A Kernel Test of Goodness of Fit

Kacper Chwialkowski* KACPER.CHWIALKOWSKI@GMAIL.COM
Heiko Strathmann* HEIKO.STRATHMANN@GMAIL.COM
Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Unit, University College London, United Kingdom

Abstract

We propose a nonparametric statistical test for goodness-of-fit: given a set of samples, the test determines how likely it is that these were generated from a target density function. The measure of goodness-of-fit is a divergence constructed via Stein's method using functions from a Reproducing Kernel Hilbert Space. Our test statistic is based on an empirical estimate of this divergence, taking the form of a V-statistic in terms of the log gradients of the target density and the kernel. We derive a statistical test, both for i.i.d. and non-i.i.d. samples, where we estimate the null distribution quantiles using a wild bootstrap procedure. We apply our test to quantifying convergence of approximate Markov Chain Monte Carlo methods, statistical model criticism, and evaluating quality of fit vs model complexity in nonparametric density estimation.

1. Introduction

Statistical tests of goodness-of-fit are a fundamental tool in statistical analysis, dating back to the test of Kolmogorov and Smirnov (Kolmogorov, 1933; Smirnov, 1948). Given a set of samples $\{Z_i\}_{i=1}^n$ with distribution $Z_i \sim q$, our interest is in whether q matches some reference or target distribution p, which we assume to be only known up to the normalisation constant. Recently, in the multivariate setting, Gorham & Mackey (2015) proposed an elegant measure of sample quality with respect to a target. This measure is a maximum discrepancy between empirical sample expectations and target expectations over a large class of test functions, constructed so as to have zero expectation over the target distribution by use of a Stein operator. This operator depends only on the derivative of the $\log q$: thus, the approach can be applied very generally, as it does not require closed-form integrals over the target distribution (or numerical approximations of such integrals). By contrast, many earlier discrepancy measures require integrals with respect to the target (see below for a review). This is problematic if the intention is to perform benchmarks for assessing Markov Chain Monte Carlo, since these integrals will certainly not be known to the practitioner.

A challenge in applying the approach of Gorham & Mackey is the complexity of the function class used, which results from applying the Stein operator to the bounded Lipschitz functions.[1] Thus, their sample quality measure requires solving an expensive linear program that arises from a complicated construction of graph Stein discrepancies and geometric spanners. Their metric furthermore requires access to nontrivial lower bounds that, despite being provided for log-concave densities, are a largely open problem otherwise, in particular for multivariate cases.

An important application of a goodness-of-fit measure is in statistical testing, where it is desired to determine whether the empirical discrepancy measure is large enough to reject the null hypothesis (that the sample arises from the target distribution). One approach is to establish the asymptotic behaviour of the test statistic, and to set a test threshold at a large quantile of the asymptotic distribution. The asymptotic behaviour of the Wasserstein-based Stein discrepancy remains a challenging open problem, due to the complexity of the function class used. It is not clear how one would compute p-values for this statistic, or determine when the goodness of fit test would allow us to accept the null hypothesis (at the user-specified test level).

The key contribution of this work is to define a statistical test of goodness-of-fit, based on a Stein discrepancy computed in a Reproducing Kernel Hilbert Space (RKHS). To

[1] The bounded Lipschitz functions give rise to the Wasserstein integral probability metric. By contrast, the Kolmogorov-Smirnov test uses functions of bounded variation 1 (Müller, 1997). Multivariate generalisations of the K-S test exist, however the computational cost of a consistent test rapidly becomes prohibitive with increasing dimension (Justel et al., 1997).

construct our test statistic, we apply the Stein operator to our chosen set of RKHS functions, and define our measure of goodness of fit as the largest discrepancy over this set between empirical sample expectations and target expectations (the latter being zero, due to the effect of the Stein operator). This approach is a natural extension to goodness-of-fit testing of the earlier two-sample tests (Gretton et al., 2012) and independence tests (Gretton et al., 2007) based on the maximum mean discrepancy, which is an integral probability metric. As with these earlier tests, our statistic is a simple U-statistic, and can be computed in closed form and in quadratic time; moreover, it is an unbiased estimate of the corresponding population discrepancy. As with all Stein-based discrepancies, only the gradient of the log-density of the target density is needed; we do not require integrals with respect to the target density – including the normalisation constant. Given that our test statistic is a V-statistic, we may make use of the extensive literature on asymptotics of V-statistics to formulate a hypothesis test (Serfling, 1980; Leucht & Neumann, 2013). We are able to provide statistical tests even in the case of correlated samples, which is essential if the test is to be used in assessing the quality of output of an MCMC procedure.

Several alternative approaches exist in the statistics literature to goodness-of-fit testing. A first strategy is to partition the space, and to conduct the test on a histogram estimate of the distribution (Barron, 1989; Beirlant et al., 1994; Györfi & van der Meulen, 1990; Györfi & Vajda, 2002). Such space partitioning approaches can have attractive theoretical properties (e.g. distribution-free test thresholds) and work well in low dimensions, however they are much less powerful than alternatives once the dimensionality increases (Gretton & Gyorfi, 2010). A second popular approach has been to use the smoothed L_2 distance between the empirical characteristic function of the sample, and the characteristic function of the target density. This dates back to the test of Gaussianity of Baringhaus & Henze (1988), who used a squared exponential smoothing function (see Eq. 2.1 in their paper). For this choice of smoothing function, their statistic is identical to the maximum mean discrepancy (MMD) with the squared exponential kernel, which can be shown using the Bochner representation of the kernel (compare with Sriperumbudur et al. 2010, Corollary 4). It is essential in this case that the target distribution be Gaussian, since the convolution with the kernel (or in the Fourier domain, the smoothing function) must be available in closed form. An L_2 distance between Parzen window estimates can also be used (Bowman & Foster, 1993), giving the same expression again, although the optimal choice of bandwidth for consistent Parzen window estimates may not be a good choice for testing (Anderson et al., 1994). A different smoothing scheme in the frequency domain results in an energy distance statistic (this likewise being an MMD with a particular choice of kernel; see Sejdinovic et al., 2013), which can be used in a test of normality (Székely & Rizzo, 2005). The key point is that the required integrals are again computable in closed form for the Gaussian, although the reasoning may be extended to certain other families of interest, e.g. (Rizzo, 2009). The requirement of computing closed-form integrals with respect to the test distribution severely restricts this testing strategy. Finally, a problem related to goodness-of-fit testing is that of model criticism (Lloyd & Ghahramani, 2015). In this setting, samples generated from a fitted model are compared via the maximum mean discrepancy with samples used to train the model, such that a small MMD indicates a good fit. There are two limitation to the method: first, it requires samples from the model (which might not be easy if this requires a complex MCMC sampler); second, the choice of number of samples from the model is not obvious, since too few samples cause a loss in test power, and too many are computationally wasteful. Neither issue arises in our test, since we do not require model samples.

