

Testing goodness-of-fit via rate distortion

Peter Harremoës

Centrum voor Wiskunde en Informatica (CWI)

Amsterdam, NL-1090 GB

The Netherlands

P.Harremoes@cwi.nl

Abstract—A framework is developed using techniques from rate distortion theory in statistical testing. The idea is first to do optimal compression according to a certain distortion function and then use information divergence from the compressed empirical distribution to the compressed null hypothesis as statistic. Only very special cases have been studied in more detail, but they indicate that the approach can be used under very general conditions.

I. INTRODUCTION

There are many well-known examples of a fruitful interplay between information theory and statistics. It started with [1] and [2] and is well described in [3]. Information divergence or Kullback Leibler information plays a central role in measuring the distance between probability distributions. Statistical testing is often delicate if the sample size is small compared with size of the alphabet (sample space). If the alphabet is a continuous set the normal approach in statistics is to discretize the alphabet, but information is lost during discretization, and often it is not clear how one should discretize the space.

Rate distortion theory was developed as a theoretical framework for lossy compression. An obvious example is image compression, but rate distortion theory often fails for this kind of application for three reasons. First of all it is often very difficult to specify an appropriate distortion function. Secondly, the statistics of the source is often not known. Thirdly, in most cases it is impossible to calculate the rate distortion function exactly and even a numerical calculation may be very involved due to the number of variables.

Although rate distortion theory was developed for lossy compression we claim that the ideas are very useful for statistical analysis.

II. LIKELIHOOD RATIO TESTING

On a finite sample space of size k one can use information divergence as statistics for testing goodness of fit. This is the likelihood ratio test. We want to test a null hypothesis $H_0 : P = P_0$. An iid sample ω from P of size n is made and the null hypothesis is accepted if $D(\text{Emp}_n(\omega) \| P_0)$ is smaller than some value and rejected if it exceeds this value. The critical value is determined by the significance level. If P_0 is the uniform distribution then $H(\text{Emp}_n(\omega)) = \log k - D(\text{Emp}_n(\omega) \| P_0)$ so in this case it makes no difference whether one uses entropy or information divergence. Using large deviation theory one will see that no other test is more Bahadur efficient than the likelihood ratio test. The

distribution of $2nD(\text{Emp}_n(\omega) \| P_0)$ will converge to a χ^2 distribution with k degrees of freedom, so determining the values correspond to different significance levels is simple.

This method cannot be used directly if the sample space is infinite and P_0 is continuous. If P_0 is a distribution on \mathbb{R} with continuous distribution function F then a popular method for testing goodness of fit is the divide \mathbb{R} into k bins of equal probability. As one want to keep points together if they are close on the real axis the bins should be chosen of the form $[F^{-1}(\frac{i-1}{k}); F^{-1}(\frac{i}{k})[$. If f maps a point into its bin then P_0 is mapped into a uniform distribution so we can use the entropy $H(f(\text{Emp}_n(\omega)))$ as statistic to test goodness of fit. The idea is then to increase the number of bins slowly as n increases. Recently it was proved that entropy is more Bahadur efficient than other power statistics if k is increased so slowly that the mean number of samples per bin n/k tends to infinity for $n \rightarrow \infty$, see [4], [5] and references in there. This condition will hold if for instance $k = n^{1/2}$ and this choice of number of bins will also ensure that distribution of entropy will be asymptotically Gaussian.

It is easy to divide \mathbb{R} into k bins of equal probability for a continuous distribution but it is not obvious how to do the same for distributions on \mathbb{R}^2 or in higher dimensions. Even in one dimension it is far from obvious why the bins should be of equal probability. Maybe a different choice of bins would sometimes give a test that in one or another sense is more efficient. To get better founded criteria for how to choose bins we need a distortion function.

III. THE RATE DISTORTION TEST

Consider a distribution Q on a set Ω with a distortion function $d : \Omega \times \Omega \rightarrow \mathbb{R}$. For a distortion level d_0 the optimal coupling at distortion level d_0 is given by a Markov kernel $\Psi_{d_0} : \Omega \rightarrow M_+^1(\Omega)$. We shall use Ψ_{d_0} to smooth the empirical distribution so that we can compare it with the null hypothesis H_0 , i.e. we shall use $D(\Psi_{d_0}(\text{Emp}_n(\omega)) \| \Psi_{d_0}(Q))$ as statistic for testing goodness of fit. There are various ways to approximate $D(\Psi_{d_0}(\text{Emp}_n(\omega)) \| \Psi_{d_0}(Q))$ numerically. We shall not discuss this problem. In general the rate distortion function and Ψ_{d_0} cannot be calculated exactly but using iterative methods like the Arimoto Blahut algorithm they can be approximated. We shall discuss three examples where the rate distortion function and Ψ_{d_0} are given by explicit formulas.

