Journal of Machine Learning Research 26 (2025) 1-60 Submitted 2/24; Revised 1/25; Published 2/25

Composite Goodness-of-fit Tests with Kernels

Oscar Key[*] OSCAR.KEY.20@UCL.AC.UK
Centre for Artificial Intelligence, University College London

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, University College London

François-Xavier Briol F.BRIOL@UCL.AC.UK
Department of Statistical Science, University College London

Tamara Fernandez[*] TAMARA.FERNANDEZ@UAI.CL
Faculty of Engineering and Science, Adolfo Ibañez University

Editor: Pierre Alquier

Abstract

We propose kernel-based hypothesis tests for the challenging composite testing problem, where we are interested in whether the data comes from *any* distribution in some *parametric family*. Our tests make use of minimum distance estimators based on kernel-based distances such as the maximum mean discrepancy. As our main result, we show that we are able to estimate the parameter and conduct our test on the same data (without data splitting), while maintaining a correct test level. We also prove that the popular wild bootstrap will lead to an overly conservative test, and show that the parametric bootstrap is consistent and can lead to significantly improved performance in practice. Our approach is illustrated on a range of problems, including testing for goodness-of-fit of a non-parametric density model, and an intractable generative model of a biological cellular network.

Keywords: bootstrap, hypothesis testing, kernel methods, minimum distance estimation, Stein's method

1. Introduction

Most statistical or machine learning algorithms are based on assumptions about the distribution of the observed data, of auxiliary variables, or of certain estimators. Crucially, the validity of such assumptions directly impacts the performance of these algorithms. To verify whether such assumptions are reasonable, one approach is *goodness-of-fit testing*, which considers the problem of rejecting, or not, the hypothesis that some data was generated by a *fixed* distribution. More precisely, given a distribution P and some observed data $x_1, \ldots, x_n$ from some distribution Q, goodness-of-fit tests compare the null hypothesis $H_0 : P = Q$ against the alternative hypothesis $H_1 : P \neq Q$.

∗. Equal contribution

There is a vast literature on goodness-of-fit testing, and the reader is referred to Chapter 14 in Lehmann and Romano (2005) for a detailed introduction. In the case of univariate real observations, a popular family of tests uses the cumulative distribution function (CDF) as the test statistic, for example the Kolmogorov-Smirnov and Anderson-Darling tests (Kolmogorov, 1933; Anderson and Darling, 1954). Various approaches have been taken to support multivariate observations, including extending these CDF tests (Justel et al., 1997), developing a statistic based on the empirical characteristic function (Jiménez-Gamero et al., 2009), or partitioning the observation space into bins and evaluating a test statistic on the resulting discrete empirical measures (Gyorfi and van der Meulen, 1991; Beirlant et al., 1994; Györfi and Vajda, 2002; Beirlant et al., 2001). Unfortunately, the applicability of all of these tests is limited, because they rely on statistics that are often difficult to compute for complex models or high-dimensional observation spaces.

To overcome these issues, a class of kernel-based hypothesis tests has been proposed, which construct the statistic through either the maximum mean discrepancy (MMD) (Lloyd and Ghahramani, 2015; Zhu et al., 2019; Balasubramanian et al., 2021; Kellner and Celisse, 2019), the kernel Stein discrepancy (KSD) (Liu et al., 2016; Chwialkowski et al., 2016), the energy distance (Székely and Rizzo, 2005, 2013), or the Hilbert-Schmidt independence criterion (Sen and Sen, 2014). These tests can be applied to a wide-range of data types by choosing a test statistic based on an appropriate kernel which can be easily evaluated in the setting considered. Additionally, many of these tests are straightforward to apply even for intractable generative models for which no density exists, or for models with unnormalised likelihoods, making them particularly versatile.

In this paper, we consider the more complex question of whether our data comes from *any element of some parametric family* of distributions. Let $\{P_\theta\}_{\theta\in\Theta}$ denote a parametric family indexed by a parameter θ in some space Θ. Our test compares the null hypothesis

$$H_0^C : \exists\theta_0 \in \Theta \text{ such that } P_{\theta_0} = Q,$$

against the alternative hypothesis

$$H_1^C : Q \notin \{P_\theta\}_{\theta\in\Theta}.$$

Such tests are known as *composite goodness-of-fit tests*, and can be much more challenging to construct since θ_0 is usually unknown. They have the potential to answer important questions relating to model misspecification, allowing the user to confirm whether the parametric model they have selected is appropriate for their analysis. Once again there is a significant literature tackling this problem, including tests which are specific to simple parametric families such as Gaussians (Shapiro and Wilk, 1965; Lilliefors, 1967), or composite versions of some of the non kernel-based tests mentioned above (Durbin, 1975; Neyman, 1967). Unfortunately, existing these tests all suffer from the same drawback as their simple goodness-of-fit counterparts, in that they are only applicable in a limited range of settings.

Our paper fills an important gap in the literature by proposing the first set of comprehensive *kernel-based composite hypothesis tests*. Given a kernel-based discrepancy D, our test statistics take the form $\Delta = \min_{\theta\in\Theta} n\,\mathrm{D}(P_\theta, Q_n)$. That is, we use the smallest discrepancy between Q_n and any element of the parametric family to determine whether H_0^C should be rejected or not. In this work, we primarily consider the case where $\mathrm{D} = \mathrm{MMD}^2$, and include

full theoretical and empirical analysis of the test's behaviour. This analysis is challenging because, in contrast to existing kernel-based tests, the data now enters the test statistic twice: once to select the closest element of the parametric family, and a second time to estimate the discrepancy. We also include encouraging empirical results for $\mathrm{D} = \mathrm{KSD}^2$, but leave the extension of our theoretical framework to this test for future work.

Our tests extend the advantages of kernel-based tests to the composite setting. In particular, they can be directly applied to both generative or unnormalised models, two wide classes which cannot be tackled in full generality with classical composite tests. Within the kernel testing literature, general composite tests have not yet been proposed. Existing tests have either been limited to specific parametric families, such as multivariate Gaussians (Kellner and Celisse, 2019) or survival models with a specific survival function (Fernández et al., 2020), or required splitting the data set thus reducing test power (Chwialkowski et al., 2016; Liu et al., 2016). Wolfer and Alquier (2022) also proposed a promising non-asymptotic composite test based on the MMD, but did not study it for generative models nor consider bootstrapping algorithms. We also show in our experiments that this approach leads to a more conservative test.

The remainder of the paper is as follows. Section 2: we recall existing work on testing and parameter estimation with kernels. Section 3: we propose novel kernel-based composite hypothesis tests, and show that the MMD test statistic has well-behaved limiting distributions despite the reuse of the data for both estimation and testing. Section 4: we compare two choices of bootstrap method that we can use to implement these tests in practice. Section 5: we demonstrate these algorithms on a range of problems including an intractable generative model and non-parametric density estimation. Section 6: we discuss limitations of our work. Section 7: we highlight connections with existing tests. The code to reproduce our experiments, and implement the tests for your own models, is available at: `https://github.com/oscarkey/composite-tests`

2. Background

We begin by reviewing the use of kernel discrepancies for testing and estimation. We denote by $\mathcal{X}$ the data space and $\mathcal{P}(\mathcal{X})$ the set of Borel probability distributions on $\mathcal{X}$.

2.1 Kernel-based Discrepancies

To measure the similarity of two distributions $P, Q \in \mathcal{P}(\mathcal{X})$, we can use a discrepancy $\mathrm{D} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty)$. We consider discrepancies related to integral pseudo-probability metrics (IPMs) (Muller, 1997). An IPM indexed by the set of functions $\mathcal{F}$ has the form

$$\mathrm{D}_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]| \,, \tag{1}$$

where $\mathcal{F}$ is sufficiently rich so that $\mathrm{D}_{\mathcal{F}}$ is a *statistical divergence*; i.e. $\mathrm{D}_{\mathcal{F}}(P, Q) = 0 \Leftrightarrow P = Q$.

A first discrepancy is the *maximum mean discrepancy* (MMD) (Gretton et al., 2007, 2012). Denote by $\mathcal{H}_K$ the reproducing kernel Hilbert space (RKHS) associated to the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Berlinet and Thomas-Agnan, 2004). The MMD is obtained by taking $\mathcal{F}_{\mathrm{MMD}} = \{f \in \mathcal{H}_K \,|\, \|f\|_{\mathcal{H}_K} \leq 1\}$, which is a convenient choice because it allows the supremum

in Equation 1 to be evaluated in closed form if $\mathbb{E}_{X\sim P}[\sqrt{K(X,X)}] < \infty \; \forall P \in \mathcal{P}(\mathcal{X})$:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{X,X'\sim P}[K(X,X')] - 2\mathbb{E}_{X\sim P, X'\sim Q}[K(X,X')] + \mathbb{E}_{X,X'\sim Q}[K(X,X')]. \quad (2)$$

Assuming we have access to independent and identically distributed realisations $\{\tilde{x}_i\}_{i=1}^m \overset{iid}{\sim} P$ and $\{x_j\}_{j=1}^n \overset{iid}{\sim} Q$, we can define $P_m = \frac{1}{m}\sum_{i=1}^m \delta_{\tilde{x}_i}$ and $Q_n = \frac{1}{n}\sum_{j=1}^n \delta_{x_j}$ (where δ_x is a Dirac measure at $x \in \mathcal{X}$) and we get the following estimate:

$$\mathrm{MMD}^2(P_m, Q_n) = \frac{1}{m^2}\sum_{i,j=1}^m K(\tilde{x}_i,\tilde{x}_j) - \frac{2}{nm}\sum_{i=1}^m\sum_{j=1}^n K(\tilde{x}_i, x_j) + \frac{1}{n^2}\sum_{i,j=1}^n K(x_i,x_j).$$

Furthermore, when $n = m$, this simplifies to

$$\mathrm{MMD}^2(P_n, Q_n) = \frac{1}{n^2}\sum_{i,j=1}^n h_{\mathrm{MMD}}((x_i,\tilde{x}_i),(x_j,\tilde{x}_j)),$$

where $h_{\mathrm{MMD}}((x_i,\tilde{x}_i),(x_j,\tilde{x}_j)) = K(\tilde{x}_i,\tilde{x}_j) + K(x_i,x_j) - K(\tilde{x}_i,x_j) - K(\tilde{x}_j,x_i)$. This is known as a V-statistic (Gretton et al., 2012), and h_{MMD} is known as the core of this statistic. The statistic is biased, but has smaller variance than alternative U-statistics for this quantity.

We also consider the *kernel Stein discrepancy* (KSD) (Oates et al., 2017; Chwialkowski et al., 2016; Liu et al., 2016), which is obtained by applying Stein operators to functions in some RKHS; see Anastasiou et al. (2023) for a recent review. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{H}_K^d = \mathcal{H}_K \times \ldots \times \mathcal{H}_K$ denote the d-dimensional tensor product of $\mathcal{H}_K$. The most common instance is the Langevin KSD for which $\mathcal{F}_{\mathrm{KSD}} = \{\mathcal{S}_P[f] \,|\, \|f\|_{\mathcal{H}_K^d} \leq 1\}$ where $\mathcal{S}_P[g](x) = g(x)\cdot\nabla_x \log p(x) + \nabla_x \cdot g(x)$ is the Langevin Stein operator, p is the Lebesgue density of P and $\nabla_x = (\partial/\partial x_1, \ldots, \partial/\partial x_d)^\top$. With this choice, we obtain the discrepancy:

$$\mathrm{KSD}^2(P,Q) = \mathbb{E}_{X,X'\sim Q}[h_{\mathrm{KSD}}(X,X')],$$

$$\text{where} \qquad h_{\mathrm{KSD}}(x,x') = K(x,x')\nabla_x \log p(x)\cdot\nabla_{x'}\log p(x') + \nabla_x \log p(x)\cdot\nabla_{x'}K(x,x') + \nabla_{x'}\log p(x')\cdot\nabla_x K(x,x') + \nabla_x\cdot\nabla_{x'}K(x,x').$$

Given $Q_n = \frac{1}{n}\sum_{j=1}^n \delta_{x_j}$, we obtain a V-statistic $\mathrm{KSD}^2(P,Q_n) = \frac{1}{n^2}\sum_{i,j=1}^n h_{\mathrm{KSD}}(x_i,x_j)$.

We choose the MMD and the KSD to implement our tests because they are applicable in complementary settings. The MMD can be estimated whenever samples from P_θ are available. This makes the corresponding test particularly suitable for simulator-based models (also called generative models); i.e. models which can be represented through a pair $(\mathbb{U}, G_\theta)$ such that sampling $x \sim P_\theta$ involves sampling $u \sim \mathbb{U}$, and setting $x = G_\theta(u)$. This class of models covers many models widely used in machine learning, including variational autoencoders (Kingma and Welling, 2014) and generative adversarial networks (Li et al., 2015; Dziugaite et al., 2015), but also in the sciences including in synthetic biology (Bonassi et al., 2011) and telecommunication engineering (Bharti et al., 2021) to name just a few. The interested reader is referred to Cranmer et al. (2020) for a recent review. On the other hand, the KSD requires pointwise evaluations of $\nabla \log p$ and the KSD-based test will therefore be particularly suitable when the density of P_θ is tractable. This includes cases where the density can be only evaluated up to normalisation constant, such as for deep energy models or nonparametric

density estimation models (Canu and Smola, 2006; Fukumizu, 2009; Sriperumbudur et al., 2017; Matsubara et al., 2022), exponential random graphs (Robins et al., 2007) and large spatial models based on lattice structures (Besag, 1974).

Throughout this paper, we will assume that our kernel-based divergences are valid statistical divergences. This can be guaranteed for the MMD by assuming the kernel K is characteristic (Sriperumbudur et al., 2010). For the KSD we require that K is strictly integrally positive definite and $\mathbb{E}_{X\sim Q}[|\nabla_x \log p(X) - \nabla_x \log q(X)|] < \infty$ (see for example Chwialkowski et al. (2016, Theorem 2.1) and Barp et al. (2019, Proposition 1)). Examples of kernels on $\mathbb{R}^d$ which satisfy these assumptions include the Gaussian, inverse multiquadric and Matérn kernels.

While these conditions ensure that the MMD and KSD can distinguish any pair of distributions when $n, m \to \infty$, for finite samples this is not necessarily the case. The ability of the divergences to distinguish distributions is primarily dependent whether an appropriate kernel has been selected for the data in question. Additionally, the KSD suffers from a failure mode when applied to multi-modal distributions, common for methods based on $\nabla_x \log p$: if Q contains isolated modes with areas of near-zero density between them, the KSD cannot distinguish this from P that is identical except having a different weighting of the modes (Wenliang and Kanagawa, 2021). We discuss the impacts of these failure modes on ours tests in Section 6.

2.2 Goodness-of-fit Testing with Kernels

We now recall how divergences lend themselves naturally to (standard) goodness-of-fit testing. Let $\mathrm{D} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be some statistical divergence, and recall that we are testing $H_0 : P = Q$. A natural approach is to compute $\mathrm{D}(P, Q)$ and check whether it is zero (in which case H_0 holds) or not (in which case H_1 holds). Since we only have access to independent realisations $\{x_j\}_{j=1}^n$ instead of Q itself, this idealised procedure is replaced by the evaluation of $\mathrm{D}(P, Q_n)$. The question then becomes whether or not $\mathrm{D}(P, Q_n)$ is further away from zero than we would expect under H_0 given a data set of size n.

Kernel-based discrepancies can be computed in a wide range of scenarios including for intractable models. These divergences also have favourable sample complexity in that $\mathrm{D}(P, Q_n)$ converges to $\mathrm{D}(P, Q)$ at a square-root n rate for both KSD (Matsubara et al., 2022, Theorem 4) and MMD (Briol et al., 2019, Lemma 1). The MMD was first used by Gretton et al. (2007, 2012) for two-sample testing, and then studied for goodness-of-fit testing by Lloyd and Ghahramani (2015). MMD tests are closely related to many classical tests; see Sejdinovic et al. (2013). Chwialkowski et al. (2016); Liu et al. (2016) introduced an alternate test based on $\mathrm{KSD}^2(P, Q_n)$.

To determine when H_0 should be rejected, we select an appropriate threshold $c_\alpha \in \mathbb{R}$, which will depend on the level of the test $\alpha \in [0, 1]$. More precisely, c_α should be the $(1 - \alpha)$-quantile of the distribution of the test statistic under H_0. This distribution will usually be unknown, but can be approximated using a bootstrap method. A common example is the wild bootstrap (Shao, 2010; Leucht and Neumann, 2013), which was specialised for kernel tests by Chwialkowski et al. (2014). We will return to this point in Section 4.

2.3 Minimum Distance Estimation with Kernels

A statistical divergence D can also be used for parameter estimation through *minimum distance estimation* (Wolfowitz, 1957; Parr and Schucany, 1980). Given a parametric family $\{P_\theta\}_{\theta\in\Theta} \subset \mathcal{P}(\mathcal{X})$ indexed by $\Theta \subseteq \mathbb{R}^p$ and some data $\{x_i\}_{i=1}^n \overset{iid}{\sim} Q$, a natural estimator is

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q_n),$$

In practice, the minimum is often obtained using numerical optimisation algorithms. Under regularity conditions on D and P_θ, the estimator approaches $\theta^* := \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q)$ as $n \to \infty$. In particular, when the model is well-specified (i.e. H_0^C holds), then $\theta^* = \theta_0$.

The MMD was first used for parameter estimation for neural networks by Dziugaite et al. (2015); Li et al. (2015); Sutherland et al. (2017); Li et al. (2017); Bińkowski et al. (2018), then studied more broadly as a statistical estimator by Briol et al. (2019); Chérief-Abdellatif and Alquier (2022, 2020); Niu et al. (2021); Dellaporta et al. (2022); Bharti et al. (2023). They closely relate to minimum scoring rule estimators based on the kernel-scoring rule (Dawid, 2007). The use of KSD for parameter estimation was first proposed by Barp et al. (2019), and later extended by Betsch et al. (2020); Grathwohl et al. (2020); Gong et al. (2021); Matsubara et al. (2022). These estimators are also closely related to score-matching estimators (Hyvärinen, 2006); see Barp et al. (2019, Theorem 2) for details.

3. Methodology: The Test Statistic

We are now ready to present our composite goodness-of-fit tests. The first section describes our general framework and studies the asymptotic distribution of the MMD test statistic. In the next section, we then describe how to implement the tests through bootstrapping.

3.1 Composite Goodness-of-fit Testing with Kernels

The high-level idea behind our test-statistics is to use the smallest kernel-based discrepancy value between the data Q_n and any element of the parametric family:

$$\Delta := \min_{\theta\in\Theta} n\,\mathrm{D}(P_\theta, Q_n) = \min_{P\in\{P_\theta\}_{\theta\in\Theta}} n\,\mathrm{D}(P, Q_n).$$

To implement the tests, we propose a two-stage approach:

Stage 1 (Estimate): Compute $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q_n)$, so that $\Delta = n\,\mathrm{D}(P_{\hat{\theta}_n}, Q_n)$.

Stage 2 (Test): Compute or estimate a test threshold c_α from the distribution of test statistic under H_0^C. If $\Delta > c_\alpha$ reject H_0^C, else do not reject.

In this paper, we will study two other instances of this framework. Our primary focus will be on the *MMD composite goodness-of-fit test*, where $\mathrm{D} = \mathrm{MMD}^2$. The next section provides regularity conditions so that as $n \to \infty$, Δ for this test converges to an infinite sum of χ^2 random variables under H_0^C but diverges under H_1^C, thus allowing us to distinguish between the two hypotheses. The MMD test is widely applicable since it only requires that we can sample from P_θ.

(a) Non-composite test

(b) Composite test

Figure 1: Illustration of the sources of error when estimating the test statistic under the null hypothesis. Existing non-composite tests use the test statistic $n\,\mathrm{D}(P, Q_n)$, where D is a statistical divergence, and thus encounter only "test" error. The composite test we introduce uses the statistic $n\,\mathrm{D}(P_{\hat{\theta}_n}, Q_n)$, and thus encounters both "estimation" and "test" error.

The second instance will be the *KSD composite goodness-of-fit test*, where $\mathrm{D} = \mathrm{KSD}^2$. Although we do not provide theoretical guarantees for this case, we do conjecture that similar guarantees exist. This test is particularly interesting since it only requires a tractable (possibly unnormalised) density function for P_θ.

In both tests, stage 1 requires computing the solution to the optimisation problem $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q_n)$. The best method for this depends on the model and should be selected alongside it. In this paper, for the MMD test, we use gradient-based stochastic optimisation and automatic differentiation. For the KSD test, we use the closed-form expression for the minimiser, which is available when P_θ is a member of the exponential family of distributions (see Section D.1). However, stochastic optimisation could also be used in order to work with more general families of models.

3.2 Theoretical Analysis of the MMD Composite Goodness-of-fit fest

To present our analysis of the MMD test, we first define some additional notation. Since we will be interested in asymptotic distributions, we will now consider the data to be random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Q$, rather than the realisations $x_1, \ldots, x_n$ introduced so far. We then overload $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ to define a random measure. This is done in order to simplify notation, and whether we refer to the random or realised Q_n can be inferred from the context.

The results we present in this paper are based on the asymptotic distribution of $n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n)$ as $n \to \infty$. This is similar to standard kernel-based goodness-of-fit tests which use $n\,\mathrm{MMD}^2(P, Q_n)$, the main difference being that our test statistic depends on the data through both $\hat{\theta}_n$ and Q_n. In other words, we are now using the data twice, once for estimation and once for testing. We note that this creates some non-trivial dependence and makes the derivation of theoretical guarantees significantly more challenging. We first state and discuss our assumptions, then present all results in the following subsection.

3.2.1 Assumptions

We will write $\mathcal{C}^m(\mathcal{X})$ for the set of m-times continuously differentiable functions on $\mathcal{X}$, and $\mathcal{C}_b^m(\mathcal{X})$ for the subset of $\mathcal{C}^m(\mathcal{X})$ where all of the m derivatives are bounded. For bivariate

functions, we will denote by $\mathcal{C}_b^{m,m}(\mathcal{X} \times \mathcal{X})$ the set of function which are $\mathcal{C}_b^m(\mathcal{X} \times \mathcal{X})$ with respect to each input. We can now present our first assumption.

Assumption 1. *The observations $\{x_i\}_{i=1}^n$ are independent and identically distributed realisations from Q.*

This assumption is common, and covers a very large class of applications, including all our examples in Section 5. It could be relaxed to dependent data but this would require the development of specialised tools and notions of dependence which are beyond the scope of this paper; see for example Chérief-Abdellatif and Alquier (2022).

Assumption 2. *$\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{X}$ is separable, and Θ is a compact and convex subset of $\mathbb{R}^p$.*

The assumptions are relatively mild and needed for the consistency of our estimators. Even when working with models where the assumptions on Θ are not satisfied, it is often possible to reparameterise the model so that the assumption holds.

Our next assumption defines regularity assumptions on the model considered by the MMD test, which will have to be verified on a per-model basis.

Assumption 3. *$\{P_\theta\}_{\theta \in \Theta}$ is an identifiable parametric family of models. Each element $P_\theta \in \mathcal{P}$ has density p_θ with respect to a common measure λ which satisfies:*

- *For each fixed $x \in \mathcal{X}$, $p.(x) \in \mathcal{C}^3(\Theta)$ and $p_\theta : \mathcal{X} \to \mathbb{R}$ is measurable for almost all $\theta \in \Theta$.*
- *For any collection $\ell, i, j \in \{1, \ldots, p\}$, the following holds true:*

$$\int_{\mathcal{X}} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_\ell} p_\theta(x) \right| \lambda(dx) < \infty,$$
$$\int_{\mathcal{X}} \sup_{\theta \in \Theta} \left| \frac{\partial^2}{\partial \theta_\ell \partial \theta_i} p_\theta(x) \right| \lambda(dx) < \infty,$$
$$\text{and } \int_{\mathcal{X}} \sup_{\theta \in \Theta} \left| \frac{\partial^3}{\partial \theta_\ell \partial \theta_i \partial \theta_j} p_\theta(x) \right| \lambda(dx) < \infty.$$

The smoothness assumptions with respect to the parameter are required to ensure convergence of $\hat{\theta}_n$ to θ_0 in an appropriate sense under H_0^C. Note that this assumption, together with the fact that Θ is compact, allows us to deduce $\int_{\mathcal{X}} \sup_{\theta \in \Theta} p_\theta(x) \lambda(dx) < \infty$. On the other hand, the smoothness assumptions with respect to $\mathcal{X}$ are needed for the divergences to be well defined.

