathbb{E}_{y,y'}[k(y,y')] + \mathbb{E}_{x,x',y,y'}[k(x,y)k(x',y')] \\
&\quad - 4\mathbb{E}_{x,x',y}[k(x,y)k(x',y)] - 4\mathbb{E}_{y,y',x}[k(x,y)k(x,y')] - \text{MMD}_k(P,Q)^4
\end{aligned}$$

The following integral is the third term in the above and by setting $\mu = 0$ or replacing γ^2 by $2\gamma^2$ can recover all but the negative terms in the above expression, after discarding terms which have powers higher than γ^4 in the denominator.

$$\begin{aligned}
\mathbb{E}_{x,y}[k(x,y)^2] &= \int\int e^{-\frac{1}{\gamma^2}\|x-y\|^2} dP(x)dQ(y) \\
&\approx \int\int 1 - \frac{\|x-y\|^2}{\gamma^2} + \frac{\|x-y\|^4}{2\gamma^4} dP(x)dQ(y) \\
&= 1 - \frac{2}{\gamma^2}\text{Tr}(\Sigma) - \frac{\|\mu\|^2}{\gamma^2} + \frac{2}{\gamma^4}\text{Tr}(\Sigma^2) + \frac{2}{\gamma^4}\text{Tr}(\Sigma^2) + \frac{4}{\gamma^4}\mu^T\Sigma\mu \\
&\quad + \frac{2}{\gamma^4}\text{Tr}(\Sigma)^2 + \frac{2\text{Tr}(\Sigma)\|\mu\|^2}{\gamma^4} + \frac{\|\mu\|^4}{2\gamma^4}
\end{aligned}$$

comparing this to the analogous calculation [48, Subsection D.2]

$$1 - \frac{2d\sigma^2}{\gamma^2} - \frac{\|\mu\|^2}{\gamma^2} + \frac{4d\sigma^4}{\gamma^4} + \frac{4\sigma^2\|\mu\|^2}{\gamma^4} + \frac{2d^2\sigma^4}{\gamma^4} + \frac{2d\sigma^2\|\mu\|^2}{\gamma^4} + \frac{\|\mu\|^4}{2\gamma^4}$$

its clear that if $\Sigma = \sigma^2 I$ then the two agree. In [48, Subsection D.6] its shown that only constants multiplied by the terms $d\sigma^4/\gamma^4, \sigma^2\|\mu\|^2/\gamma^4$, which correspond to

$\mathrm{Tr}(\Sigma^2)/\gamma^4, \mu^T\Sigma\mu/\gamma^4$, remain in after cancellation in the approximation of ξ_2. Since all other calculations are analogous it follows that making this replacement completes the approximation. For completeness however, we give full expressions of the terms which need to be combined for the expression of ξ_2 full derivations of all can be obtained using the Gaussian integral identities above and following the cancellations made in [48, Section 5]

$$\mathbb{E}_{x,x'}[k(x,x')^2] = \mathbb{E}_{y,y'}[k(y,y')^2] \approx 1 - \frac{2}{\gamma^2}\mathrm{Tr}(\Sigma) + \frac{4}{\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)^2$$

$$\mathbb{E}_{x,y}[k(x,y)^2] \approx 1 - \frac{2}{\gamma^2}\mathrm{Tr}(\Sigma) - \frac{\|\mu\|^2}{\gamma^2} + \frac{4}{\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{4}{\gamma^4}\mu^T\Sigma\mu + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)^2 + \frac{2\mathrm{Tr}(\Sigma)\|\mu\|^2}{\gamma^4} + \frac{\|\mu\|^4}{2\gamma^4}$$

$$\mathbb{E}_{x,x'}[k(x,x')]\mathbb{E}_{y,y'}[k(y,y')] \approx 1 - \frac{2}{\gamma^2}\mathrm{Tr}(\Sigma) + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)^2$$

$$\mathbb{E}_{x,y}[k(x,y)]^2 \approx 1 - \frac{2}{\gamma^2}\mathrm{Tr}(\Sigma) - \frac{\|\mu\|^2}{\gamma^2} + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{2}{\gamma^4}\mu^T\Sigma\mu + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)^2 + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)\|\mu\|^2 + \frac{\|\mu\|^4}{2\gamma^4}$$

$$\begin{aligned}
&\mathbb{E}_{x,x',y}[k(x,x')k(x,y)] = \mathbb{E}_{y,y',x}[k(y,y')k(y,x)] \\
&\approx 1 - \frac{2}{\gamma^2}\mathrm{Tr}(\Sigma) - \frac{\|m\|^2}{2\gamma^2} + \frac{5}{2\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{2}{\gamma^4}\mathrm{Tr}(\Sigma)^2 \\
&+ \frac{1}{\gamma^4}\mu^T\Sigma\mu + \frac{1}{\gamma^4}\mathrm{Tr}(\Sigma)\|\mu\|^2 + \frac{\|m\|^4}{8\gamma^4}
\end{aligned}$$

Putting all these together gives the required approximation:

$$\xi_2 \approx \frac{4}{\gamma^4}\mathrm{Tr}(\Sigma^2) + \frac{4}{\gamma^4}\mu^T\Sigma\mu$$

□

Proof of Lemma 2. We use the standard fact that for a real valued random variables ξ the following inequality holds

$$|\mathbb{E}[\xi] - \mathrm{Median}[\xi]| \leq \mathrm{Var}[\xi]^{\frac{1}{2}}$$

and we will be using this inequality on $\|x-y\|^2$ where $x \sim P, y \sim Q$. First note that $\mathbb{E}[\|x-y\|^2] = \mathrm{Tr}(\Sigma_1+\Sigma_2) + \|\mu_1-\mu_2\|^2 \asymp d$ by the assumptions. It remains to show that $\mathrm{Var}[\|x-y\|^2]^{\frac{1}{2}} \lesssim d$. For economy of notation let $\mu = \mu_1 - \mu_2$ and $\Sigma = \Sigma_1 + \Sigma_2$.