Example 1 (Test of uniformity): We consider a set A with l elements. The set has no particular structure so we use

Hamming distortion as distortion function. Our null hypothesis is $P = U$ where u denotes the uniform distribution on A . In this case the Markov kernel Ψ_{d_0} has the form

$$\Psi_{d_0} : x \rightarrow \alpha \delta_x + (1 - \alpha)U$$

for some value $\alpha \in [0; 1]$ determined by d_0 . The Markov kernel maps the uniform distribution into the uniform distribution. Therefore the statistic of the rate distortion test has the form

$$D(\alpha \text{Emp}_n(\omega) + (1 - \alpha)U \| U).$$

This statistic is closely related to the idea of *local alternatives* often studied in statistics.

Example 2 (Normality test): We consider the real numbers with squared Euclidian distance as distortion function. Our null hypothesis is $P = \Phi$ where Φ denotes the standard Gaussian distribution. The optimal Markov kernel for the rate distortion problem sends x into the distribution of $\alpha x + (1 - \alpha^2)^{1/2} Z$ where Z is a standard Gaussian random variable. We see that the Gaussian distribution is mapped into it self. Thus the statistic of the rate distortion test is

$$D(\alpha X + (1 - \alpha^2)^{1/2} Z \| \Phi)$$

where we have identified the random variable

$$\alpha X + (1 - \alpha^2)^{1/2} Z$$

with its distribution. This Markov kernel can be rewritten as

$$\begin{aligned} D(\alpha X + (1 - \alpha^2)^{1/2} Z \| \Phi) \\ = D\left(X + \left(\frac{1}{\alpha^2} - 1\right)^{1/2} Z \| \Phi(0, \alpha^2)\right) \end{aligned}$$

so the Markov kernels essentially smooth data by adding an independent Gaussian random variable with variance $\alpha^{-2} - 1$. The idea of smoothing data is well-known in statistics.

Example 3 (Test of uniformity of angular data): In this example we consider data with values on the circle s_1 that we can identify with $\mathbb{R}/2\pi\mathbb{Z}$. See [6] for references. As distortion function we shall use $4 \cos^2\left(\frac{\theta_2 - \theta_1}{2}\right)$, i.e. squared Euclidean distance between points on a circle. We shall test the hypothesis $P = U$ where U denotes the uniform distribution on the circle. The optimal Markov kernel is a smoothing by adding a von Mises distribution

$$\frac{\exp(\kappa \cos(\theta))}{2\pi I_0(\kappa)}$$

where I_0 is the modified Bessel function of order 0 with parameter κ determined by the distortion level [7], [8]. The Markov kernel maps the uniform distribution into the uniform distribution.

IV. LIMITS FOR EXTREME VALUES OF β

Often the rate distortion curve is parametrized by its slope β . Here we shall discuss the effect of choosing very small or very large values of β when the sample is kept fixed. We shall go through our three main examples from this point of view.

Example 4 (Test of uniformity continued): Small or large values of β corresponds to small or large values of α . For $\alpha = 1$ we get the statistic

$$D(\text{Emp}_n(\omega) \| U)$$

which is the likelihood ratio test. For α close to 0 we use that information divergence is an f -divergence with $f(x) = x \ln x$ so that

$$\begin{aligned} \frac{d}{d\alpha} D(\alpha \text{Emp}_n(\omega) + (1 - \alpha)U \| U) \\ = \frac{d}{d\alpha} \sum_{i=1}^l \frac{1}{l} f\left(\frac{\alpha \hat{p}(i) + (1 - \alpha)\frac{1}{l}}{\frac{1}{l}}\right) \\ = \sum_{i=1}^l \frac{1}{l} \left(\hat{p}(i) - \frac{1}{l}\right) f'\left(\frac{\alpha \hat{p}(i) + (1 - \alpha)\frac{1}{l}}{\frac{1}{l}}\right) \end{aligned}$$

and

$$\begin{aligned} \frac{d}{d\alpha} D(\alpha \text{Emp}_n(\omega) + (1 - \alpha)U \| U) \\ = \sum_{i=1}^l \frac{1}{l} \left(\hat{p}(i) - \frac{1}{l}\right)^2 f''\left(\frac{\alpha \hat{p}(i) + (1 - \alpha)\frac{1}{l}}{\frac{1}{l}}\right). \end{aligned}$$