In our experiments, a particular focus is on applying our goodness-of-fit test to certify the output of approximate Markov Chain Monte Carlo (MCMC) samplers (Korattikara et al., 2014; Welling & Teh, 2011a; Bardenet et al., 2014). These methods use modifications to Markov transition kernels that improve mixing speed at the cost of worsening the asymptotic bias. The bias-variance trade-off can usually be tuned with parameters of the sampling algorithms. It is therefore important to test whether for a particular parameter setting and run-time, the samples are of the desired quality. This question cannot be answered with classical MCMC convergence statistics, such as the widely used potential scale reduction factor (R-factor) (Gelman & Rubin, 1992) or the effective sample size, since these assume that the Markov chain reaches its equilibrium distribution. By contrast, our test exactly quantifies the asymptotic bias of approximate MCMC.

Paper outline We begin our presentation in the section 2 with a high-level construction of the RKHS-based Stein discrepancy and associated statistical test. In Section 3, we provide additional details and prove the main results. Section 4 contains experimental illustrations on synthetic examples, statistical model criticism, bias-variance trade-offs in approximate MCMC, and convergence in non-parametric density estimation.

2. Test Definition: Statistic and Threshold

We begin with a high-level construction of our divergence discrepancy and the statistical test. While this section aims to communicate the main ideas, we provide details and proofs in Section 3.

2.1. Stein Operator in RKHS

Our goal is to write the maximum discrepancy between target distribution p and observed sample distribution q in a RKHS. Denote by $\mathcal{F}$ the RKHS of real-valued functions with reproducing kernel k, and by $\mathcal{F}^d$ the product RKHS consisting of elements $f := (f_1, \ldots, f_d)$ with $f_i \in \mathcal{F}$, and with a standard inner product. Similarly to Stein (1972); Gorham & Mackey (2015), we begin by defining a Stein operator T acting on $f \in \mathcal{F}^d$

$$Tf := \sum_{i=1}^{d} \frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i}.$$

Suppose a random variable Z is distributed according to a measure[2] q and X is distributed according to the target measure p. As we will see, the operator can be expressed by defining a function that depends on gradients of the log-density and the kernel,

$$\xi(x, \cdot) := \left[\nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot)\right], \tag{1}$$

whose inner product with f gives exactly the expected value of the Stein operator

$$\mathbb{E}Tf(Z) = \langle f, \mathbb{E}\xi(Z)\rangle_{\mathcal{F}^d} = \sum_{i=1}^{d} \langle f_i, \mathbb{E}\xi_i(Z)\rangle_{\mathcal{F}},$$

c.f. Lemma 3.3. For X from the target measure, we have $\mathbb{E}(Tf)(X) = 0$, which can be seen using integration by parts, c.f. Lemma 5.1 in the supplement. We can now define a Stein discrepancy and express it in the RKHS,

$$\begin{aligned} S(Z) &:= \sup_{\|f\|<1} \mathbb{E}(Tf)(Z) - \mathbb{E}(Tf)(X) \\ &= \sup_{\|f\|<1} \langle f, \mathbb{E}\xi(Z) - \mathbb{E}\xi(X)\rangle \\ &= \sup_{\|f\|<1} \langle f, \mathbb{E}\xi(Z)\rangle \\ &= \|\mathbb{E}\xi(Z)\|, \end{aligned}$$

c.f. Lemma 3.4. This makes it clear why $\mathbb{E}(Tf)(X) = 0$ is a desirable property: we can compute $S(Z)$ by computing $\|\mathbb{E}\xi(Z)\|$, without the need to access X in the form of samples from p. We arrive at our first main result, which states that the above discrepancy can be used to distinguish two distributions $p, q \in \mathcal{P}$.

Theorem 2.1. *Let $q, p \in \mathcal{P}$ where the derivatives of elements of $\mathcal{P}$ satisfy assumption (ii) in Section 3.1, and let $Z \sim q$. Let the RKHS $\mathcal{F}$ satisfy properties (iii) and (v) in Section 3.1, which include the requirement that $\mathcal{F}$ be cc-universal (Carmeli et al., 2010, Definition 4.1). Then $S(Z) = 0$ if and only if $p = q$.*

[2] Throughout the article, all occurrences of Z, e.g. Z', Z_i, $Z_\heartsuit$, are understood to be distributed according to q.

Section 3.1 contains the formal statements of the assumptions on $\mathcal{P}$ and $\mathcal{F}$, and a proof. The following theorem gives a simple closed form expression.

Theorem 2.2. *Let*

$$\begin{aligned} h(x, y) := &\nabla \log p(x)^\top \nabla \log p(y) k(x, y) \\ &+ \nabla \log p(y)^\top \nabla_x k(x, y) \\ &+ \nabla \log p(x)^\top \nabla_y k(x, y) \\ &+ \sum_{i=1}^{d} \frac{\partial^2 k(x, y)}{\partial x_i \partial y_i}. \end{aligned}$$

The squared Stein discrepancy is $S(Z)^2 = \mathbb{E}h(Z, Z')$.

We now proceed with constructing an estimator for $S(Z)^2$, and outline its asymptotic properties.

2.2. Wild Bootstrap Testing

It is straightforward to estimate the squared Stein discrepancy $S(Z)^2$ from samples $\{Z_i\}_{i=1}^n$: a quadratic time estimator is a V-Statistic, and takes the form

$$V_n = \frac{1}{n^2} \sum_{i,j=1}^{n} h(Z_i, Z_j).$$

The asymptotic null distribution of the normalised V-Statistic nV_n, however, has no computable closed form. Furthermore, care has to be taken when the Z_i exhibit correlation structure, as the null distribution significantly changes, impacting test significance. The wild bootstrap technique (Shao, 2010; Leucht & Neumann, 2013; Fromont et al., 2012) addresses both problems. First, it allows to simulate from the null distribution to compute test thresholds. Second, it accounts for correlation structure in the Z_i by mimicking it with an auxiliary random process: a Markov chain taking values in $\{-1, 1\}$, starting from $W_{1,n} = 1$,

$$W_{t,n} = \mathbf{1}(U_t > a_n) W_{t-1,n} - \mathbf{1}(U_t < a_n) W_{t-1,n},$$

where the U_t are uniform i.i.d. random variables and a_n is the probability of $W_{t,n}$ changing sign (for i.i.d. data we may set $a_n = 0.5$). This leads to a bootstrapped V-statistic

$$B_n = \frac{1}{n^2} \sum_{i,j=1}^{n} W_{i,n} W_{j,n} h(Z_i, Z_j).$$

Proposition 3.6 establishes that, under the null hypothesis, nB_n is a good approximation of nV_n, so it is possible to approximate quantiles of the null distribution by sampling from it. Under the alternative, however, V_n dominates B_n – resulting in almost sure rejection of the null hypothesis.