In many cases, when applying the MMD test, we will not have access to a computable density function and will only be able to sample from the model. For example, many models are defined in terms of a distribution $\mathbb{U}$ on some latent space, and a generator function $G_\theta : \mathcal{U} \to \mathcal{X}$, where a sample from the model $\tilde{x} = G_\theta(u)$ with $u \sim \mathbb{U}$. In this case, the results could be updated to instead place assumptions on $\mathbb{U}$ and G_θ rather than on the density, as was done in Briol et al. (2019).

We make an additional assumption about the model family under the null hypothesis.

Assumption 4. *Under H_0^C, we have that θ_0, the parameter value under the null hypothesis, belongs to the interior of Θ.*

For most models this assumption will be met, or could be met by reparameterising the model. Our next assumption relates to the kernel underlying the MMD.

Assumption 5. *The kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded, continuous and characteristic.*

This assumption is mild and will be satisfied by kernels commonly used in the literature on minimum MMD estimators, such as the Gaussian and Matérn kernels of sufficient smoothness.

Our final assumption concerns the regularity of the divergences around the minimisers. This is necessary to guarantee convergence of our minimum distance estimators under H_0^C, and is a mild assumption which is common in the literature (see e.g. Theorem 2 in Briol et al. (2019) and Theorem 4 in Barp et al. (2019)).

Assumption 6. *The Hessian matrix $\mathbf{H} \in \mathbb{R}^{p \times p}$ given by $\mathbf{H}_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \mathrm{MMD}^2(P_\theta, P_{\theta_0})\big|_{\theta=\theta_0}$ for any $i, j \in \{1, \ldots, p\}$ is positive definite.*

Remark 1. *We could extend the results to the KSD test by noting that the KSD is equivalent to an MMD with a specific kernel (Barp et al., 2019). However, this kernel is often unbounded and depends on p_θ. Thus, additional assumptions on p_θ would be required. We leave this extension to future work. In our experiments we assume that the KSD is a valid statistical divergence, which holds under the conditions discussed in Section 2.1.*

3.2.2 Behaviour of the Test Statistic Under H_0^C and H_1^C

We now present results on the asymptotic distribution of our test statistic, and the power (i.e. probability of correctly rejecting H_0^C when it does not hold) of the associated tests, as $n \to \infty$.

Theorem 2 (Convergence under the null hypothesis)**.** *Under H_0^C and Assumptions 1 to 6:*

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \qquad \text{as } n \to \infty, \tag{3}$$

where $(Z_i)_{i\geq 1}$ are a collection of i.i.d. standard normal random variables and $(\lambda_i)_{i\geq 1}$ are constants, which depend on the choice of kernel, where $\sum_{i=1}^{\infty} \lambda_i < \infty$.

Theorem 2 guarantees that our test statistic has a well-behaved limiting distribution under the null hypothesis H_0^C. Unfortunately, this distribution is not tractable, but knowing that it exists and is well-behaved is still necessary for guaranteeing that we can approximate quantiles with a bootstrapping method, as will be discussed in Section 4. The results are analogous to Theorem 12 in Gretton et al. (2012), though this result is only valid for testing H_0 vs. H_1 and not H_0^C vs. H_1^C. Our second result concerns the power of the test: when $\mathcal{H}_1^{\mathrm{C}}$ holds, we show that our test is able to reject $\mathcal{H}_0^{\mathrm{C}}$ when n is large enough.

Theorem 3 (Consistency under the alternative hypothesis)**.** *Under Assumptions 1 to 6, we have that H_1^C holds if and only if $\liminf_{n\to\infty} \mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) > 0$.*

As our test statistic, $n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n)$, is scaled by n, the above result implies that the test statistic diverges as $n \to \infty$.

4. Methodology: Practical Implementation via Bootstraps

To implement the tests, we require an estimate of the rejection threshold c_α, which should be set to the $(1-\alpha)$-quantile of the distribution of $\Delta = n\,\mathrm{D}(P_{\hat{\theta}_n}, Q_n)$ under H_0^C. Since this distribution is unknown, we estimate the threshold using bootstrap algorithms. In following section we present two possible choices: the wild bootstrap and the parametric bootstrap, and analyse their behaviour when applied to the MMD test.

4.1 The Wild Bootstrap

A natural first approach is to use the wild bootstrap, since that is the current state-of-the-art for non-composite kernel goodness-of-fit tests (Chwialkowski et al., 2016; Liu et al., 2016; Schrab et al., 2022). The algorithm is based on bootstrapped versions of the test statistics, defined for non-composite tests as

$$n\,\mathrm{MMD}_W^2(P_n, Q_n) := \frac{1}{n}\sum_{i,j=1}^{n} W_i W_j h_{\mathrm{MMD}}((x_i, \tilde{x}_i), (x_j, \tilde{x}_j)),$$

$$n\,\mathrm{KSD}_W^2(P, Q_n) := \frac{1}{n}\sum_{i,j=1}^{n} W_i W_j h_{\mathrm{KSD},P}(x_i, x_j),$$

where $W_1, \ldots, W_n$ are i.i.d. Rademacher random variables (i.e. random variables taking value -1 and 1 with probability 1/2 each). Note that the bootstrapped version of the MMD test requires realisations of both P and Q (i.e. P_n and Q_n) whereas the bootstrapped version of the KSD test requires evaluations of the score of the distribution P and samples from Q (denoted through the empirical measure Q_n).

Given fixed observations Q_n, it is possible to draw samples $\Delta^{(1)}, \ldots, \Delta^{(b)}$ of the bootstrapped test statistics by taking fresh realisations of $W_1, \ldots, W_n$ for each sample. If we consider an MMD test, under H_0, Chwialkowski et al. (2014, Theorem 1) show that the distributions of $n\,\mathrm{MMD}^2(P, Q_n)$ and $n\,\mathrm{MMD}_W^2(P, Q_n)$ converge to the same distribution as $n \to \infty$. Thus, we can estimate the threshold c_α during a non-composite test by computing the $(1-\alpha)$-quantile of $\{\Delta^{(k)}\}_{k=1}^b$. Under H_1, Chwialkowski et al. (2014, Proposition 3.2) show that $n\,\mathrm{MMD}_W^2(P_n, Q_n)$ will converge to a finite quantity, while Theorem 3 shows that $n\,\mathrm{MMD}^2(P_n, Q_n) \to \infty$. This means that $\mathbb{P}(\mathrm{MMD}^2(P_n, Q_n) > c_\alpha) \to 1$ as $n \to \infty$ and the test rejects H_0 correctly.

We can extend the wild bootstrap to the composite case by first estimating the parameter and then applying the wild bootstrap to the estimated distribution. Thus, the bootstrapped test statistics are $n\,\mathrm{MMD}_W^2(P_{\hat{\theta}_n,n}, Q_n)$ and $n\,\mathrm{KSD}_W^2(P_{\hat{\theta}_n}, Q_n)$, where $P_{\hat{\theta}_n,n} = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{X}_i}$ with $\bar{X}_i \sim P_{\hat{\theta}_n}$. Algorithm 1 describes this approach for the MMD, and Algorithm 2 for the KSD. Note that the KSD version of the test does not require an additional sampling step since the KSD can be estimated with a single sample, whereas two different samples are needed to estimate the MMD.

It is not clear from the existing results in the literature if applying the wild bootstrap to the composite test in this fashion will result in a test with a type I error rate (i.e. probability of wrongly rejecting H_0^C when it holds) smaller than or equal to α. This is because the composite test statistic contains an additional source of error, in comparison to

Algorithm 1: Wild bootstrap (MMD)

Input: $P_{\hat{\theta}_n}$, Q_n, α, b

$P_{\hat{\theta}_n,n} = \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{x}_i}$ where $\tilde{x}_1, \ldots, \tilde{x}_n \overset{iid}{\sim} P_{\hat{\theta}_n}$;

for $k \in \{1, \ldots, b\}$ **do**

  $w^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)}) \overset{iid}{\sim}$ Rademacher;

  $\Delta^{(k)} = n\,\mathrm{MMD}^2_{w^{(k)}}(P_{\hat{\theta}_n,n}, Q_n)$;

$c_\alpha = \mathrm{quantile}(\{\Delta^{(1)}, \ldots, \Delta^{(b)}\}, 1-\alpha)$;

Algorithm 2: Wild bootstrap (KSD)

Input: $P_{\hat{\theta}_n}$, Q_n, α, b

for $k \in \{1, \ldots, b\}$ **do**

  $w_1^{(k)}, \ldots, w_n^{(k)} \overset{iid}{\sim}$ Rademacher;

  $\Delta^{(k)} = n\,\mathrm{KSD}^2_{w^{(k)}}(P_{\hat{\theta}_n}, Q_n)$;

$c_\alpha = \mathrm{quantile}(\{\Delta^{(1)}, \ldots, \Delta^{(b)}\}, 1-\alpha)$;

the non-composite statistic. $\mathrm{MMD}^2(P_n, Q_n)$ contains one source of error, due to estimating $\mathrm{MMD}^2(P, Q_n)$ through samples. However, $\mathrm{MMD}^2(P_{\hat{\theta}_n,n}, Q_n)$ contains a second source of error because P_{θ_0}, the true data generating distribution under H_0^C, is estimated with $P_{\hat{\theta}_n}$. Figure 1 illustrates these sources of error in the non-composite and composite test statistics under the null hypothesis. If we fix an estimate of $\hat{\theta}_n$ and then apply the wild bootstrap, we fail to take into account this additional source of error.

Our main result for this section shows that applying the wild bootstrap does lead to a type I error rate smaller than or equal to α in the MMD test. However, our result shows the (asymptotic) type I error rate is strictly below α, which is surprising as usually resampling strategies yield asymptotically-exact type-I error rates. This result suggests that our test is conservative and could potentially be improved further.

Theorem 4. *Let $\alpha \in (0, 1)$, and define*

$$q_\alpha = \inf\left\{\gamma : \limsup_{n\to\infty} \mathbb{P}(n\,\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n) > \gamma) \leq \alpha\right\}.$$

Under the null hypothesis and Assumptions 1 to 6, it holds

$$\limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) \geq q_\alpha\right) < \alpha.$$

The root cause of the above result is that, under both H_0^C and H_1^C, $\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) \leq \mathrm{MMD}^2(P_{\theta_0}, Q_n)$, because $\hat{\theta}_n$ is chosen as the minimiser of $\theta \to \mathrm{MMD}^2(P_\theta, Q_n)$. The impact of this on the wild bootstrap is that, for all $x \in \mathbb{R}$,

$$\limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) \geq x\right) < \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n) \geq x\right)$$

as we show in the proof of Theorem 4. This means that every quantile of the asymptotic distribution of the wild bootstrap statistic is shifted away from zero compared to the test statistic, as demonstrated in Figure 2. Thus, the threshold computed from the samples of the bootstrapped test statistic is too large, and the resulting test is conservative. We conjecture that similar behaviour occurs under H_1^C, resulting in the loss of power we observe in the experiments (see e.g. Figure 3b). To achieve better performance, we can use an alternative bootstrap which does take account of estimation error.

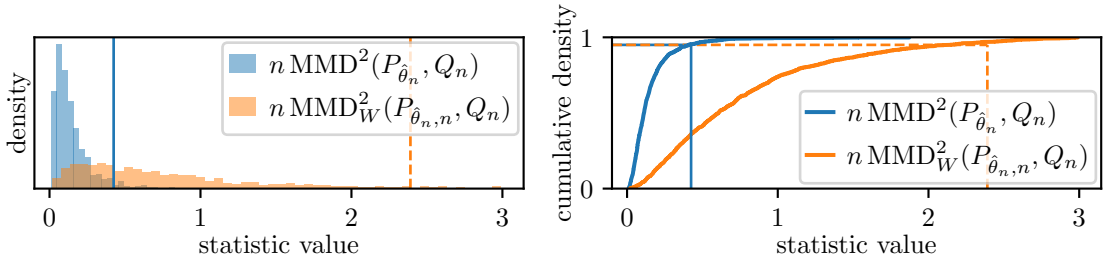


Figure 2: Distribution of $\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n)$ and $\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n)$ under H_0^C, obtained by simulation. | and ⋮ show the respective $(1-\alpha)$ quantiles, demonstrating that the wild bootstrap estimates a conservative threshold.

4.2 The Parametric Bootstrap

We now consider the parametric bootstrap (Stute et al., 1993) presented in Algorithm 3, which is commonly used in the composite testing literature (for example Kellner and Celisse (2019)). To approximate the distribution of $n\,\mathrm{D}(P_{\hat{\theta}_n}, Q_n)$ under H_0^C, this approach first fits the parameter to the observations. It then repeatedly resamples the observations, re-estimates the parameter and recomputes the test statistic. By repeatedly re-estimating the parameter, the parametric bootstrap takes account of the estimation error.

Algorithm 3: Parametric bootstrap

Input: D, P_θ, Q_n, α, b (num bootstrap samples)
$\hat{\theta}_n = \arg\min_\theta \mathrm{D}(P_\theta, Q_n)$;
for $k \in \{1, \ldots, b\}$ **do**
  $Q_n^{(k)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{(k)}}$, $\{x_i^{(k)}\}_{i=1}^n \overset{iid}{\sim} P_{\hat{\theta}_n}$;
  $\hat{\theta}_n^{(k)} = \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q_n^{(k)})$;
  $\Delta^{(k)} = n\,\mathrm{D}(P_{\hat{\theta}_n^{(k)}}, Q_n^{(k)})$;
$c_\alpha = \mathrm{quantile}(\{\Delta^{(1)}, \ldots, \Delta^{(b)}\}, 1-\alpha)$;

Algorithm 3 generates samples $\Delta^{(k)}$ of the bootstrapped test statistic $n\,\mathrm{D}(P_{\hat{\theta}_n^\star}, Q_n^\star)$, where $\hat{\theta}_n^\star = \arg\min_{\theta\in\Theta} \mathrm{D}(P_\theta, Q_n^\star)$, and $Q_n^\star = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^\star}$ with $X_i^\star \sim P_{\hat{\theta}_n}$. We show that the parametric bootstrap is valid for the MMD test, which once again requires that the distribution of the test statistic $n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n)$ and the distribution of the bootstrapped test statistic $n\,\mathrm{MMD}^2(P_{\hat{\theta}_n^\star}, Q_n^\star)$ converge to the same distribution as $n \to \infty$. To present the result we require some additional notation, which is defined formally in Section A.1.4: considering random variables A_n and A, we write $A_n \overset{\mathcal{D}_D}{\to} A$ to indicate that A_n converges to A in distribution conditioned on the data, $\{X_i\}_{i=1}^n$.

Theorem 5 (Parametric bootstrap under the null). *Under H_0^C and Assumptions 1 to 6, we have that $n\,\mathrm{MMD}^2(P_{\theta_n^\star}, Q_n^\star) \xrightarrow{\mathcal{D}_D} \Gamma_\infty$, where Γ_∞ is a random variable with distribution given by Equation 3, i.e. the distribution of the test statistic under H_0^C.*

Note that the above convergence holds conditioned on any sequence of observations, thus holds in general. In Section 5 we find empirically that applying the parametric bootstrap to our tests results in a good type I error rate and a better performance than the wild bootstrap when n is small. However, the parametric bootstrap is substantially more computationally intensive because it requires repeatedly computing a minimum distance estimator (and hence the kernel matrix) on fresh data, whereas the wild bootstrap only uses a single minimum distance estimator. Assuming we are computing the minimum distance estimators through T steps of a numerical optimiser based on function evaluations or gradients, the computational complexity of the wild bootstrap algorithms is $O((T+b)n^2)$, whilst the parametric bootstrap is $O(Tbn^2)$. We therefore expect the parametric bootstrap to be significantly more expensive than the wild bootstrap. However, this extra computation is likely an issue only for large n, and in this high data regime we find in our experiments that the wild bootstrap achieves similar power to the parametric bootstrap. Thus, in a high data regime there is little penalty to using the cheaper wild bootstrap. Note that the cost could be further reduced to be linear in n through alternative estimates of the MMD (see e.g. Lemma 14 of Gretton et al. (2012)), but this would likely lead to substantially lower test power.

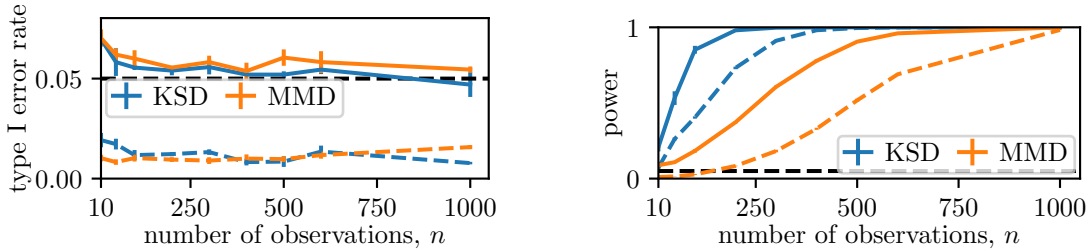
5. Empirical Results

In the following section, we now study the performance of our tests in a range of settings. We consider three examples: (i) a multivariate Gaussian location and scale model, (ii) a generative model of the interaction of genes through time called a toggle-switch model, and (iii) an unnormalised nonparametric density model. For all experiments we set the test level $\alpha = 0.05$, and give full implementation details in Section D.

5.1 Gaussian Model

We begin by exploring the different configurations of our test, in regard to the underlying discrepancy and the bootstrap method, for various Gaussian models. Unless otherwise stated, we use a Gaussian kernel with lengthscale computed using the median heuristic (Gretton et al., 2007).

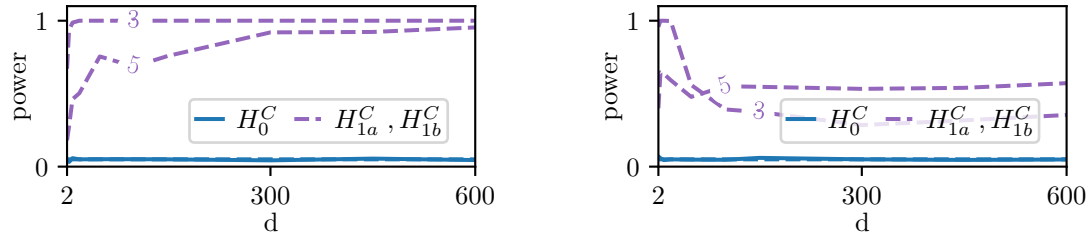
We test Gaussianity by considering $H_0^C : P_\theta = \mathcal{N}(\mu, \Sigma)$ with unknown parameters $\theta = \{\mu, \Sigma\}$, for $\mu \in \mathbb{R}^d$ and $\Sigma \in (\mathbb{R}^+)^{d\times d}$. We start with $d = 1$, and test against a particular alternative where $Q =$ Students-t with 2 degrees of freedom. Figure 3 shows the type I error rate and power of the MMD and KSD tests using both bootstraps as n increases. Considering the MMD test under H_0^C, the test using the wild bootstrap converges to a conservative type I error rate as n becomes large, while the test using the parametric has the correct rate. This is consistent with our analysis of the bootstraps in Section 4. Under H_1^C, the parametric bootstrap has higher power, once again due to the wild bootstrap being conservative, but both bootstraps achieve a power of one once n is large enough. For these results, m, the number of samples from P_θ, was set to $m = n$ as a balance between computational cost and



(a) Type I error rate under H_0^C, where $Q = \mathcal{N}(\mu_0, 1^2)$.

(b) Power under H_1^C, where $Q =$ Students-t with 2 degrees of freedom.

Figure 3: Performance of the MMD and KSD tests as n increases, where $H_0^C : P = \mathcal{N}(\mu, 1)$. = show the parametric bootstrap, :: show the wild bootstrap, -- shows the level. The error bars show one standard error across 4 random seeds.



(a) MMD test

(b) KSD test

Figure 4: Power of the parametric bootstrap tests as d increases. $H_0^C : P = \mathcal{N}(\mu, \sigma^2)$, H_{1a}^C and H_{1b}^C indicate $Q =$ Students-t, with 3 and 5 degrees of freedom respectively. -- shows the level.

power. Better power could be achieved by increasing m, although this would increase the computational cost.

The figure also shows the same set of results for the KSD test. It has similar convergence behaviour, suggesting that our theoretical results may also hold for the KSD. Comparing the MMD and KSD tests under H_1^C we note that the KSD test has higher power for any given n. This is both because the KSD does not require sampling from P_θ, and because the MMD estimation process requires stochastic optimisation, whereas for the KSD we use a closed-form estimator.

Second, we examine the performance of the parametric bootstrap tests as d, the dimensionality of the observations, increases. We test against the family of multivariate Gaussians with known covariance, $H_0^C : P = \mathcal{N}(\mu_d, I_{d\times d})$ with unknown parameter $\theta = \mu_d \in \mathbb{R}^d$. We consider two alternatives, H_{1a}^C and H_{1b}^C, where we generate data from multivariate t-distributions with $\nu = 3$ and $\nu = 5$ degrees of freedom respectively. As ν becomes large, the observed data become indistinguishable from a Gaussian, hence we expect the tests to have

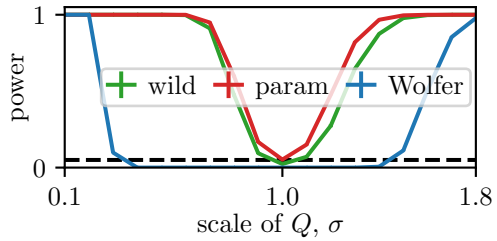


Figure 5: Comparison of the MMD tests against Wolfer and Alquier (2022). $H_0^C : P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(\mu_0, \sigma^2)$. $n = 200$. **- -** shows the level. The error bars show one standard error across 4 random seeds.

lower power against larger ν. We set the scale parameter, $\Sigma = I_{d\times d}(\nu - 2)/\nu$, so that the samples have covariance I and the distance between the samples does not grow more rapidly than under H_0^C as d increases. Figure 4 shows that both tests maintain the type I error rate of 0.05 under H_0^C. Under H_{1a}^C and H_{1b}^C we find that the power of the MMD test increases with d, while the power of the KSD test decreases with d.

Next, we further examine the power of the MMD tests when using the wild and parametric bootstraps. We also include the non-asymptotic MMD test introduced by Wolfer and Alquier (2022), for comparison. We take H_0^C to be a family of one dimensional Gaussians with known variance, specifically $P = \mathcal{N}(\mu, \sigma^2)$, where $\theta = \mu \in \mathbb{R}$ and $\sigma^2 = 1$. We generate data from $\mathcal{N}(\mu_0, \sigma^2)$, where $\mu_0 \in \mathbb{R}$ is an arbitrary mean, and compute the power of the tests for different values of σ^2. An ideal test would have power close to $\alpha = 0.05$ when $\sigma^2 = 1$, and power close to one otherwise. Figure 5 shows the results, with the parametric bootstrap having higher power than the wild bootstrap once again. The test introduced by Wolfer and Alquier (2022) achieves lower power than both our tests with either bootstrap.