Using the Gaussian integral identities in the proof of Lemma 1 we obtain

$$\begin{aligned}\text{Var}[\|x-y\|^2] &= \mathbb{E}[\|x-y\|^4] - \mathbb{E}[\|x-y\|^2]^2 \\ &= 2\text{Tr}(\Sigma^2) + \text{Tr}(\Sigma)^2 + 4\mu^T(\Sigma)\mu + \|\mu\|^4 + 2\text{Tr}(\Sigma)\|\mu\|^2 \\ &\quad - \text{Tr}(\Sigma)^2 - \|\mu\|^4 - 2\text{Tr}(\Sigma)\|\mu\|^2 \\ &= 2\text{Tr}(\Sigma^2) + 4\mu^T\Sigma\mu \\ &\leq 2\text{Tr}(\Sigma)^2 + 4\text{Tr}(\Sigma)\|\mu\|^2 \\ &\lesssim d^2\end{aligned}$$

therefore we may conclude that Median$[\|x-y\|^2] \asymp d$. □

9.3 Proofs for Section 5

Lemma 3. *The SE-T function is a kernel.*

Proof of Lemma 3. It is shown in [52, Theorem 3] that if $k(x,y) = \phi(\|x-y\|_{\mathcal{Y}}^2)$ for ϕ a completely monotone function then k is positive definite on $\mathcal{Y}$ and it is well known that e^{-ax} is such a function for $a > 0$ therefore k_I is a kernel. Now take k_T to be the SE-T kernel then for any $N \in \mathbb{N}, \{a_n\}_{n=1}^N \subset \mathbb{R}, \{x_n\}_{n=1}^N \subset \mathcal{X}$

$$\sum_{n,m=1}^{N} a_n a_m k_T(x_n, x_m) = \sum_{n,m=1}^{N} a_n a_m k_I(T(x_n), T(x_m)) \geq 0$$

□

Proof of Theorem 1. This proof uses the argument of [36, Theorem 1]. The plan is to first show the function space stated in the theorem is an RKHS and that the kernel is the SE-T kernel so by uniqueness of kernel for RKHS we are done. This is done using the Aronszajn theorem [36, Theorem 9] which identifies the kernel as an infinite sum of basis functions, see also [42, Theorem 2.4].

First we prove that $\mathcal{H}_{k_T}(\mathcal{X})$ is a separable Hilbert space. That it is an inner product space is clear from the definition of the inner product and the assumption that $\lambda_n > 0$ for all $n \in \mathbb{N}$. The definition of the inner product means completeness of $\mathcal{H}_{k_T}(\mathcal{X})$ equivalent to completeness of the weighted l^2 space given by

$$l^2_\lambda(\Gamma) = \left\{ (w_\gamma)_{\gamma\in\Gamma} \colon \|(w_\gamma)_{\gamma\in\Gamma}\|^2_{l^2_\lambda(\Gamma)} = \sum_{\gamma\in\Gamma} \frac{\gamma!}{\lambda^\gamma} w_\gamma^2 \right\}$$

which can be easily seen to be complete since Γ is countable and $\gamma!/\lambda^\gamma$ is positive for all $\gamma \in \Gamma$. To see that $\mathcal{H}_{k_T}(\mathcal{X})$ is separable observe that from the definition of the inner product, the countable set of functions

$$\phi_\gamma(x) = \sqrt{\frac{\lambda^\gamma}{\gamma!}} e^{-\frac{1}{2}\|Tx\|^2} x^\gamma$$

is orthonormal and spans $\mathcal{H}_{k_T}(\mathcal{X})$ hence is an orthonormal basis.

Next we prove that $H_{k_T}(\mathcal{X})$ is an RKHS. Expanding the kernel through the exponential function gives

$$\begin{aligned} k_T(x,y) &= e^{-\frac{1}{2}\|Tx\|^2} e^{-\frac{1}{2}\|Tx\|^2} e^{\langle Tx, Ty\rangle} \\ &= e^{-\frac{1}{2}\|Tx\|^2} e^{-\frac{1}{2}\|Tx\|^2} \sum_{n=0}^{\infty} \frac{\langle Tx, Ty\rangle^n}{n!} \end{aligned}$$

and by the assumption on T

$$\langle Tx, Ty\rangle^n = \left(\sum_{m=1}^{\infty} \lambda_m x_m y_m\right)^n = \sum_{|\gamma|=n} \frac{n!}{\gamma!} \lambda^\gamma x^\gamma y^\gamma$$

where $x_m = \langle x, e_m\rangle$, similarly for y_m, therefore

$$k_T(x,y) = e^{-\frac{1}{2}\|Tx\|^2} e^{-\frac{1}{2}\|Tx\|^2} \sum_{|\gamma|\geq 0} \frac{\lambda^\gamma}{\gamma!} x^\gamma y^\gamma = \sum_{|\gamma|\geq 0} \phi_\gamma(x)\phi_\gamma(y)$$

So for any $F \in \mathcal{H}_{k_T}(\mathcal{X})$ we have

$$\langle F, k_T(\cdot, x)\rangle_{\mathcal{H}_{k_T}(\mathcal{X})} = \sum_{\gamma\in\Gamma} \langle F, \phi_\gamma\rangle_{\mathcal{H}_{k_T}(\mathcal{X})} \phi_\gamma(x) = F(x)$$

so k_T is a reproducing kernel of $\mathcal{H}_{k_T}(\mathcal{X})$ so by uniqueness of reproducing kernels we have that $\mathcal{H}_{k_T}(\mathcal{X})$ is the RKHS of k_T. □