We propose the following test procedure for testing the null hypothesis that the Z_i are distributed according to the target distribution p.

- Calculate the test statistic V_n.
- Obtain wild bootstrap samples $\{B_n\}_{i=1}^D$ and estimate the $1-\alpha$ empirical quantile of these samples.
- If V_n exceeds the quantile, reject.

3. Proofs of the Main Results

We now prove the claims made in the previous Section.

3.1. Stein Operator in RKHS

We make the following assumptions. Let $\mathcal{P}$ be a family of distributions on a real coordinate space, where its elements $p \in \mathcal{P}$ satisfy two conditions:

(i) $\nabla \log p(x)$ is Lipschitz continuous.

(ii) $\mathbb{E}\|\nabla \log p(Z)\|^2 \leq \infty$ for any random variable.

The kernels k considered in this work satisfy

(iii) $\mathbb{E}\left(\frac{\partial^2 k(Z,Z)}{\partial x_i \partial x_{i+d}}\right)^2 \leq \infty$.

(iv) $\nabla_x k(x,y)$ is Lipschitz continuous.

(v) k is bounded, symmetric and cc-universal (Carmeli et al., 2010).

Requirements (i) and (iv) are used in Proposition 3.6 regarding the wild-bootstrap procedure, (ii) and (iii) are needed for Bocher integrability of ξ in Lemma 3.2, and (v) is needed in the proof of Theorem 2.1.

We show in Lemma 5.1 in the Appendix that the expected value of the Stein operator is zero on the target measure.

The following lemmas are useful in proving our main results, Theorems 2.2 and 2.1.

Lemma 3.1. *$\xi(x,\cdot)$ (see Eq. (1)) is an element of the reproducing kernel Hilbert space $\mathcal{F}^d$.*

Proof. We use the proof of Steinwart & Christmann (2008, Corollary 4.36) to see that for all $x \in R^d$ each entry of $\nabla k(x,\cdot)$ belongs to $\mathcal{F}$. $\frac{\partial \log p(x)}{\partial x_i} k(x,\cdot) \in \mathcal{F}$, since $k(x,\cdot) \in \mathcal{F}$ and $\frac{\partial \log p(x)}{\partial x_i}$ is a scalar. □

The following lemma shows that the expected value of ξ is well defined – it is needed for establishing a link between Stein operator Tf and ξ.

Lemma 3.2. *For any random variable Z, expected value of $\xi(Z)$ is is element of $\mathcal{F}^d$ (ξ is Bochner integrable wrt the measure of Z).*

Proof. It is sufficient to check that coefficients of ξ are Bochner integrable (Steinwart & Christmann, 2008, Definition A.5.20). First we check that for any random variable Z,

$$\mathbb{E}\left\|\frac{\partial \log p(Z)}{\partial x_i} k(Z,\cdot)\right\|^2 < C\mathbb{E}\|\nabla \log p(X)\|^2 < \infty,$$

for some constant C, which follows from assumption (i) and boundedness of the kernel. Next we check that

$$\mathbb{E}\left\|\frac{\partial k(Z,\cdot)}{\partial x}\right\|^2 = \mathbb{E}\left(\frac{\partial^2 k(Z,Z)}{dx_i dx_{i+d}}\right)^2 < \infty,$$

which follows from assumption (iii). □

We can now show that the expected value of the Stein operator can be expressed as an inner product with an element of $\mathcal{F}^d$, where this element is the expected value of ξ.

Lemma 3.3. *For any random variable Z, the expected value of the Stein operator coincides with the inner product of f and the expected value of $\xi(Z)$,*

$$\mathbb{E}Tf(Z) = \langle f, \mathbb{E}\xi(Z)\rangle_{\mathcal{F}^d} = \sum_{i=1}^d \langle f_i, \mathbb{E}\xi_i(Z)\rangle_{\mathcal{F}}.$$

Proof. We write

$$\begin{aligned}
&\langle f_i, \mathbb{E}\xi_i(Z)\rangle_{\mathcal{F}} \\
&= \left\langle f_i, \mathbb{E}\left[\frac{\partial \log p(Z)}{\partial x_i} k(Z,\cdot) + \frac{\partial k(Z,\cdot)}{\partial x_i}\right]\right\rangle_{\mathcal{F}} \\
&= \mathbb{E}\left\langle f_i, \frac{\partial \log p(Z)}{\partial x_i} k(Z,\cdot) + \frac{\partial k(Z,\cdot)}{\partial x_i}\right\rangle_{\mathcal{F}} \\
&= \mathbb{E}\left[\frac{\partial \log p(Z)}{\partial x_i} f_i(Z) + \frac{\partial k(Z,\cdot)}{\partial x_i}\right].
\end{aligned}$$

The second equality follows from the fact that a linear operator $\langle f_i, \cdot\rangle_{\mathcal{F}}$ can be interchanged with the Bochner integral, and the fact that ξ is Bochner integrable (Lemma 3.2). The last equality is an application of the reproducing property. □

From the inner product representation, we get

Lemma 3.4. *The discrepancy $S(Y,\mathcal{F},p)$ is maximized by the expected value of ξ, i.e. $S(Y,\mathcal{F},p) = \|\mathbb{E}\xi(Y)\|$.*

Proof. By the Lemma 3.3, $\mathbb{E}Tf(Y) = \langle f, \mathbb{E}\xi(Y)\rangle$ and therefore, $S(Y,\mathcal{F},p)$ is maximized by $\frac{\mathbb{E}\xi(Y)}{\|\mathbb{E}\xi(Y)\|}$. □

We are now ready for the proof of the closed form formula for $S(Y,\mathcal{F},p)^2$.