We also consider the time taken by each bootstrap method. For $n = 100$ and $d = 1$, we found that each wild bootstrap test took approximately 1ms, while each parametric bootstrap test approximately 5ms. These figures depend heavily on the implementation and hardware details (as given in Section D). In particular, we execute all bootstrap samples in parallel on a GPU, which is particularly beneficial for the parametric bootstrap where each sample has a high computational cost. However, it will not be possible for larger models because it requires keeping b copies of the model parameters in memory. Additionally, it would be possible to further optimise the bootstrap implementations. Despite this, these approximate numbers illustrate that the parametric bootstrap test is much more expensive. Thus, for large n where the two bootstraps achieve similar power (see Figure 3b) it may be better to use the wild bootstrap.

Figure 6 explores the robustness of the test to corrupted observations under Huber's contamination model (Huber and Ronchetti, 2011), as previously explored in the testing setting by Liu and Briol (2024); Schrab and Kim (2024). Once again we consider $H_0^C : P = \mathcal{N}(\mu_d, I_{d\times d})$, for $\mu_d \in \mathbb{R}^d$. We assume that H_0^C holds, but some observations have been corrupted with a large amount of noise. Thus, we generate data from $Q = (1 - \epsilon)\mathcal{N}(\mathbf{0}_d, I_{d\times d}) + \epsilon\mathcal{N}(\mathbf{10}_d, 0.2I_{d\times d})$, where $0 \leq \epsilon \leq 1$ controls the percentage of observations

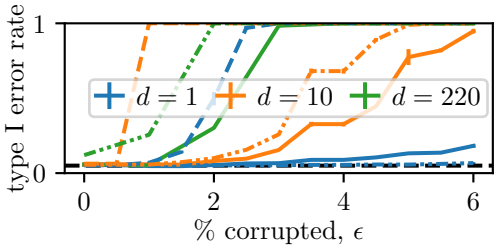


Figure 6: Type I error rate of the tests under corrupted observations. $H_0^C : P = \mathcal{N}(\mu_d, \Sigma_{d\times d})$ and $Q = (1-\epsilon)P + \epsilon\mathcal{N}(10_d, 0.2I_{d\times d})$. $n = 200$. show the MMD, the KSD, the KSD with robust kernel, and -- the level. The error bars show one standard error across 4 random seeds.

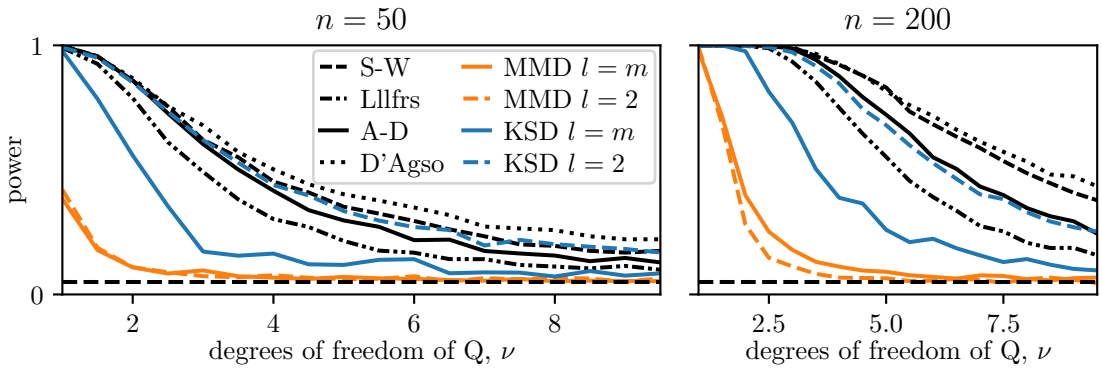


Figure 7: Comparison of the MMD and KSD tests (parametric bootstrap), and the Shapiro-Wilks (S-W), Lilliefors (Lllfrs), Anderson-Darling (A-D) and D'Agostino's (D'Agso) tests. $H_0 : P = \mathcal{N}(\mu, \sigma^2)$, and Q = a series of Student's t-distributions. Two kernel lengthscales are shown for our tests $l = m$ (the median heuristic) and $l = 2$ (selected via grid search). -- shows the level.

that are corrupted. A robust test would maintain the correct type I error rate as ϵ grows. We evaluate the MMD test and KSD test with a Gaussian kernel (as defined in Appendix C.2). We also include a robust KSD test using the tilted kernel $K'(x, x') = w(x)K(x, x')w(x')$, where K is the Gaussian kernel, $w(x) = (1 + a^2 \|x\|_2^2)^{-b}$, and we choose $a = 1$, $b = \frac{1}{2}$ (Barp et al., 2019; Liu and Briol, 2024). The results on one-dimensional observations, $d = 1$, show that the MMD test is robust up to 2-3% corruption, the KSD test with Gaussian kernel is not robust, and the KSD test with tilted kernel is robust to $> 6\%$ corruption. All the tests become less robust as d increases.

To complete the experiments on Gaussian models, we compare the performance of our tests to existing specialised tests of Gaussianity. We consider three nonparametric tests: the

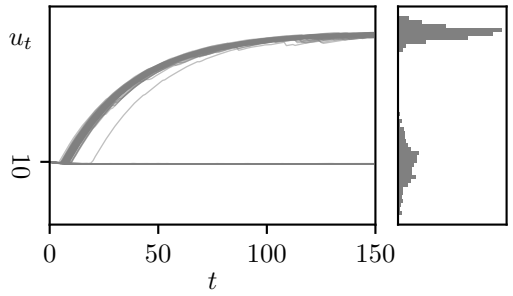


Figure 8: *Left:* Evolution of the u_t variable of the toggle switch model, starting with initial state $u_t = 10$. Each line shows a different random evolution of the model. *Right:* Histogram of the resulting values of y.

Anderson-Darling, Lilliefors (the Kolomogorov-Smirnov test specialised to Gaussians) and Shapiro-Wilks tests. We also consider D'Agostino's K^2 parametric test (D'Agostino and Pearson, 1973). For this comparison we return to testing $H_0^C : P_\theta = \mathcal{N}(\mu, \sigma^2)$, for unknown parameters $\theta = \{\mu, \sigma^2\}$, in one dimension. We perform the test against data generated from a Student's t-distribution, and vary the degree of freedom ν. Once again, we expect the test power to fall to $\alpha = 0.05$, the specified level of the test, as ν becomes large. Figure 7 shows the results of this experiment, for both tests using the parametric bootstrap and for two choices of lengthscale. The MMD test has substantially lower power than the other tests for both lengthscales. The KSD test is an improvement, having performance closer to the specialised tests when setting the lengthscale with the median heuristic. Increasing n improves the power of each test, but does not change the ordering. These results reveal that we should prefer tests specifically designed for the model in question where possible, as opposed to our generally applicable tests. The advantage of our tests is that they are applicable beyond Gaussianity, including for unnormalised or generative models. They are also applicable to multivariate data, which is not the case for the Anderson-Darling or Lilliefors tests. In this plot we also demonstrate the importance of choosing the kernel. We include the power of our test for a more optimal choice of $l = 2$, which we found by sampling additional observations and performing a grid search. For this kernel, the KSD test has comparable power to the specialised tests. However, in practice we cannot use this method to select the kernel, as we are not able to sample additional observations. Future work could look at better strategies than the median heuristic for selecting the bandwidth parameter in composite tests.

5.2 Toggle-Switch Model

We now consider a 'toggle switch' model, which is a generative model with unknown likelihood and hence suitable for our composite MMD tests. The model describes how the expression level of two genes u and v interact in a sample of cells (Bonassi et al., 2011; Bonassi and West, 2015). Sampling from the model involves two coupled discretised-time equations. Denote by $P^{\text{ts}}_{\theta,T}$ a model with parameters $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \mu, \sigma, \gamma)$ with a discretisation consisting of

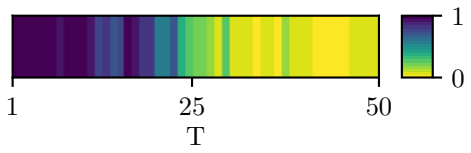


Figure 9: Fraction out of 20 repeats for which the MMD test (parametric bootstrap) rejects $H_0^C : P = P^{\text{ts}}_{\theta,T}$, comparing against data generated from $Q = P^{\text{ts}}_{\theta_0,300}$.

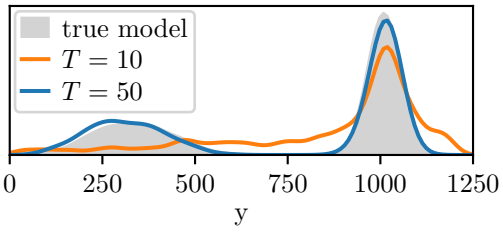


Figure 10: Fit of the toggle switch model for $T = 10$ (for which H_0^C is rejected) and $T = 50$ (for which it is not). The densities are generated using kernel density estimation on 500 samples.

T steps. To sample $y \sim P^{\text{ts}}_{\theta,T}$, we sample $y \sim \widetilde{\mathcal{N}}(\mu + u_T, \mu\sigma/u_T^\gamma)$, where

$$u_{t+1} \sim \widetilde{\mathcal{N}}(\mu_{u,t}, 0.5), \qquad \mu_{u,t} = u_t + \frac{\alpha_1}{\left(1 + v_t^{\beta_1}\right)} - (1 + 0.03u_t),$$
$$v_{t+1} \sim \widetilde{\mathcal{N}}(\mu_{v,t}, 0.5), \qquad \mu_{v,t} = v_t + \frac{\alpha_2}{\left(1 + u_t^{\beta_2}\right)} - (1 + 0.03v_t),$$

and $u_0 = v_0 = 10$. In the above $\widetilde{\mathcal{N}}$ indicates a Gaussian distribution truncated to give non-negative realisations. Figure 8 demonstrates how this sampling process works, showing the evolution of u_t as t increases. To fit the model, we use stochastic optimisation; see Section D.3.3.

Existing work considers the model to have converged for $T = 300$, but using such a large value can be computationally expensive, thus a practitioner may wonder if it is possible to use a smaller value. We apply our parametric bootstrap test to this scenario by generating $n = 500$ observations from $P^{\text{ts}}_{\theta_0,300}$, and then testing the $H_0^C : \exists\tilde{\theta}_0$ such that $P^{\text{ts}}_{\tilde{\theta}_0,T} = P^{\text{ts}}_{\theta_0,300}$ for values of $1 \leq T \leq 50$. To generate the observations we follow Bernton et al. (2019) and use $\theta_0 = (22, 12, 4, 4.5, 325, 0.25, 0.15)$. Figure 9 shows the result of the test, revealing that the test is unlikely to reject H_0^C for $T \geq 40$. In Figure 10 we give examples of the fit of the model for small and large T.

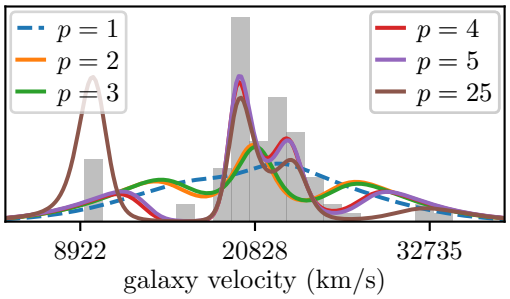


Figure 11: Fit of the kernel exponential family model on the galaxies data set. The grey histogram shows the data set, while the coloured lines show the density of the fitted model with varying p. The dashed lines indicate that the wild bootstrap test rejected H_0^C, while the solid lines indicate that it failed to reject.

5.3 Density Estimation with a Kernel Exponential Family Model

We now consider the kernel exponential family $p_\theta(x) \propto q_{\text{kef}}(x) \exp(f(x))$ where f is an element of some RKHS $\mathcal{H}_\kappa$ with Gaussian kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with lengthscale l_{kef}, and q_{kef} is a reference density (Canu and Smola, 2006). Following Matsubara et al. (2022), we will work with a finite-rank approximation $f(x) = \sum_{i=1}^{p} \theta_i \phi_i(x)$ where $\theta \in \Theta = \mathbb{R}^p$ and

$$\phi_i = \sqrt{\frac{2^i x^{2i}}{l_{\text{kef}}^{2i} i!}} \exp\left(-\frac{x^2}{l_{\text{kef}}^2}\right), \text{ for all } i,$$

are basis functions. Matsubara et al. (2022) use this model for a density estimation task on a data set of velocities of 82 galaxies (Postman et al., 1986; Roeder, 1990), and use the approximation with $p = 25$ for computational convenience. However, an open question is whether $p = 25$ is sufficient to achieve a good fit to the data.

We propose to use our composite KSD test to answer this question. Figure 11 shows the fit of the model for increasing values of p, and whether the test rejected H_0^C. We use a sum of IMQ kernels; see Section D.2 for a discussion of this choice. We find that the test rejects H_0^C for $p = 1, 2, 3$, but does not reject for $p = 4, 5, 25$, with the wild and parametric bootstraps producing similar results. This suggests that $p = 25$ is a suitable choice, though it would also be reasonable to use a smaller value of p which would further decrease the computational cost of inference.

6. Limitations

While we find that our composite tests allow us to answer much more complex questions than existing kernel tests, we did also observe some limitations that we discuss below.

The performance of any kernel-based test depends heavily on the choice of kernel, as this determines the ability of the discrepancy to distinguish between distributions when observations are finite. In our composite tests, this issue has a larger effect on the power of the test since we use the discrepancy not only as a test statistic but also for parameter

estimation. For example, in Appendix B, we demonstrate how a very poor choice of kernel can result in biased estimates of the parameter and thus a type I error rate larger than the level. In practice, we should therefore consider our method as a test of both the model and the performance of the estimator, rather than of the model alone. We consider this to be a desirable behaviour, since, in practice, a parametric model is only ever as useful as our ability to estimate its parameters.

In the non-composite case, several techniques have been developed to address kernel selection, including the power-maximizing kernel (Sutherland et al., 2017), Sup-MMD (Fukumizu et al., 2009), and the aggregated procedure of Schrab et al. (2023, 2022). Future work could investigate how these procedures could be adapted for composite tests. As an initial trial, in Appendix C we apply the aggregated procedure to our composite KSD test, finding that it sometimes achieves a small increase in power, and sometimes a small decrease.

Future work could also consider kernels which are specialised to structured data such as time-series, graphs, and images.

A further issue that can arise is the numerical optimiser implementing the estimator failing to converge to a global optimum (for example, because the objective is non-convex and the optimiser only finds local minimums). This failure mode is also illustrated in Section B. Once again, this means we should consider our method as a test of the fitting method, in addition to the model itself.

All tests based on the KSD can suffer from type I errors due to its inability to distinguish between multi-modal distributions that only differ by the weight of each mode, and this limitation also applies to our test. The limitation can be observed in Figure 11 where, for $p = 25$, the model allocates too much weight to the left mode, yet the test still fails to reject H_0^C. However, as $n \to \infty$ this error will no longer occur.

A final limitation of the tests is due to the difficulty of approximating c_α for models which are expensive to fit due to a large number of parameters and/or the size of the associated data sets. In those cases, the wild bootstrap might not provide a good enough approximation of c_α, whilst the high computational cost of the parametric bootstrap (which requires fitting the model many times) may be prohibitive. Alternative bootstrap algorithms could be developed for this task; see for example the work of Kojadinovic and Yan (2012).

7. Connections with Existing Tests

Connections between (non-composite) kernel-based and classical tests are well-studied; see Sejdinovic et al. (2013) for an extensive discussion showing that MMD tests are equivalent to tests with the energy distance when using an appropriate kernel. Naturally, similar results also hold for composite tests, and we sketch out some of these connections below.

7.1 Likelihood-ratio Tests

The first connection is with likelihood-ratio tests (Lehmann and Romano, 2005, Section 12.4.4), which compare a null model p_{θ_0} and an alternative parametric model q_γ. Interestingly, we will show in this section that our tests are intimately linked to likelihood ratio tests where the alternative model is parameterised in some reproducing kernel Hilbert space.

Likelihood-ratio are based on the test statistic $S_n(\gamma; f) = \frac{\partial}{\partial \gamma} \frac{1}{n} \sum_{i=1}^n \log q_\gamma(x_i)$ which is optimised over γ via the maximum likelihood estimator (MLE). Now consider the alternative

model given by $q_{f,\gamma}(x) \propto \exp(\gamma f(x))p_{\theta_0}(x)$ for some $\gamma \in \mathbb{R}$. This model is indexed by the function f, and consists of a perturbation of p_{θ_0} of size γ in the sense that $q_{f,\gamma} = p_{\theta_0}$ when $\gamma = 0$. In this case,

$$S_n(\gamma; f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{\int_{\mathcal{X}} f(x)\exp(\gamma f(x))p_{\theta_0}(x)dx}{\int_{\mathcal{X}} \exp(\gamma f(x))p_{\theta_0}(x)dx},$$

and $S_n(0; f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{\mathcal{X}} f(x)p_{\theta_0}(x)dx$. If $\gamma = 0$, the MLE of γ will be close to zero, and thus $S_n(0) \approx 0$. Note that $|S_n(\gamma; f)| > 0$ implies that for some value $\gamma > 0$, the perturbation model is a better fit for Q_n than P_{θ_0}. This choice of alternative model allows us to recover the MMD and KSD by considering suitably defined perturbations in some RKHS:

$$\sup_{f\in\mathcal{F}_{\mathrm{MMD}}} S_n(0; f) = \mathrm{MMD}(P_{\theta_0}, Q_n), \qquad \sup_{f\in\mathcal{F}_{\mathrm{KSD}}} S_n(0; f) = \mathrm{KSD}(P_{\theta_0}, Q_n).$$

where $\mathcal{F}_{\mathrm{MMD}}$ and $\mathcal{F}_{\mathrm{KSD}}$ were defined in Section 2. Therefore, MMD and KSD tests are likelihood-ratio tests with an alternative model parametrised in some RKHS. Of course, the argument above also holds if θ_0 is replaced by $\hat{\theta}_n$. As a result, our composite tests can be thought of as a composite version of likelihood-ratio tests where $\hat{\theta}_n$ is a minimum distance estimator based on the MMD and KSD.

7.2 Tests Based on Characteristic Functions

The composite goodness-of-fit tests in this paper are also closely connected with hypothesis tests based on distances between characteristic functions. Recall that the characteristic function of some distribution Q is defined as $\phi_Q(\omega) = \mathbb{E}_{X\sim Q}[\exp(iX^\top\omega)]$. The following results shows that using the MMD is equivalent to comparing the (weighted) L^2 distance between characteristic functions, where the weight depends on the choice of kernel.

Theorem 6 (Theorem 3 and Corollary 4, Sriperumbudur et al. (2010))**.** *Let $\mathcal{X} = \mathbb{R}^d$ and suppose $K(x, y) = \psi(x - y)$ where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a bounded and continuous positive function. Following Bochner's theorem, ψ is the Fourier transform of a finite nonnegative Borel measure Λ: $\psi(x) = \int_{\mathbb{R}^d} \exp(-ix^\top\omega)\Lambda(d\omega)$ and hence the MMD can be expressed as*

$$\mathrm{MMD}^2(P, Q) = \int_{\mathbb{R}^d} |\phi_P(\omega) - \phi_Q(\omega)|^2\Lambda(d\omega).$$

There are a number of composite tests based on weighted distances between characteristic functions. One example is the work of Henze and Zirkler (1990), which consider testing for multivariate Gaussians with unknown mean and covariance. This test therefore uses a distance equivalent to the MMD with a specific choice of kernel. However, this paper makes use of maximum likelihood estimation rather than of a minimum MMD estimator. Interestingly, this means their test is equivalent to that in Kellner and Celisse (2019), but the latter paper does not make the connection and propose to use the test by Henze and Zirkler (1990) as a competitor in their numerical experiments. Note that this correspondence extends to the functional data case; see Wynne and Nagy (2021) for more details.

Another related work is that of Koutrouvelis and Kellermeier (1981), who focus on (possibly non-Gaussian) univariate distributions and consider an approximation of the

weighted L^2 distance between empirical characteristic functions. Here, the approximation consists of discretising the integral in the definition of L^2 distance, and this discrepancy can therefore be thought of as an approximation of the MMD with a specific choice of kernel. Their approach does however have an additional level of approximation due to the discretisation of the integral which is unnecessary when noting the connection with the MMD. Note that Koutrouvelis and Kellermeier (1981) used the same for both discrepancy for estimation and testing.

The connection between tests based on kernels and characteristic functions can also be extended to the case of the KSD, albeit through modified characteristic functions. To state the result, define $\mathcal{S}^l_{P,x}g = g(x)(\partial \log p(x)/\partial x_l) + (\partial g(x)/\partial x_l)$ for some sufficiently regular scalar-valued test function g and $l \in \{1, \ldots, d\}$. For some fixed l, the modified characteristic function of some distribution Q will take the form

$$\tilde{\phi}^l_Q(\omega) = \mathbb{E}_{X\sim Q}\left[\mathcal{S}^l_{P,x}\left[\exp(iX^\top\omega)\right]\right].$$

Theorem 7. *[Simplified version of Theorem 4.1, Wynne et al. (2022)] Let $\mathcal{X} = \mathbb{R}^d$ and suppose $K(x, y) = \psi(x - y)$ where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a bounded and continuous positive function so that $\psi(x) = \int_{\mathbb{R}^d} \exp(-ix^\top\omega)\Lambda(d\omega)$. Then, the KSD can be expressed as*

$$\mathrm{KSD}^2(P, Q) = \int_{\mathbb{R}^d} \sum_{l=1}^{d} \left|\tilde{\phi}^l_Q(\omega)\right|^2 \Lambda(d\omega) = \int_{\mathbb{R}^d} \sum_{l=1}^{d} \left|\tilde{\phi}^l_Q(\omega) - \tilde{\phi}^l_P(\omega)\right|^2 \Lambda(d\omega).$$

Ebner (2021) proposed a test of normality which uses a test statistic exactly of the form in Theorem 7. Their test implicitly assumes that Λ has a density, and they study the impact of the choice of density on the performance of the test. Through the theorem above, we can see that this is equivalent to varying the choice of kernel K indexing the KSD.

8. Conclusion

This paper proposes and studies two kernel-based tests to verify whether a given data set is a realisation from any element of some parametric family of distributions. This was achieved by combining existing kernel hypothesis tests with recently developed minimum distance estimators, using either the MMD or KSD. We also studied two bootstrap algorithms to implement these tests: a wild bootstrap algorithm which is suitable in large-data regimes, and a parametric bootstrap which is suitable in smaller data regimes.

A number of limitations were mentioned in Section 6, and these could each be addressed in future work. For example, the issue that the tests perform poorly when the minimum distance estimators provide a poor parameter estimate could be mitigated by allowing for the use of alternative estimators which are specialised for models at hand. This is straightforward to do in practice, but would require an extension of our theoretical results. Relatedly, the performance of the composite tests is dependent on the choice of an appropriate divergence, usually through a choice of kernel. This issue could be mitigated by using the aggregate approach of Schrab et al. (2023, 2022), which removes the need to carefully select a kernel, within our composite tests.

Additionally, our paper focused on Euclidean data, our MMD and KSD tests could straightforwardly be generalised to more complex settings. For example, the cases of discrete

(Yang et al., 2018), manifold (Xu and Matsuda, 2020), censored (Fernández and Gretton, 2019) or time-to-event (Fernández et al., 2020) data could all be covered through our general methodology. Furthermore, models such as point processes (Yang et al., 2019), latent variable models (Kanagawa et al., 2019) or exponential random graph models (Xu and Reinert, 2021) could also be tackled in this way.

Finally, we also envisage that our approach could be extended to more complex testing questions. For example, the framework could be extended to relative tests, where instead of testing whether data comes from a fixed parametric model, the relative fit of several parametric models would be compared (Bounliphone et al., 2016; Jitkrittum et al., 2018; Lim et al., 2019). Alternatively, we could extend our methodology to construct composite tests for conditional distributions (Jitkrittum et al., 2020).

Acknowledgments

We thank Antonin Schrab for his helpful instructions on implementing the aggregated version of test, as presented in Section C.

OK and FXB acknowledge support from the Engineering and Physical Sciences Research Council with grant numbers EP/S021566/1 and EP/Y011805/1 respectively. TF was supported by the Data Observatory Foundation - ANID Technology Center No. DO210001 and ANID FONDECYT grant No. 11221143. Arthur Gretton acknowledges support from the Gatsby Charitable Foundation.