Proof of Theorem 2. If $k_T(x,y) = \hat{\mu}(x-y)$ for some Borel measure on $\mathcal{X}$ then Equation 2 lets us write

$$\begin{aligned} &\mathrm{MMD}_{k_T}(P,Q)^2 \\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} k_T(x,y) d(P-Q)(x) d(P-Q)(y) \\ &= \int_{\mathcal{X}}\int_{\mathcal{X}}\int_{\mathcal{X}} e^{i\langle h, x-y\rangle} d\mu(h) d(P-Q)(x) d(P-Q) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} e^{i\langle h,x\rangle} d(P-Q)(x)\right)\left(\int_{\mathcal{X}} e^{-i\langle h,y\rangle} d(P-Q)(y)\right) d\mu(h) \qquad (17) \\ &= \int_{\mathcal{X}} \left|\hat{P}(h) - \hat{Q}(h)\right|^2 d\mu(h) \end{aligned}$$

where Equation 17 is obtained by using Fubini's theorem to swap the integrals which is permitted since $|e^{i\langle h, x-y\rangle}| = 1$ and is therefore integrable with respect to P and Q.

Fourier transforms of finite Borel measures on $\mathcal{X}$ are uniformly continuous [1, Proposition 2.21] therefore if μ has full support, meaning that $\mu(U) > 0$ for every open $U \subset \mathcal{X}$, then we may immediately conclude that $\hat{P} = \hat{Q}$ and since the Fourier transform of finite Borel measures on $\mathcal{X}$ is injective [44, Proposition 1.7] we may conclude that $P = Q$ meaning that Φ_{k_T} is injective.

Assume T is non-degenerate. If k_T is the SE-T kernel then [44, Proposition 1.25] shows that $\mu = N_{T^2}$ has full support since T^2 is also non-denegerate. Therefore k_T is characteristic.

For the converse direction we use the contrapositive and assume that T is degenerate meaning there exists $x^* \in \mathcal{X}$ with $x^* \neq 0$ and $Tx = 0$. Set $P = \delta_0$ and $Q = \delta_{x^*}$ the Dirac measures on $0, x^*$ then

$$\Phi_{k_T}(P) = k_T(0, \cdot) = k_T(x^*, \cdot) = \Phi_{k_T}(Q)$$

Therefore Φ_{k_T} is not injective so k_T is not characteristic. □

Proof of Theorem 3. The result is a corollary of the next theorem which is where Theorem 9 is employed.

Theorem 10. *Let $\mathcal{X}$ be a real, separable Hilbert space and $T = I$ then the SE-T kernel is characteristic.*

Proof. The idea of this proof is to use the contrapositive and assume $P \neq Q$ and conclude that $\text{MMD}_{k_I}(P, Q) > 0$. The main tool shall be Theorem 9 since $P \neq Q$ implies $\hat{P} \neq \hat{Q}$ and Theorem 9 implies that these Fourier transforms vary slowly in some sense so there will be a set of big enough measure such that $\hat{P}(x) \neq \hat{Q}(x)$ for x in this set, which will allow us to deduce $\text{MMD}_{k_I}(P, Q) > 0$.

Suppose P, Q are Borel probability measures on $\mathcal{X}$ with $P \neq Q$ then $\hat{P} \neq \hat{Q}$ [44, Proposition 1.7] so there exists $x^* \in \mathcal{X}, \varepsilon > 0$ such that $|\hat{P}(x^*) - \hat{Q}(x^*)| > \varepsilon$. By Theorem 9 there exists $S, R \in L_1^+(\mathcal{X})$ such that $\hat{P}$ (respectively $\hat{Q}$) is continuous with respect to the norm induced by S (repectively R).

For $r > 0$ let $B_S(x^*, r) = \{x \in \mathcal{X} \colon \langle S(x - x^*), x - x^* \rangle < r^2\}$ be the ball based at x^* of radius r with respect to the norm induced by S, and $B_R(x^*, r)$ will denote the analogous ball with respect to the norm induced by R. By the continuity properties of $\hat{P}, \hat{Q}$ there exists $r > 0$ such that

$$\begin{aligned} x \in B_S(x^*, r) &\implies |\hat{P}(x) - \hat{P}(x^*)| < \frac{\varepsilon}{4} \\ x \in B_R(x^*, r) &\implies |\hat{Q}(x) - \hat{Q}(x^*)| < \frac{\varepsilon}{4} \end{aligned}$$

The set $B_S(x^*, r) \cap B_R(x^*, r)$ is non-empty since it contains x^* and if $x \in B_S(x^*, r) \cap B_R(x^*, r)$ then by reverse triangle inequality

$$\begin{aligned} |\hat{P}(x) - \hat{Q}(x)| &= |\hat{P}(x) - \hat{P}(x^*) + \hat{P}(x^*) - \hat{Q}(x^*) - \hat{Q}(x)| \\ &\geq |\hat{P}(x^*) - \hat{Q}(x^*)| - |\hat{P}(x) - \hat{P}(x^*)| - |\hat{Q}(x) - \hat{Q}(x^*)| \\ &> \varepsilon - \frac{\varepsilon}{4} - \frac{\varepsilon}{4} \\ &= \frac{\varepsilon}{2} \end{aligned} \tag{18}$$

Now define the operator $U = S + R$ which is positive, symmetric and trace class since both S and R have these properties. Note that $B_U(x^*, r) \subset B_S(x^*, r) \cap B_R(x^*, r)$ because

$$\begin{aligned} \|x - x^*\|_U^2 = \langle U(x - x^*), x - x^* \rangle &= \langle S(x - x^*), x - x^* \rangle + \langle R(x - x^*), x - x^* \rangle \\ &= \|x - x^*\|_S^2 + \|x - x^*\|_R^2 \end{aligned}$$