Thus a second order Taylor expansion gives

$$\begin{aligned} D(\alpha \text{Emp}_n(\omega) + (1 - \alpha)U \| U) &\approx \frac{f''(1)}{2} \sum_{i=1}^l \frac{1}{l} \left(\hat{p}(i) - \frac{1}{l}\right)^2 \\ &= \frac{\chi^2(\text{Emp}_n(\omega), U)}{2l^2}. \end{aligned}$$

Thus using a small value of α approximately corresponds to replace the likelihood ratio test with a χ^2 test.

Example 5 (Normality test continued): Small or large values of β corresponds to small or large values of α . If the i 'th observation is denoted x_i then

$$\begin{aligned} D(\Psi_{d_0}(\text{Emp}_n(\omega)) \| \Psi_{d_0}(\Phi)) \\ = D\left(\Psi_{d_0}\left(\frac{1}{n} \sum_{i=1}^n \delta_{x_i}\right) \| \Psi_{d_0}(\Phi)\right) \\ = D\left(\frac{1}{n} \sum_{i=1}^n \Psi_{d_0}(\delta_{x_i}) \| \Psi_{d_0}(\Phi)\right) \\ = \frac{1}{n} \sum_{i=1}^n D(\Psi_{d_0}(\delta_{x_i}) \| \Psi_{d_0}(\Phi)) \\ + \frac{1}{n} \sum_{i=1}^n D\left(\Psi_{d_0}(\delta_{x_i}) \| \frac{1}{n} \sum_{j=1}^n \Psi_{d_0}(\delta_{x_j})\right) \\ = \frac{1}{n} \sum_{i=1}^n \frac{x_i^2}{2} \\ + \frac{1}{n} \sum_{i=1}^n D\left(\Psi_{d_0}(\delta_{x_i}) \| \frac{1}{n} \sum_{j=1}^n \Psi_{d_0}(\delta_{x_j})\right). \end{aligned}$$

For large values of α we only smooth a little so the different observations smoothed are approximately singular. Thus

$$D(\Psi_{d_0}(Emp_n(\omega)) \|\Psi_{d_0}(\Phi)) \approx \frac{\frac{1}{n} \sum_{i=1}^n x_i^2}{2} + \log n.$$

In this case the use of rate distortion statistic is approximately equivalent to the use of the statistic $\frac{1}{n} \sum_{i=1}^n x_i^2$. This statistic is sufficient for alternatives in the exponential family $\Phi(0, \sigma^2)$, $\sigma > 0$.

For small values of α we use a different expansion. We use that $D(\alpha X + (1 - \alpha^2)^{1/2} Z \|\Phi)$ has a leading term determined by the mean value of X . Therefore the statistic essentially reduces to $\frac{1}{n} \sum_{i=1}^n x_i^2$. This statistic is sufficient for alternatives in the exponential family $\Phi(\mu, 1)$.

Example 6 (Uniformity of angular data continued): Small or large values of β corresponds to small or large values of κ . For small values of κ we have

$$\frac{\exp(\kappa \cos(\theta))}{2\pi I_0(\kappa)} \approx 1 + \kappa \cos(\theta).$$

For observations $\theta_1, \theta_2, \dots, \theta_n$ the smoothed distribution approximately has density

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (1 + \kappa \cos(\theta - \theta_1)) \\ &= 1 + \frac{\kappa}{n} \sum_{i=1}^n (\cos(\theta - \theta_i)) \\ &= 1 + \frac{\kappa}{n} \sum_{i=1}^n \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} \cdot \begin{pmatrix} \cos \theta_i \\ \sin \theta_i \end{pmatrix} \\ &= 1 + \kappa \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} \cdot \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \cos \theta_i \\ \sin \theta_i \end{pmatrix}. \end{aligned}$$

The rate distortion statistic will approximately be given by $\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \cos \theta_i \\ \sin \theta_i \end{pmatrix}$. By rotational symmetry the information divergence does not depend on the direction of the vector $\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \cos \theta_i \\ \sin \theta_i \end{pmatrix}$. Thus the use of the rate distortion statistic is essentially equivalent to the use of the statistic $\left\| \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \cos \theta_i \\ \sin \theta_i \end{pmatrix} \right\|_2^2$. This is the most used statistic for testing uniformity of angular data.