Appendix

Contents

Appendix A. Theoretical Results for MMD

A.1 Definitions

A.1.1 General Notation

In this section we make some additional definitions which will be used in the proofs. We denote by Q_n the empirical measure based on the sample $X_1, \dots, X_n \overset{i.i.d.}{\sim} Q$, and we denote by $Q_n^\star$ the empirical measure based on the parametric bootstrap samples $X_1^\star, \dots, X_n^\star \overset{i.i.d}{\sim} P_{\hat{\theta}_n}$ given $\hat{\theta}_n$.

1. We define the functions $L_n : \Theta \to \mathbb{R}$ and $L : \Theta \to \mathbb{R}$ by

$$L_n(\theta) = \mathrm{MMD}^2(P_\theta, Q_n) \qquad \text{and} \qquad L(\theta) = \mathrm{MMD}^2(P_\theta, Q). \tag{4}$$

   Similarly, for the parametric Bootstrap, we define $L_n^\star : \Theta \to \mathbb{R}$ by

$$L_n^\star(\theta) = \mathrm{MMD}^2(P_\theta, Q_n^\star). \tag{5}$$

   Note then that $\hat{\theta}_n = \arg\min_{\theta\in\Theta} L_n(\theta)$ and $\theta_n^\star = \arg\min_{\theta\in\Theta} L_n^\star(\theta)$.

2. We denote by $\mathbf{H} \in \mathbb{R}^{p\times p}$ the Hessian matrix given by $\mathbf{H}_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L(\theta_0)$ for any $i, j \in \{1, \dots, p\}$.

3. We define the function $g : \mathcal{H} \times \Theta \to \mathbb{R}$ by $g(\omega, \theta) = \mathbb{E}_{X\sim P_\theta}(\omega(X))$.

4. We denote by μ_θ, μ_n and $\mu_n^\star \in \mathcal{H}$, respectively, the kernel mean embeddings of the distribution function P_θ and the empirical distribution functions Q_n and $Q_n^\star$, given by

$$\mu_\theta(x) = \int_{\mathcal{X}} K(x,y) p_\theta(y) \lambda(dy), \quad \mu_n(x) = \frac{1}{n}\sum_{i=1}^n K(x, X_i), \quad \mu_n^\star(x) = \frac{1}{n}\sum_{i=1}^n K(x, X_i^\star).$$

5. We define $\boldsymbol{\phi} : \mathcal{X} \times \Theta \to \mathbb{R}^p$ by

$$\boldsymbol{\phi}(x,\theta) = \nabla_\theta \mu_\theta(x) - \int_{\mathcal{X}} (\nabla_\theta \mu_\theta(y)) p_\theta(y) \lambda(dy), \tag{6}$$

where $\nabla_\theta \mu_\theta : \mathcal{X} \to \mathbb{R}^p$ is such that $(\nabla_\theta \mu_\theta(x))_i = \frac{\partial}{\partial \theta_i} \mu_\theta(x)$ for $i \in \{1, \ldots, p\}$.

6. We define by $\eta : \mathcal{H} \times \mathcal{X} \times \Theta \to \mathbb{R}$ by

$$\eta(\omega, x, \theta) = \omega(x) - g(\omega, \theta) - 2 \left\langle \nabla_\theta g(\omega, \theta_0), \mathbf{H}^{-1} \boldsymbol{\phi}(x, \theta) \right\rangle, \tag{7}$$

where $\nabla_\theta g : \mathcal{H} \times \Theta \to \mathbb{R}^p$ is such that $(\nabla_\theta g(\omega, \theta))_i = \frac{\partial}{\partial \theta_i} g(\omega, \theta)$ for $i \in \{1, \ldots, p\}$.

7. We define the functionals $S_n : \mathcal{H} \to \mathbb{R}$ and $S_n^\star : \mathcal{H} \to \mathbb{R}$ by

$$S_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta(\omega, X_i, \theta_0), \qquad \text{and} \qquad S_n^\star(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \eta(\omega, X_i^\star, \hat{\theta}_n), \tag{8}$$

where $\eta : \mathcal{H} \times \mathcal{X} \times \Theta \to \mathbb{R}$ is defined in Equation 7.

Remark 8 (*Bound for RKHS functions*)**.** *Observe that for any $\omega \in \mathcal{H}$ and $x \in \mathcal{X}$,*

$$|\omega(x)| = |\langle \omega, K_x \rangle_{\mathcal{H}}| \leq \|\omega\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} = \|\omega\|_{\mathcal{H}} \sqrt{K(x,x)},$$

where the inequality is due to the Cauchy-Schwarz's inequality. By Assumptions 1 to 6, we have that the kernel is bounded by some constant $C > 0$. Thus $\sup_{x \in \mathcal{X}} |\omega(x)| \leq \|\omega\|_{\mathcal{H}} \sqrt{C}$.

Remark 9 (*Interchange of integration and differentiation*)**.** *Observe that for any fixed $\omega \in \mathcal{H}$, $g(\omega, \cdot) \in \mathcal{C}^3(\Theta)$ and, moreover*

$$\mathcal{D}_\theta g(\omega, \theta) = \int_{\mathcal{X}} \omega(x) \mathcal{D}_\theta p_\theta(x) \lambda(dx),$$

where $\mathcal{D}_\theta$ is any derivative up to order three. The previous result holds by an application of Lebesgue's dominated convergence theorem since $p_\theta(x) \in \mathcal{C}^3(\Theta)$ for all $x \in \mathcal{X}$, and by the integrability conditions of Assumption 3. Similarly $\mu_\theta(x) \in \mathcal{C}^3(\Theta)$ for all $x \in \mathcal{X}$, and $\mathcal{D}_\theta \mu_\theta(x) = \int_{\mathcal{X}} K(x,y) \mathcal{D}_\theta p_\theta(y) \lambda(dy)$.

Moreover, by the integrability conditions of Assumption 3 and since the kernel is bounded by a constant $C > 0$ (Assumption 5), it holds

$$\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \left| \int_{\mathcal{X}} K(x,y) \mathcal{D}_\theta p_\theta(y) \lambda(dy) \right| \leq C \int_{\mathcal{X}} \sup_{\theta \in \Theta} |\mathcal{D}_\theta p_\theta(y)| \, \lambda(dy) < \infty. \tag{9}$$

A.1.2 Test Statistic

In our proofs we consider an alternative form of the definition of the MMD, which is equivalent to that in Equation 2:

$$n\,\mathrm{MMD}^2(P_\theta, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega(X_i) - g(\omega,\theta)\right)^2.$$

A.1.3 Wild Bootstrap

In order to analyse the asymptotic behaviour of the wild bootstrap test-statistic, we write it in terms of a supremum as follows

$$\begin{aligned} n\,\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n) &:= \frac{1}{n}\sum_{i,j=1}^{n} W_i W_j h_{\mathrm{MMD}}((X_i,\tilde{X}_i),(X_j,\tilde{X}_j)) \\ &= \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\omega(X_i) - \omega(\tilde{X}_i))\right)^2, \end{aligned}$$

where recall that $X_1,\ldots,X_n \overset{i.i.d.}{\sim} Q$, $\tilde{X}_1,\ldots,\tilde{X}_n \overset{i.i.d.}{\sim} P_{\hat{\theta}_n}$ given $\hat{\theta}_n$, and $W_1,\ldots,W_n \overset{i.i.d.}{\sim}$ Rademacher are independent of everything.

A.1.4 Parametric bootstrap

To obtain the results for the parametric bootstrap, we condition on the data $X_1,\ldots,X_n$. To this end, we use the following notation:

- Define

$$\begin{aligned} &\mathbb{P}_D(\cdot) = \mathbb{P}(\cdot|X_1,\ldots,X_n), \quad \mathbb{E}_D(\cdot) = \mathbb{E}(\cdot|X_1,\ldots,X_n), \\ &\mathbb{V}\mathrm{ar}_D(\cdot) = \mathbb{V}\mathrm{ar}(\cdot|X_1,\ldots,X_n), \text{ and } \mathbb{C}\mathrm{ov}_D(\cdot,\cdot) = \mathbb{C}\mathrm{ov}(\cdot,\cdot|X_1,\ldots,X_n). \end{aligned}$$

- Define $a_n = o_{p_D}(1)$ if, for each $\varepsilon > 0$, it holds $\mathbb{P}_D(|a_n| \geq \varepsilon) \xrightarrow{\mathbb{P}} 0$ when $n \to \infty$.
- Define $a_n = O_{p_D}(1)$ if for any $\varepsilon > 0$ there exists $M > 0$ such that

$$\mathbb{P}\left(\{\mathbb{P}_D(|a_n| > M) \leq \varepsilon\}\right) \to 1.$$

- Given a random variable a, define $a_n \xrightarrow{\mathcal{D}_D} a$ (i.e. a_n converges in distribution to a given the data points $X_1,\ldots,X_n$) if and only if $|\mathbb{E}_D(f(a_n) - f(a))| \xrightarrow{\mathbb{P}} 0$, for any bounded, uniformly continuous real-valued f.

A.2 Preliminary Results

We state some preliminary results that will be used in our proofs. The first corresponds to Theorem 1 of Fernández and Rivera (2022), and it will be used to prove Theorems 2 and 5.

Theorem 10. *Let $(S_n)_{n\geq 1}$ be a sequence of bounded linear test-statistics satisfying conditions G_0-G_2 (stated below). Then*

$$\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega) \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where $Z_1, Z_2, \ldots$ are i.i.d. standard normal random variables, and $\lambda_1, \lambda_2, \ldots$ are the eigenvalues of the operator $T_\sigma : \mathcal{H} \to \mathcal{H}$ defined by $(T_\sigma \omega)(x) = \sigma(\omega, K_x)$, where $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is the bilinear form of Condition G_0.

We state conditions G_0-G_2.

Condition G_0 *There exists a continuous bilinear form $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that for any $m \in \mathbb{N}$, the bounded linear test-statistic $S_n(w_1 + \cdots + w_m)$ converges in distribution to a normal random variable with mean 0 and variance given by $\sum_{i=1}^m \sum_{j=1}^m \sigma(w_i, w_j)$.*

Condition G_1 *For some orthonormal basis $(\psi_i)_{i\geq 1}$ of $\mathcal{H}$ we have $\sum_{i\geq 1} \sigma(\psi_i, \psi_i) < \infty$.*

Condition G_2 *For some orthonormal basis $(\psi_i)_{i\geq 1}$ of $\mathcal{H}$, and for any $\varepsilon > 0$, we have that*

$$\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k=i+1}^{\infty} S_n^2(\psi_k) \geq \varepsilon \right) = 0.$$

The following result appears in Nickl (2012).

Theorem 11 (Theorem 1, Nickl (2012))**.** *Suppose that $\Theta \subset \mathbb{R}^p$ is compact (i.e. bounded and closed). Assume that $L : \Theta \to \mathbb{R}$ is a deterministic function with $L \in \mathcal{C}^0(\Theta)$, and that θ_{opt} is the unique minimiser of L. If*

$$\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\mathbb{P}} 0$$

as n grows to infinity, then any solution $\hat{\theta}_n$ of $\min_{\theta\in\Theta} L_n(\theta)$ satisfies

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_{opt} \quad as \quad n \to \infty.$$

A.3 Auxiliary Results

The following collection of lemmas and propositions will be used in the proofs of our main results. Their proofs are provided in Section A.5.

Lemma 12. *Under Assumptions 1 to 6 and H_0^C it holds that*

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) + o_p(1),$$

where $\boldsymbol{\phi} : \mathcal{X} \times \Theta \to \mathbb{R}^p$ is defined in Equation 6, and $\mathbf{H}$ is the Hessian matrix defined in Section A.1.1.

Proposition 13 (Normal distribution for the Wild Bootstrap statistic)**.**
Define $Z_n : \mathcal{H} \to \mathbb{R}$ *by*

$$Z_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(\omega(X_i) - \omega(\tilde{X}_i)),$$

where $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Q$ *and* $\tilde{X}_1, \ldots, \tilde{X}_n \overset{i.i.d.}{\sim} P_{\hat{\theta}_n}$ *given* $\hat{\theta}_n$*, and* $W_1, \ldots, W_n \overset{i.i.d.}{\sim}$ *Rademacher that are independent of everything else.*

Then, $Z_n(\omega) \overset{\mathcal{D}_D}{\to} N(0, \nu(\omega, \omega))$*, where*

$$\nu(\omega, \omega') = 2 \int (\omega(x) - g(\omega, \theta_0))(\omega'(x) - g(\omega', \theta_0)) p_{\theta_0}(x) \lambda(dx). \tag{10}$$

Proposition 14. *Under Assumptions 1 to 6 and the null hypothesis it holds that* $\hat{\theta}_n \overset{\mathbb{P}}{\to} \theta_0$.

Proposition 15.

$$\nabla_\theta L_n(\theta) = -\frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(X_i, \theta), \qquad \text{and} \qquad \nabla_\theta L_n^\star(\theta) = -\frac{2}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(X_i^\star, \theta).$$

Moreover, under Assumptions 1 to 6 the following items hold.

i) $\mathbb{E}_{X \sim P_\theta}(\boldsymbol{\phi}(X, \theta)) = \mathbf{0}$*, and* $\mathbb{E}_D(\boldsymbol{\phi}(X^\star, \hat{\theta}_n)) = \mathbf{0}$*, where* $X^\star \sim P_{\hat{\theta}_n}$ *given* $\hat{\theta}_n$*,*

ii) $\sup_{\theta \in \Theta} \sup_{x \in \mathbb{R}^d} \|\boldsymbol{\phi}(x, \theta)\| < \infty$*, and*

iii) $\sqrt{n} \|\nabla_\theta L_n(\theta_0)\| = O_p(1)$ *holds under the null hypothesis, and* $\sqrt{n} \|\nabla_\theta L_n^\star(\hat{\theta}_n)\| = O_p(1)$.

Proposition 16. *Under Assumptions 1 to 6 the following items hold.*

i) For any fixed $\theta \in \Theta$ *and* $\omega \in \mathcal{H}$*,* $\mathbb{E}_{X \sim P_\theta}(\eta(\omega, X, \theta)) = 0$*,*

ii) there exists constants $C_1 > 0$ *and* $C_2 > 0$ *such that for any fixed* $\theta \in \Theta$ *and* $\omega \in \mathcal{H}$*,*

$$\mathbb{E}_{X \sim P_\theta}(\eta^2(\omega, X, \theta)) \leq C_1 \mathbb{E}_{X \sim P_\theta}\left(\omega^2(X)\right) + C_2 \left\|\nabla_\theta g(\omega, \theta_0)\right\|^2,$$

iii) there exists constants $C_1 > 0$ *and* $C_2 > 0$ *that do not depend on* ω*,* x *nor* θ*, such that for all* $\omega \in \mathcal{H}$*, it holds* $\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} |\eta(\omega, x, \theta)| < C_1 \|\omega\|_{\mathcal{H}} + C_2 \|\omega\|_{\mathcal{H}}^2 < \infty$.

Lemma 17. *Under Assumptions 1 to 6 and the null hypothesis, it holds that for any* $\ell, j \in \{1, \ldots, p\}$ *and* $\varepsilon > 0$

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial^2}{\partial \theta_\ell \partial \theta_j} \mu_\theta(X_i) - \mathbb{E}_{X \sim P_{\theta_0}} \left(\frac{\partial^2}{\partial \theta_\ell \partial \theta_j} \mu_\theta(X) \right) \right) \right| \overset{\mathbb{P}}{\to} 0, \tag{11}$$

and,

$$\mathbb{P}_D \left(\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial^2}{\partial \theta_\ell \partial \theta_j} \mu_\theta(X_i^\star) - \mathbb{E}_{X \sim P_{\theta_0}} \left(\frac{\partial^2}{\partial \theta_\ell \partial \theta_j} \mu_\theta(X) \right) \right) \right| \geq \varepsilon \right) \overset{\mathbb{P}}{\to} 0, \tag{12}$$

as n *grows to infinity.*

Proposition 18. *Define* $\mathbf{H}_n$ *and* $\mathbf{H}_n^\star$ *by*

$$(\mathbf{H}_n)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L_n(\tilde{\theta}^j) \qquad \text{and} \qquad (\mathbf{H}_n^\star)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L_n^\star(\tilde{\theta}_\star^j)$$

for any $i, j \in \{1, \ldots, p\}$*, where* L_n *and* $L_n^\star$ *are defined in Equations* 4 *and* 5*, respectively,* $\tilde{\theta}^1, \ldots, \tilde{\theta}^p$ *lie between* θ_0 *and* $\hat{\theta}_n$*, and* $\tilde{\theta}_\star^1, \ldots, \tilde{\theta}_\star^p$ *lie between* $\hat{\theta}_n$ *and* $\hat{\theta}_n^\star$*.*

Then, under Assumptions 1 to 6 and the null hypothesis it holds

$$\|\mathbf{H}_n - \mathbf{H}\| = o_p(1) \qquad \text{and} \qquad \|\mathbf{H}_n^\star - \mathbf{H}\| = o_{p_D}(1)$$

as n *grows to infinity.*

Proposition 19. *Under Assumptions 1 to 6 and the null hypothesis, the following holds.*

i) $\sqrt{n}(\theta_0 - \hat{\theta}_n) = \sqrt{n}\mathbf{H}^{-1}\nabla_\theta L_n(\theta_0) + o_p(1)$*, and*

ii) $\sqrt{n}(\hat{\theta}_n - \theta_n^\star) = \sqrt{n}\mathbf{H}^{-1}\nabla_\theta L_n^\star(\hat{\theta}_n) + o_p(1)$*.*

Lemma 20. *Under Assumptions 1 to 6, it holds*

i) $\sqrt{n}(\hat{\theta}_n - \theta_n^\star) = O_p(1)$*, and*

ii) $\theta_n^\star - \theta_0 = o_p(1)$ *holds under the null hypothesis.*

Proposition 21. *Under the null hypothesis and Assumptions 1 to 6, we have that*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \omega(X_i) - g(\omega, \hat{\theta}_n) = S_n(\omega) + o_p(1),$$

holds for every $\omega \in \mathcal{H}$*, where* $S_n : \mathcal{H} \to \mathbb{R}$ *is defined in Equation* 8 *and the error term does not depend on* ω*.*

Proposition 22. *Consider* $\boldsymbol{S}_n = (S_n(\omega_1), \ldots, S_n(\omega_m))$*, where* $\omega_\ell \in \mathcal{H}$ *for each* $\ell \in [m]$ *and fixed* $m \in \mathbb{N}$*. Then, under the null hypothesis, and Assumptions 1 to 6, we have* $\boldsymbol{S}_n \xrightarrow{\mathcal{D}} N_m(0, \boldsymbol{\Sigma})$ *as* n *grows to infinity, where for any* $i, j \in [m]$ *we have*

$$\boldsymbol{\Sigma}_{ij} = \sigma(\omega_i, \omega_j) := \mathbb{E}_{X \sim P_{\theta_0}} \left(\eta(\omega_i, X, \theta_0) \eta(\omega_j, X, \theta_0) \right), \tag{13}$$

and the function η *is defined in Equation* 7*.*

The next two results are analogous to Proposition 21 and Proposition 22 but for the parametric bootstrap test-statistic. Note that these results are obtained conditioned on the data $X_1, \ldots, X_n$.

Proposition 23. *Under the null hypothesis and Assumptions 1 to 6, it holds that*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \omega(X_i^\star) - g(\omega, \theta_n^\star) = S_n^\star(\omega) + o_p(1),$$

where $S_n^\star : \mathcal{H} \to \mathbb{R}$ *is defined in Equation* 8*.*

Proposition 24. *Consider* $\boldsymbol{S}_n^\star = (S_n^\star(\omega_1), \ldots, S_n^\star(\omega_m))$*, where* $\omega_\ell \in \mathcal{H}$ *for each* $\ell \in [m]$ *and* $m \in \mathbb{N}$ *is fixed. Then, under the null hypothesis, and Assumptions 1 to 6, we have* $\boldsymbol{S}_n^\star \xrightarrow{\mathcal{D}} N_m(0, \boldsymbol{\Sigma})$ *as* n *grows to infinity, where* $\boldsymbol{\Sigma}$ *is defined in Equation* 13

Proposition 25. *Let* $(\psi_i)_{i \geq 1}$ *be any orthonormal basis of* $\mathcal{H}$*. Then under Assumptions 1 to 6 it holds that* $\sum_{i \geq 1} \|\nabla_\theta g(\psi_i, \theta_0)\|^2 < \infty$*.*

A.4 Proofs of Main Results

This section contains proofs of results given in the main paper.

Proof of Theorem 2 By the definition of the maximum mean discrepancy, it holds

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \gamma_n^2(\omega), \quad \text{where} \quad \gamma_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \omega(X_i) - g(\omega, \hat{\theta}_n).$$

Proposition 21 yields $\gamma_n(\omega) = S_n(\omega) + o_p(1)$, where $S_n : \mathcal{H} \to \mathbb{R}$ is the functional defined in Equation 8, and the error term does not depend on ω. Then, by using the fact that $\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega)$ converges in distribution (which will be proved next) we obtain

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega) + o_p(1).$$

By Slutsky's theorem, we deduce that $n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n)$ and $\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega)$ have the same asymptotic null distribution.

It is not difficult to verify that $S_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta(\omega, X_i, \theta_0)$ is bounded and linear in $\mathcal{H}$. The linearity follows from the fact that for any fixed $x \in \mathcal{X}$ and $\theta \in \Theta$, $\eta(\omega, x, \theta)$ is linear on its argument ω. Boundedness follows from item iii of Proposition 16 since $\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}} |\eta(\omega, x, \theta)| < C_1 \|\omega\|_{\mathcal{H}} + C_2 \|\omega\|_{\mathcal{H}}^2$ where $C_1 > 0$ and $C_2 > 0$ are constants that do not depend on ω. Then, we can use Theorem 10 to study the asymptotic null distribution of $\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n(\omega)^2$ and thus obtain the result. To apply Theorem 10 we need to verify conditions G_0, G_1 and G_2, which are stated in Section A.2.