Since U is a positive, compact, symmetric operator there exists a decomposition into its eigenvalues $\{\lambda_n\}_{n=1}^{\infty}$, a non-negative sequence converging to zero, and eigenfunctions $\{e_n\}_{n=1}^{\infty}$ which form an orthonormal basis of $\mathcal{X}$. We will later need to associate a non-degenerate Gaussian measure with U. To this end define V to be the positive, symmetric, trace class operator with eigenvalues $\{\rho_n\}_{n=1}^{\infty}$ where $\rho_n = \lambda_n$ if $\lambda_n > 0$ otherwise $\rho_n = n^{-2}$, and eigenfunctions $\{e_n\}_{n=1}^{\infty}$ inherited from U. Then by construction V is injective, positive, symmetric and trace class. The V induced norm dominates the U induced norm therefore $B_V(x^*, r) \subset B_U(x^*, r)$ so for $x \in B_V(x^*, r)$ we have $|\hat{P}(x) - \hat{Q}(x)| > \varepsilon/2$.

Now we construct an operator which will approximate I, define the operator $I_m x = \sum_{n=1}^{\infty} \omega_n^{(m)} \langle x, e_n \rangle e_n$ where $\omega_n^{(m)} = 1$ for $n \leq m$ and $\omega_n^{(m)} = n^{-2}$ for $n > m$ and $\{e_n\}_{n=1}^{\infty}$ is the eigenbasis of V, then $I_m \in L_1^+(\mathcal{X})$ for every $m \in \mathbb{N}$. It is easy to see $k_{I_m^{1/2}}$ converges pointwise to k_I as $m \to \infty$ since $e^{-\frac{1}{2}x}$ is a continuous function on $\mathbb{R}$ and $\|x\|_{I_m^{1/2}}^2 \to \|x\|_I^2 = \|x\|^2$ however clearly $I_m^{\frac{1}{2}}$ does not converge in operator norm to I since I is not a compact operator.

Since $k_{I_m} \leq 1$ for all m we may use the bounded convergence theorem to obtain

$$\begin{aligned}
\text{MMD}_{k_I}(P, Q)^2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} k_I(x, y) d(P - Q)(x) d(P - Q)(y) \\
&= \lim_{m \to \infty} \int_{\mathcal{X}} \int_{\mathcal{X}} k_{I_m^{1/2}}(x, y) d(P - Q)(x) d(P - Q)(y) \\
&= \lim_{m \to \infty} \int_{\mathcal{X}} |\hat{P}(x) - \hat{Q}(x)|^2 dN_{I_m}(x) \qquad (19)
\end{aligned}$$

where Equation 19 is by the same reasoning as in the proof of Theorem 2. In light of the lower bound we derived earlier over $B_V(x^*, r)$ of the integrand in Equation 18

$$\begin{aligned}
\text{MMD}_{k_I}(P, Q)^2 &= \lim_{m \to \infty} \int_{\mathcal{X}} |\hat{P}(x) - \hat{Q}(x)|^2 dN_{I_m}(x) \\
&\geq \lim_{m \to \infty} \int_{B_V(x^*, r)} \frac{\varepsilon^2}{4} dN_{I_m}(x)
\end{aligned}$$

so if we can lower bound $N_{I_m}(B_V(x^*, r))$ by a positive constant indepdendent of m then we are done. This set is the ball with respect to $V \in L_1^+(\mathcal{X})$ which is a somehow large set, see the discussion after Theorem 9, and we will use a push-forward of measure argument.

Define $T(x) = x - x^*$ then $N_{I_m}(B_V(x^*, r)) = T_{\#} N_{I_m}(B_V(0, r))$ and [44, Proposition 1.17] tells us $T_{\#} N_{I_m}(B_V(0, r)) = N_{-x^*, I_m}(B_V(0, r))$. Next we note that $N_{-x^*, I_m}(B_V(0, r)) = V_{\#}^{\frac{1}{2}} N_{-x^*, I_m}(B(0, r))$ and [44, Proposition 1.18] tells us that

$$V_{\#}^{\frac{1}{2}} N_{-x^*, I_m}(B(0, r)) = N_{-V^{\frac{1}{2}} x^*, V^{\frac{1}{2}} I_m V^{\frac{1}{2}}}(B(0, r))$$

For ease of notation let $y^* = -V^{\frac{1}{2}} x^*$ and since we choose to construct I_m from the eigenbasis of V we have $V_m x \coloneqq V^{\frac{1}{2}} I_m V^{\frac{1}{2}} x = \sum_{n \in \mathbb{N}} \rho_n^{(m)} \langle x, e_n \rangle e_n$ where $\rho_n^{(m)} = \rho_n$ for $n \leq m$ and $\rho_n^{(m)} = \rho_n n^{-2}$ for $n > m$ so $V_m \in L_1^+(\mathcal{X})$ and is injective

for every $m \in \mathbb{N}$. We follow the proof of [44, Proposition 1.25] and define the sets

$$A_l = \left\{ x \in \mathcal{X} \colon \sum_{n=1}^{l} \langle x, e_n \rangle^2 \leq \frac{r^2}{2} \right\}$$
$$B_l = \left\{ x \in \mathcal{X} \colon \sum_{n=l+1}^{\infty} \langle x, e_n \rangle^2 < \frac{r^2}{2} \right\}$$

