

Entropy Testing is Efficient

Peter Harremoës
 Centrum voor Wiskunde en Informatica
 P.O. 94079, 1090 GB Amsterdam
 The Netherlands
 P.Harremoës@cwi.nl

Igor Vajda
 Inform. Theory and Automation
 Czech Acad. Sciences, Prague
 Czech Republic
 vajda@utia.caz.cz

Abstract—This paper compares the power divergence statistics of orders $\alpha > 1$ with the information divergence statistic in the problem of testing the uniformity of a distribution. In this problem the information divergence statistic is equivalent to the entropy statistic. Extending some previously established results about information diagrams, it is proved that in this problem the information divergence statistic is more efficient in the Bahadur sense than any power divergence statistic of order $\alpha > 1$. This means that the entropy provides in a certain sense the most efficient way of characterizing the uniformity of a distribution.

I. POWER DIVERGENCE STATISTICS

Let $M(k)$ denote the set of all discrete probability distributions of the form $P = (p_1, \dots, p_k)$ and $M(k|n)$ the subset of possible types. One of the fundamental problems of mathematical statistics can be described as follows. Consider n balls distributed into boxes $1, \dots, k$ independently according to an unknown probability law $P_n \in M(k)$, which possibly depends on the number of balls n . This results in frequency counts X_{n1}, \dots, X_{nk} the vector of which $\mathbf{X}_n = (X_{n1}, \dots, X_{nk}) \in \{0, 1, \dots\}^k$ is multinomially distributed with parameters k, n and P_n . The problem is to decide on the basis of observations \mathbf{X}_n whether the unknown law P_n is equal to a given $Q = (q_1, \dots, q_k) \in M(k)$ or not.

The observations \mathbf{X}_n are represented by the empirical distribution

$$\hat{P}_n = (\hat{p}_{n1} \triangleq X_{n1}/n, \dots, \hat{p}_{nk} \triangleq X_{nk}/n) \in M(k|n) \quad (1)$$

and procedure \mathcal{T} on accepting or rejecting a hypothesis based on \hat{P}_n is called a test. The test uses a statistic $T_n(\hat{P}_n, Q)$ which characterizes the goodness-of-fit between the distributions \hat{P}_n and Q . The test \mathcal{T} rejects the hypothesis $P_n = Q$ if $T = T_n(\hat{P}_n, Q)$ exceeds a certain level $r_n \in \mathbb{R}$.

The goodness-of-fit statistic is usually one of the *power divergence statistics*

$$T_\alpha = T_{\alpha,n} = 2n D_\alpha(\hat{P}_n, Q), \quad \alpha \in \mathbb{R}. \quad (2)$$

where $D_\alpha(P, Q)$ denotes the so-called α -divergence (*power divergence* of order α) of distributions $P, Q \in M(k)$ defined by

$$D_\alpha(P, Q) = \sum_{j=1}^k q_j \phi_\alpha \left(\frac{p_j}{q_j} \right), \quad \alpha \in \mathbb{R}, \quad (3)$$

for the power function ϕ_α of order $\alpha \in \mathbb{R}$ given on the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \text{when } \alpha(\alpha-1) \neq 0 \quad (4)$$

and by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1, \quad \phi_1(t) = t \ln t - t + 1. \quad (5)$$

For details about definition (3) and the properties of power divergences, see [1] and [2]. Important examples of statistics based on power divergences are the Pearson statistic ($\alpha = 2$),

the Neyman statistic ($\alpha = -1$), the log-likelihood ratio ($\alpha = 1$), the reversed log-likelihood ratio ($\alpha = 0$) and the Freeman-Tukey statistic ($\alpha = 1/2$). In what follows we focus on the statistic $D_\alpha(\hat{P}_n, Q)$ rather than the one-one related $T_\alpha = T_{\alpha,n}$.

In this paper we deal with the question of which of the power divergence statistics $T_\alpha, \alpha \in \mathbb{R}$ is preferable for testing the hypothesis that the true distribution is uniform, i.e. the hypothesis $\mathcal{H} : P_n = U \triangleq (1/k, \dots, 1/k) \in M(k|n)$. Then

$$\mathbf{X}_n \sim \text{Multinomial}_k(n, U) \quad \text{under } \mathcal{H}. \quad (6)$$

The alternative to the hypothesis \mathcal{H} is denoted by \mathcal{A}_n . Thus

$$\mathbf{X}_n \sim \text{Multinomial}_k(n, P_n) \quad \text{under } \mathcal{A}_n \quad (7)$$

for $P_n \in M(k)$.

