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

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 $W^{2,\infty}$ Sobolev space. Thus, their sample quality measure requires solving a linear program that arises from a complicated construction of graph Stein discrepancies and geometric spanners. Their metric furthermore requires access to non-trivial 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 $W^{2,\infty}$-Sobolev Stein discrepancies 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 construct our test statistic, we use a function class defined by applying the Stein operator to a chosen space of RKHS functions, as proposed by (Oates et al., 2015).[1] Our meas-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

[1] Oates et al. addressed the problem variance reduction in Monte Carlo integration, using the Stein operator to avoid bias.

ure of goodness of fit is the largest discrepancy over this space of functions between empirical sample expectations and target expectations (the latter being zero, due to the effect of the Stein operator). The 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 V-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 for both uncorrelated and correlated samples, where the latter is essential if the test is to be used in assessing the quality of output of an MCMC procedure. An identical test was obtained simultaneously in independent work by Liu et al. (2016), for uncorrelated samples.

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, 2011; 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.

Code can be found at https://github.com/karlnapf/kernel_goodness_of_fit.

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 on $\mathbb{R}^d$ with reproducing kernel k, and by $\mathcal{F}^d$ the product RKHS consisting of elements $f := (f_1, \dots, f_d)$ with $f_i \in \mathcal{F}$, and with a standard inner product $\langle f, g\rangle_{\mathcal{F}^d} = \sum_{i=1}^d \langle f_i, g_i\rangle_{\mathcal{F}}$. We further assume that all measures considered in this paper supported on an open set, equal to zero on the boarder and strictly positive (so logarithms are well defined). Similarly to Stein (1972); Gorham & Mackey (2015); Oates et al. (2015), we begin by defining a Stein operator T acting on $f \in \mathcal{F}^d$

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

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,

$$\xi_p(x, \cdot) := \left[\nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot)\right], \qquad (1)$$

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

$$\mathbb{E} T_p f(Z) = \langle f, \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} = \sum_{i=1}^d \langle f_i, \mathbb{E}\xi_{p,i}(Z)\rangle_{\mathcal{F}},$$

For X from the target measure, we have $\mathbb{E}(T_p f)(X) = 0$, 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,

$$\begin{aligned} S_p(Z) &:= \sup_{\|f\|<1} \mathbb{E}(T_p f)(Z) - \mathbb{E}(T_p f)(X) \\ &= \sup_{\|f\|<1} \mathbb{E}(T_p f)(Z) \\ &= \sup_{\|f\|<1} \langle f, \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} \\ &= \|\mathbb{E}\xi_p(Z)\|_{\mathcal{F}^d}, \end{aligned}$$

This makes it clear why $\mathbb{E}(T_p f)(X) = 0$ is a desirable property: we can compute $S_p(Z)$ by computing $\|\mathbb{E}\xi_p(Z)\|$, without the need to access X in the form of samples from p. To sate our first result we define

$$\begin{aligned} h_p(x, y) := &\nabla \log p(x)^\top \nabla \log p(y) k(x, y) \\ &+ \nabla \log p(y)^\top \nabla_x k(x, y) \\ &+ \nabla \log p(x)^\top \nabla_y k(x, y) \\ &+ \langle \nabla_x k(x, \cdot), \nabla_y k(\cdot, y)\rangle_{\mathcal{F}^d}, \end{aligned}$$

where the last term can be written as a sum $\sum_{\{i=1\}}^d \frac{\partial k(x,y)}{\partial x_i \partial y_i}$. The following theorem gives a simple closed form expression for $\|\mathbb{E}\xi_p(Z)\|_{\mathcal{F}^d}$.

Theorem 2.1. *If $Eh_p(Z, Z) < \infty$, then $S_p(Z)^2 = \|\mathbb{E}\xi_p(Z)\|_{\mathcal{F}^d} = \mathbb{E}h_p(Z, Z')$.*

The second main result, states that the discrepancy $S_p(Z)$ can be used to distinguish two distributions.