Proof of Theorem 2.2. We use the notation

$$\nabla_x k(x,y) = \left(\frac{\partial k(x,y)}{\partial x_1}, \cdots, \frac{\partial k(x,y)}{\partial x_d}\right)$$
$$\nabla_y k(x,y) = \left(\frac{\partial k(x,y)}{\partial y_1}, \cdots, \frac{\partial k(x,y)}{\partial y_d}\right),$$

giving

$$\begin{aligned}
&S(X,\mathcal{F},p)^2 = \langle \xi, \xi \rangle_{\mathcal{F}^d} \\
&= \langle \mathbb{E}\left[\nabla \log p(X) k(X,\cdot) + \nabla_x k(X,\cdot)\right], \\
&\quad \mathbb{E}\left[\nabla \log p(X) k(X,\cdot) + \nabla_x k(X,\cdot)\right] \rangle \\
&= \mathbb{E}\langle \nabla \log p(X) k(X,\cdot) + \nabla_x k(X_1,\cdot), \\
&\quad \nabla \log p(X) k(\cdot,X) + \nabla_y k(\cdot,X) \rangle \\
&= \mathbb{E}\nabla \log p(X)^\top \nabla \log p(X') k(X,X') \\
&\quad + \mathbb{E}\nabla \log p(X) \nabla_x k(X,X') \\
&\quad + \mathbb{E}\nabla \log p(X)^\top \nabla_y k(X,X') \\
&\quad + \mathbb{E}\sum_{i=1}^{d} \frac{\partial^2 k(X,X')}{\partial x_i \partial x_{i+d}}.
\end{aligned}$$

□

Finally, we prove that the discrepancy S discriminates different probability measures.

Proof of Theorem 2.1. If $p = q$ then $S(Y,\mathcal{F},p)$ is 0 by Lemma (5.1). Suppose $p \neq q$, but $S(Y,\mathcal{F},p) = 0$. If $S(Y,\mathcal{F},p) = 0$ then $\mathbb{E}\xi(Y) = 0$. For each dimension of $\mathbb{E}\xi(Y)$, we add and subtract $\log q(Y)$

$$\begin{aligned}
&\mathbb{E}\left(\frac{\partial}{\partial x_i}\log p(Y) k(Y,\cdot) + \frac{\partial}{\partial x_i} k(Y,\cdot)\right) \\
&= \mathbb{E}\left(\frac{\partial}{\partial x_i}(\log q(Y)) k(Y,\cdot) + \frac{\partial}{\partial x_i} k(Y,\cdot)\right) \\
&\quad + \mathbb{E}\left(\frac{\partial}{\partial x_i}(\log p(Y) - \log q(Y)) k(Y,\cdot)\right).
\end{aligned}$$

We have used Lemma 5.1 to see that

$$\mathbb{E}\left(\frac{\partial}{\partial x_i}(\log q(Y)) k(Y,\cdot) + \frac{\partial}{\partial x_i} k(Y,\cdot)\right) = 0.$$

We recognise that the expected value of $\frac{\partial}{\partial x_i}(\log p(Y) - \log q(Y)) k(Y,\cdot)$ is the mean embedding of a function $g(y) = \frac{\partial}{\partial x_i}\left(\log \frac{p(y)}{q(y)}\right)$ with respect to the measure q. By assumption (ii) Function g is square integrable ($\mathbb{E}(\frac{\partial}{\partial x_i}\log p(Y))^2 \leq \infty$ and $\mathbb{E}(\frac{\partial}{\partial x_i}\log q(Y))^2 \leq \infty$) . Therefore, since the kernel k is cc-universal, by Carmeli et al. (2010, Theorem 4.4 c) this embedding is zero if and only if $g = 0$, which implies that

$$\nabla \log \frac{p(y)}{q(y)} = (0, \cdots, 0).$$

A constant vector field of derivatives can only be generated by a constant function, so $\log \frac{p(y)}{q(y)} = C$, for some C, which implies that $p(y) = e^C q(y)$. Since p and q both integrate to one, $C = 0$ and so $p = q$ – a contradiction. □

3.2. Wild Bootstrap Testing

The two concepts required to derive the distribution of the test statistic are: τ-mixing (Dedecker et al., 2007; Leucht & Neumann, 2013), and V-statistics Serfling (1980).

τ-mixing is a notion of dependence within the observations, weak enough for most practical applications. Trivially, i.i.d. observations are τ-mixing. As for the Markov Chains, whose convergence we study in the experiments, the property of geometric ergodicity implies τ-mixing (given that the stationary distribution has a finite moment of some order: see Appendix B of Chwialkowski & Gretton (2014)). For further details on τ-mixing, see Dedecker & Prieur (2005); Dedecker et al. (2007). For this work we will assume a technical condition $\sum_{t=1}^{\infty} t^2 \sqrt{\tau(t)} \leq \infty$.

A direct application of Theorem 2.1 (Leucht, 2012) characterizes the limiting behavior of nV_n for τ-mixing processes,

Proposition 3.5. Under the null hypothesis nV_n converges weakly to some distribution.

The proof, which is a simple verification of the assumptions, can be found in the Appendix. Although a formula for a limit distribution of V_n can be derived explicitly (Theorem 2.1 (Leucht, 2012)), we do not provide it here. To our knowledge there are no methods of obtaining quantiles of a limit of V_n in closed form. The common solution is to estimate quantiles by a resampling method, as described in Section 2. The validity of this resampling method is guaranteed by the following proposition (which follows from Theorem 2.1 (Leucht, 2012) and modification of the Lemma 5 Chwialkowski et al. (2014)) , proved in the supplement.

Proposition 3.6. Let $f(W_{1,n}, \cdots, W_{t,n}) = \sup_x |P(nB_n > x | W_{1,n}, \cdots, W_{t,n}) - P(nV_n > x)|$ be a difference between quantiles. Under the null hypothesis, $f(W_{1,n}, \cdots, W_{t,n})$ converges to zero in probability. Under the alternative hypothesis, B_n converges to zero, while V_n converges to a positive constant.

As a consequence, if the null hypothesis is true, we can approximate any quantile; while under the alternative hypothesis, all quantiles of B_n collapse to zero while $P(V_n > 0) \to 1$.

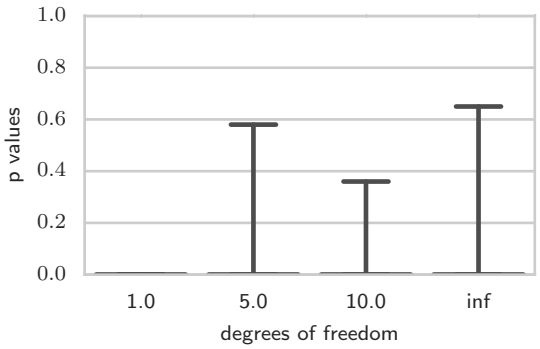

Figure 1. Large autocovariance, unsuitable bootstrap. The parameter a_n is too large and the bootstrapped V-statistics B_n are, on average, too low. Therefore it is very likely that $V_n > B_n$ and the test is too conservative.