Example 1: Let μ, ν be probability measures on the Borel line $(\mathbb{R}, \mathcal{B})$ with absolutely continuous distribution functions F, G and Y_1, \dots, Y_n an i.i.d. sample from the probability space $(\mathbb{R}, \mathcal{B}, \mu)$. Consider a statistician who knows neither the probability measure μ governing the random sample (Y_1, \dots, Y_n) nor this sample itself. Nevertheless, he observes the frequencies $\mathbf{X}_n = (X_{n1}, \dots, X_{nk})$ of the samples Y_1, \dots, Y_n in an interval partition $\mathcal{P}_n = \{A_{n1}, \dots, A_{nk}\}$ of \mathbb{R} chosen by him. Using \mathbf{X}_n he has to decide about the hypothesis \mathcal{H} that the unknown probability measure on $(\mathbb{R}, \mathcal{B})$ is the given ν . Thus for a partition $\mathcal{P}_n = \{A_{n1}, \dots, A_{nk}\}$ under his control he obtains the observations generated by $P_n = (\mu(A_{n1}), \dots, \mu(A_{nk}))$ and his task is to test the hypothesis $\mathcal{H} : \mu = \nu$. Knowing ν , he can use the quantile function G^{-1} of ν or, more precisely, the quantiles $G^{-1}(j/k)$ of the orders j/k for $1 \leq j \leq k$ cutting \mathbb{R} into a special system of intervals $\mathcal{P}_n = \{A_{n1}, \dots, A_{nk}\}$ with the property

$\nu(A_{nj}) = 1/k$ for $1 \leq j \leq k$. Hence for this special partition we get

$$P_n = U = (1/k, \dots, 1/k) \in M(k|n) \quad \text{under } \mathcal{H} \quad (8)$$

and

$$P_n = (\mu(A_{n1}), \dots, \mu(A_{nk})) \in M(k) \quad \text{under } \mathcal{A}_n. \quad (9)$$

We see from (8) and (9) that the partitions \mathcal{P}_n generated by quantiles lead exactly to the situation assumed in (6) - (7).

The formulas for divergences $D_\alpha(P, G)$ simplify when $Q = U$, e.g.,

$$D_1(P, U) = \ln k - H(P) \quad \text{for } P \in M(k) \quad (10)$$

where $H(P)$ denotes the Shannon entropy

$$H(P) = - \sum_{j=1}^k p_j \ln p_j.$$

Similarly, (3) and (4) imply for all $\alpha > 1$ and $P \in M(k)$

$$D_\alpha(P, U) = \frac{k^{\alpha-1} IC_\alpha(P) - 1}{\alpha(\alpha-1)}. \quad (11)$$

Here

$$IC_\alpha(P) = \sum_{j=1}^k p_j^\alpha \quad \text{for } P \in M(k) \quad (12)$$

is the *index of coincidence* of P of order $\alpha > 1$ introduced by [3], taking on values between $k^{1-\alpha}$ and 1.

From (10) we see that the information divergence statistic $T_{1,n} = 2nD_1(\hat{P}_n, U)$ is one-one related to the entropy statistic $2nH(\hat{P}_n)$, and from (11) we see that for each $\alpha > 1$ the power divergence statistic $T_{\alpha,n} = 2nD_\alpha(\hat{P}_n, U)$ is one-one related to the corresponding IC -statistic $2nIC_\alpha(\hat{P}_n)$. The entropy $H(\hat{P}_n)$ as well as the indices of coincidence $IC_\alpha(\hat{P}_n)$, $\alpha > 1$ characterize the uniformity of the distribution \hat{P}_n . We are interested in the characterization, that is most efficient from the statistical point of view.

The rest of this paper is organized as follows. In Section II we introduce the concept of Bahadur efficiency that shall be used to compare different statistics. In Section III conditions under which the different statistics are consistent are characterized. In Section IV contains a sketch of the steps leading to the main result that is stated in Section V together with a short discussion. Many technical details have been omitted in this paper and the interested reader should consult [4] for a more detailed presentation.

II. BAHADUR EFFICIENCY

In this short report we focus on the typical situation where $k = k_n$ satisfies a condition of the type

$$\lim_{n \rightarrow \infty} \frac{k^2}{n} = 0. \quad (13)$$

This condition implies that $(T_\alpha - k)/\sqrt{2k}$ is asymptotically Gaussian under the hypothesis \mathcal{H} [5] so that it is easy to

calculate for which values of the statistic T_α the hypothesis should be accepted or rejected at a specified asymptotic significance level.