We have

$$\begin{aligned} & D(\Psi_{d_0}(Emp_n(\omega)) \|U) \\ &= \frac{1}{n} \sum_{i=1}^n D(\Psi_{d_0}(\delta_{\theta_i}) \|U) \\ &+ \frac{1}{n} \sum_{i=1}^n D\left(\Psi_{d_0}(\delta_{\theta_i}) \left\| \frac{1}{n} \sum_{j=1}^n \Psi_{d_0}(\delta_{x_j})\right.\right) \\ &= D(\Psi_{d_0}(\delta_0) \|U) \\ &+ \frac{1}{n} \sum_{i=1}^n D\left(\Psi_{d_0}(\delta_{\theta_i}) \left\| \frac{1}{n} \sum_{j=1}^n \Psi_{d_0}(\delta_{x_j})\right.\right). \end{aligned}$$

For large values of κ the term

$$\frac{1}{n} \sum_{i=1}^n D\left(\Psi_{d_0}(\delta_{\theta_i}) \left\| \frac{1}{n} \sum_{j=1}^n \Psi_{d_0}(\delta_{x_j})\right.\right)$$

will be dominated by the pair (θ_i, θ_j) , $i \neq j$ for which $\cos(\theta_i - \theta_j)$ is maximal.

V. HODGE AND LEHMAN EFFICIENCY

For testing uniformity with Hamming distortion we see that if we do not compress data ($\alpha = 1$) the rate distortion test gives the statistic $D(Emp_n(\omega) \|U)$ which is known to be Bahadur efficient for testing uniformity. This is in general not the case. For a rate distortion test of normality little compression gives a statistic that is efficient for Gaussian alternatives with mean zero and variance different from 1, but it is obviously not efficient against other alternatives with mean 0 and variance 1. Similarly the rate distortion test of uniformity of angular data depends of the maximal value of $\cos(\theta_i - \theta_j)$, $i \neq j$ but not on values of all other observed angles which is obviously not efficient. So the question is how much one should to compress in order to get an efficient test against any alternative.

There are several ways of measuring efficiency among which the following are most important. In this short note it is neither possible to give all definitions nor proofs in details.

Hodge and Lehman efficiency An alternative hypothesis and a significance level are fixed. One is interested in the sample size that is needed to achieve a certain large power of the test.

Bahadur efficiency An alternative hypothesis and a power level are fixed. One is interested in the sample size that is needed to achieve a certain small significance level of the test.

Pitman efficiency The alternative is moved closer when the sample size is increased. This is done in a way so that the power of the test is constant. One is interested in the sample size that is needed to achieve a certain fixed significance level of the test.

The Hodge and Lehman efficiency is often the easiest to calculate but most tests are equally efficient in this sense. More tests can be distinguished by their Pitman efficiency. The Bahadur efficiency is often the most sensitive and at the same time often the hardest to calculate.

Theorem 7: Assume that the space Ω is compact and that the distortion function is continuous. Let d_n denote a decreasing sequence of distortion values. Assume that Q generates data. Then

$$Q(D(\Psi_{d_n}(Emp_n(\omega)) \|\Psi_{d_n}(Q)) \geq \varepsilon) \rightarrow 0 \text{ for } n \rightarrow \infty$$

if d_n tends to 0 sufficiently slowly.

Proof: It is sufficient to show that

$$Q(D(\Psi_{\delta}(Emp_n(\omega)) \|\Psi_{\delta}(Q)) \geq \varepsilon) \rightarrow 0$$

for any fixed distortion level $\delta > 0$. Weak convergence means that $Emp_n(\omega)$ converges to Q in the Wasserstein sense. Continuity of the distortion function d implies that Ψ_{δ} is weak continuous on the set of probability measures. ■

Theorem 8: Let d_n denote a decreasing sequence of distortion values. Assume that Q generates data. Then

$$\liminf D(\Psi_{d_n}(Emp_n(\omega)) \| \Psi_{d_n}(P)) \geq D(Q \| P)$$

almost surely.