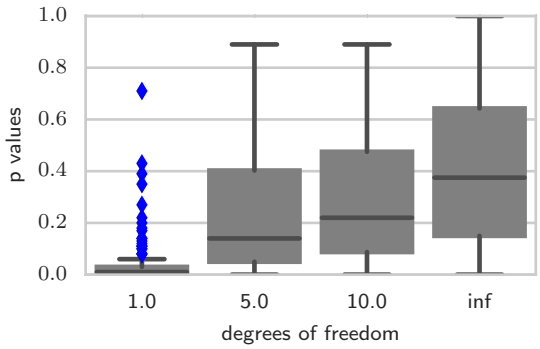

Figure 2. Large autocovariance, suitable bootstrap. The parameter $a_n$is chosen suitably, but due to a large autocorrelation withing the samples, the power of the test is small (effective sample size is small).

4. Experiments

We provide a number of experimental applications for our test. We begin with a simple check to establish correct test calibration on non-i.i.d. data, followed by a demonstration of statistical model criticism for Gaussian Process (GP) regression. We then apply the proposed test to quantify bias-variance trade-offs in MCMC, and demonstrate how to use the test to verify whether MCMC samples are drawn from the desired stationary distribution. In the final experiment, we move away from the MCMC setting, and use the test to evaluate the convergence of a nonparametric density estimator.

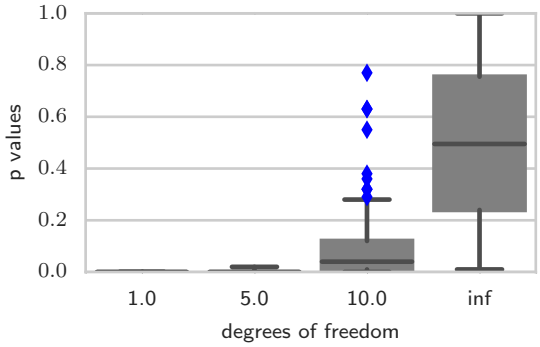

Figure 3. Thinned sample, suitable bootstrap. Most of the autocorrelation within the sample is canceled by thinning. To guarantee that the remaining autocorrelation is handled properly, the flip probability is set at 0.1.

STUDENT'S T VS NORMAL

In our first task, we modify experiment 4.1 from Gorham & Mackey 2015. The null hypothesis is that the observed samples come from a standard normal distribution. We study the power of the test against samples from a Student's t distribution. We expect to observe low p-values when testing against a Student's t distribution with few degrees of freedom. We considered 1, 5, 10 or ∞ degrees of freedom, where ∞ is equivalent to sampling from a standard normal distribution. For a fixed number of degrees of freedom we drew 1400 samples and calculated the p-value. This procedure was repeated one hundred times, and the bar plots of p-values are shown in Figures 1,2,3.

The twist on the original experiment 4.1 by Gorham & Mackey 2015 is that in our case, the draws from the Student's t distribution were given temporal correlation. The samples were generated using a Metropolis–Hastings algorithm, with a Gaussian random walk (variance equal to 0.5). We emphasize the need for an appropriate choice of the wild bootstrap process parameter, a_n, which indicates the probability of a sign flip. In Figure 1 we plot p-values for a_n being set to 0.5. Such a high value of a_n is suitable for iid observations, but results in p-values that are too conservative for temporally correlated observations. In Figure 2, $a_n = 0.02$, which gives a well calibrated distribution of the p-values under the null hypothesis (see box plot for an infinite number degrees of freedom), however the power of the test is reduced. Indeed, p-values for five degrees of freedom are already large. The solution that we recommend is a mixture of thinning and adjusting a_n, as presented in the Figure 3. We have thinned the observations by a factor of 20 and set $a_n = 0.1$, thus preserving both good statistical power and correct calibration of p-values under the null hypothesis.

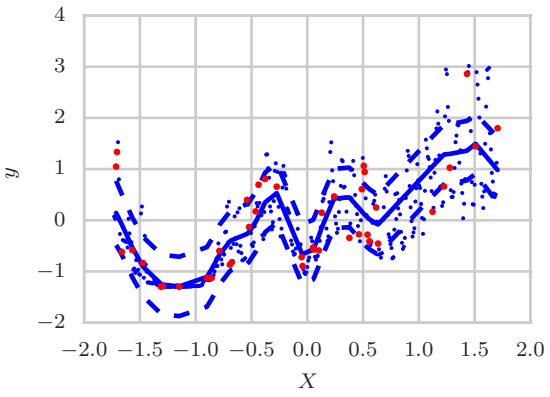

Figure 4. Fitted GP and data used to fit (blue) and to apply test (red).

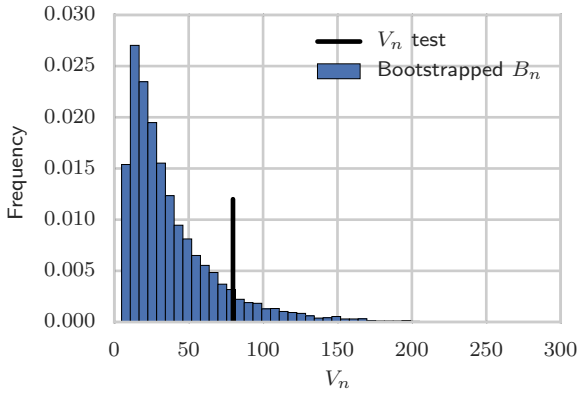

Figure 5. Bootstrapped B_n distribution with the test statistic V_n marked.