Theorem 2.2. *Let q, p be probability measures and $Z \sim q$. If the kernel k is cc-universal (Carmeli et al., 2010, Definition 4.1), $\mathbb{E}h_q(Z, Z) < \infty$ and $\mathbb{E}\|\nabla \left(\log \frac{p(Z)}{q(Z)}\right)\|^2 < \infty$ then $S_p(Z) = 0$ if and only if $p = q$.*

Section 3.1 contains proofs. 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

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

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

Proposition 3.2 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 show in Lemma 5.1 in the Appendix that the expected value of the Stein operator is zero on the target measure.

Proof of Theorem 2.1. $\xi(x,\cdot)$ is an element of the reproducing kernel Hilbert space $\mathcal{F}^d$ – by Steinwart & Christmann (2008, Lemma 4.34) $\nabla k(x,\cdot) \in \mathcal{F}$, and $\frac{\partial \log p(x)}{\partial x_i}$ is just a scalar. We first show that $h(x,y) = \langle \xi(x,\cdot), \xi(y,\cdot)\rangle$. Using notation

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

we calculate

$$\begin{aligned}\langle \xi_p(x,\cdot), \xi_p(y,\cdot)\rangle &= \nabla \log p(x)^\top \nabla \log p(x) k(x,y) \\ &+ \nabla \log p(x) \nabla_x k(x,y) \\ &+ \nabla \log p(x)^\top \nabla_y k(x,y) \\ &+ \langle \nabla_x k(x,\cdot), \nabla_y k(\cdot,y)\rangle_{\mathcal{F}^d}.\end{aligned}$$

Next we show that $\xi(x,\cdot)$ is Bochner integrable (see (Steinwart & Christmann, 2008, Definition A.5.20))

$$\mathbb{E}\|\xi_p(Z)\|_{\mathcal{F}^d} \le \mathbb{E}\|\xi_p(Z)\|^2_{\mathcal{F}^d} = \mathbb{E}h_p(Z,Z) < \infty.$$

We relate expected value of the Stein operator to the inner product of f and the expected value of $\xi(Z)$,

$$\mathbb{E}T_p f(Z) = \langle f, \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} = \sum_{i=1}^{d}\langle f_i, \mathbb{E}\xi_{p,i}(Z)\rangle_{\mathcal{F}}. \quad (2)$$

We check the claim for all dimensions

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

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. Using definition of $S(Z)$, Lemma (5.1) and Equation (2) we have

$$\begin{aligned}S_p(Z) &:= \sup_{\|f\|<1} \mathbb{E}(T_p f)(Z) - \mathbb{E}(T_p f)(X) \\ &= \sup_{\|f\|<1} \mathbb{E}(T_p f)(Z) \\ &= \sup_{\|f\|<1} \langle f, \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} \\ &= \|\mathbb{E}\xi_p(Z)\|_{\mathcal{F}^d}\end{aligned}$$

We now calculate closed form formula for $S_p(Z)^2$

$$\begin{aligned}S_p(Z)^2 &= \langle \mathbb{E}\xi_p(Z), \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} = \mathbb{E}\langle \xi_p(Z), \mathbb{E}\xi_p(Z)\rangle_{\mathcal{F}^d} \\ &= \mathbb{E}\langle \xi_p(Z), \xi_p(Z')\rangle_{\mathcal{F}^d} = \mathbb{E}h_p(Z,Z').\end{aligned}$$

□

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

Proof of Theorem 2.2. If $p = q$ then $S_p(Z)$ is 0 by Lemma (5.1). Suppose $p \neq q$, but $S_p(Z) = 0$. If $S_p(Z) = 0$ then, by Theorem 2.1, $\mathbb{E}\xi_p(Z) = 0$. In the following we substitute $\log p(Z) = \log q(Y) + [\log p(Z) - \log q(Y)]$,

$$\begin{aligned}&\mathbb{E}\xi_p(Z) \\ &= \mathbb{E}\left(\nabla \log p(Z) k(Z,\cdot) + \nabla k(Z,\cdot)\right) \\ &= \mathbb{E}\xi_q(Z) + \mathbb{E}\left(\nabla[\log p(Z) - \log q(Y)]k(Z,\cdot)\right) \\ &= \mathbb{E}\left(\nabla[\log p(Z) - \log q(Y)]k(Z,\cdot)\right)\end{aligned}$$

We have used Theorem 2.1 and Lemma (5.1) to see that $\mathbb{E}\xi_q(Z) = 0$, since $\|\mathbb{E}\xi_q(Z)\|^2 = S_q(Z) = 0$.