Proof: This follows by lower semi-continuity of information because $\Psi_{d_n}(Emp_n(\omega))$ tends to Q and $\Psi_{d_n}(P)$ tends to P in the weak topology. ■

If P denotes an alternative to a nul-hypothesis Q then according to Sanov's theorem for a fixed significance level the best achievable type 2 error decreases like $\exp(-nD(Q \| P))$. The two previous theorems together implies that the rate-distortion test on a compact set with a continuous distortion function achieves the same exponential decrease in type 2 error. Hence, the rate distortion test is efficient in the sense of Hodge and Lehman.

VI. BAHADUR EFFICIENCY

We shall analyze this question in the case of testing uniformity of angular data because this is of particular simplicity because angles can be identified with elements of $SO(2)$.

Theorem 9: Let d_n denote a decreasing sequence of distortion values. Assume that P generates data. Then

$$\liminf D(\Psi_{d_n}(Emp_n(\omega)) \| U) \geq D(P \| U)$$

almost surely.

Proof: The proof is essentially the same as the proof of Theorem 8. ■

The theorem implies that for any $K < D(P \| U)$ we have $D(\Psi_{d_n}(Emp_n(\omega)) \| U) \geq K$ eventually almost surely so if P is the distribution of the alternative hypothesis then and the power of the test is kept fixed, then the acceptance regions of alternative P in the rate distortion test must have the form $D(\Psi_{d_n}(Emp_n(\omega)) \| U) \geq K_n$ for $K_n \rightarrow D(P \| U)$. In order to determine the Bahadur efficiency we have to bound the probability of $D(\Psi_{d_n}(Emp_n(\omega)) \| U) \geq K_n$ under the null hypothesis that data are generated by a uniform distribution. Now partition the set of angles $[0; 2\pi[$ into k_n intervals of length $2\pi/k_n$. We choose k_n such that $\frac{n}{k \log k} \rightarrow \infty$ for $n \rightarrow \infty$. Let F_n denote the σ -algebra generated by these intervals. Then

$$\lim -\frac{1}{n} \Pr \left(D \left(Emp_n(\omega)_{|F_n} \| U_{|F_n} \right) \geq K_n \right) = D(P \| U).$$

We are interested in

$$D(\Psi_{d_n}(Emp_n(\omega)) \| U) = D(\Psi_{d_n}(Emp_n(\omega)) \| \Psi_{d_n}(U))$$

and not $D(Emp_n(\omega)_{|F_n} \| U_{|F_n})$ but each subinterval has length $2\pi/k_n$ so

$$\begin{aligned} & \left| \log \frac{d\Psi_{d_n}(Emp_n(\omega))}{d\Psi_{d_n}(U)} - \log \frac{dEmp_n(\omega)_{|F_n}}{dU_{|F_n}} \right| \\ & \leq \left| \log \frac{\frac{\exp(\kappa_n \cos(0))}{2\pi I_0(\kappa)}}{\frac{\exp(\kappa_n \cos(\theta))}{2\pi I_0(\kappa)}} \right| \\ & = \kappa_n |\cos(0) - \cos(2\pi/k_n)|. \end{aligned}$$

Therefore

$$\lim -\frac{1}{n} \Pr(D(\Psi_{d_n}(Emp_n(\omega)) \| U) \geq K_n) = D(P \| U)$$

if the the test is Bahadur efficient if

$$\frac{\kappa_n |\cos(0) - \cos(2\pi/k_n)|}{n} \rightarrow 0$$

for $n \rightarrow \infty$. An expansion of cosine around 0 shows that the condition is equivalent to

$$\frac{\kappa_n}{nk_n^2} \rightarrow 0 \text{ for } n \rightarrow \infty.$$

If we choose $k_n = n^\gamma$ where $\gamma < 1$ we get the sufficient condition

$$\frac{\kappa_n}{n^{1+2\gamma}} \rightarrow 0 \text{ for } n \rightarrow \infty.$$

This leads us to the following theorem.