Statistical Model Criticism on Gaussian Processes

We next apply our test to the problem of statistical model criticism for GP regression. Our presentation and approach are similar to the non i.i.d. case of Lloyd & Ghahramani (2015, Section 6). We used the Solar dataset, consisting of a 1D regression problem with $N = 402$ pairs (X, y). We fit $N_{\text{train}} = 361$ data using a GP with a squared exponential kernel and a Gaussian noise model, and performed standard maximum likelihood II on the hyperparameters (length-scale, overall scale, noise-variance). We then applied our test to the remaining $N_{\text{test}} = 41$ data. Our test attempts to falsify the null hypothesis that the Solar dataset was generated from the predictive distribution (conditioned on training data and predicted position) of the GP. Lloyd & Ghahramani (2015) refer to this setup as non i.i.d., since the predictive distribution is a different univariate Gaussian for every predicted point. Note that in contrast to their MMD-based method, our test does *not* need to simulate from the multiple predictive distributions. Our particular $N_{\text{train}}, N_{\text{test}}$ were chosen to make sure the GP fit has stabilised, i.e. adding more data did not cause further model refinement. Figure 4 shows training and testing data, and the fitted GP. Clearly, the Gaussian noise model is a poor fit for this particular dataset, e.g. around $X = -1$. Figure 5 shows the distribution over $D = 10000$ bootstrapped V-statistics B_n with $n = N_{\text{test}}$. The test statistic lies in an upper quantile of the bootstrapped null distribution, indicating (correctly) that it is unlikely the test points were generated by the fitted GP model, even for the low number of test data observed, $N_{\text{test}} = 41$.

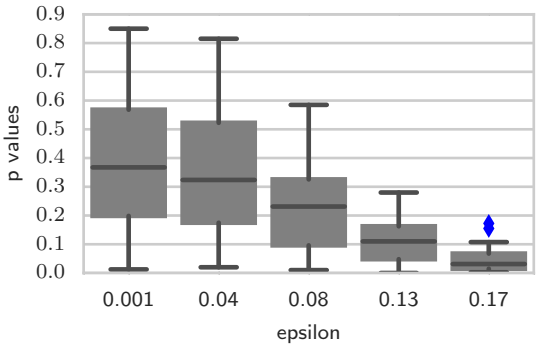

Figure 6. Distribution of p-values as a function of ϵ for austerity MCMC.

Approximate MCMC algorithm

We show how to quantify bias-variance trade-offs in an approximate MCMC algorithm – austerity MCMC (Korat-

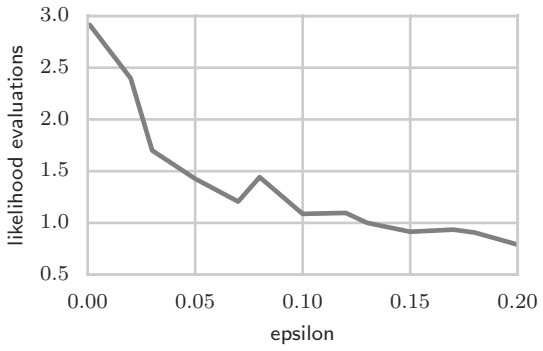

Figure 7. Average number of likelihood evaluations a function of ϵ for austerity MCMC (the y-axis is in millions of evaluations).

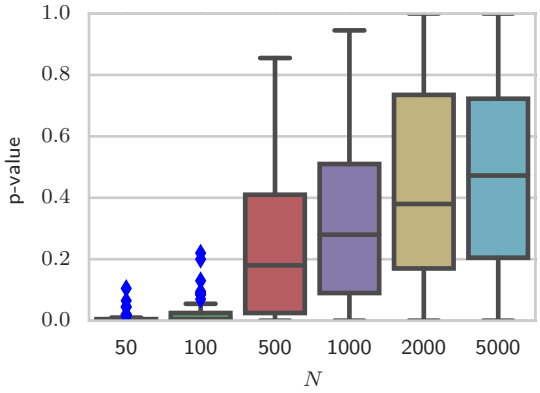

Figure 8. Density estimation: P-values for an increasing number of data N for the non-parametric model.

tikara et al., 2013). For the purpose of illustration we use a simple generative model from Gorham & Mackey (2015); Welling & Teh (2011b),

$$\theta_1 \sim N(0, 10); \theta_2 \sim N(0, 1)$$
$$X_i \sim \frac{1}{2}N(\theta_1, 4) + \frac{1}{2}N(\theta_2, 4).$$

Austerity MCMC is a Monte Carlo procedure designed to reduce the number of likelihood evaluation in the acceptance step of the Metropolis-Hastings algorithm. The crux of method is to look at only a subset of the data, and make an acceptance/rejection decision based on this subset. The probability of making a wrong decision is proportional to a parameter $\epsilon \in [0, 1]$. This parameter influences the time complexity of Austerity MCMC: when ϵ is larger, i.e., when there is a greater tolerance for error, the expected computational cost is lower. We simulated $\{X_i\}_{1 \leq i \leq 400}$ points from the model with $\theta_1 = 0$ and $\theta_2 = 1$. In this setting there were two modes in the posterior distribution: one at $(0, 1)$ and the other at $(1, -1)$. We ran the Austerity algorithm with ϵ varying over the range $[0.001, 0.2]$. For each ϵ we calculated an individual thinning factor, such that correlation between consecutive samples from the chains was smaller than 0.5 (greater ϵ generally required more thinning). For each ϵ we tested the hypothesis that $\{\theta_i\}_{1 \leq i \leq 500}$ were drawn from the true stationary posterior, using our goodness of fit test. We generated 100 p-values for each ϵ , as shown in Figure 6. It is clear that $\epsilon = 0.09$ yields a good approximation of the true stationary distribution, while being parsimonious in terms of likelihood evaluations, as shown in Figure 7.

Convergence in non-parametric density estimation

In our final experiment, we apply our goodness of fit test to measuring quality-of-fit in nonparametric density estimation. We evaluate two density models: the infinite dimensional exponential family (Sriperumbudur et al., 2014), and a recent approximation to this model using random Fourier features (Strathmann et al., 2015). Our implementation of the model assumes the log density to take the form $f(x)$, where f lies in an RKHS induced by a Gaussian kernel with bandwidth 1. We fit the model using N observations drawn from a standard Gaussian, and performed our quadratic time test on a separate evaluation dataset of fixed size, $N_{\text{test}} = 500$. Our goal was to identify N sufficiently large that the goodness of fit test did not reject the null hypothesis (i.e., the model had learned the density sufficiently well, bearning in mind that it is guaranteed to converge for sufficiently large N). Figure 8 shows how the distribution of p-values evolves as a function of N; this distribution is uniform for $N = 5000$, but at $N = 500$, the null hypothesis would very rarely be rejected.