We are interested in the relative asymptotic efficiencies of the power divergence statistics T_{α_1} and T_{α_2} for $1 \leq \alpha_1 < \alpha_2 < \infty$. The condition (13) implies that the *Pitman asymptotic relative efficiencies* of all statistics T_α , $\alpha \in \mathbb{R}$ coincide (see [2]). In this situation preferences between these statistics must be based on the Bahadur efficiencies $BE(T_{\alpha_1} | T_{\alpha_2})$. We use the general definition of the Bahadur efficiency presented by Quine and Robinson [6] who extended the original concept of [7] (see also [8]). Quine and Robinson demonstrated that $BE(T_1 | T_2) = \infty$ so that the log-likelihood ratio statistic T_1 is more Bahadur efficient than the Pearson statistic T_2 . Using the results from [9], this first achievement was extended by [10] who proved that the Bahadur efficiencies of the reversed log-likelihood ratio statistic T_0 and the Neyman statistic T_{-1} coincide and these statistics are less Bahadur efficient than Pearson's T_2 .

A problem left open in the previous literature is to evaluate the Bahadur efficiencies of the remaining statistics T_α , $\alpha \in \mathbb{R}$, in particular to confirm or reject the conjecture that the log-likelihood ratio statistic is most Bahadur efficient in the class of all power divergence statistics T_α , $\alpha \in \mathbb{R}$. In this paper we present the solution of this problem for $\alpha \geq 1$. Our solution is based on the results on indices of coincidence derived in [3]. Before defining the Bahadur efficiency, we introduce some important auxiliary concepts.

Definition 2: For $\alpha \in \mathbb{R}$ we say that

- 1) the model satisfies the *Bahadur condition* if there exists $\Delta_\alpha > 0$ such that under the alternatives \mathcal{A}_n

$$\lim_{n \rightarrow \infty} D_\alpha(P_n, U) = \Delta_\alpha. \quad (14)$$

- 2) the statistic $D_\alpha(\hat{P}_n, U)$ is *consistent* if the Bahadur condition holds and for $n \rightarrow \infty$

$$ED_\alpha(\hat{P}_n, U) \rightarrow 0 \quad \text{under } \mathcal{H} \quad (15)$$

$$\text{while } D_\alpha(\hat{P}_n, U) \xrightarrow{p} \Delta_\alpha \quad \text{under } \mathcal{A}_n.$$

The Bahadur condition (14) means that in term of the statistic $D_\alpha(\hat{P}_n, U)$, the alternatives \mathcal{A}_n are neither too near to nor too far from the hypothesis \mathcal{H} . It can be deduced from [11] that the Bahadur condition holds for the model of Example 1. The consistency of $D_\alpha(\hat{P}_n, U)$ introduced in Definition 2 means that the $D_\alpha(\hat{P}_n, U)$ -based test of the hypothesis $\mathcal{H} : U$ against the alternative $\mathcal{A}_n : P_n$ of any fixed asymptotic significance level has a power tending to 1. Indeed, under \mathcal{H} we have $D_\alpha(\hat{P}_n, U) \xrightarrow{p} 0$ so that the rejection level of the $D_\alpha(\hat{P}_n, U)$ -based test of any asymptotic significance level $s \in]0; 1[$ tends to 0 for $n \rightarrow \infty$ while under \mathcal{A}_n we have $D_\alpha(\hat{P}_n, U) \xrightarrow{p} \Delta_\alpha > 0$.

The above considered Bahadur efficiency $BE(T_{\alpha_1} | T_{\alpha_2})$ is defined under the condition that for $\alpha = \alpha_1$ and $\alpha = \alpha_2$ the statistic $D_\alpha(\hat{P}_n, U)$ is consistent and admits the so-called

Bahadur function. In the sequel $P(B_n)$ shall denote the probability of events B_n depending on the random observations \mathbf{X}_n (see (6) and (7)) and E the corresponding expectation.