Theorem 10: The rate distortion test of uniformity of angular data has smoothing by a von Mises distribution with parameter κ_n . If $\kappa_n \rightarrow \infty$ for $n \rightarrow \infty$ and there exist $\eta \in [1; 3[$ such that

$$\frac{\kappa_n}{n^\eta} \rightarrow 0 \text{ for } n \rightarrow \infty,$$

then the rate distortion test is Bahadur efficient.

The method sketched here can be extended to prove Bahadur efficiency of rate distortion test of the uniformity on compact groups [7], [8].

VII. DISCUSSION

A new statistical test is proposed. It is based on a rate distortion function. By specifying the distortion function one does not have to divide the data into bins as this is build into the test. We have discussed the test in detail for a few examples. The example with testing uniformity of angular data can be extended to compact groups. There is no standard procedure for testing uniformity on a group, but there are many competing tests for the Gaussian distribution. In [9], [10] and [11] it has been shown by simulations that tests based on estimation entropy are more powerful than many other test for normality that one can find in the literature. The author has done some simulation to compare these tests with the test proposed here. These simulations indicates that the rate distortion test has a good power, but these results are still preliminary and will not be presented in this short note.

We saw that the rate distortion test has good Bahadur efficiency for angular data. We conjecture that the proposed test has high Bahadur efficiency in any case where it can be applied. It is not clear how to formulate this conjecture precisely, and it may be hard to prove because the rate distorting function normally cannot be calculated exactly.

A nice feature about the rate distortion test is that one can get a clear understanding of the effect of very small or very large compression. In our examples very small or very large compression in the rate distortion test corresponds to other familiar test like χ^2 -testing, and this may actually be used to give new interpretations of these tests. This is in contrast

with the common approach via discretizations. It is simply difficult to analyze the effect of discretize data into very few bins because 2 gives an absolute lower bound on how many bins one can use if the analysis should not become trivial.

Another conjecture that has been supported by numerical calculations is that the rate distortion statistics is asymptotically Gaussian. As it is now we have to Monte Carlo simulate the rate distortion statistics, and each simulation involves a numerical calculation of the rate distortion function. If it can be proved that the distribution of the rate distortion statistics is asymptotically Gaussian it means that the number of simulations can be reduced significantly because one just has to estimate mean and variance in order to be able to calculate the critical value for a specified significance level.

In this paper some simple examples where the rate distortion function can be calculated exactly, have been discussed. There are other examples than these where the rate distortion function can be calculated exactly. One interesting example is the Poisson process discussed in [12]. The setup is slightly different than the one presented here and therefore we cannot discuss it in this short paper. Nevertheless the ideas presented in this paper can be used to construct a test of whether a random process is a Poisson process. Contrary to the examples discussed in this paper this test of the Poisson process is completely new in the sense that it does not relate to any established statistical test.

Acknowledgement The authors want to thank Peter Grünwald, Igor Vajda, and Nisheeth Srivastava for useful discussions.

REFERENCES

- [1] A. Wald, *Sequential Analysis*. Wiley, 1947.
- [2] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
- [3] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory, Now Publishers Inc., 2004.
- [4] P. Harremoës and I. Vajda, "On the Bahadur-efficient testing of uniformity by means of the entropy," *IEEE Trans. Inform Theory*, vol. 54, pp. 321–331, Jan. 2008.
- [5] P. Harremoës and I. Vajda, "Efficiency of entropy testing," in *International Symposium on Information Theory*, pp. 2639–2643, IEEE, July 2008.
- [6] E. Batschelet, *Circular Statistics in Biology*. Mathematics in Biology, Academic Press, 1981.
- [7] P. Harremoës, "Maximum entropy on compact groups," in *Proceedings of International Symposium on Information Theory*, (Seattle, USA), pp. 108–112, IEEE, 9-14. July 2006.
- [8] P. Harremoës, "Maximum entropy on compact groups." Submitted to *Entropy*, 2009.
- [9] M. A. Stephens, "Edf statistics for goodness of fit and some comparisons," *Journal of the the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974.
- [10] I. Arizono and H. Ohta, "A test for normality based on kullback-leibler information," *The American Statistician*, vol. 43, no. 1, pp. 20–22, 1989.
- [11] Y. Steinberg and O. Zeitouni, "On tests of normality," *IEEE Transactions on Information Theory*, vol. 38, no. 6, pp. 1779–1787, 1992.
- [12] S. Verdu, "The exponential distribution in information theory," *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 1779–1787, 1996.