Since V_m is non-degenerate for every $m \in \mathbb{N}$ we have that for every $l \in \mathbb{N}$ the events A_l, B_l are independent under N_{y^*,V_m} [44, Example 1.22] meaning $\forall m, l \in \mathbb{N}$ we have

$$N_{y^*,V_m}(B(0,r)) \geq N_{y^*,V_m}(A_l \cap B_l) = N_{y^*,V_m}(A_l) N_{y^*,V_m}(B_l)$$

and by the measure theoretic Chebyshev inequality, for every $l \in \mathbb{N}$

$$\begin{aligned} N_{y^*,V_m}(B_l) &\geq 1 - N_{y^*,V_m}(B_l^c) \\ &\geq 1 - \frac{2}{r^2} \sum_{n=l+1}^{\infty} \int_{\mathcal{X}} \langle x, e_n \rangle^2 dN_{y^*,V_m} \\ &= 1 - \frac{2}{r^2} \left(\sum_{n=l+1} \rho_n^{(m)} + \langle y^*, e_n \rangle^2 \right) \\ &\geq 1 - \frac{2}{r^2} \left(\sum_{n=l+1}^{\infty} \rho_n + \langle y^*, e_n \rangle^2 \right) \end{aligned}$$

As the final line involves the tail of a finite sum with no dependcy on m there exists an $L \in \mathbb{N}$ such that $N_{y^*,V_m}(B_L) > \frac{1}{2}$ for every $m \in \mathbb{N}$ and $l \geq L$. Note that for $m > L$ we have $N_{y^*,V_m}(A_L) = N_{y^*,V}(A_L)$ since A_L depends only on the first L coordinates and $\rho_n^{(m)} = \rho_n$ for $n \leq L$ if $m > L$. So for $m > L$

$$N_{y^*,V_m}(A_L) = N_{y^*,V}(A_L) \geq N_{y^*,V}\left(B\left(0, \frac{r}{\sqrt{2}}\right)\right) > c > 0$$

for some c since V is non-degenerate [44, Proposition 1.25].

Overall we have shown that there exists an $L \in \mathbb{N}$ such that for $m > L$ we have $N_{I_m}(B_V(x^*, r)) = N_{y^*,V_m}(B(0,r)) > \frac{c}{2}$. Therefore, by substituting back into the lower bound for $\text{MMD}_{k_I}(P,Q)^2$

$$\begin{aligned} \text{MMD}_{k_I}(P,Q)^2 &= \lim_{m \to \infty} \int_{\mathcal{X}} |\hat{P}(x) - \hat{Q}(x)|^2 dN_{I_m}(x) \\ &\geq \lim_{m \to \infty} \int_{B_V(x^*,r)} \frac{\varepsilon^2}{4} dN_{I_m}(x) \\ &> \frac{\varepsilon^2 c}{8} > 0 \end{aligned}$$

which implies by contrapositive that k_I is characteristic. □

With Theorem 10 proved we proceed to prove Theorem 3. By Equation 2

$$\begin{aligned}
\mathrm{MMD}_{k_T}(P,Q)^2 &= \int\int k_T(x,x')dP(x)dP(x') + \int\int k_T(y,y')dQ(y)dQ(y') \\
&\quad - 2\int\int k_T(x,y)dQ(x)dP(y) \\
&= \int\int k_I(x,x')dT_{\#}P(x)dT_{\#}P(x') + \int\int k_I(y,y')dT_{\#}Q(y)dT_{\#}Q(y') \\
&\quad - 2\int\int k_I(x,y)dT\#P(x)dT_{\#}Q(y) \qquad (20) \\
&= \mathrm{MMD}_{k_I}(T_{\#}P, T_{\#}Q)^2
\end{aligned}$$

Using Theorem 10 we know that if $\mathrm{MMD}_{k_T}(P,Q) = 0$ then $T_{\#}P = T_{\#}Q$ and all that is left to show is that the assumption on T implies $P = Q$. By the definition of push-forward measure we know that for every $B \in \mathcal{B}(\mathcal{Y})$ we have the equality $P(T^{-1}B) = Q(T^{-1}B)$. By the assumptions on $\mathcal{X}, \mathcal{Y}, T$ we know that $T(A) \in \mathcal{B}(\mathcal{Y})$ for every $A \in \mathcal{B}(\mathcal{X})$ [32, Theorem 15.1]. Hence for any $A \in \mathcal{B}(\mathcal{X})$ take $B = T(A)$ then $P(A) = P(T^{-1}B) = Q(T^{-1}B) = Q(A)$, which shows $P = Q$. □

Proof of Proposition 1. For any $N \in \mathbb{N}$ take any $\{a_n\}_{n=1}^N \subset \mathbb{R}, \{x_n\}_{n=1}^N \subset \mathcal{X}$ then

$$\sum_{n,m=1}^{N} k_{C_{k_0}}(x_n, x_m) = \int_{\mathcal{X}} \sum_{n,m=1}^{N} k_0(\langle x_n, h\rangle, \langle x_m, h\rangle) dN_C(h) \geq 0$$

since this is the integral of a non-negative quantity as k_0 is a kernel. Symmetry of $k_{C_{k_0}}$ follows since k_0 is symmetric meaning k_{C,k_0} is a kernel.

Expanding k_0 using its spectral measure we have

$$k_{C,k_0}(x,y) = \int_{\mathcal{X}}\int_{\mathbb{R}} e^{i\langle x-y, rh\rangle} d\mu(r) dN_C(h) = \hat{\nu}(x-y)$$

where $\nu(A) = \int_{\mathcal{X}}\int_{\mathbb{R}} \mathbb{1}_A(rh) d\mu(r) dN_C(h)$ for all $A \in \mathcal{B}(\mathcal{X})$. This is the law of the $\mathcal{X}$ valued random variable ξX where $\xi \sim \mu$ and $X \sim N_C$ are independent. We will show that ν has full support on $\mathcal{X}$ from which is follows that k_{C,k_0} is characteristic by following the proof of Theorem 2.