We next consider the random fourier feature approximation to this model, where the log pdf, f, is approximated using a finite dictionary of random Fourier features (Rahimi & Recht, 2007). The natural question when using this approximation is: "How many random features do it I need?" Using the same test power $N_{\text{test}} = 500$ as above, and a large number of available samples $N = 5 \cdot 10^4$, Figure 9 shows the distributions of p-values for an increasing number of random features m. From about $m = 50$, the null hypothesis would rarely be rejected, given a reasonable choice of test level. Note, however, that the p-values do *not* have a uniform distribution, even for a large number of random features. This subtle effect is caused by over-smoothing due to the regularisation approach taken in (Strathmann et al., 2015, KMC finite), which would not otherwise have been detected.

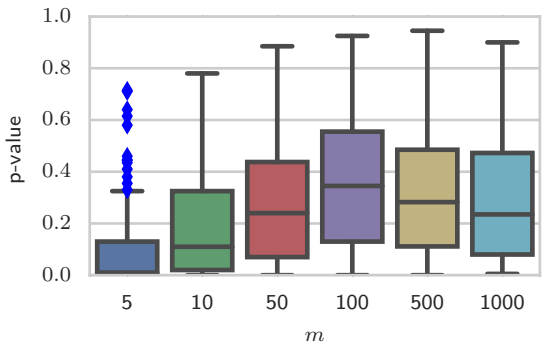

Figure 9. Approximate density estimation: P-values for an increasing number of random features m.

References

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

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up Markov Chain Monte Carlo: an adaptive subsampling approach. In *ICML*, pp. 405–413, 2014.

Baringhaus, L. and Henze, N. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35:339–348, 1988.

Barron, A. R. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17:107–124, 1989.

Beirlant, J., Györfi, L., and Lugosi, G. On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318, 1994.

Bowman, A.W. and Foster, P.J. Adaptive smoothing and density based tests of multivariate normality. *J. Amer. Statist. Assoc*, 88:529–537, 1993.

Carmeli, Claudio, De Vito, Ernesto, Toigo, Alessandro, and Umanitá, Veronica. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Chwialkowski, K. and Gretton, A. A kernel independence test for random processes. In *ICML*, 2014.

Chwialkowski, Kacper, Ramdas, Aaditya, Sejdinovic, Dino, and Gretton, Arthur. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pp. 1972–1980, 2015.

Chwialkowski, Kacper P, Sejdinovic, Dino, and Gretton, Arthur. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pp. 3608–3616, 2014.

Dedecker, J., Doukhan, P., Lang, G., Louhichi, S., and Prieur, C. *Weak dependence: with examples and applications*, volume 190. Springer, 2007.

Dedecker, Jérôme and Prieur, Clémentine. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

Fromont, M., Laurent, B, Lerasle, M, and Reynaud-Bouret, P. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *COLT*, pp. 23.1–23.22, 2012.

Gelman, A. and Rubin, D.B. Inference from iterative simulation using multiple sequences. *Statistical science*, pp. 457–472, 1992.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *NIPS*, pp. 226–234, 2015.

Gretton, A. and Gyorfi, L. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

Gretton, A., Fukumizu, K., Teo, C, Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. In *NIPS*, volume 20, pp. 585–592, 2007.

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

Györfi, L. and Vajda, I. Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 56:57–67, 2002.

Györfi, L. and van der Meulen, E. C. A consistent goodness of fit test based on the total variation distance. In Roussas, G. (ed.), *Nonparametric Functional Estimation and Related Topics*, pp. 631–645. Kluwer, Dordrecht, 1990.

Justel, A., na, D. Pe and Zamar, R. A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics and Probability Letters*, 35(3):251–259, 1997.

Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91, 1933.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. In *ICML*, pp. 181–189, 2014.

Korattikara, Anoop, Chen, Yutian, and Welling, Max. Austerity in mcmc land: Cutting the metropolis-hastings budget. *arXiv preprint arXiv:1304.5299*, 2013.

Leucht, A. Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli*, 18(2):552–585, 2012.

Leucht, A. and Neumann, M.H. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2013.03.003. URL http://www.sciencedirect.com/science/article/pii/S0047259X13000304.

Lloyd, James R and Ghahramani, Zoubin. Statistical model criticism using kernel two sample tests. In *NIPS*, pp. 829–837, 2015.

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

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184, 2007.

Rizzo, M. L. New goodness-of-fit tests for pareto distributions. *ASTIN Bulletin: Journal of the International Association of Actuaries*, 39(2):691–715, 2009.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41 (5):2263–2702, 2013.

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

Shao, X. The dependent wild bootstrap. *J. Amer. Statist. Assoc.*, 105(489):218–235, 2010.

Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

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

Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., and Hyvärinen, A. Density Estimation in Infinite Dimensional Exponential Families. *arXiv preprint arXiv:1312.3516*, 2014.

Stein, Charles. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pp. 583–602, Berkeley, Calif., 1972. University of California Press. URL http://projecteuclid.org/euclid.bsmsp/1200514239.

Steinwart, I. and Christmann, A. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008. ISBN 978-0-387-77241-7.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In *NIPS*, 2015.

Székely, G. J. and Rizzo, M. L. A new test for multivariate normality. *J. Multivariate Analysis*, 93(1):58–80, 2005.

Welling, M. and Teh, Y.W. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *ICML*, pp. 681–688, 2011a.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011b.

Appendix

5. Proofs

Lemma 5.1. *If a random variable X is distributed according to p, then for all $f \in \mathcal{F}$, the expected value of T is zero, i.e.* $\mathbb{E}(Tf)(X) = 0$.

Proof. First, we show that the functions $g_i := pf_i$ vanish at infinity, i.e. for all dimensions j,

$$\lim_{x_j \to \infty} g_i(x_1, \cdots, x_d) = 0.$$

The density function p vanishes at infinity. The function f is bounded, which is implied by the Cauchy-Schwarz inequality – $|f(x)| \leq \|f\| \sqrt{k(x,x)}$. This implies that the function g vanishes at infinity. To show that the expected value $\mathbb{E}(T)f(X)$ is zero, it is sufficient to show that for all dimensions i, the expected value of $\frac{\partial \log p(X)}{\partial x_i} f_i(X) + \frac{\partial f_i(X)}{\partial x_i}$ is zero,

$$\begin{aligned}
&\mathbb{E}\left(\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i}\right) \\
&= \int_{R_d} \left[\frac{\partial \log p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i}\right] q(x) dx \\
&= \int_{R_d} \left[\frac{1}{p(x)} \frac{\partial q(x)}{\partial x_i} f(x) + \frac{\partial f(x)}{\partial x_i}\right] q(x) dx \\
&= \int_{R_d} \left[\frac{\partial p(x)}{\partial x_i} f_i(x) + \frac{\partial f_i(x)}{\partial x_i} q(x)\right] dx \\
&\overset{(a)}{=} \int_{R_{d-1}} \left(\lim_{R \to \infty} p(x) f_i(x) \Big|_{x_i=-R}^{x_i=R}\right) dx \\
&= \int_{R_{d-1}} 0 dx = 0
\end{aligned}$$