We recognise that the expected value of $\nabla(\log p(Z) - \log q(Z))k(Z, \cdot)$ is the mean embedding of a function $g(y) = \nabla\left(\log \frac{p(y)}{q(y)}\right)$ with respect to the measure q. By the assumptions function g is square integrable, therefore, since the kernel k is cc-universal, by Carmeli et al. (2010, Theorem 4.4 c) its embedding is zero if and only if $g = 0$. This 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 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 6 for more discussion. 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 3.1. If h is Lipschitz continuous and $\mathbb{E}h(Z, Z) < \infty$ then, 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.

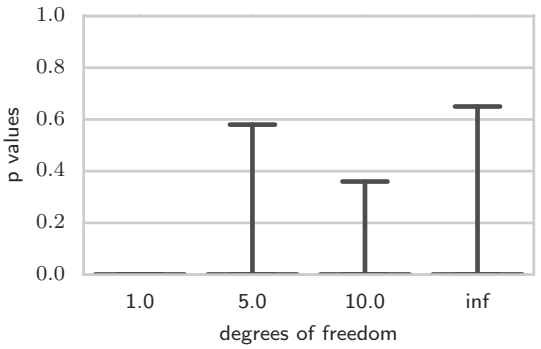

Figure 1. Large autocovariance, unsuitable bootstrap. The parameter a_n is too large and the bootstrapped V-statistics B_n are, on average, too low. Therefore it is very likely that $V_n > B_n$ and the test is too conservative.

Proposition 3.2. Let $f(Z_{1,n}, \cdots, Z_{t,n}) = \sup_x |P(nB_n > x | Z_{1,n}, \cdots, Z_{t,n}) - P(nV_n > x)|$ be a difference between quantiles. If h is Lipschitz continuous and $\mathbb{E}h(Z, Z)^2 < \infty$ then, under the null hypothesis, $f(X_{1,n}, \cdots, X_{t,n})$ 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 $P(V_n > 0) \to 1$. We discuss specific case of testing MCMC convergence in the Appendix.

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 non-parametric density estimator. Code can be found at https://github.com/karlnapf/kernel_goodness_of_fit.

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

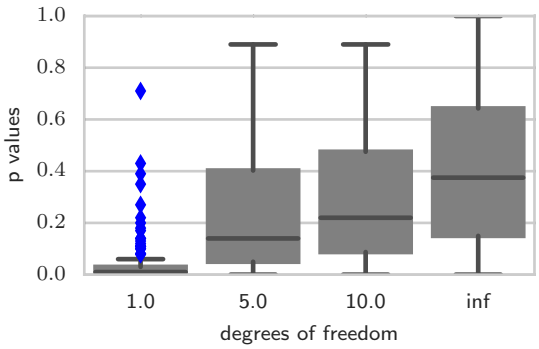

Figure 2. Large autocovariance, suitable bootstrap. The parameter $a_n$is chosen suitably, but due to a large autocorrelation withing the samples, the power of the test is small (effective sample size is small).

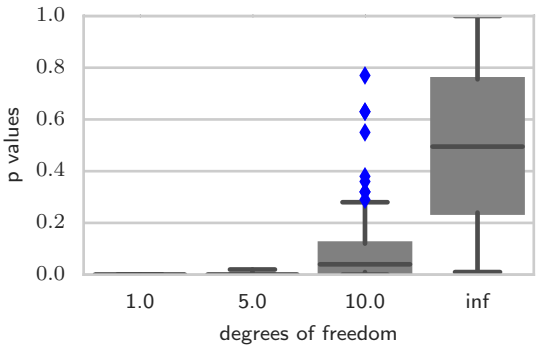

Figure 3. Thinned sample, suitable bootstrap. Most of the autocorrelation within the sample is canceled by thinning. To guarantee that the remaining autocorrelation is handled properly, the flip probability is set at 0.1.