Definition 3: For $\alpha \in \mathbb{R}$ we say that the Bahadur function for the statistic $T_\alpha = 2nD_\alpha(\hat{P}_n, U)$ exists if there exists a sequence $c_{\alpha,n} > 0$ and a continuous function $g_\alpha :]0; \infty[\rightarrow]0; \infty[$ such that under \mathcal{H}

$$\lim_{n \rightarrow \infty} -\frac{c_{\alpha,n}}{n} \ln P(D_\alpha(\hat{P}_n, U) \geq \Delta) = g_\alpha(\Delta). \quad (16)$$

Next follows the basic definition of the present paper where $\Delta_{\alpha_1}, \Delta_{\alpha_2}$ are the limits from the Bahadur condition and $g_{\alpha_1}, g_{\alpha_2}$ and $c_{\alpha_1,n}, c_{\alpha_2,n}$ are the functions and sequences from the definition of the Bahadur function.

Definition 4: Assume that the statistics $D_{\alpha_1}(\hat{P}_n, U)$ and $D_{\alpha_2}(\hat{P}_n, U)$ are consistent and that the corresponding Bahadur functions g_{α_1} and g_{α_2} exist. Then the *Bahadur efficiency* $BE(T_{\alpha_1} | T_{\alpha_2})$ of the corresponding power divergence T_{α_1} with respect to T_{α_2} is defined by

$$BE(T_{\alpha_1} | T_{\alpha_2}) = \frac{g_{\alpha_1}(\Delta_{\alpha_1})}{g_{\alpha_2}(\Delta_{\alpha_2})} \lim_{n \rightarrow \infty} \frac{c_{\alpha_1,n}}{c_{\alpha_2,n}} \quad (17)$$

provided the limit exists in $[0, \infty]$.

Assume that the statistics $D_{\alpha_i}(\hat{P}_n, U)$ are consistent for $i \in \{1, 2\}$ and there exist Bahadur functions g_{α_i} satisfying (16) for some sequences $c_{\alpha_i,n} > 0$. Then the definition of consistency implies that both the T_{α_i} -tests of the uniformity hypothesis $\mathcal{H} : U$ will achieve identical powers

$$\pi = P(D_{\alpha_i}(\hat{P}_n, U) \geq r_{n,i}) \quad \text{for } \pi \in]0, 1[\quad \text{and } i = 1, 2$$

under \mathcal{A}_n if and only if $r_{n,i} \downarrow \Delta_{\alpha_i}$ for $i \in \{1, 2\}$ as $n \rightarrow \infty$. The convergence $r_{n,i} \downarrow \Delta_{\alpha_i}$ leads to the approximate T_{α_i} -test significance levels

$$s_{n,i} \triangleq P(D_{\alpha_i}(\hat{P}_n, U) \geq \Delta_{\alpha_i}) \approx P(D_{\alpha_i}(\hat{P}_n, U) \geq r_{n,i})$$

for $i = 1, 2$ under \mathcal{H} where $s_{n,i} \rightarrow 0$ as $n \rightarrow \infty$ for $i = 1, 2$ under \mathcal{H} . By (16), the T_{α_i} -tests need different sample sizes

$$n_i = \frac{c_{\alpha_i,n}}{g_{\alpha_i}(\Delta_{\alpha_i})} \ln \frac{1}{s_n}, \quad i \in \{1, 2\} \quad (18)$$

to achieve the same approximate test significance levels $s_n = s_{n,1} = s_{n,2}$ when n is here playing the role of a formal parameter that increases to ∞ .

III. CONSISTENCY

The following theorem presents consistency conditions for all statistics $D_\alpha(\hat{P}_n, U)$, $\alpha \geq 1$.

Theorem 5: Let for all $\alpha \geq 1$ the Bahadur condition (14) hold. Then $D_\alpha(\hat{P}_n, U)$ is consistent if

$$\alpha \in [1; 2] \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{k}{n} = 0, \quad (19)$$

or

$$\alpha > 2 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{k^{\alpha-1}}{n} = 0. \quad (20)$$

Proof: Under \mathcal{H} we have $D_\alpha(P_n, U) = D_\alpha(U, U) = 0$. Hence it suffices to prove that under both \mathcal{H} and \mathcal{A}_n

$$\lim_{n \rightarrow \infty} E \left| D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U) \right| = 0 \quad \text{for all } \alpha \geq 1.$$

Put for brevity $D_{\alpha,n} = D_\alpha(P_n, U)$ and $\Lambda_{\alpha,n} = D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)$.