Fix any open ball $B = B(h,r) \subset \mathcal{X}$ then given the assumption on μ by intersecting with $(0,\infty)$ or $(-\infty, 0)$ we may assume that a, b have the same sign. Assume that $a, b > 0$, the proof for when $a, b < 0$ is analogous. We first treat the case $h \neq 0$. Set $\delta = \min(\frac{1}{2}(\frac{b}{a} - 1), \frac{r}{2\|h\|})$ so that $(a, a(1+\delta)) \subset (a,b)$. Now consider the ball $B' = B(\frac{h}{a(1+\delta)}, \frac{r}{4a(1+\delta)})$, take any $c \in (a, a(1+\delta))$ and any $x \in B'$ then

$$\begin{aligned}
\|cx - h\| &\leq \left\| \xi x - \frac{\xi h}{a(1+\delta)} \right\| + \left\| \frac{\xi h}{a(1+\delta)} - h \right\| \\
&\leq \frac{cr}{4a(1+\delta)} + \|h\| \left(1 - \frac{c}{a(1+\delta)}\right) \\
&< \frac{r}{4} + \|h\| \left(1 - \frac{1}{1+\delta}\right) \\
&< \frac{r}{4} + \frac{r}{2} < r
\end{aligned}$$

Therefore for any $c \in (a, a(1+\delta))$ we have $cB' \subset B$. Hence $\mathbb{P}(\xi X \in B) = \nu(B) \geq \mu\big((a, a(1+\delta))\big) N_C(B') > 0$ by the assumptions on μ and the way N_C is non-degenerate.

The case $h = 0$ is analogous, take $B' = B(0, \frac{r}{2b})$ then for every $c \in (a, b)$ we have $cB' \subset B$ and we again conclude $\nu(B) > 0$. □

Proof of Corollary 1. The idea of the proof is to represent the IMQ-T kernel as an integral of the SE-T kernel then use the same limit argument as in the proof of Theorem 10 and push-forward argument of Theorem 3.

Throughout this proof k_T^{IMQ} and k_T^{SE} will denote the IMQ-T and SE-T kernels respectively. By the same proof technique as Theorem 3 is suffices to prove that k_I^{IMQ} is characteristic. Let $\{e_n\}_{n=1}^{\infty}$ be any orthonormal basis of $\mathcal{X}$ and let $I_m x = \sum_{n=1}^{m} \langle x, e_n \rangle e_n$ so that I_m converges to I pointwise. Then by the same limiting argument in the proof of Theorem 10 using bounded convergence theorem, letting N_1 be the $\mathcal{N}(0,1)$ measure on $\mathbb{R}$ we have

$$
\begin{aligned}
&\text{MMD}_{k_I^{\text{IMQ}}}(P,Q)^2 \\
&\qquad = \lim_{m\to\infty} \int_{\mathcal{X}} \int_{\mathcal{X}} k_{I_m}^{\text{IMQ}}(x,y) d(P-Q)(x) d(P-Q)(y) \\
&\qquad = \lim_{m\to\infty} \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathbb{R}} k_{I_m}^{\text{SE}}(zx, zy) dN_1(z) d(P-Q)(x) d(P-Q)(y) \qquad (21) \\
&\qquad = \int_{\mathbb{R}} \text{MMD}_{k_I^{\text{SE}}}(z_{\#}P, z_{\#}Q)^2 dN_1(z) \qquad (22)
\end{aligned}
$$

Equation 21 is from the integral representation of $k_{I_m}^{\text{IMQ}}$ in Equation 13 which can be used since $I_m \in L_1^+(\mathcal{X})$, Equation 22 is obtained by using Fubini's theorem and bounded convergence theorem and $z_{\#}P$ denotes the push-forward of the measure under the liner map from $\mathcal{X}$ to $\mathcal{X}$ defined as multiplication by the scalar z. The integrand in Equation 22 can be rewritten as

$$
\begin{aligned}
\text{MMD}_{k_I^{\text{SE}}}(z_{\#}P, z_{\#}Q)^2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} k_I^{\text{SE}}(x,y) d(z_{\#}P - z_{\#}Q)(x) d(z_{\#}P - z_{\#}Q)(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k_I^{\text{SE}}(zx, zy) d(P-Q)(x) d(P-Q)(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} e^{-\frac{z^2}{2}\|x-y\|^2} d(P-Q)(x) d(P-Q)(y) \\
&= \text{MMD}_{k_{z^2 I}^{\text{SE}}}(P,Q)^2
\end{aligned}
$$

which makes it clear that $\text{MMD}_{k_I^{\text{SE}}}(z_{\#}P, z_{\#}Q)^2$ is a continuous, non-negative function of z and equals 0 if and only if $z = 0$. Using this we deduce

$$
\begin{aligned}
\text{MMD}_{k_I^{\text{IMQ}}}(P,Q)^2 &= \int_{\mathbb{R}} \text{MMD}_{k_I^{\text{SE}}}(z_{\#}P, z_{\#}Q)^2 dN_{\sigma^2}(z) \\
&= \int_{\mathbb{R}} \text{MMD}_{k_{z^2 I}^{\text{SE}}}(P,Q)^2 dN_{\sigma^2}(z) \\
&> 0
\end{aligned}
$$

since N_{σ^2} has strictly positive density. This proves that the IMQ-I kernel is characteristic and by the same push-forward argument as in the proof of Theorem 3 we may conclude that the IMQ-T kernel is characteristic too. □