Condition G_0 follows directly from Proposition 22, where $\sigma : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ is given by $\sigma(\omega, \omega') := \mathbb{E}_{X\sim P_{\theta_0}}(\eta(\omega, X, \theta_0)\eta(\omega', X, \theta_0))$ for any $\omega, \omega' \in \mathcal{H}$. To verify condition G_1 consider an orthonormal basis $(\psi_i)_{i\geq 1}$ of $\mathcal{H}$ and observe that by Proposition 16.ii. there exists constants $C_1 > 0$ and $C_2 > 0$ such that for all $i \geq 1$ it holds

$$\sigma(\psi_i, \psi_i) = \mathbb{E}_{X\sim P_{\theta_0}}(\eta^2(\psi_i, X, \theta_0)) < C_1 \mathbb{E}_{X\sim P_{\theta_0}}\left(\psi_i{}^2(X)\right) + C_2 \left\|\nabla_\theta g(\psi_i, \theta_0)\right\|^2.$$

Then,

$$\sum_{i\geq 1} \sigma(\psi_i, \psi_i) \leq C_1 \sum_{i\geq 1} \mathbb{E}_{X\sim P_{\theta_0}}\left(\psi_i{}^2(X)\right) + C_2 \sum_{i\geq 1} \|\nabla_\theta g(\psi_i, \theta_0)\|^2 < \infty. \tag{14}$$

where the last inequality holds since $K(x, x) = \sum_{i\geq 1} \psi_i^2(x)$, and the kernel is bounded by our assumptions, and by Proposition 25. Finally, we proceed to verify condition G_2. For this, let $\varepsilon > 0$, and note

$$\begin{aligned}\mathbb{P}\left(\sum_{k\geq i} S_n^2(\psi_k) \geq \varepsilon\right) \leq \sum_{k\geq i} \varepsilon^{-1}\mathbb{E}\left(S_n^2(\psi_k)\right) &= \varepsilon^{-1}\sum_{k\geq i} \mathbb{E}_{X\sim P_{\theta_0}}(\eta^2(\psi_k, X, \theta_0)) \\ &= \varepsilon^{-1}\sum_{k\geq i} \sigma(\psi_k, \psi_k),\end{aligned}$$

where first inequality is due to Markov's inequality, the subsequent equality follows from the definition of $S_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \eta(\omega, X_i, \theta_0)$ given in Equation 8, and the last equality is due

to the definition of σ. Finally, notice that

$$\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k\geq i} S_n(\psi_k)^2 \geq \varepsilon\right) \leq \varepsilon^{-1} \lim_{i\to\infty} \sum_{k\geq i} \sigma(\psi_k, \psi_k) = 0,$$

where the limit is due to Equation 14. ■

Proof of Theorem 3 Observe that

$$\mathrm{MMD}(P_{\hat{\theta}_n}, Q_n) = \left\|\mu_{\hat{\theta}_n} - \mu_n\right\|_{\mathcal{H}} \geq \left\|\mu_{\hat{\theta}_n} - \mu_Q\right\|_{\mathcal{H}} - \|\mu_Q - \mu_n\|_{\mathcal{H}},$$

where μ_Q is the kernel mean embedding of Q given by $\mu_Q(x) = \int K(x,y)q(y)\lambda(dy)$. Note that $\lim_{n\to\infty} \|\mu_Q - \mu_n\|_{\mathcal{H}} = 0$ a.s. by the standard law of large numbers for RKHS in Chen and Zhu (2011, Theorem 1). Next, we will prove that

$$\liminf_{n\to\infty} \left\|\mu_{\hat{\theta}_n} - \mu_Q\right\|_{\mathcal{H}} > 0 \quad a.s.$$

We proceed by contradiction. Suppose that $\liminf_{n\to\infty} \left\|\mu_{\hat{\theta}_n} - \mu_Q\right\|_{\mathcal{H}} = 0$ a.s., then there exists a collection of indices $(a(n))_{n\geq 1}$ such that the subsequence $\mu_{\widehat{\theta}_{a(n)}}$ satisfies

$$\lim_{n\to\infty} \left\|\mu_{\widehat{\theta}_{a(n)}} - \mu_Q\right\|_{\mathcal{H}}^2 = 0 \quad a.s. \tag{15}$$

Additionally, note that since the set Θ is compact, there exists a subsequence $(\widehat{\theta}_{a(b(n))})_{n\geq 1}$ and $\theta^\star \in \Theta$ such that $\lim_{n\to\infty} \|\widehat{\theta}_{a(b(n))} - \theta^\star\| = 0, a.s.$ By Assumptions 3 and 5, the kernel K is bounded, $p_\theta(x) \in \mathcal{C}^3(\Theta)$ for each fixed $x \in \mathcal{X}$, and $\int_{\mathcal{X}} \sup_{\theta\in\Theta} p_\theta(x)\lambda(x)dx < \infty$. Therefore, by the Lebesgue's dominated convergence theorem we deduce

$$\begin{aligned}
&\left\|\mu_{\widehat{\theta}_{a(b(n))}} - \mu_{\theta^\star}\right\|_{\mathcal{H}}^2 \\
&\quad = \int_{\mathcal{X}}\int_{\mathcal{X}} K(x,y)\left(p_{\widehat{\theta}_{a(b(n))}}(x) - p_{\theta^\star}(x)\right)\left(p_{\widehat{\theta}_{a(b(n))}}(y) - p_{\theta^\star}(y)\right)\lambda(x)\lambda(y)dxdy \\
&\quad \to 0
\end{aligned} \tag{16}$$

when n grows to infinity. Then, by the triangle inequality and Equations 15 and 16 we have

$$\|\mu_{\theta^\star} - \mu_Q\|_{\mathcal{H}} \leq \left\|\mu_{\theta^\star} - \mu_{\widehat{\theta}_{a(b(n))}}\right\|_{\mathcal{H}} + \left\|\mu_{\widehat{\theta}_{a(b(n))}} - \mu_Q\right\|_{\mathcal{H}} \to 0 \quad a.s.$$

Thus $\|\mu_{\theta^\star} - \mu_Q\|_{\mathcal{H}} = \mathrm{MMD}(Q, P_{\theta^\star}) = 0$, and since the kernel is characteristic by Assumption 5, we conclude $Q = P_{\theta^\star}$, but this is a contradiction since under the alternative hypothesis $Q \notin \mathcal{P}_\theta$. ■

Proof of Theorem 4 Observe that

$$n\,\mathrm{MMD}_W^2(P_{\hat{\theta}_n,n}, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} Z_n^2(\omega), \tag{17}$$

where $Z_n : \mathcal{H} \to \mathbb{R}$ is defined $Z_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\omega(X_i) - \omega(\tilde{X}_i))$. It is not difficult to see that Z_n is linear on its argument ω, and that Z_n is bounded since Assumptions 1 to 6 yield $\sup_{x\in\mathcal{X}} |\omega(x)| \leq \|\omega\|_{\mathcal{H}} \sup_{x\in\mathcal{X}} \sqrt{K(x,x)} < \infty$. Then, the asymptotic distribution of the wild bootstrap test-statistic $n\,\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n)$ can be obtained from Theorem 10 and Equation 17, by verifying that Z_n satisfies conditions G_0-G_2, which we proceed to verify.

Observe that Condition G_0 follows directly from Proposition 13, where we obtain that

$$Z_n(\omega_1 + \ldots + \omega_m) \overset{\mathcal{D}_D}{\to} N\left(0, \sum_{i=1}^m \sum_{j=1}^m \nu(\omega_i, \omega_j)\right),$$

for any $m \in \mathbb{N}$ and $\omega_1, \ldots, \omega_m \in \mathcal{H}$, where $\nu : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is given by

$$\nu(\omega, \omega') = 2 \int (\omega(x) - g(\omega, \theta_0))(\omega'(x) - g(\omega', \theta_0)) p_{\theta_0}(x) \lambda(dx).$$

To verify Condition G_1, we consider some orthonormal basis $(\psi_i)_{i\geq 1}$ of $\mathcal{H}$, and verify that $\sum_{i\geq 1} \nu(\psi_i, \psi_i) < \infty$ holds. Note that

$$\begin{aligned}\sum_{i\geq 1} \nu(\psi_i, \psi_i) &= \sum_{i\geq 1} 2 \int (\psi_i(x) - g(\psi_i, \theta_0))^2 p_{\theta_0}(x) \lambda(dx) \\ &\leq 2 \int \sum_{i\geq 1} \psi_i(x)^2 p_{\theta_0}(x) \lambda(dx) = 2 \int K(x,x) p_{\theta_0}(x) \lambda(dx) < \infty,\end{aligned}$$

where the first inequality follows from the fact that the variance of a random variable is upper bounded by its second moment, and the last inequality holds from the fact that $\sum_{i\geq 1} \psi_i^2(x) = K(x,x)$, and that the kernel is bounded by Assumption 5.

Finally, note that $\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}_D\left(\sum_{k\geq i} Z_n^2(\psi_k) \geq \varepsilon\right) = 0$ verifies Condition G_2. By Markov's inequality

$$\mathbb{P}_D\left(\sum_{k\geq i} Z_n^2(\psi_k) \geq \varepsilon\right) \leq \frac{1}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D(Z_n^2(\psi_k)),$$

and note that

$$\begin{aligned}\frac{1}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D(Z_n^2(\psi_k)) &= \frac{1}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D\left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n W_j(\psi_k(X_j) - \psi_k(\tilde{X}_j))\right)^2\right) \\ &= \frac{1}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D\left(\mathbb{E}\left(\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n W_j(\psi_k(X_j) - \psi_k(\tilde{X}_j))\right)^2 \middle| (\tilde{X}_\ell)_{\ell=1}^n\right)\right) \\ &= \frac{1}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D\left(\frac{1}{n} \sum_{j=1}^n (\psi_k(X_j) - \psi_k(\tilde{X}_j))^2\right) \\ &\leq \frac{2}{\varepsilon} \sum_{k\geq i} \left(\frac{1}{n} \sum_{j=1}^n \psi_k(X_j)\right)^2 + \frac{2}{\varepsilon} \sum_{k\geq i} \mathbb{E}_D\left(\psi_k^2(\tilde{X}_1)\right)\end{aligned}$$

where the first equality follows from replacing Z_n by its definition, the third equality follows from noticing that conditioned on the data $(X_i, \tilde{X}_i)_{i=1}^n$, Z_n is the sum of independent random variables, $\mathbb{E}(W_i|(X_i, \tilde{X}_i)_{i=1}^n) = 0$ and $W_i^2 = 1$. Finally, note that the last inequality holds from noticing that $(a-b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$, and $(\tilde{X}_i)_{i=1}^n$ are i.i.d. give the data $(X_i)_{i=1}^n$.

By using the previous set of equations, we obtain

$$\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}_D \left(\sum_{k\leq i} Z_n^2(\psi_k) \geq \varepsilon \right)$$
$$\leq \lim_{i\to\infty} \limsup_{n\to\infty} \frac{2}{\varepsilon} \sum_{k\geq i} \left(\frac{1}{n} \sum_{j=1}^n \psi_k(X_j) \right)^2 + \lim_{i\to\infty} \limsup_{n\to\infty} \frac{2}{\varepsilon} \int \sum_{k\geq i} \psi_k^2(x) p_{\hat{\theta}_n}(x) \lambda(dx)$$

We shall prove the previous equation is zero. For the first term observe that

$$\begin{aligned} \lim_{i\to\infty} \limsup_{n\to\infty} \frac{2}{\varepsilon} \sum_{k\geq i} \left(\frac{1}{n} \sum_{j=1}^n \psi_k(X_j) \right)^2 &\leq \lim_{i\to\infty} \limsup_{n\to\infty} \frac{2}{n\varepsilon} \sum_{j=1}^n \sum_{k\geq i} \psi_k^2(X_j) \\ &= \lim_{i\to\infty} \frac{2}{\varepsilon} \int \sum_{k\geq i} \psi_k^2(x) p_{\theta_0}(x) \lambda(dx) \\ &= 0, \end{aligned}$$

where the first inequality follows by Jensen's inequality, the first equality follows from the law of large numbers since $\sum_{k\geq i} \psi_k^2(x) \leq \sum_{k\geq 1} \psi_k^2(x) = K(x,x)$ and the kernel is bounded by Assumption 5. The same argument, combined with Lebesgue's dominated convergence theorem, explains the last equality.

For the second term, the Lebesgue's dominated convergence theorem yields

$$\lim_{i\to\infty} \limsup_{n\to\infty} \frac{2}{\varepsilon} \int \sum_{k\geq i} \psi_k^2(x) p_{\hat{\theta}_n}(x) \lambda(dx) = \lim_{i\to\infty} \frac{2}{\varepsilon} \int \sum_{k\geq i} \psi_k^2(x) p_{\theta_0}(x) \lambda(dx) = 0.$$

Since conditions G_0-G_2 are satisfied, then Theorem 10 yields that

$$n \, \mathrm{MMD}_W^2(P_{\hat{\theta}_n,n}, Q_n) \xrightarrow{\mathcal{D}_D} \sum_{i=1}^\infty \lambda_i' Z_i'^2, \tag{18}$$

where $Z_1', \ldots, Z_n'$ are standard normal i.i.d. random variables, and $\lambda_1', \lambda_2', \ldots,$ are the eigenvalues of the operator $T_\nu : \mathcal{H} \to \mathcal{H}$ given by $(T_\nu \omega)(x) = \nu(\omega, K_x)$.

On the other hand, observe that the wild bootstrap for the goodness-of-fit problem satisfies that

$$n \, \mathrm{MMD}_W^2(P_{\theta_0}, Q_n) \xrightarrow{\mathcal{D}_D} \sum_{i=1}^\infty \lambda_i Z_i^2, \tag{19}$$

where $Z_1, \ldots, Z_n$ are standard normal i.i.d. random variables, and $\lambda_1, \lambda_2, \ldots,$ are the eigenvalues of the operator $T_\sigma : \mathcal{H} \to \mathcal{H}$ given by $(T_\sigma \omega)(x) = \sigma(\omega, K_x)$ where $\sigma = \frac{1}{2}\nu$ (hence the eigenvalues are half of those in Equation 18). Therefore, we deduce that for every $x \in \mathbb{R}$,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(n\,\mathrm{MMD}_W^2(P_{\hat{\theta}_n,n}, Q_n) \geq x) &= \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}_W^2(P_{\theta_0}, Q_n) \geq \frac{x}{2}\right) \\ &> \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}_W^2(P_{\theta_0}, Q_n) \geq x\right) \qquad (20)\end{aligned}$$

Replacing $x = q_\alpha$ in the previous equation yields

$$\begin{aligned}\alpha = \limsup_{n\to\infty} \mathbb{P}(n\,\mathrm{MMD}_W^2(P_{\hat{\theta}_n,n}, Q_n) \geq q_\alpha) &> \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}_W^2(P_{\theta_0}, Q_n) \geq q_\alpha\right) \\ &= \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) \geq q_\alpha\right), \qquad (21)\end{aligned}$$

where the first equality is due to the definition of q_α, the first inequality holds by Equation 20 and the last equality holds since $n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n)$ and $n\,\mathrm{MMD}_W^2(P_{\theta_0}, Q_n)$ have the same asymptotic null distribution (Chwialkowski et al., 2014, Theorem 1).

Finally note that

$$\begin{aligned}&\limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) \geq q_\alpha\right) \\ &\quad= \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) - \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \geq q_\alpha\right) \\ &\quad< \limsup_{n\to\infty} \mathbb{P}\left(n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) \geq q_\alpha\right) < \alpha,\end{aligned}$$

where the first equality holds due to Lemma 12, the first inequality holds since both $\mathbf{H}$ and $\mathbf{H}^{-1}$ are positive definite matrices, and thus $\sum_{i=1}^n \sum_{j=1}^n \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \geq 0$, and the last inequality holds by Equation 21. ■

Proof of Theorem 5 By the definition of the maximum mean discrepancy, we have

$$n\,\mathrm{MMD}^2(P_{\theta_n^\star}, Q_n^\star) = \sup_{\omega \in \mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (\gamma_n^\star(\omega))^2, \quad \text{where} \quad \gamma_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \omega(X_i^\star) - g(\omega, \theta_n^\star).$$

Proposition 23 yields $\gamma_n^\star(\omega) = S_n^\star(\omega) + o_p(1)$, where $S_n^\star : \mathcal{H} \to \mathbb{R}$ is the functional defined in Equation 8, and the error term does not depend on ω. Then, by using the fact that $\sup_{\omega \in \mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (S_n^\star(\omega))^2$ converges in distribution (which is proved next) we obtain

$$n\,\mathrm{MMD}^2(P_{\theta_n^\star}, Q_n^\star) = \sup_{\omega \in \mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (S_n^\star(\omega))^2 + o_p(1).$$

By Slutsky's theorem, we conclude that $n\,\mathrm{MMD}^2(P_{\theta_n^\star}, Q_n^\star)$ and $\sup_{\omega \in \mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (S_n^\star(\omega))^2$ have the same asymptotic null distribution.

We use Theorem 10 to obtain the asymptotic null distribution of $\sup_{\omega \in \mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (S_n^\star(\omega))^2$. Observe however that to apply this theorem, we need to check that

$$S_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta(\omega, X_i^\star, \hat{\theta}_n)$$

is a bounded linear functional. The linearity follows from the fact that for any fixed $x \in \mathcal{X}$ and $\omega \in \Theta$, $\eta(\omega, x, \theta)$ is a linear function of ω. Boundedness follows from item iii of Proposition 16 since $\sup_{\theta\in\Theta} \sup_{x\in\mathcal{X}} |\eta(\omega, x, \theta)| < C_1 \|\omega\|_{\mathcal{H}} + C_2 \|\omega\|_{\mathcal{H}}^2$ where $C_1 > 0$ and $C_2 > 0$ are constants that do not depend on ω.

We proceed to check Conditions G_0, G_1 and G_2 of Theorem 10. Condition G_0 follows directly from Proposition 24, where we identify $\sigma(\omega, \omega') = \mathbb{E}_{X\sim P_{\theta_0}}(\eta(\omega, X, \theta_0)\eta(\omega', X, \theta_0))$ for any $\omega, \omega' \in \mathcal{H}$. Indeed, note that it is the same bilinear form of Equation 13, and thus Condition G_1 was already proved in Equation 14 in the proof of Theorem 2.

Finally, to check condition G_2 observe that

$$\begin{aligned}\mathbb{P}_D\left(\sum_{k\geq i}(S_n^\star(\psi_k))^2 \geq \varepsilon\right) \leq \frac{1}{\varepsilon}\sum_{k\geq i}\mathbb{E}_D\left((S_n^\star(\psi_k))^2\right) &= \varepsilon^{-1}\sum_{k\geq i}\mathbb{E}_D(\eta^2(\psi_k, X^\star, \hat{\theta}_n)) \\ &= \varepsilon^{-1}\sum_{k\geq i}\int_{\mathcal{X}}\eta^2(\psi_k, z, \hat{\theta}_n)p_{\hat{\theta}_n}(z)\lambda(dz)\end{aligned} \tag{22}$$

where $X^\star \sim P_{\hat{\theta}_n}|\hat{\theta}_n$. The first inequality follows from Markov's inequality, and the first equality follows since $S_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta(\omega, X_i^\star, \hat{\theta}_n)$, and recall that $(X_i^\star)_{i=1}^n$ are generated i.i.d. from $P_{\hat{\theta}_n}|\hat{\theta}_n$.

From Proposition 16.ii., there exists constants $C_1 > 0$ and $C_2 > 0$ such that for all $k \geq 1$ it holds

$$\int_{\mathcal{X}}\eta^2(\psi_k, z, \hat{\theta}_n)p_{\hat{\theta}_n}(z)\lambda(dz) \leq C_1\int_{\mathcal{X}}\psi_k^2(z)p_{\hat{\theta}_n}(z)\lambda(dz) + C_2\|\nabla_\theta g(\psi_k, \theta_0)\|^2. \tag{23}$$

Finally, by combining Equations 22 and 23, we get

$$\begin{aligned}&\lim_{i\to\infty}\limsup_{n\to\infty}\mathbb{P}_D\left(\sum_{k\geq i}(S_n^\star(\psi_k))^2 \geq \varepsilon\right) \\ &\leq \lim_{i\to\infty}\limsup_{n\to\infty}\sum_{k\geq i}\varepsilon^{-1}C_1\int_{\mathcal{X}}\psi_k^2(z)p_{\hat{\theta}_n}(z)\lambda(dz) + \varepsilon^{-1}C_2\lim_{i\to\infty}\sum_{k\geq i}\|\nabla_\theta g(\psi_k, \theta_0)\|^2.\end{aligned}$$

Note that the second limit of the previous equation converges to zero by Proposition 25. For the first limit, observe

$$\lim_{i\to\infty}\limsup_{n\to\infty}\sum_{k\geq i}\int_{\mathcal{X}}\psi_k^2(z)p_{\hat{\theta}_n}(z)\lambda(dz) \leq \lim_{i\to\infty}\sum_{k\geq i}^{\infty}\int_{\mathcal{X}}\psi_k^2(z)\sup_{\theta\in\Theta}p_\theta(z)\lambda(dz) = 0$$

since $\sum_{k=1}^\infty \psi_k^2(y) = K(y, y)$, and $\int_{\mathcal{X}}\sup_{\theta\in\Theta}p_\theta(z)\lambda(dz) < \infty$ under Assumptions 1 to 6. ■

A.5 Proof of Auxiliary Results

We proceed to prove the auxiliary results.

Proof of Lemma 12 By the definition of the MMD, it holds that

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \gamma_n^2(\omega), \quad \text{where} \quad \gamma_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \omega(X_i) - g(\omega, \hat{\theta}_n).$$

Proposition 21 yields $\gamma_n(\omega) = S_n(\omega) + o_p(1)$, where $S_n : \mathcal{H} \to \mathbb{R}$ is the functional defined in Equation 8, and the error term does not depend on ω. Then, by using the fact that $\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega)$ converges in distribution- which is proved in Theorem 2- we obtain

$$n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^2(\omega) + o_p(1).$$

Note that $S_n(\omega) = A_n(\omega) + B_n(\omega)$, where $A_n : \mathcal{H} \to \mathbb{R}$ and $B_n : \mathcal{H} \to \mathbb{R}$ are given by

$$A_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\omega(X_i) - g(\omega, \theta_0)) \quad \text{and} \quad B_n(\omega) = -\frac{2}{\sqrt{n}}\sum_{i=1}^n \langle \nabla_\theta g(\omega, \theta_0), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0)\rangle.$$

Thus

$$\begin{aligned} n\,\mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) &= \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} (A_n(\omega) + B_n(\omega))^2 + o_p(1) \\ &= (A_n^1 + B_n^1)(A_n^2 + B_n^2)K + o_p(1) \\ &= n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) + A_n^1 B_n^2 K + B_n^1 A_n^2 K + B_n^1 B_n^2 K + o_p(1), \end{aligned} \tag{24}$$

where $A_n^i K$ with $i \in \{1, 2\}$ denotes the application of A_n to the i-the coordinate of the kernel, and the same notation is used for B_n^i. The second equality is due to Fernández and Rivera (2022), and the third equality holds since $n\,\mathrm{MMD}^2(P_{\theta_0}, Q_n) = A_n^1 A_n^2 K$.