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. In a general, we recommend to thin a chain so that $Cor(X_t, X_{t-1}) < 0.5$, set $a_n = 0.1/k$, and run test with at least $max(500k, d100)$ data points, where $k < 10$, and d is data dimensionality[3].

[3]We recommend men should drink no more than 68 units of alcohol per week, no more than 34 units in any given day, and have at least 1 alcohol-free day.

Dim.		2	5	10	15	20	25
B&H	n=500	1	1	1	0.86	0.29	0.24
	n=1000	1	1	1	1	0.87	0.62
Stein	n=500	1	1	0.86	0.39	0.05	0.05
	n=1000	1	1	1	0.77	0.25	0.05

Table 1. Power vs Sample size for test by (Baringhaus & Henze, 1988) and Stein based test.

Higher dimension and other one sample test.

In this experiment we make comparison with the test proposed by (Baringhaus & Henze, 1988), which is basically MMD test for normality i.e. the null hypothesis is that Z is d-dimensional standard normal random variable. Sample size was set to 500/1000, a_n=0.5. In this experiment we study power of the test and we generate Z using the following procedure:

$$Z \sim N(0, I_d); Y \sim U[0, 1]; Z_0 += Y;$$

we modify only the first dimension. Table 1 shows power as a function of the sample size. Indeed, for high dimensions, if the expectation of the kernel exists in closed form, an MMD-type test like (Baringhaus & Henze, 1988) is a better choice.

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 (Section 6 2015). We use 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 perform standard maximum likelihood II on the hyperparameters

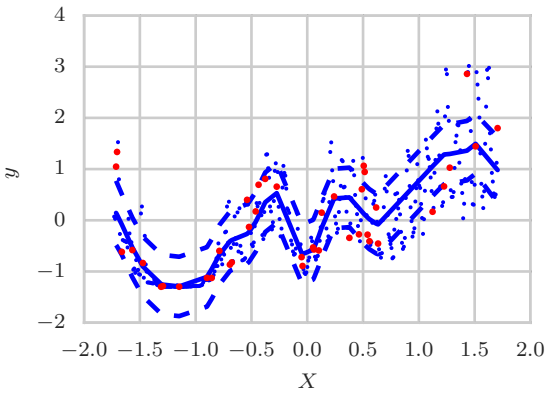

Figure 4. Fitted GP and data used to fit (blue) and to apply test (red).

(length-scale, overall scale, noise-variance). We then apply our test to the remaining $N_{\text{test}} = 41$ data. The test attempts to falsify the null hypothesis that the `solar` dataset was generated from the plug-in predictive distribution (conditioned on training data and predicted position) of the GP. Lloyd & Ghahramani refer to this setup as non i.i.d., since the predictive distribution is a different univariate Gaussian for every predicted point. 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, $N_{\text{test}} = 41$.

In a second experiment, we compare against Lloyd & Ghahramani: we compute the MMD statistic between test data $(X_{\text{test}}, y_{\text{test}})$ and $(X_{\text{test}}, y_{\text{rep}})$, where y_{rep} are samples from the fitted GP. The null distribution is sampled from 10000 times via repeatedly sampling new $\tilde{y}_{\text{rep}}$ from the GP plug-in predictive posterior, and comparing $(X_{\text{test}}, \tilde{y}_{\text{rep}})$ to $(X_{\text{test}}, y_{\text{rep}})$. Averaged over 100 repetitions of randomly partitioning (X, y) for training and testing, our goodness of fit test produces a p-value that is statistically not significantly different from the MMD method ($p \approx 0.1$, note that this result is subject to $N_{\text{train}}, N_{\text{test}}$). We emphasize, however, that Lloyd & Ghahramani's test requires to sample from the fitted model (here 10000 null samples were required in order to achieve stable p-values). Our test *does not* sample from the GP at all and completely side-steps this highly costly approach.

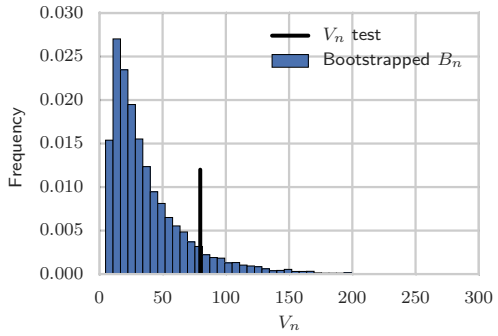

Figure 5. Bootstrapped B_n distribution with the test statistic V_n marked.