9.4 Proofs for Section 6

Proof of Proposition 2. First note that the functions $e^{-\frac{x^2}{2}}$ and $(x^2+1)^{-\frac{1}{2}}$ have Lipschitz constants $L_{SE} = e^{-\frac{1}{2}}$ and $L_{IMQ} = \frac{2}{3\sqrt{3}}$ respectively. Therefore for any $x, y \in \mathcal{X}$

$$\begin{aligned}
|k_T(x,y) - k_T(\mathcal{R}\tilde{x}, \mathcal{R}\tilde{y})| &\leq L|\|T(x) - T(y)\|_{\mathcal{Y}} - \|T(\mathcal{R}\tilde{x}) - T(\mathcal{R}\tilde{y})\|_{\mathcal{Y}}| \\
&\leq L\|T(x) - T(y) - T(\mathcal{R}\tilde{x}) + T(\mathcal{R}\tilde{y})\|_{\mathcal{Y}} \qquad (23) \\
&\leq L\big(\|T(\mathcal{R}\tilde{x}) - T(x)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}) - T(y)\|_{\mathcal{Y}}\big) \qquad (24)
\end{aligned}$$

where L may be set to L_{SE}, L_{IMQ} if k_T is the SE-T or IMQ-T kernel. Equation 23 uses the reverse triangle inequality and Equation 24 uses the triangle inequality. Using Equation 3 gives

$$\begin{aligned}
&\left|\widehat{\mathrm{MMD}}_{k_T}(X_n, Y_n)^2 - \widehat{\mathrm{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2\right| \\
&\qquad \leq \frac{1}{n(n-1)} \sum_{i \neq j}^{n} |h(z_i, z_j) - h(\mathcal{R}\tilde{z}_i, \mathcal{R}\tilde{z}_j)| \\
&\qquad \leq \frac{2L}{n(n-1)} \sum_{i \neq j}^{n} \|T(\mathcal{R}\tilde{x}_i) - T(x_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{x}_j) - T(x_j)\|_{\mathcal{Y}} \\
&\qquad\qquad + \|T(\mathcal{R}\tilde{y}_i) - T(y_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_j) - T(y_j)\|_{\mathcal{Y}} \qquad (25) \\
&\qquad = \frac{4L}{n} \sum_{i=1}^{n} \|T(\mathcal{R}\tilde{x}_i) - T(x_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_i) - T(y_i)\|_{\mathcal{Y}} \qquad (26)
\end{aligned}$$

where Equation 25 follows from expanding using the definition of h and using the triangle inequality and Equation 26 follows from counting the number of pairs of indices in the sum. Substituing in L_{SE} or L_{IMQ} completes the proof. □

Proof of Theorem 4. We adopt the notation for L from the proof of Proposition 2 for the Lipschitz constants of the kernels. Define the two random variables

$$\begin{aligned}
A_n &= n^{\frac{1}{2}}\left(\widehat{\mathrm{MMD}}_{k_T}(\mathcal{R}\tilde{X}_n, \mathcal{R}\tilde{Y}_n)^2 - \widehat{\mathrm{MMD}}_{k_T}(X_n, Y_n)^2\right) \\
B_n &= n^{\frac{1}{2}}\left(\widehat{\mathrm{MMD}}_{k_T}(X_n, Y_n)^2 - \mathrm{MMD}_{k_T}(P, Q)^2\right)
\end{aligned}$$

It is known $B_n \to \mathcal{N}(0, \xi)$ [22, Corollay 16] so the proof is complete by Slutsky's

theorem if $A_n \xrightarrow{\mathbb{P}} 0$. Fix any $\varepsilon > 0$ then by Proposition 2

$$\begin{aligned}
\mathbb{P}(|A_n| > \varepsilon) &\leq \mathbb{P}\left(\frac{4L}{n^{\frac{1}{2}}}\sum_{i=1}^{n}\|T(\mathcal{R}\tilde{x}_{i,u(n)}) - T(x_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_{i,v(n)}) - T(y_i)\|_{\mathcal{Y}} > \varepsilon\right) \\
&\leq \frac{4L}{\varepsilon n^{\frac{1}{2}}}\mathbb{E}\left[\sum_{i=1}^{n}\|T(\mathcal{R}\tilde{x}_{i,u(n)}) - T(x_i)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_{i,v(n)}) - T(y_i)\|_{\mathcal{Y}}\right] \qquad (27) \\
&= \frac{4Ln^{\frac{1}{2}}}{\varepsilon}\mathbb{E}[\|T(\mathcal{R}\tilde{x}_{u(n)}) - T(x)\|_{\mathcal{Y}} + \|T(\mathcal{R}\tilde{y}_{v(n)}) - T(y)\|_{\mathcal{Y}}] \qquad (28) \\
&\to 0 \qquad (29)
\end{aligned}$$

where Equation 27 is by Markov's inequality, Equation 28 is by the assumption that the samples from P, Q and disctretisation U, V is i.i.d. across samples and Equation 29 is by assumption. □

Proof of Theorem 5. Suppose $P_n \xrightarrow{w} P$ then by [57, Lemma 10], which holds in our case since the key intermediate result [2, Theorem 3.3] only requires $\mathcal{X}$ to be a Hausdorff space, we have $\text{MMD}(P_n, P) \to 0$.

Suppose $\text{MMD}(P_n, P) \to 0$, by Prokhorov's theorem [5, Section 5] we know that $\{P_n\}_{n=1}^{\infty}$ is relatively compact. Since k is characteristic we know that $\mathcal{H}_k$ is a separating set in the sense of [15, Chapter 4] and $\text{MMD}(P_n, P) \to 0$ implies that for every $F \in \mathcal{H}_k$ that $\lim_{n\to\infty}\int F dP_n = \int F dP$ therefore [15, Lemma 3.4.3] applies and we may conclude that $P_n \xrightarrow{w} P$. □

Proof of Theorem 6. Note that $\Phi_{k_T}(N_{a,S})(x) = \Phi_{k_I}(N_{Ta,TST})(Tx)$ [44, Proposition 1.18]. The proof simply uses [13, Proposition 1.2.8] to calculate the Gaussian integrals.