We proceed to analyse $A_n^1 B_n^2 K$, $B_n^1 A_n^2 K$ and $B_n^1 B_n^2$. Observe that

$$B_n\omega = -\frac{2}{\sqrt{n}}\sum_{i=1}^n \left\langle \int_{\mathcal{X}} \omega(x)\nabla_\theta p_{\theta_0}(x)\lambda(dx), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0)\right\rangle$$

where the equality follows by Lebesgue's dominated convergence theorem (see Remark 9). For all $x \in \mathcal{X}$, define $K_x : \mathcal{X} \to \mathbb{R}$ by $K_x(\cdot) = K(x, \cdot)$. Then, $B_n^2 K \in \mathcal{H}$ is given by

$$(B_n^2 K)(x) = B_n K_x = -\frac{2}{\sqrt{n}}\sum_{i=1}^n \left\langle \int_{\mathcal{X}} K_x(y)\nabla_\theta p_{\theta_0}(y)\lambda(dy), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0)\right\rangle. \tag{25}$$

By linearity, we obtain

$$A_n^1 B_n^2 K = -\frac{2}{\sqrt{n}}\sum_{i=1}^n \left\langle \int_{\mathcal{X}} (A_n^1 K)(y)\nabla_\theta p_{\theta_0}(y)\lambda(dy), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0)\right\rangle \tag{26}$$

Note that for any fixed $y \in \mathcal{X}$, it holds

$$(A_n^1 K)(y) = \frac{1}{\sqrt{n}}\sum_{j=1}^n \left(K(X_j, y) - \int_{\mathcal{X}} K(x, y)p_{\theta_0}(x)\lambda(dx)\right).$$

Then, a simple computation shows that

$$\int_{\mathcal{X}} (A_n^1 K)(x) \nabla_\theta p_{\theta_0}(x) \lambda(dx) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left(\nabla_\theta \mu_{\theta_0}(X_j) - \int_{\mathcal{X}} (\nabla_\theta \mu_{\theta_0}(x)) p_{\theta_0}(x) \lambda(dx) \right)$$
$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \boldsymbol{\phi}(X_j, \theta_0).$$

By substituting the previous equation into Equation 26, we obtain

$$A_n^1 B_n^2 K = -\frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle \boldsymbol{\phi}(X_j, \theta_0), \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \right\rangle = -\frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \phi(X_i, \theta_0). \tag{27}$$

A similar computation shows that $B_n^1 A_n^2 K = A_n^1 B_n^2 K$. For the term $B_n^1 B_n^2 K$, we have

$$\begin{aligned}
&B_n^1 B_n^2 K \\
&= -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \left\langle \int_{\mathcal{X}} (B_n^1 K)(x) \nabla_\theta p_{\theta_0}(x) \lambda(dx), \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \right\rangle \\
&= \frac{4}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle \int_{\mathcal{X}} \left(\left\langle \int_{\mathcal{X}} K_x(y) \nabla_\theta p_{\theta_0}(y) \lambda(dy), \mathbf{H}^{-1} \boldsymbol{\phi}(X_j, \theta_0) \right\rangle \right) \nabla_\theta p_{\theta_0}(x) \lambda(dx), \right. \\
&\qquad \left. \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \right\rangle \\
&= \frac{4}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\mathcal{X}} \int_{\mathcal{X}} K(x, y) \left\langle \nabla_\theta p_{\theta_0}(y), \mathbf{H}^{-1} \boldsymbol{\phi}(X_j, \theta_0) \right\rangle \left\langle \nabla_\theta p_{\theta_0}(x), \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) \right\rangle \lambda(dy) \lambda(dx) \\
&= \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{\ell=1}^{p} \sum_{m=1}^{p} (\mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0))_\ell (\mathbf{H}^{-1} \boldsymbol{\phi}(X_j, \theta_0))_m \mathbf{H}_{m,\ell} \\
&= \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0),
\end{aligned} \tag{28}$$

where the second equality holds from replacing $(B_n^2 K)(x)$ by the expression given in Equation 25, and the fourth equality holds since

$$\mathbf{H}_{m,\ell} = 2 \int_{\mathcal{X}} \int_{\mathcal{X}} K(x, y) \left(\frac{\partial}{\partial \theta_m} p_{\theta_0}(y) \right) \left(\frac{\partial}{\partial \theta_\ell} p_{\theta_0}(x) \right) \lambda(dy) \lambda(dx).$$

By substituting Equation 27 and Equation 28 into Equation 24, we obtain

$$n \, \mathrm{MMD}^2(P_{\hat{\theta}_n}, Q_n) = n \, \mathrm{MMD}^2(P_{\theta_0}, Q_n) - \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\phi}(X_j, \theta_0)^\intercal \mathbf{H}^{-1} \boldsymbol{\phi}(X_i, \theta_0) + o_p(1).$$

■

Proof of Proposition 13 Note that conditioned on the data $\widetilde{D} = (X_i, \tilde{X}_i)_{i=1}^n$, $Z_n(\omega)$ is a sum of i.i.d. random variables with mean given by

$$\mathbb{E}_{\widetilde{D}}\left(Z_n(\omega)\right) = \mathbb{E}_{\widetilde{D}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(\omega(X_i) - \omega(\tilde{X}_i))\right) = 0,$$

and variance given by $\mathbb{V}\mathrm{ar}_{\widetilde{D}}(Z_n(\omega)) = \frac{1}{n}\sum_{i=1}^n(\omega(X_i) - \omega(\tilde{X}_i))^2$. We claim that

$$\frac{1}{n}\sum_{i=1}^n(\omega(X_i) - \omega(\tilde{X}_i))^2 \xrightarrow{\mathbb{P}} 2\int(\omega(x) - g(\omega, \theta_0))^2 p_{\theta_0}(x)\lambda(dx) = \nu(\omega, \omega).$$

Thus, by Lyapunov's Central limit theorem, we obtain $Z_n(\omega) \xrightarrow{\mathcal{D}_D} N(0, \nu(\omega, \omega))$ which yields the desired result.

To prove the claim, consider $\xi_n = \frac{1}{n}\sum_{i=1}^n(\omega(X_i) - \omega(\tilde{X}_i))^2$. We will show that $\mathbb{E}(\xi_n) = \nu(\omega, \omega)$ and that $\mathbb{V}\mathrm{ar}(\xi_n) \to 0$. Therefore the convergence in probability follows from Markov's inequality.

Consider the data $D = (X_i)_{i\geq 1}$, and define $\mathbb{E}_D(\cdot) = \mathbb{E}(\cdot|(X_i)_{i=1}^n)$. Then, observe that

$$\begin{aligned}\mathbb{E}\left(\xi_n\right) &= \mathbb{E}\left(\left(\omega(X_1) - \omega(\tilde{X}_1)\right)^2\right) = \mathbb{E}\left(\mathbb{E}_D\left(\left(\omega(X_1) - \omega(\tilde{X}_1)\right)^2\right)\right) \\ &= \mathbb{E}\left(\omega^2(X_1) - 2\omega(X_1)\mathbb{E}_D(\omega(\tilde{X}_1)) + \mathbb{E}_D(\omega^2(\tilde{X}_1))\right), \end{aligned} \tag{29}$$

where

$$\mathbb{E}_D\left(\omega(\tilde{X}_1)\right) = \int\omega(z)p_{\hat{\theta}_n}(z)\lambda(dz) \quad \text{and} \quad \mathbb{E}_D\left(\omega^2(\tilde{X}_1)\right) = \int\omega^2(z)p_{\hat{\theta}_n}(z)\lambda(dz).$$

Recall that $\int\sup_{\theta\in\Theta}p_\theta(x)\lambda(dx) < \infty$ by Assumption 3, and that $\omega \in \mathcal{H}$ is bounded by Assumption 5. Then, under the null hypothesis, and by applying Lebesgue's dominated convergence theorem twice, it holds

$$\lim_{n\to\infty}\mathbb{E}\left(\mathbb{E}_D\left(\omega(\tilde{X}_1)\right)\right) = \mathbb{E}\left(\lim_{n\to\infty}\int\omega(z)p_{\hat{\theta}_n}(z)\lambda(dz)\right) = \int\omega(z)p_{\theta_0}(z)\lambda(dz).$$

By using the same arguments we can show that $\lim_{n\to\infty}\mathbb{E}(\mathbb{E}_D(\omega^2(\tilde{X}_1))) = \int\omega^2(z)p_{\theta_0}(z)\lambda(dz)$.

The previous equations together with Equation 29, yield

$$\lim_{n\to\infty}\mathbb{E}\left(\xi_n\right) = 2\left(\int\omega^2(z)p_{\theta_0}(z) - \left(\int\omega(z)p_{\theta_0}(z)\right)^2\right) = \nu(\omega, \omega).$$

Finally, we proceed to verify that $\lim_{n\to\infty}\mathbb{V}\mathrm{ar}(\xi_n) = 0$. Note that

$$\mathbb{V}\mathrm{ar}\left(\xi_n\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}_D(\xi_n)\right) + \mathbb{E}(\mathbb{V}\mathrm{ar}_D\left(\xi_n\right)),$$

and by the previous arguments $\lim_{n\to\infty}\mathbb{V}\mathrm{ar}\left(\mathbb{E}_D(\xi_n)\right) = 0$ since $\mathbb{E}_D(\xi_n)$ converges to a constant. Note that conditioned on the data D, the random variables $(\omega(X_i) - \omega(\tilde{X}_i))_{i=1}^n$ are independent. Thus

$$\mathbb{V}\mathrm{ar}_D\left(\xi_n\right) = \frac{1}{n^2}\sum_{i=1}^n\mathbb{V}\mathrm{ar}_D\left(\left(\omega(X_i) - \omega(\tilde{X}_i)\right)^2\right) \to 0,$$

where the limit holds since the ω is bounded. Hence we conclude that $\lim_{n\to\infty} \mathbb{V}\mathrm{ar}(\xi_n) = 0$. ■

Proof of Proposition 14 To prove this result we use Theorem 11. Note that by Assumption 2 the set Θ is bounded and closed, by Assumption 3 we have $L(\theta) \in \mathcal{C}^0(\Theta)$, and by Assumption 5 the kernel K is characteristic, which means that the maximum mean discrepancy is a distance between probability measures. Thus, since the parametric family $\{P_\theta\}_{\theta\in\Theta}$ is identifiable (Assumption 3), we deduce that $L(\theta) = \mathrm{MMD}^2(P_\theta, P_{\theta_0})$ has a unique minimiser under the null hypothesis.

Observe that

$$\begin{aligned}
\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| &= \sup_{\theta\in\Theta} \left| \|\mu_\theta - \mu_n\|_{\mathcal{H}}^2 - \|\mu_\theta - \mu_{\theta_0}\|_{\mathcal{H}}^2 \right| \\
&\leq \sup_{\theta\in\Theta} \left| 2\left\langle \mu_\theta, \mu_{\theta_0} - \mu_n \right\rangle_{\mathcal{H}} \right| + \left| \|\mu_n\|_{\mathcal{H}}^2 - \|\mu_{\theta_0}\|_{\mathcal{H}}^2 \right| \\
&\leq 2 \sup_{\theta\in\Theta} \|\mu_\theta\|_{\mathcal{H}} \|\mu_{\theta_0} - \mu_n\|_{\mathcal{H}} + \left| \|\mu_n\|_{\mathcal{H}}^2 - \|\mu_{\theta_0}\|_{\mathcal{H}}^2 \right| \\
&\leq C \|\mu_{\theta_0} - \mu_n\|_{\mathcal{H}} + \left| \|\mu_n\|_{\mathcal{H}}^2 - \|\mu_{\theta_0}\|_{\mathcal{H}}^2 \right| \to 0
\end{aligned}$$

in probability, where the last result holds by the standard law of large numbers for RKHS Berlinet and Thomas-Agnan (2004, Section 9.1.1-9.1.2). Then, by Theorem 11, we deduce that $\hat{\theta}_n \to \theta_0$ in probability. ■

Proof of Proposition 15 Recall the definition $\boldsymbol{\phi}$ in Equation 6, then

$$\begin{aligned}
\nabla_\theta L_n(\theta) = \nabla_\theta \|\mu_n - \mu_\theta\|_{\mathcal{H}}^2 &= -2\left\langle \nabla_\theta \mu_\theta, \mu_n - \mu_\theta \right\rangle_{\mathcal{H}} \\
&= -\frac{2}{n} \sum_{i=1}^n \left(\nabla_\theta \mu_\theta(X_i) - \int_{\mathcal{X}} (\nabla_\theta \mu_\theta(x))\, p_\theta(x) \lambda(dx) \right) \\
&= -\frac{2}{n} \sum_{i=1}^n \boldsymbol{\phi}(X_i, \theta).
\end{aligned}$$

By doing the same computations, we obtain $\nabla_\theta L_n^\star(\theta) = \nabla_\theta \|\mu_n^\star - \mu_\theta\|_{\mathcal{H}}^2 = -\frac{2}{n}\sum_{i=1}^n \boldsymbol{\phi}(X_i^\star, \theta)$, where recall that $\mu_n^\star$ is the mean kernel embedding of the empirical distribution $Q_n^\star$ obtained from the parametric bootstrap samples $X_1^\star, \ldots, X_n^\star$.

We proceed to prove item i). Observe that clearly,

$$\mathbb{E}_{X\sim P_\theta}(\boldsymbol{\phi}(X,\theta)) = \mathbb{E}_{X\sim P_\theta}\left(\nabla_\theta \mu_\theta(X) - \int_{\mathcal{X}} (\nabla_\theta \mu_\theta(y)) p_\theta(y) \lambda(dy) \right) = 0,$$

and similarly $\mathbb{E}_D(\boldsymbol{\phi}(X_i^\star, \hat{\theta}_n)) = 0$. We continue by checking ii). For any fixed $x \in \mathcal{X}$ and $\theta \in \Theta$, we have

$$\|\boldsymbol{\phi}(x,\theta)\|^2 = \sum_{i=1}^p \left(\frac{\partial}{\partial\theta_i}\mu_\theta(x) - \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta_i}\mu_\theta(y) \right) p_\theta(y)\lambda(dy) \right)^2 \leq 4 \sum_{i=1}^p \left(\sup_{x\in\mathcal{X}} \left| \frac{\partial}{\partial\theta_i}\mu_\theta(x) \right| \right)^2. \tag{30}$$

Additionally, by similar arguments as those shown in Remark 9 we have that for any $i \in \{1, \ldots, p\}$, it holds

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\frac{\partial}{\partial\theta_i}\mu_\theta(x)\right| = \sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\int_{\mathcal{X}} K(x,y)\frac{\partial}{\partial\theta_i}p_\theta(y)\lambda(dy)\right| < \infty.$$

We finish by verifying item iii). Recall that $\nabla_\theta L_n(\theta_0) = -\frac{2}{n}\sum_{i=1}^n \boldsymbol{\phi}(X_i,\theta_0)$, where $\boldsymbol{\phi}(X_1,\theta_0),\ldots,\boldsymbol{\phi}(X_n,\theta_0)$ are i.i.d. random variables. Moreover, under the null hypothesis, the items i) and ii) of this proposition deduce $\mathbb{E}(\boldsymbol{\phi}(X_i,\theta_0)) = 0$ and $\mathbb{E}(\|\boldsymbol{\phi}(X_i,\theta_0)\|^2) < \infty$. Thus, $\mathbb{E}(\|\sqrt{n}\nabla_\theta L_n(\theta_0)\|^2) = 4\mathbb{E}(\|\boldsymbol{\phi}(X_1,\theta_0)\|^2) < \infty$, and we obtain $\|\sqrt{n}\nabla_\theta L_n(\theta_0)\| = O_p(1)$.

By using the same arguments we deduce that for almost all sequences $(X_i)_{i\geq 1}$,

$$\mathbb{E}_D(\|\sqrt{n}\nabla_\theta L_n^\star(\hat{\theta}_n)\|^2) = 4\mathbb{E}_D(\|\boldsymbol{\phi}(X_1^\star,\hat{\theta}_n)\|^2) < \infty.$$

Thus $\|\sqrt{n}\nabla_\theta L_n^\star(\hat{\theta}_n)\| = O_p(1)$. ■

Proof of Proposition 16 We start with i). Observe that

$$\mathbb{E}_{X\sim P_\theta}\left(\eta(\omega,X,\theta)\right) = \mathbb{E}_{X\sim P_\theta}\left(\omega(X) - g(\omega,\theta)\right) - 2\left\langle \nabla_\theta g(\omega,\theta_0), \mathbf{H}^{-1}\mathbb{E}_{X\sim P_\theta}\left(\boldsymbol{\phi}(X,\theta)\right)\right\rangle = 0,$$

where the first equality follows from the linearity of expectations, and the last equality is due to $g(\omega,\theta) = \mathbb{E}_{X\sim P_\theta}(\omega(X))$ and $\mathbb{E}_{X\sim P_\theta}(\boldsymbol{\phi}(X,\theta)) = 0$ from Proposition 15, item i).

We continue by verifying items ii) and iii). Observe that for any fixed $\omega \in \mathcal{H}$, $x \in \mathcal{X}$ and $\theta \in \Theta$, we have that

$$\begin{aligned}
|\eta(\omega,x,\theta)| &= \left|\omega(x) - g(\omega,\theta) - 2\left\langle \nabla_\theta g(\omega,\theta_0), \mathbf{H}^{-1}\boldsymbol{\phi}(x,\theta)\right\rangle\right| \\
&\leq |\omega(x) - g(\omega,\theta)| + 2\left\|\nabla_\theta g(\omega,\theta_0)\right\| \left\|\mathbf{H}^{-1}\right\|_F \left\|\boldsymbol{\phi}(x,\theta)\right\| \\
&\leq |\omega(x) - g(\omega,\theta)| + C'\left\|\nabla_\theta g(\omega,\theta_0)\right\| \qquad (31)
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $C' > 0$ is a constant that does not depend on ω, x, nor θ. In the previous set of equations, the first inequality follows from the triangle inequality and the Cauchy-Schwarz's inequality, and the second inequality holds by taking $C' = 2\|\mathbf{H}^{-1}\|_F \sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\|\boldsymbol{\phi}(x,\theta)\| < \infty$ since $\|\mathbf{H}^{-1}\|_F < \infty$, and $\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\|\boldsymbol{\phi}(x,\theta)\| < \infty$ by item ii) of Proposition 15.

We proceed to verify item ii). From Equation 31, we have that for any fixed $\omega \in \mathcal{H}$ and $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}_{X\sim P_\theta}\left(\eta^2(\omega,X,\theta)\right) &\leq 2\mathbb{E}_{X\sim P_\theta}\left(|\omega(X) - g(\omega,\theta)|^2\right) + 2C'^2\left\|\nabla_\theta g(\omega,\theta_0)\right\|^2 \\
&\leq C_1\mathbb{E}_{X\sim P_\theta}\left(\omega^2(X)\right) + C_2\left\|\nabla_\theta g(\omega,\theta_0)\right\|^2,
\end{aligned}$$

where $C_1 = 2$ and $C_2 = 2C'^2$. The first inequality holds from the fact that for any $a, b \in \mathbb{R}$ we have that $(a+b)^2 \leq 2(a^2+b^2)$, and the second inequality follows from the fact that the variance of random variable is upper bounded by its second moment.

We continue with item iii). Note that for each $\omega \in \mathcal{H}$, we have $\sup_{x\in\mathcal{X}}|\omega(x)| \leq \|\omega\|_{\mathcal{H}}\sqrt{C}$ (see Remark 8), and thus

$$\sup_{\theta\in\Theta}|g(\omega,\theta)| \leq \sup_{\theta\in\Theta}\int_{\mathcal{X}}|\omega(x)|p_\theta(x)\lambda(dx) \leq \|\omega\|_{\mathcal{H}}\sqrt{C}. \qquad (32)$$

Additionally, note that (see Remark 9)

$$\sup_{\theta\in\Theta}\|\nabla_\theta g(\omega,\theta)\|^2 = \sup_{\theta\in\Theta}\sum_{i=1}^{p}\left(\int_{\mathcal{X}}\omega(x)\frac{\partial}{\partial\theta_i}p_\theta(x)\lambda(dx)\right)^2$$
$$\leq \|\omega\|_{\mathcal{H}}^2\, C\sum_{i=1}^{p}\left(\int_{\mathcal{X}}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}p_\theta(y)\right|\lambda(dy)\right)^2 < \infty, \tag{33}$$

where the equality holds under Assumption 3 by the Lebesgue's dominated convergence theorem (see Remark 9), the first inequality is due to the fact that $\sup_{x\in\mathcal{X}}|\omega(x)| \leq \|\omega\|_{\mathcal{H}}\sqrt{C}$, and the last inequality holds by Assumption 3. Then, by combining Equations 31 to 33, we get

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}|\eta(\omega,x,\theta)| \leq 2\|\omega\|_{\mathcal{H}}\sqrt{C} + \|\omega\|_{\mathcal{H}}^2\, C\sum_{i=1}^{p}\left(\int_{\mathcal{X}}\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}p_\theta(y)\right|\lambda(dy)\right)^2$$
$$\leq C_1\|\omega\|_{\mathcal{H}} + C_2\|\omega\|_{\mathcal{H}}^2 < \infty,$$

where $C_1 > 0$ and $C_2 > 0$ are constants that do not depend on ω, x or θ.

■

Proof of Lemma 17 We begin by defining $Z_n(\theta) = |Y_n(\theta) - \gamma(\theta)|$ and $Z_n^\star(\theta) = |Y_n^\star(\theta) - \gamma(\theta)|$, where

$$Y_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X_i), \quad \gamma(\theta) = \mathbb{E}_{X\sim P_{\theta_0}}\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X)\right), \quad \text{and}$$
$$Y_n^\star(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X_i^\star).$$

Observe that Equation 11 is equivalent to $\sup_{\theta\in\Theta} Z_n(\theta) \xrightarrow{\mathbb{P}} 0$ as n grows to infinity, and Equation 12 is equivalent to $\mathbb{P}_D\left(\sup_{\theta\in\Theta} Z_n^\star(\theta) \geq \varepsilon\right) \xrightarrow{\mathbb{P}} 0$ holds for any $\varepsilon > 0$.

To verify Equation 11 we use Theorem 21.9 of Davidson (1994), from which we just need to check the following properties: (i) $Z_n(\theta) \xrightarrow{\mathbb{P}} 0$ for each fixed $\theta \in \Theta$, and (ii) $Z_n(\cdot)$ is stochastically equicontinuous. Analogously, for Equation 12 we need to check (i') $\mathbb{P}_D\left(Z_n^\star(\theta) \geq \varepsilon\right) \xrightarrow{\mathbb{P}} 0$, and (ii') for almost all sequences $(X_i)_{i\geq 1}$, $Z_n^\star(\cdot)$ is stochastically equicontinuous.

We start proving (i) and (i'). Note that (i) follows from the law of large numbers since

$$\mathbb{E}_{X\sim P_{\theta_0}}\left(\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X)\right|\right) = \mathbb{E}_{X\sim P_{\theta_0}}\left(\left|\int_{\mathcal{X}}K(X,y)\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}p_\theta(y)\lambda(dy)\right|\right) < \infty,$$

holds under Assumption 3 (see Remark 9). For item (i'), note that for any $\varepsilon > 0$ we have

$$\mathbb{P}_D\left(Z_n^\star(\theta) \geq \varepsilon\right) = \mathbb{P}_D\left(|\gamma(\theta) - Y_n^\star(\theta)| \geq \varepsilon\right)$$
$$\leq \mathbb{P}_D\left(|\gamma_n(\theta) - Y_n^\star(\theta)| \geq \varepsilon/2\right) + \mathbb{P}_D\left(|\gamma(\theta) - \gamma_n(\theta)| \geq \varepsilon/2\right),$$

where $\gamma_n(\theta) = \mathbb{E}_{X\sim P_{\hat{\theta}_n}}\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X)\right)$. The last two terms converge to zero. To see this observe that

$$\mathbb{P}_D\left(|\gamma_n(\theta) - Y_n^\star(\theta)| \geq \frac{\varepsilon}{2}\right) \leq 4\frac{\mathbb{E}_D(|\gamma_n(\theta) - Y_n^\star(\theta)|^2)}{\varepsilon^2} \leq \frac{4}{\varepsilon^2 n}\mathbb{E}_D\left(\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X_1^\star)\right)^2\right) \to 0,$$

where the first inequality is due to Chebyshev's inequality, and the second inequality holds since $X_1^\star, \ldots, X_n^\star \overset{i.i.d.}{\sim} P_{\hat{\theta}_n}|\hat{\theta}_n$. For the second term

$$\begin{aligned}|\gamma_n(\theta) - \gamma(\theta)| &= \left|\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(x)(p_{\hat{\theta}_n}(x) - p_{\theta_0}(x))\lambda(dx)\right| \\ &\leq \sup_{x\in\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(x)\right| \int_{\mathcal{X}} |(p_{\hat{\theta}_n}(x) - p_{\theta_0}(x))|\lambda(dx)\end{aligned}$$

now, by Taylor's theorem, for each $x \in \mathcal{X}$, it exists θ_x in the line between θ_0 and $\hat{\theta}_n$ such that $p_{\hat{\theta}_n}(x) - p_{\theta_0}(x) = \nabla_\theta p_{\theta_x}(x)\cdot(\hat{\theta}_n - \theta_0)$. By the previous equality and by Holder's inequality

$$\begin{aligned}|\gamma_n(\theta) - \gamma(\theta)| &\leq \sup_{x\in\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(x)\right| \left(\int_{\mathcal{X}} \|\nabla_\theta p_{\theta_x}(x)\|_1 \lambda(dx)\right) \|\hat{\theta}_n - \theta_0\|_\infty \\ &\leq \sup_{x\in\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(x)\right| \left(\int_{\mathcal{X}} \sup_{\theta\in\Theta}\|\nabla_\theta p_\theta(x)\|_1 \lambda(dx)\right) \|\hat{\theta}_n - \theta_0\|_\infty \to 0,\end{aligned}$$

since $|\hat{\theta}_n - \theta_0| = o_p(1)$ by Proposition 14, and since the other terms are finite due to Assumptions 3 and 5 and Remark 9.