where in (a) we used integration by parts and that g_i vanishes at infinity. □

Proof of proposition 3.5. We check assumptions of the Theorem 2.1 (Leucht, 2012). The condition A1, $\sum_{t=1}^{\infty} \sqrt{\tau(t)} \leq \infty$, is implied by assumption $\sum_{t=1}^{\infty} t^2 \sqrt{\tau(t)} \leq \infty$ in Section 3. Condition A2 (iv), Lipschitz continuity of h, follows from assumption (iv). Conditions A2 i), ii) positive definiteness, symmetry and degeneracy of h follow from the proof of Theorem (2.1). Indeed

$$h(x,y) = \langle [\nabla \log p(x) k(x,\cdot) + \nabla_1 k(x,\cdot)], [\nabla \log p(y) k(y,\cdot) + \nabla_1 k(y,\cdot)] \rangle_{\mathcal{F}^d}$$

so the statistic is an inner product and hence positive definite. Degeneracy under the null follows from the fact that for any t, by Lemma (5.1),$\mathbb{E}(\nabla \log p(x) k(x,t) + \nabla_1 k(x,t)) = 0$. Finally, condition A2 (iii), $\mathbb{E}h(X,X) \leq \infty$ follows from assumptions (ii), (iii) and boundedness of the kernel. □

Proof of proposition 3.6. We use Theorem 2.1 (Leucht, 2012) to see that, under the null hypothesis, $f(W_{1,n}, \cdots, W_{t,n})$ converges to zero in probability. We have checked assumptions A1,A2 in the proof of the proposition 3.5. Assumption B1 is identical to our assumption from Section 3 . Finally we check assumption B2 (bootstrap assumption): $\{W_{t,n}\}_{1 \leq t \leq n}$ is a row-wise strictly stationary triangular array independent of all Z_t such that $\mathbb{E}W_{t,n} = 0$ and $\sup_n \mathbb{E}|W_{t,n}^{2+\sigma}| = 1 < \infty$ for some $\sigma > 0$. The auto-covariance of the process is given by $\mathbb{E}W_{s,n}W_{t,n} = (1-2p_n)^{-|s-t|}$, so the function $\rho(x) = \exp(-x)$, and $l_n = \log(1-2p_n)^{-1}$. We verify that $\lim_{u \to 0} \rho(u) = 1$. If we set $p_n = w_n^{-1}$, such that $w_n = o(n)$ and

$\lim_{n\to\infty} w_n = \infty$, then $l_n = O(w_n)$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = \frac{1-(1-2p_n)^{n+1}}{p_n} = O(w_n) = O(l_n)$. Finally we show that B_n converges to zero under the alternative. It is sufficient to check that $\mathbb{E}B_n \to 0, \mathbb{E}B_n^2 \to 0$.

$$\begin{aligned}
\mathbb{E}B_n &= \frac{1}{n^2}\sum_{i,j} \mathbb{E}W_i W_j \mathbb{E}h(Z_i, Z_j) \\
&= \frac{1}{n^2}\sum_{i\in N^m} \rho(|j-i|/l_n)\mathbb{E}h(Z_j, Z_i) \\
&\leq \frac{1}{n^2}\sum_{i\in N^m} \rho(|j-i|/l_n)C \\
&\to 0
\end{aligned}$$

for some constant $C = \mathbb{E}h(Z_1, Z_2)$, whose existence follows from assumptions i) and iii). As for $\mathbb{E}B_n^2$, we have

$$\begin{aligned}
\mathbb{E}B_n^2 &= \frac{1}{n^4}\sum_{i,j,k,l} \mathbb{E}W_i W_j W_k W_l \mathbb{E}h(Z_i, Z_j)h(Z_k, Z_l) \\
&\leq \frac{1}{n^4}\sum_{i\neq j, i\neq k, i\neq l, j\neq k, j\neq l, k\neq l} \mathbb{E}W_i W_j W_k W_l \mathbb{E}h(Z_i, Z_j)^2 \mathbb{E}h(Z_k, Z_l)^2 + C'\frac{6n^3}{n^4} \\
&\leq \frac{1}{n^4}\sum_{i\neq j, i\neq k, i\neq l, j\neq k, j\neq l, k\neq l} \mathbb{E}W_i W_j W_k W_l C' + \frac{6C'}{n} \\
&= \frac{6C'}{n} \to 0,
\end{aligned}$$

where $C' = \mathbb{E}h(Z_i, Z_j)^2 \mathbb{E}h(Z_k, Z_l)^2$. □

5.1. Linear time test

We may use similar reasoning for the quadratic time test to define a linear time test, based on the two-sample test of Chwialkowski et al. (2015). For some fixed location y and a random variable X, define a random variable $s(X, y)$ as

$$s(X, y) = \nabla \log p(X) g(X, y) - \nabla g(X, y). \tag{2}$$

For some number of random locations Y_1, Y_J and a random variable X define a random vector Z_i

$$Z_i = (s(X_i, Y_1), \cdots, s(X_i, Y_J)) \in \mathbf{R}^J. \tag{3}$$

Let W_n be a mean of Z_i's $W_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and Σ_n its covariance matrix $\Sigma_n = \frac{1}{n} Z Z^T$. The test statistic is

$$S_n = n W_n \Sigma_n^{-1} W_n. \tag{4}$$

The computation of S_n requires inversion of a $J \times J$ matrix Σ_n, but this is fast and numerically stable: J will typically be small, and is less than 10 in our experiments. The next proposition demonstrates the use of S_n as a one-sample test.

Proposition 5.2 (Asymptotic behavior of S_n)**.** *If $\mathbb{E}s(X, y) = 0$ for all y, then the statistic S_n is a.s. asymptotically distributed as a χ^2-random variable with Jd degrees of freedom, where d is X dimensionality (as $n \to \infty$ with d fixed). If $\mathbb{E}s(X, y) \neq 0$ for almost all y then a.s. for any fixed r, $\mathbb{P}(S_n > r) \to 1$ as $n \to \infty$.*

One sample test Calculate S_n. Choose a threshold r_α corresponding to the $1-\alpha$ quantile of a χ^2 distribution with J degrees of freedom, and reject the null hypothesis whenever S_n is larger than r_α.