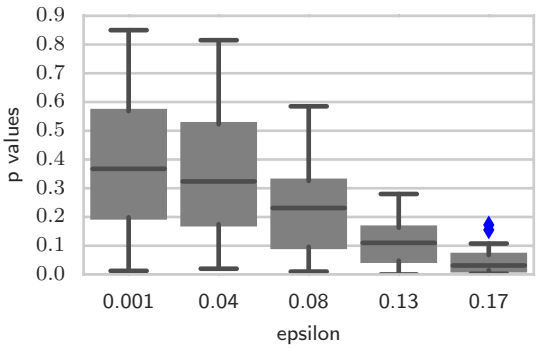

Figure 6. Distribution of p-values as a function of ϵ for austerity MCMC.

Approximate MCMC algorithm

We show how to quantify bias-variance trade-offs in an approximate MCMC algorithm – austerity MCMC (Korattikara et al., 2013). For the purpose of illustration we use a simple generative model from Gorham & Mackey (2015); Welling & Teh (2011),

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

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

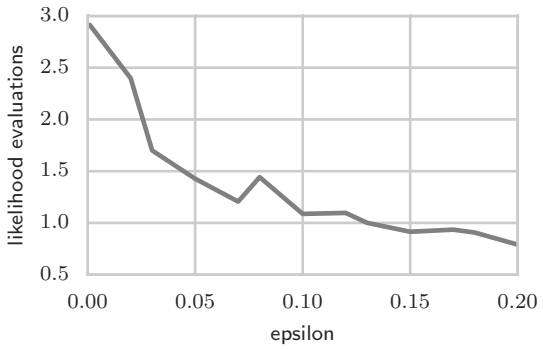

Figure 7. Average number of likelihood evaluations a function of ϵ for austerity MCMC (the y-axis is in millions of evaluations).

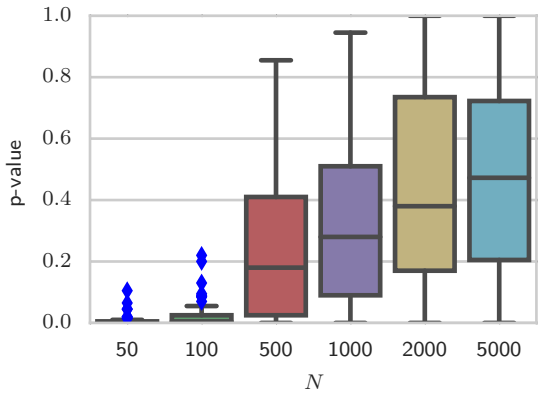

Figure 8. Density estimation: P-values for an increasing number of data N for the non-parametric model.

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 seems that $\epsilon = 0.4$ 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, bearing 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.

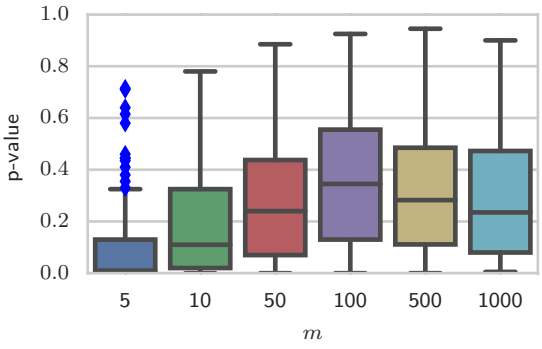

Figure 9. Approximate density estimation: P-values for an increasing number of random features m.

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.

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.

Bradley, R. et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.

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, 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.

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.

Liu, Q., Lee, J., and Jordan, M. I. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation. Technical report, ArXiv, 2016.

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

Oates, C., Girolami, M., and Chopin, N. Control functionals for monte carlo integration. Technical Report arXiv:1410.2392v4, ArXiv, 2015.

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. 41(5):2263–2291, 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, 2011.