We continue by checking (ii) and (ii'). To check ii) we use Theorem 21.10 of Davidson (1994) from which it is enough to verify that $|Z_n(\theta) - Z_n(\theta')| \leq B\|\theta - \theta'\|_\infty$ holds for all $\theta, \theta' \in \Theta$ and for all n, where B is a constant which does not depend of n, θ and θ'. To check condition ii') we will verify that there exists $B^\star > 0$ such that $|Z_n^\star(\theta) - Z_n^\star(\theta')| \leq B^\star\|\theta - \theta'\|_\infty$, holds for every sequence $(X_i)_{i\geq 1}$. We only verify that $|Z_n(\theta) - Z_n(\theta')| \leq B\|\theta - \theta'\|_\infty$ for some constant B since the analogous result for parametric bootstrap follows from the same arguments.

By the triangle inequality, we obtain

$$|Z_n(\theta) - Z_n(\theta')| \leq \left|\gamma(\theta) - \gamma(\theta')\right| + \left|Y_n(\theta) - Y_n(\theta')\right|.$$

Since the kernel is bounded by a constant $C > 0$, it holds

$$\begin{aligned}\left|\gamma(\theta) - \gamma(\theta')\right| &= \left|\mathbb{E}_{X\sim P_{\theta_0}}\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_\theta(X) - \frac{\partial^2}{\partial\theta_\ell\partial\theta_j}\mu_{\theta'}(X)\right)\right| \\ &= \left|\mathbb{E}_{X\sim P_{\theta_0}}\left(\int_{\mathcal{X}} K(X,y)\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\lambda(dy)\right)\right| \\ &\leq C\int_{\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\right|\lambda(dy), \qquad (34)\end{aligned}$$

and

$$\left|Y_n(\theta) - Y_n(\theta')\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}} K(X_i, y)\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\lambda(dy)\right|$$
$$\leq C\int_{\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\right|\lambda(dy), \tag{35}$$

By combining Equation 34 and Equation 35, we get

$$|Z_n(\theta) - Z_n(\theta')| \leq 2C\int_{\mathcal{X}}\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\right|\lambda(dy),$$

By the mean value theorem, we have $\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y)) = \nabla_\theta\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}p_{\theta_y}(y)\right)\cdot(\theta - \theta')$, where θ_y is in the line between θ and θ' (which belongs to Θ by convexity of this set). By Holder's inequality, we obtain

$$\left|\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}(p_\theta(y) - p_{\theta'}(y))\right| \leq \left\|\nabla_\theta\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}p_{\theta_y}(y)\right)\right\|_1 \|\theta - \theta'\|_\infty$$
$$\leq \left(\sup_{\theta\in\Theta}\left\|\nabla_\theta\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}p_\theta(y)\right)\right\|_1\right)\|\theta - \theta'\|_\infty.$$

Then, we conclude that

$$|Z_n(\theta) - Z_n(\theta')| \leq \underbrace{2C\int_{\mathcal{X}}\sup_{\theta\in\Theta}\left\|\nabla_\theta\left(\frac{\partial^2}{\partial\theta_\ell\partial\theta_j}p_\theta(y)\right)\right\|_1\lambda(dy)}_{B_{\ell,j}}\|\theta - \theta'\|_\infty.$$

Finally, choose $B = \max_{\ell,j\in[p]} B_{\ell,j}$ and note it is finite by Assumption 3 together with Remark 9. ■

Proof of Proposition 18 We start by proving $\|\mathbf{H}_n - \mathbf{H}\| = o_p(1)$. To prove this result, it suffices to show that each component of $(\mathbf{H}_n)_{ij}$ converges in probability to $\mathbf{H}_{ij}$. Observe that

$$(\mathbf{H}_n)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}L_n(\tilde{\theta}^j) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}(L_n(\tilde{\theta}^j) - L(\tilde{\theta}^j)) + \frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\tilde{\theta}^j).$$

A simple computation shows that for each $\theta\in\Theta$ it holds

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}(L_n(\theta) - L(\theta)) = 2\left(\mathbb{E}_{X\sim P_{\theta_0}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mu_\theta(X)\right) - \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mu_\theta(X_i)\right) = o_p(1).$$

where the second equality holds by Lemma 17. Thus,

$$(\mathbf{H}_n)_{ij} = o_p(1) + \frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\tilde{\theta}^j).$$

Finally, note that $L(\theta) = \|\mu_\theta - \mu_{\theta_0}\|^2_{\mathcal{H}} \in \mathcal{C}^0(\Theta)$ since $\mu_\theta(x) \in \mathcal{C}^3(\Theta)$ for all $x \in \mathcal{X}$ (see Assumption 3 and Remark 9). Then, the continuous mapping theorem, together with the fact that $\tilde{\theta}^j_n \xrightarrow{\mathbb{P}} \theta_0$ (by Proposition 14), yields

$$(\mathbf{H}_n)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L(\theta_0) + o_p(1) = \mathbf{H}_{ij} + o_p(1).$$

We continue with the result for $\mathbf{H}^\star_n$. Similarly to what we did before, we sum and subtract the term $\frac{\partial^2}{\partial\theta_i\partial\theta_j} L(\tilde{\theta}^j_\star)$ to $(\mathbf{H}_n)^\star_{ij}$ and obtain

$$(\mathbf{H}_n)^\star_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(L^\star_n(\tilde{\theta}^j_\star) - L(\tilde{\theta}^j_\star)\right) + \frac{\partial^2}{\partial\theta_i\partial\theta_j} L(\tilde{\theta}^j_\star).$$

Observe that for any $\theta \in \Theta$, under the null hypothesis, we have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(L^\star_n(\theta) - L(\theta)\right) = 2\left(\mathbb{E}_{X\sim P_{\theta_0}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mu_\theta(X)\right) - \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta_i\partial\theta_j}\mu_\theta(X^\star_i)\right).$$

Then, the result follows from Lemma 17 and from the fact that $\hat{\theta}^\star_n - \theta_0 = o_p(1)$ and the same arguments as before. ■

Proof of Proposition 19 We start by proving item i). By Assumptions 1 to 6, we have that $L_n(\theta) = \|\mu_\theta - \mu_n\|^2_{\mathcal{H}} \in \mathcal{C}^3(\Theta)$ since $\mu_\theta(x) \in \mathcal{C}^3(\Theta)$ for all $x \in \mathcal{X}$ (see Remark 9). Then, for each $j \in \{1, \ldots, p\}$, a first order Taylor's expansion of $\frac{\partial}{\partial\theta_j} L_n(\theta)$ around $\hat{\theta}_n$ exists and is given by

$$\frac{\partial}{\partial\theta_j} L_n(\theta) = \frac{\partial}{\partial\theta_j} L_n(\hat{\theta}_n) + \left\langle \nabla_\theta\left(\frac{\partial}{\partial\theta_j} L_n(\tilde{\theta}^j)\right), \theta - \hat{\theta}_n\right\rangle = \left\langle \nabla_\theta\left(\frac{\partial}{\partial\theta_j} L_n(\tilde{\theta}^j)\right), \theta - \hat{\theta}_n\right\rangle,$$

where $\tilde{\theta}^j$ lies in the line segment between $\hat{\theta}_n$ and θ, and $\tilde{\theta}^j \in \Theta$ by convexity. The second equality holds since $\hat{\theta}_n$ is the minimiser of L_n, and belongs to the interior of Θ. By using the previous equation and by evaluating at θ_0, we obtain $\nabla_\theta L_n(\theta_0) = \mathbf{H}^\intercal_n(\theta_0 - \hat{\theta}_n)$, where $(\mathbf{H}_n)_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j} L_n(\tilde{\theta}^j)$ for each $i, j \in [p]$.

By Proposition 18 we have that $\mathbf{H}_n$ converges in probability to the Hessian matrix $\mathbf{H}$. Recall that $\mathbf{H}$ is positive definite by Assumption 6. Then, by the continuity of the determinant, we have that $\mathbf{H}_n$ is invertible in sets of arbitrarily large probability for all n large enough, and thus

$$\begin{aligned}\sqrt{n}(\theta_0 - \hat{\theta}_n) &= \sqrt{n}(\mathbf{H}^\intercal_n)^{-1}\nabla_\theta L_n(\theta_0) \\ &= \sqrt{n}(\mathbf{H}^\intercal)^{-1}\nabla_\theta L_n(\theta_0) + \sqrt{n}((\mathbf{H}^\intercal_n)^{-1} - \mathbf{H}^{-1})\nabla_\theta L_n(\theta_0).\end{aligned}$$

By Proposition 15, $\|\sqrt{n}\nabla_\theta L_n(\theta_0)\| = O_p(1)$, and since $\mathbf{H}_n \xrightarrow{\mathbb{P}} \mathbf{H}$, we have

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) = \sqrt{n}(\mathbf{H}^\intercal)^{-1}\nabla_\theta L_n(\theta_0) + o_p(1).$$

For item ii) we follow a very similar argument. Note that by Assumptions 1 to 6, $L_n^\star(\theta) \in \mathcal{C}^3(\Theta)$. Then, for each $j \in \{1, \dots, p\}$, we perform a first order Taylor's approximation of $\frac{\partial}{\partial \theta_j} L_n^\star(\theta)$ around $\theta_n^\star$ and we get

$$\frac{\partial}{\partial \theta_j} L_n^\star(\hat{\theta}_n) = \left\langle \nabla_\theta \left(\frac{\partial}{\partial \theta_j} L_n^\star(\tilde{\theta}_\star^j) \right), \hat{\theta}_n - \theta_n^\star \right\rangle,$$

where $\tilde{\theta}_\star^j$ lies in the line segment between $\theta_n^\star$ and $\hat{\theta}_n$.

By using the previous equation, we obtain $\nabla_\theta L_n^\star(\hat{\theta}_n) = \mathbf{H}_n^{\star\intercal}(\hat{\theta}_n - \theta_n^\star)$, where $(\mathbf{H}_n^\star)_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n^\star(\tilde{\theta}_\star^j)$ for each $i, j \in [p]$, and all $\tilde{\theta}_\star^1, \dots, \tilde{\theta}_\star^p$ lie in the line segment between $\hat{\theta}_n$ and $\theta_n^\star$. By Proposition 18 we have that $\mathbf{H}_n^\star$ converges in probability to the Hessian matrix $\mathbf{H}$. Then, by the continuity of the determinant, we have that $\mathbf{H}_n^\star$ is invertible in sets of arbitrarily large probability for all n large enough, and thus $\sqrt{n}(\hat{\theta}_n - \theta_n^\star) = \sqrt{n}(\mathbf{H}_n^{\star\intercal})^{-1} \nabla_\theta L_n^\star(\hat{\theta}_n)$. We finish by noting that under Assumptions 1 to 6, Proposition 15 deduces that $\|\sqrt{n}\nabla_\theta L_n^\star(\hat{\theta}_n)\| = O_p(1)$ and by Proposition 18, $(\mathbf{H}_n^{\star\intercal})^{-1}$ converges to $\mathbf{H}^{-1}$.

■

Proof of Lemma 20 We start by proving i). Proposition 19 yields $\sqrt{n}(\hat{\theta}_n - \theta_n^\star) = \sqrt{n}\mathbf{H}^{-1}\nabla_\theta L_n^\star(\hat{\theta}_n) + o_p(1)$, and Proposition 15 yields $\|\sqrt{n}\nabla_\theta L_n^\star(\hat{\theta}_n)\| = O_p(1)$, which together give us that $\sqrt{n}(\hat{\theta}_n - \theta_n^\star) = O_p(1)$.

We continue by proving ii). Observe that

$$\mathbb{P}(|\theta_n^\star - \theta_0| \geq \epsilon) \leq \mathbb{P}\left(|\theta_n^\star - \hat{\theta}_n| \geq \epsilon/2\right) + \mathbb{P}\left(|\hat{\theta}_n - \theta_0| \geq \epsilon/2\right) \to 0,$$

where the first probability tends to 0 by part i), and the second one by Proposition 14 (that can only be applied under the null hypothesis). ■

Proof of Proposition 21 Define $\gamma_n : \mathcal{H} \to \mathbb{R}$ by

$$\gamma_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\omega(X_i) - g(\omega, \hat{\theta}_n) \right). \tag{36}$$

By Assumptions 1 to 6, for each fixed $\omega \in \mathcal{H}$, we have $g(\omega, \theta) \in \mathcal{C}^3(\Theta)$ (see Remark 9). Then, a first order Taylor's expansion of $g(\omega, \theta)$ around θ_0 yields $g(\omega, \hat{\theta}_n) = g(\omega, \theta_0) + \langle \nabla_\theta g(\omega, \tilde{\theta}_n), \hat{\theta}_n - \theta_0 \rangle$, where $\tilde{\theta}_n$ lies on the line segment between θ_0 and $\hat{\theta}_n$. Note that by the convexity of Θ, it holds that $\tilde{\theta}_n \in \Theta$. Then, by replacing the previous expression in Equation 36 we obtain

$$\gamma_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\omega(X_i) - g(\omega, \theta_0) \right) - \sqrt{n} \left\langle \nabla_\theta g(\omega, \tilde{\theta}_n), \hat{\theta}_n - \theta_0 \right\rangle.$$

By assuming the null hypothesis, Proposition 14 yields $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, and Proposition 15.iii and Proposition 19.i yield $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$. Then, by the continuous mapping theorem we obtain

$$\gamma_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\omega(X_i) - g(\omega, \theta_0) \right) - \sqrt{n} \left\langle \nabla_\theta g(\omega, \theta_0), \hat{\theta}_n - \theta_0 \right\rangle + o_p(1). \tag{37}$$

By combining Proposition 15 and Proposition 19.i, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}\mathbf{H}^{-1}\nabla_\theta L_n(\theta_0) + o_p(1) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0) + o_p(1).$$

By replacing the previous expression in Equation 37 we obtain

$$\begin{aligned}\gamma_n(\omega) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\omega(X_i) - g(\omega, \theta_0) - 2\left\langle \nabla_\theta g(\omega, \theta_0), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i, \theta_0)\right\rangle\right) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta(\omega, X_i, \theta_0) + o_p(1) = S_n(\omega) + o_p(1).\end{aligned}$$

■

Proof of Proposition 22 First note that

$$\mathbf{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\eta(\omega_1, X_i, \theta_0), \ldots, \eta(\omega_m, X_i, \theta_0)),$$

where $(\eta(\omega_1, X_i, \theta_0), \ldots, \eta(\omega_m, X_i, \theta_0))_{i=1}^n$ are a collection of i.i.d. random vectors. By Proposition 16.i, we have that for any fixed $\omega \in \mathcal{H}$ it holds $\mathbb{E}_{X\sim P_{\theta_0}}(\eta(\omega, X, \theta_0)) = 0$, and thus $\mathbb{E}_{X\sim P_{\theta_0}}(\mathbf{S}_n) = \mathbf{0}$. Additionally, by Proposition 16.ii, for any $\omega \in \mathcal{H}$ we have

$$\mathbb{E}_{X\sim P_{\theta_0}}(\eta^2(\omega, X, \theta_0)) < \infty.$$

Thus, the Central Limit Theorem yields $\mathbf{S}_n \xrightarrow{\mathcal{D}} N_m(0, \boldsymbol{\Sigma})$, where for any $i, j \in [m]$ we have

$$\boldsymbol{\Sigma}_{ij} = \sigma(\omega_i, \omega_j) = \mathbb{E}_{X\sim P_{\theta_0}}(\eta(\omega_i, X, \theta_0)\eta(\omega_j, X, \theta_0)).$$

■

Proof of Proposition 23 Define $\gamma^\star : \mathcal{H} \to \mathbb{R}$ by

$$\gamma_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\omega(X_i^\star) - g(\omega, \theta_n^\star)\right). \tag{38}$$

By Assumptions 1 to 6, for each $\omega \in \mathcal{H}$, $g(\omega, \theta_n^\star) \in \mathcal{C}^3(\Theta)$ (see Remark 9). By a first order Taylor's expansion of $g(\omega, \theta)$ around $\hat{\theta}_n$ we obtain $g(\omega, \theta_n^\star) = g(\omega, \hat{\theta}_n) + \langle \nabla_\theta g(\omega, \tilde{\theta}_n^\star), \theta_n^\star - \hat{\theta}_n\rangle$ where the $\tilde{\theta}_n^\star$ belongs to the line segment between $\hat{\theta}_n$ and $\theta_n^\star$. Then, it follows that

$$\gamma_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\omega(X_i^\star) - g(\omega, \hat{\theta}_n) - \left\langle \nabla_\theta g(\omega, \tilde{\theta}_n^\star), \theta_n^\star - \hat{\theta}_n\right\rangle\right).$$

Proposition 14 yields that $\hat{\theta}_n - \theta_0 = o_p(1)$, and Lemma 20 yields $\theta_n^\star - \theta_0 = o_p(1)$ and $\sqrt{n}(\theta_n^\star - \hat{\theta}_n) = O_p(1)$. Thus, by the continuous mapping theorem it holds

$$\gamma_n^\star(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\omega(X_i^\star) - g(\omega, \hat{\theta}_n) - \left\langle \nabla_\theta g(\omega, \theta_0), \theta_n^\star - \hat{\theta}_n\right\rangle\right) + o_p(1). \tag{39}$$

Finally, by combining Propositions 15 and 19.ii we get

$$\sqrt{n}(\theta_n^\star - \hat{\theta}_n) = -\sqrt{n}\mathbf{H}^{-1}\nabla_\theta L_n^\star(\hat{\theta}_n) + o_p(1) = \frac{2}{\sqrt{n}}\sum_{i=1}^n \mathbf{H}^{-1}\boldsymbol{\phi}(X_i^\star, \hat{\theta}_n) + o_p(1),$$

and thus by replacing in Equation 39,

$$\begin{aligned}\gamma_n^\star(\omega) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\omega(X_i^\star) - g(\omega, \hat{\theta}_n) - 2\left\langle \nabla_\theta g(\omega, \theta_0), \mathbf{H}^{-1}\boldsymbol{\phi}(X_i^\star, \hat{\theta}_n)\right\rangle\right) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \eta(\omega, X_i^\star, \hat{\theta}_n) + o_p(1).\end{aligned}$$

■

Proof of Proposition 24 We will apply the Lidenberg-Feller Central Limit Theorem for multivariate triangular arrays. Define $Z_n = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_{n,i}$, where $\{\mathbf{Y}_{n,i}\}$ is a triangular array of random vectors $\mathbf{Y}_{n,i} = (\eta(\omega_1, X_{n,i}^\star, \hat{\theta}_n), \ldots, \eta(\omega_m, X_{n,i}^\star, \hat{\theta}_n))$, $\omega_1, \ldots, \omega_m \in \mathcal{H}$ and $m \in \mathbb{N}$. Observe that by Proposition 16, we have that

$$\mathbb{E}_D(\eta(\omega, X_{n,i}^\star, \hat{\theta}_n)) = \int_{\mathcal{X}} \eta(\omega, z, \hat{\theta}_n)) p_{\hat{\theta}_n}(z)\lambda(dz) = 0$$

holds for any $\omega \in \mathcal{H}$, and thus $\mathbb{E}_D(\mathbf{Y}_{n,i}) = 0$. Consider $\mathbf{V}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{V}\mathrm{ar}_D(\mathbf{Y}_{n,i})$, and observe that for any $\ell, k \in [m]$, we have that

$$\begin{aligned}(\mathbf{V}_n)_{\ell,k} = \mathbb{C}\mathrm{ov}_D(\eta(\omega_\ell, X_{n,1}^\star, \hat{\theta}_n), \eta(\omega_k, X_{n,1}^\star, \hat{\theta}_n)) &= \int_{\mathcal{X}} \eta(\omega_\ell, z, \hat{\theta}_n)\eta(\omega_k, z, \hat{\theta}_n)) p_{\hat{\theta}_n}(z)\lambda(dz) \\ &\xrightarrow{\mathbb{P}} \int_{\mathcal{X}} \eta(\omega_\ell, z, \theta_0)\eta(\omega_k, z, \theta_0)) p_{\theta_0}(z)\lambda(dz).\end{aligned} \tag{40}$$

The convergence in probability holds since $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$ by Proposition 14, and since the map $\theta \to \int_{\mathcal{X}} \eta(\omega, z, \theta)\eta(\omega', z, \theta) p_\theta(z)\lambda(dz)$ is continuous for any fixed $\omega, \omega' \in \mathcal{H}$. Thus the continuous mapping theorem yields Equation 40.

We proceed to check Linderberg's condition. Observe that for every $\varepsilon > 0$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_D\left(\|\mathbf{Y}_{n,i}\|^2 \mathbb{1}_{\{\|\mathbf{Y}_{n,i}\| \geq \varepsilon\sqrt{n}\}}\right) &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \mathbb{E}_D\left(\eta^2(\omega_j, X_{n,i}, \hat{\theta}_n)\mathbb{1}_{\{\|\mathbf{Y}_{n,i}\| \geq \varepsilon\sqrt{n}\}}\right) \\ &\leq C_1 \frac{1}{n}\sum_{i=1}^n \mathbb{P}_D\left(\|\mathbf{Y}_{n,i}\| \geq \varepsilon\sqrt{n}\right) \\ &\leq C_1 \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}_D(\|\mathbf{Y}_{n,i}\|)}{\epsilon\sqrt{n}} \to 0, \quad a.s.\end{aligned}$$

The first inequality holds since $\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}} \eta^2(\omega_j, x, \theta) < \infty$ by Proposition 16.iii.

■

Proof of Proposition 25 By Remark 9

$$\sum_{i=1}^{\infty} \|\nabla_\theta g(\psi_i, \theta_0)\|^2 = \sum_{i=1}^{\infty} \sum_{\ell=1}^{p} \left(\int \psi_i(y) \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \lambda(dy) \right)^2. \tag{41}$$

Observe that for each $\ell \in \{1, \ldots, p\}$, it holds that

$$\begin{aligned}
&\sum_{i=1}^{\infty} \left(\int_{\mathcal{X}} \psi_i(y) \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \lambda(dy) \right)^2 \\
&\quad = \sum_{i=1}^{\infty} \int_{\mathcal{X}} \int_{\mathcal{X}} \psi_i(y)\psi_i(x) \left(\frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \right) \left(\frac{\partial}{\partial \theta_\ell} p_{\theta_0}(x) \right) \lambda(dy)\lambda(dx) \\
&\quad \leq \int_{\mathcal{X}} \int_{\mathcal{X}} \sum_{i=1}^{\infty} |\psi_i(y)||\psi_i(x)| \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \right| \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(x) \right| \lambda(dy)\lambda(dx) \\
&\quad \leq \int_{\mathcal{X}} \int_{\mathcal{X}} \sqrt{\sum_{i=1}^{\infty} |\psi_i(y)|^2} \sqrt{\sum_{j=1}^{\infty} |\psi_j(x)|^2} \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \right| \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(x) \right| \lambda(dy)\lambda(dx) \\
&\quad = \left(\int_{\mathcal{X}} \sqrt{\sum_{i=1}^{\infty} |\psi_i(y)|^2} \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \right| \lambda(dy) \right)^2 \\
&\quad = \left(\int_{\mathcal{X}} \sqrt{K(y,y)} \left| \frac{\partial}{\partial \theta_\ell} p_{\theta_0}(y) \right| \lambda(dy) \right)^2 \leq \left(\sqrt{C} \int_{\mathcal{X}} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_\ell} p_\theta(y) \right| \lambda(dy) \right)^2 < \infty
\end{aligned} \tag{42}$$

where the second inequality is due to the Cauchy-Schwarz inequality, the second equality holds since $\sum_{i=1}^{\infty} \psi_i^2(y) = K(y,y)$, the third inequality follows from the fact that the kernel is bounded (Assumption 5), and the last inequality is due to Assumption 3. By combining Equations 41 and 42 we conclude that $\sum_{i=1}^{\infty} \|\nabla_\theta g(\psi_i, \theta_0)\|^2 < \infty$.