$$\begin{aligned}
\Phi_{k_I}(N_{a,S})(x) &= \int_{\mathcal{X}} e^{-\frac{1}{2}\langle x-y, x-y\rangle} dN_{a,S}(y) \\
&= e^{-\frac{1}{2}\langle a-x, a-x\rangle}\int_{\mathcal{X}} e^{-\frac{1}{2}\langle y,y\rangle} e^{-\langle y, a-x\rangle} dN_S(y) \\
&= \|I+S\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle a-x, a-x\rangle} e^{\frac{1}{2}\|(I+S)^{-\frac{1}{2}} S^{\frac{1}{2}}(a-x)\|^2} \qquad (30) \\
&= \|I+S\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle a-x, a-x\rangle} e^{\frac{1}{2}\langle S^{\frac{1}{2}}(I+S)^{-1}S^{\frac{1}{2}}(a-x), a-x\rangle} \\
&= \|I+S\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I-S^{\frac{1}{2}}(I+S)^{-1}S^{\frac{1}{2}})(a-x), (a-x)\rangle} \\
&= \|I+S\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I+S)^{-1}(x-a), x-a\rangle} \qquad (31)
\end{aligned}$$

where Equation 30 is due to [13, Proposition 1.2.8]. The last equality is due to the Sherman-Morrison-Woodbury identity for operators [29, Theorem 3.5.6]. Substituting in Ta for a, STS for S and Tx for x gives the desired expression, as discussed at the start of the proof. □

Proof of Theorem 7. The idea of the proof is that $\text{MMD}_{k_T}(P, Q)^2$ is simply double integrals of the kernel with respect to Gaussian measures. One integral was completed in Theorem 6 and we apply [13, Proposition 1.2.8] again. Note

$\text{MMD}_{k_T}(N_{a,S}, N_{b,R})^2 = \text{MMD}_{k_I}(N_{Ta,TST}, N_{Tb,TRT})^2$ so it suffices to do the calculations for k_I and substitute the other values in. Also since k_T is translation invariant we may without loss of generality assume $a = 0$ and replace b with $a - b$ at the end .

$$\begin{aligned}
&\int_{\mathcal{X}}\int_{\mathcal{X}} k_I(x,y)dN_S(x)dN_{b,R}(y) \\
&= \|I+S\|^{-\frac{1}{2}} \int_{\mathcal{X}} e^{-\frac{1}{2}\langle (I+S)^{-1}(y-b),y-b\rangle} dN_R(y) && (32)\\
&= \|I+S\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I+S)^{-1}b,b\rangle} \int_{\mathcal{X}} e^{-\frac{1}{2}\langle (I+S)^{-1}y,y\rangle} e^{\langle y,(I+S)^{-1}b\rangle} dN_R(y) \\
&= \|I+S\|^{-\frac{1}{2}} \|I+R^{\frac{1}{2}}(I+S)^{-1}R^{\frac{1}{2}}\|^{-\frac{1}{2}} && (33)\\
&\qquad \times e^{-\frac{1}{2}\langle (I+S)^{-1}b,b\rangle} e^{-\frac{1}{2}\langle (I+S)^{-1}R^{\frac{1}{2}}(I+R^{\frac{1}{2}}(I+S)^{-1}R^{\frac{1}{2}})^{-1}R^{\frac{1}{2}}(I+S)^{-1}b,b\rangle} \\
&= \|I+S\|^{-\frac{1}{2}} \|I+R^{\frac{1}{2}}(I+S)^{-1}R^{\frac{1}{2}}\|^{-\frac{1}{2}} && (34)\\
&\qquad \times e^{-\frac{1}{2}\langle (I-(I+S)^{-1}R^{\frac{1}{2}}(I+R^{\frac{1}{2}}(I+S)^{-1}R^{\frac{1}{2}})^{-1}R^{\frac{1}{2}})(I+S)^{-1}b,b\rangle} \\
&= \|I+S\|^{-\frac{1}{2}} \|I+R^{\frac{1}{2}}(I+S)^{-1}R^{\frac{1}{2}}\|^{-\frac{1}{2}} e^{-\frac{1}{2}\langle (I+S+R)^{-1}b,b\rangle} && (35)
\end{aligned}$$

where Equation 32 is obtained by substituting the result of Theorem 6, Equation 33 is applying [13, Proposition 1.2.8], Equation 34 is just rearranging terms and Equation 35 is using the Sherman-Morrison-Woodbury identity for operators [29, Theorem 3.5.6]. The proof is completed by using the expression of MMD in terms of three double integrals and substituting in the appropriate values of S, R, b inline with the description at the start of the proof. In particular when $b = 0$ and $S = R$

$$\begin{aligned}
\|I+S\|\|I+S^{\frac{1}{2}}(I+S)^{-1}S^{\frac{1}{2}}\| &= \|(I+S)(I+(I+S)^{-1}S)\| \\
&= \|I+2S\|
\end{aligned}$$

by the Sherman-Morrison-Woodbury identity for operators. □

Proof of Corollary 2. If C, S, R commute then

$$\begin{aligned}
\|I+S^{\frac{1}{2}}CS^{\frac{1}{2}}\|\|I+R^{\frac{1}{2}}C(I+CS)^{-1}R^{\frac{1}{2}}\| &= \|I+CS\|\|I+RC(I+CS)^{-1}\| \\
&= \|(I+CS)(I+RC(I+CS)^{-1})\| \\
&= \|I+C(S+R)\|
\end{aligned}$$

□

9.5 Proof for Section 7

Proof of Proposition 3. Suppose k_0 is ISPD and T_{k_0} isn't injective. Then there exists non-zero $x \in L^2(\Omega)$ such that $T_{k_0}x = 0$ so $\int_\Omega \int_\Omega x(s)k_0(s,t)x(t)dsdt = \langle x, T_{k_0}x\rangle_{L^2(\Omega)} = \langle x, 0\rangle_{L^2(\Omega)} = 0$ contradicting k_0 being ISPD. Combining [61, Proposition 5] and [62, Theorem 9] shows that if μ_{k_0} has full support then k_0 is ISPD. □