■

Appendix B. Illustration of Limitations

We illustrate two of the limitations introduced in Section 6. The first occurs when $\mathrm{D}(P, Q_n)$ is poor at distinguishing between probability distributions. For example, this could occur when using a Gaussian kernel with a very small or large lengthscale. Figure 12 shows the type I error rate of the composite KSD test, using the parametric bootstrap, as the lengthscale of the Gaussian kernel varies. We consider the null hypothesis case where H_0^C = the data is Gaussian and the data is sampled from $Q = \mathcal{N}(0.4, 1.4^2)$. The figure shows that the type I error rate of the composite test can exceed the level for very small lengthscales, and goes to zero for large lengthscales. The figure also shows the type I error rate of a non-composite test with $H_0 : P = Q$ as defined above, which reveals that the non-composite test does not suffer

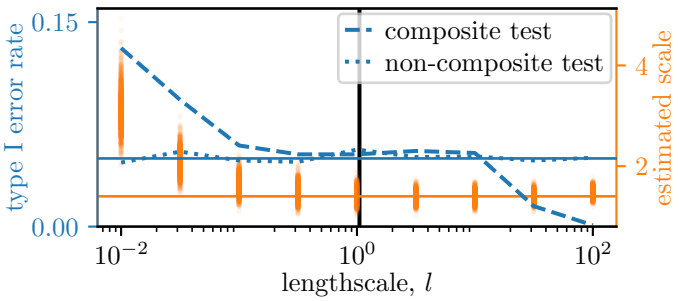


Figure 12: Comparison of the type I error rate of a composite and non-composite test, as we vary the lengthscale of the kernel. The left axis, blue, shows the type I error rate of the two tests. The right axis, orange, shows the estimates of the scale parameter of the model made by the composite test over 2000 repeats. The solid blue horizontal line shows the level of the test, 0.05. The solid orange horizontal line shows the true value of the standard deviation, 1.4. The solid black vertical line shows the lengthscale selected by the median heuristic.

from this problem. To investigate the issue the figure also shows the estimates of the scale parameter. For very small lengthscales, the parameter estimates are not centered around the true parameter value, thus the test is comparing the data to a poor choice of model from $\{P_\theta\}_\theta$, leading to type I errors. For large lengthscales, we suggest that the type I error rate of the composite test goes to zero faster than that of the non-composite test because the small amount of variance in the estimated parameters makes distinguishing the null and alternative hypotheses more difficult. Importantly, we note that the range of lengthscales for which the composite test has good performance is fairly wide (note the log x-axis scale), and the median heuristic selects a value in the middle of this range. The same behaviour can also occur for the MMD test, although, in this case, we have the additional reassurance of theoretical results which show that the test will behave correctly as n becomes large.

The second limitation we will illustrate occurs when there is large variance in the estimate of the parameter, if either the minimum distance estimator has high variance or the optimisation process fails. Figure 13 demonstrates this limitation for the MMD test, testing H_0^C that $Q = \mathcal{N}(\theta, 1^2)$ for $\theta \in \mathbb{R}$ against the alternative $\mathcal{N}(0.5, 1.5^2)$. If the optimiser is configured correctly, the figure shows that the estimates of the parameter are centered around the true parameter value and have low variance, and there is clear separation between the distributions of the test statistic under H_0^C and H_1^C. Thus, the test is able to distinguish H_0^C and H_1^C and achieve a high power. If the optimiser is configured poorly—we use too large a learning rate—we can see that the parameter estimates have higher variance, and the distributions of the test statistic under H_0^C and H_1^C overlap completely. Thus, no matter where the threshold is set it will not be possible to distinguish the two hypotheses, and the test will have very low power.

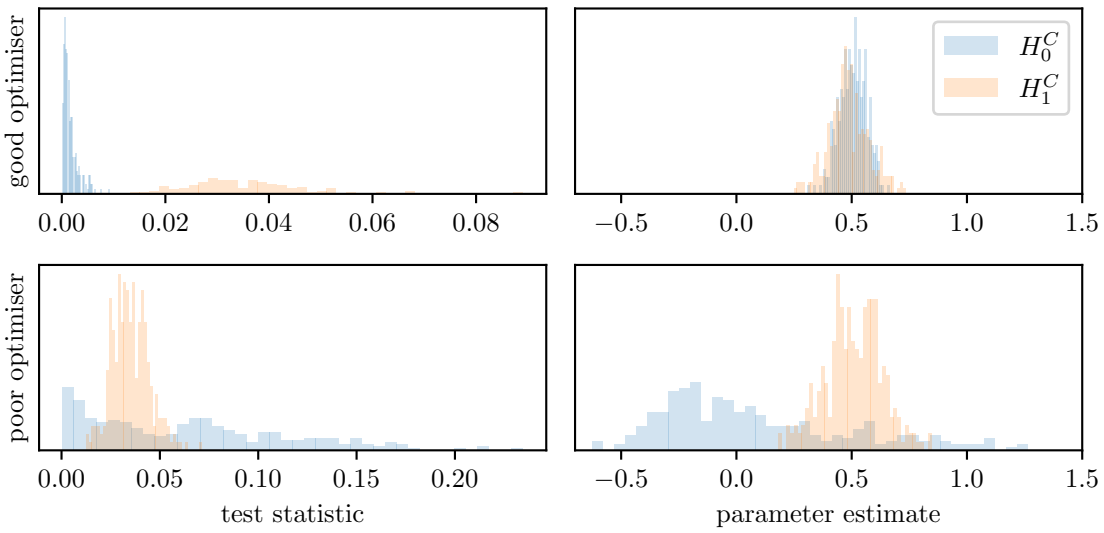


Figure 13: Demonstration of a failure mode where the parameter estimate has high variance due to a poorly configured optimiser. The first row shows the distribution of the test statistic (left) and that of the parameter estimate (right) for a well-configured optimiser, and the second row the distributions for a poorly-configured optimiser.

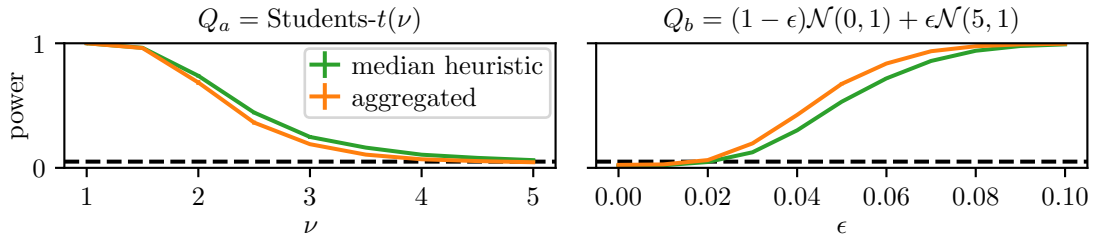


Figure 14: Power of the KSD test with wild bootstrap, comparing setting the lengthscale set using the median heuristic and aggregating tests using 8 different lengthscales. We set $H_0^C : P = \mathcal{N}(\mu, \sigma^2)$, with the left and right figure showing data generated from different distributions, Q_a and Q_b. **- -** shows the level.

Appendix C. Aggregated Composite Test

Figure 14 shows an attempt at combining the aggregated test method introduced by Schrab et al. (2023, 2022) with our composite KSD test, using the wild bootstrap. The aggregated test is the combination of 8 composite tests using the Gaussian kernel, where the lengthscales are set as the $\{25, 32, 39, 46, 54, 61, 68, 75\}$th percentiles of the pairwise distances between the observations. The non-aggregated test uses a single lengthscale set using the median heuristic, i.e. at the 50th percentile of the pairwise distances. The null hypothesis of the

test is $H_0^C : P = \mathcal{N}(\mu, \sigma^2)$, for $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. To evaluate the performance of the test, we consider the two alternatives specified in the figure. Against a Student's t-distribution, we find that the median heuristic has slightly better performance than the aggregated test, while, against a mixture of Gaussians, the aggregated test has a small advantage. We would expect the aggregated test to have a bigger advantage in the case of the mixture alternative, as here the spread of lengthscales will bring a bigger benefit than in the unimodal case of the Student's t alternative. It is somewhat unexpected that the aggregated test has slightly lower performance than the median heuristic against the Student's t alternative. Future work could investigate why this happens, and whether any modifications can be made to the aggregated testing procedure to improve the performance in the composite case.

Appendix D. Experiment Details

For all experiments we use a test level of $\alpha = 0.05$. The experiments are implemented in JAX, and executed on an Nvidia RTX 3090 GPU.

D.1 Closed-Form KSD Estimator

For any P_θ in the exponential family, we can calculate $\hat{\theta}_n = \arg\min_\theta \mathrm{KSD}^2(P_\theta, Q_n)$ in closed-form. Here we state the expression for this estimator, which we use in our Gaussian and kernel exponential family experiments. The detailed derivation of this result can be found in Barp et al. (2019, Appendix D3) or Matsubara et al. (2022).

Let the density of a model in the exponential family be $p_\theta(x) = \exp(\eta(\theta)\cdot t(x) - a(\theta) + b(x))$, with $\eta : \Theta \to \mathbb{R}^k$ an invertible map, $t : \mathbb{R}^d \to \mathbb{R}^k$ any sufficient statistic for some $k \in \mathbb{N}_1$, $a : \Theta \to \mathbb{R}$ and $b : \mathbb{R}^d \to \mathbb{R}$. Then the KSD estimator is given by $\hat{\theta}_n := \eta^{-1}\left(-\frac{1}{2}\Lambda_n^{-1} v_n\right)$, where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as $\Lambda_n := \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \Lambda(x_i, x_j)$ and $\nu_n := \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \nu(x_i, x_j)$. These are based on functions $\Lambda : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{k\times k}$ and $\nu : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^k$ which depend on the specific model, defined as $\Lambda(x, x') := K(x, x')\nabla_x t(x) \nabla_x t(x')^\top$ and

$$\begin{aligned}\nu(x, x') := {} & K(x, x')\nabla_x b(x)\nabla_x t(x')^\top + \nabla_x t(x)\nabla_{x'} K(x, x') \\ & + K(x, x')\nabla_x b(x')\nabla_x t(x)^\top + \nabla_x t(x')\nabla_x K(x', x)\,.\end{aligned}$$

D.2 Kernels

We define the Gaussian kernel as $K(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2 / 2l^2)$, and the IMQ kernel as $K(x_1, x_2) = (1 + \|x_1 - x_2\|_2^2 / 2l^2)^{-\frac{1}{2}}$, where $l \in \mathbb{R}^+$ is the lengthscale, and the exponent of $-\frac{1}{2}$ follows Matsubara et al. (2022). Where we set the lengthscale using the median heuristic, this is defined as $l_{\text{med}} = \sqrt{\text{median}(\|y_i - y_j\|_2^2 / 2)}$, where the median is taken over all pairs of observations, y_i and y_j.

Where we use a sum kernel, this is defined as $K(x_1, x_2) = \frac{1}{L}\sum_{i=1}^L K_{l_i}(x_1, x_2)$, where $K_{l_1}, \ldots, K_{l_L}$ are a set of L Gaussian or IMQ kernels with lengthscales $l_1, \ldots, l_L$. Prior work has found that using a sum of kernels with a range of lengthscales, rather than attempting to choose the lengthscale of a single kernel, produces good empirical results (Li et al., 2015; Ren et al., 2016; Sutherland et al., 2017). Note that a sum of characteristic kernels is a

characteristic kernel (Sriperumbudur et al., 2010), so this choice satisfies the assumptions we make in our theoretical results.

D.3 Details of Specific Figures

D.3.1 Distribution of Wild Bootstrap Statistic (Figure 2)

$H_0^C : P = \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, with the estimator configured the same as in the Gaussian experiments below. We sample observations from $Q = \mathcal{N}(0.3, 1.0)$. We use a Gaussian kernel with lengthscale 1.0. To compute the distribution of the test statistic, we take 1000 sets of $n = 1000$ samples from Q, and for each estimate $P_{\hat{\theta}_n}$ and compute $\mathrm{MMD}^2(P_{\hat{\theta}_n,n}, Q_n)$. To compute the distribution of the wild bootstrap statistic, we take $n = 1000$ samples from Q, estimate $P_{\hat{\theta}_n}$, sample 1000 sets of Rademacher variables, and thus compute 1000 samples of $\mathrm{MMD}^2_W(P_{\hat{\theta}_n,n}, Q_n)$.

D.3.2 Gaussian Model (Figures 3 to 7)

repeats to compute power (per seed)	1000
K	Gaussian, $l = l_{\text{med}}$
wild bootstrap samples	500
parametric bootstrap samples	300

MMD test details For the experiments using the MMD, we minimize the minimum distance estimator using the Adam optimiser (Kingma and Ba, 2015). We use a learning rate of 0.05 and 100 iterations. When $\theta = \mu_d$ (Figures 4 to 6) we sample the initial value of μ_d from $\mathcal{N}(\mathbf{0}_d, I_d)$. When $\theta = \{\mu, \sigma^2\}$ (Figures 3 and 7) sample the initial μ from $\mathcal{N}(0, 1)$, and sample σ from a standard normal distribution truncated between 0.5 and 1.5.

Timing We estimate the time taken by each bootstrap as follows. We run 3 tests as a warmup, then we run 2000 tests and calculate the mean time taken. The full configuration is as follows:

H_0^C	$P_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$	n	100
D	KSD	bootstrap samples	300
K	Gaussian, $l = 1.0$		

D.3.3 Toggle Switch

Figures 8 to 10 The true model parameters, θ_0, are

parameter	α_1	α_2	β_1	β_2	μ	σ	γ
true value	22.0	12.0	4.0	4.5	325.0	0.25	0.15

We use stochastic gradient descent with random restarts to estimate the parameters. The algorithm is:

1. Sample 500 initial parameter values, $\theta^1_{\text{init}}, \ldots, \theta^{500}_{\text{init}}$.

2. Select the 15 θ^i_{init} for which $\text{MMD}^2(P_{\theta^i_{\text{init}}}, Q_n)$ is smallest.

3. Run SGD 15 times, once for each of the selected θ^i_{init}, to find $\hat{\theta}^1_n, \ldots, \hat{\theta}^{15}_n$. Use learning rate 0.04 and perform 300 iterations.

4. Return $\hat{\theta}^i_n$ for which $\text{MMD}^2(P_{\hat{\theta}^i_n}, Q_n)$ is smallest.

The initial parameters are sampled from a uniform distribution with the following ranges:

parameter	α_1	α_2	β_1	β_2	μ	σ	γ
initial range	0.01 50.00	0.01 50.00	0.01 5.00	0.01 5.00	250.0 450.0	0.01 0.50	0.01 0.40

To run the test, we use the following configuration:

K	unweighted mixture of Gaussian kernels with lengthscales 20.0, 40.0, 80.0, 100.0, 130.0, 200.0, 400.0, 800.0, 1000.0
n	400
bootstrap	parametric
bootstrap samples	200

To create Figure 10 we apply kernel density estimation to samples from the model. We use a Gaussian kernel, with the bandwidth set to 0.1.

D.3.4 Nonparametric Density Estimation

Figure 11 We are testing whether the family of models considered by Matsubara et al. (2022) fits the data, thus our choices of q_{kef} and l_{kef} match the original paper.

K	unweighted sum of IMQ kernels with lengthscales 0.6, 1.0, 1.2
n	82
wild bootstrap samples	500
q_{kef}	$\mathcal{N}(0.0, 3.0^2)$
l_{kef}	$\sqrt{2}$

Following Matsubara et al. (2022), we normalise each observation in the data set as follows: $y_i^{\text{norm}} = \frac{y_i - \mu}{0.5 * \sigma}$, where μ and σ are the mean and standard deviation of the data set.

To plot the density of the model, we estimate the normalising constant of the model using the importance sampling approach suggested by Wenliang et al. (2019, Section 3.2). As a proposal we use $\mathcal{N}(0.0, 3.5^2)$, and we take 2000 samples.

References

A. Anastasiou, A. Barp, F-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, L. Mackey, C. J. Oates, G. Reinert, and Y. Swan. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 2023.

T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 1954.

K. Balasubramanian, T. Li, and M. Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 2021.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, 2019.

J. Beirlant, L. Gyorfi, and G. Lugosi. On the asymptotic normality of the L1- and L2-errors in histogram density estimation. *The Canadian Journal of Statistics*, 1994.

J. Beirlant, L. Devroye, L. Györfi, and I. Vajda. Large deviations of divergence measures on partitions. *Journal of Statistical Planning and Inference*, 2001.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2004.

Espen Bernton, Pierre E. Jacob, Mathieu Gerber, and Christian P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, April 2019.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1974.

S. Betsch, B. Ebner, and B. Klar. Minimum l^q-distance estimators for non-normalized parametric models. *Canadian Journal Of Statistics*, 2020.

Ayush Bharti, Francois-Xavier Briol, and Troels Pedersen. A general method for calibrating stochastic radio channel models with kernels. *IEEE Transactions on Antennas and Propagation*, 2021.

Ayush Bharti, Masha Naslidnyk, Oscar Key, Samuel Kaski, and François-Xavier Briol. Optimally-weighted estimators of the Maximum Mean Discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, 2023.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representation*, 2018.

Fernando V. Bonassi and Mike West. Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis*, 2015.

Fernando V. Bonassi, Lingchong You, and Mike West. Bayesian learning from marginal data in bionetwork models. *Statistical Applications in Genetics and Molecular Biology*, 2011.

W. Bounliphone, E. Belilovsky, M. B. Blaschko, I. Antonoglou, and A. Gretton. A test of relative similarity for model selection in generative models. In *International Conference on Learning Representations*, 2016.

François-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical inference for generative models with Maximum Mean Discrepancy. arXiv:1906.05944, 2019.

Stéphane Canu and Alexander J. Smola. Kernel methods and the exponential family. *Neurocomputing*, 2006.

Ying-Xia Chen and Wei-Jun Zhu. Note on the strong law of large numbers in a Hilbert space. *Gen. Math*, 2011.

B-E. Chérief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via Maximum Mean Discrepancy. In *2nd Symposium on Advances in Approximate Bayesian Inference*, 2020.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli*, 2022.

Kacper Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Neural Information Processing Systems*, 2014.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2016.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America*, 2020.

Ralph D'Agostino and E. S. Pearson. Tests for departure from normality. Empirical results for the distributions of $b2$ and $\sqrt{b1}$. *Biometrika*, 1973.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 2007.

C. Dellaporta, J. Knoblauch, T. Damoulas, and F-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, 2022.

J. Durbin. Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika*, 1975.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via Maximum Mean Discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

Bruno Ebner. On combining the zero bias transform and the empirical characteristic function to test normality. *Latin American Journal of Probability and Mathematical Statistics*, 2021.

T. Fernández and A. Gretton. A Maximum-Mean-Discrepancy goodness-of-fit test for censored data. In *Artificial Intelligence and Statistics*, 2019.

Tamara Fernández and Nicolás Rivera. A general framework for the analysis of kernel-based tests. arXiv:2209.00124, 2022.

Tamara Fernández, Nicolás Rivera, Wenkai Xu, and Arthur Gretton. Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, 2020.

K. Fukumizu. Exponential manifold by reproducing kernel Hilbert spaces. In *Algebraic and Geometric Methods in Statistics*. 2009.

Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, 2009.

W. Gong, Y. Li, and J. M. Hernández-Lobato. Sliced Kernelized Stein Discrepancy. *International Conference on Learning Representations*, 2021.

W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, and R. Zemel. Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, 2020.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

L. Györfi and I. Vajda. Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 2002.

L. Gyorfi and E. C. van der Meulen. A consistent goodness-of-fit test based on the total variation distance. In *Nonparametric Functional Estimation and Related Topics*. 1991.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 1990.

Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. John Wiley & Sons, 2011.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2006.

M. D. Jiménez-Gamero, V. Alba-Fernández, J. Muñoz García, and Y. Chalco-Cano. Goodness-of-fit tests based on empirical characteristic functions. *Computational Statistics & Data Analysis*, 2009.

W. Jitkrittum, P. Sangkloy, B. Schölkopf, H. Kanagawa, J. Hays, and A. Gretton. Informative features for model comparison. In *Neural Information Processing Systems*, 2018.

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Ana Justel, Daniel Pefia, Ruben Zamar, and Departament Of Statistics. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters*, 1997.

H. Kanagawa, W. Jitkrittum, L. MacKey, K. Fukumizu, and A. Gretton. A kernel Stein test for comparing latent variable models. arXiv:1907.00586, 2019.

Jérémie Kellner and Alain Celisse. A one-sample test for normality with kernel methods. *Bernoulli*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

I. Kojadinovic and J. Yan. Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap. *Canadian Journal of Statistics*, 2012.

A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italianodegli Attuari*, 1933.

I. A. Koutrouvelis and J. Kellermeier. A goodness-of-fit test based on the empirical characteristic function when parameters must be estimated. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 1981.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. 3 edition, 2005.

Anne Leucht and Michael H. Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 2013.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017.

Yujia Li, Kevin Swersky, and Richard S. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, 2015.

Hubert W Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 1967.

J. N. Lim, M. Yamada, B. Schölkopf, and W. Jitkrittum. Kernel Stein tests for multiple model comparison. In *Neural Information Processing Systems*, 2019.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, 2016.

X. Liu and F.-X. Briol. On the robustness of kernel goodness-of-fit tests. arXiv:2408.05854, 2024.

James Robert Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

T. Matsubara, J. Knoblauch, François-Xavier Briol, and C. J. Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society B: Statistical Methodology*, 2022.

A. Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

J. Neyman. *A Selection of Early Statistical Papers of J. Neyman*, chapter 28. University of California Press, 1967.

Richard Nickl. Statistical theory, 2012.

Z. Niu, J. Meier, and François-Xavier Briol. Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. arXiv:2106.11561, 2021.

C J Oates, M Girolami, and N Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society B: Statistical Methodology*, 2017.

W C Parr and W R Schucany. Minimum distance and robust estimation. *Journal of the American Statistical Association*, 1980.

M. Postman, J. P. Huchra, and M. J. Geller. Probes of large-scale structure in the Corona Borealis region. *The Astronomical Journal*, 1986.

Yong Ren, Jialian Li, Yucen Luo, and Jun Zhu. Conditional generative moment-matching networks. In *Advances in Neural Information Processing Systems*, 2016.

G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 2007.

Kathryn Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 1990.

A. Schrab and I. Kim. Robust kernel hypothesis testing under data corruption. arXiv:2405.19912, 2024.

Antonin Schrab, Benjamin Guedj, and Arthur Gretton. KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems*, 2022.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 2023.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 2013.

A. Sen and B. Sen. Testing independence and goodness-of-fit in linear models. *Biometrika*, 2014.

Xiaofeng Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 2010.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 1965.

B. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 2017.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 2010.

Winfried Stute, Wenceslao Gonzáles Manteiga, and Manuel Presedo Quindimil. Bootstrap based goodness-of-fit-tests. *Metrika*, 1993.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. Generative models and model criticism via optimized Maximum Mean Discrepancy. In *International Conference on Learning Representations*, 2017.

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 2005.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 2013.

Li Wenliang, Danica J. Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, 2019.

Li K. Wenliang and Heishiro Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. arXiv:2008.10087, 2021.

G. Wolfer and P. Alquier. Variance-aware estimation of kernel mean embedding. arXiv:2210.06672, 2022.

J. Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, 1957.

G. Wynne and S. Nagy. Statistical depth meets machine learning: Kernel mean embeddings and depth in functional data analysis. arXiv:2105.12778, 2021.

G. Wynne, M. Kasprzak, and A. B. Duncan. A spectral representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. arXiv:2206.04552, 2022.

W. Xu and T. Matsuda. A Stein goodness-of-fit test for directional distributions. In *Artificial Intelligence and Statistics*, 2020.

W. Xu and G. Reinert. A Stein goodness of fit test for exponential random graph models. In *Artificial Intelligence and Statistics*, 2021.

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning*, 2018.

J. Yang, V. Rao, and J. Neville. A Stein-Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

S. Zhu, B. Chen, P. Yang, and Z. Chen. Universal hypothesis testing with kernels: Asymptotically optimal tests for goodness of fit. In *International Conference on Artificial Intelligence and Statistics*, 2019.