For every $\alpha > 1$ we have

$$\Lambda_{\alpha,n} = \frac{k^{\alpha-1}}{\alpha(\alpha-1)} \sum_{j=1}^k (\hat{p}_j^\alpha - p_j^\alpha),$$

which implies

$$\begin{aligned} |\Lambda_{\alpha,n}| &\leq \sum_{j=1}^k \left(\frac{\alpha p_j^{\alpha-1} |\hat{p}_j - p_j| + (\alpha-1) p_j^{\alpha-2} (\hat{p}_j - p_j)^2}{(\alpha-1) p_j^{\alpha-2} (\hat{p}_j - p_j)^2} \right) \\ &\leq \frac{k^{\alpha-1}}{\alpha-1} \left(\sum_{j=1}^k (p_j^{\alpha/2})^2 \right)^{1/2} \left(\sum_{j=1}^k p_j^{\alpha-2} (\hat{p}_j - p_j)^2 \right)^{1/2} \\ &\quad + \frac{k^{\alpha-1}}{\alpha} \sum_{j=1}^k p_j^{\alpha-2} (\hat{p}_j - p_j)^2. \end{aligned}$$

Thus

$$\begin{aligned} E |\Lambda_{\alpha,n}| &\leq \frac{k^{\alpha-1}}{\alpha-1} \left(\sum_{j=1}^k p_j^\alpha \right)^{1/2} \left(\sum_{j=1}^k p_j^{\alpha-2} \frac{p_j}{n} \right)^{1/2} \\ &\quad + \frac{k^{\alpha-1}}{\alpha} \sum_{j=1}^k p_j^{\alpha-2} \frac{p_j}{n} \\ &= \frac{(k^{\alpha-1} \sum_{j=1}^k p_j^\alpha)^{1/2}}{\alpha-1} \left(\frac{k^{\alpha-1} \sum_{j=1}^k p_j^{\alpha-1}}{n} \right)^{1/2} \\ &\quad + \frac{k^{\alpha-1}}{\alpha n} \sum_{j=1}^k p_j^{\alpha-1}. \end{aligned}$$

Now we use that $k^{\alpha-1} \sum_{j=1}^k p_j^\alpha = \alpha(\alpha-1) D_{\alpha,n} + 1$ and that for $1 < \alpha \leq 2$ the function $\psi(t) = t^{\alpha-1}$ is concave and get

$$\begin{aligned} E |\Lambda_{\alpha,n}| &\leq \frac{[\alpha(\alpha-1) D_{\alpha,n} + 1]^{1/2}}{\alpha-1} \left(\frac{k^{\alpha-1} k \left(\frac{1}{k}\right)^{\alpha-1}}{n} \right)^{1/2} \\ &\quad + 2 \frac{k^{\alpha-1} k \left(\frac{1}{k}\right)^{\alpha-1}}{\alpha n} \\ &= \frac{[\alpha(\alpha-1) D_{\alpha,n} + 1]^{1/2} \left(\frac{k}{n}\right)^{1/2}}{\alpha-1} + \frac{k}{\alpha n}, \end{aligned}$$

which proves (19) because the sequence $D_{\alpha,n}$ is bounded.

Proofs of the case $\alpha = 1$ and $\alpha > 2$ follow similar steps as above. ■

IV. BAHADUR FUNCTIONS

In this section we present an alternative to the formula (16) for the Bahadur functions g_α , $\alpha \in \mathbb{R}$ which is based on the inequality (23). These formulas are given in terms of the Shannon entropy $H(P)$ maximized on the sets

$$A_{\alpha,\Delta}(k) = \{P \in M(k) : D_\alpha(P, U) \geq \Delta\} \quad (21)$$

and

$$A_{\alpha,\Delta}(k|n) = A_{\alpha,\Delta}(k) \cap M(k|n) \quad (22)$$

or, equivalently (see (10)), in terms of the information divergence $D_1(\hat{P}_n, U)$ minimized on these sets.

As observed in [9, Lemma 1], for every subset $A \subseteq M(k)$ intersecting $M(k|n)$ the divergence $D_1(P, U)$ satisfies the inequality

$$\left| \inf_{P \in A \cap M(k|n)} D_1(P, U) + \frac{1}{n} \ln \mathbb{P}(\hat{P}_n \in A) \right| \leq \frac{k \ln(n+1)}{n}. \quad (23)$$

Approximation of $-\frac{c_{\alpha,n}}{n} \ln \mathbb{P}(\hat{P}_n \in A)$ appearing in (16) by means of $c_{\alpha,n}$ times the infimum from (23) is possible under the restriction

$$\lim_{n \rightarrow \infty} c_{\alpha,n} \frac{k \ln n}{n} = 0 \quad (24)$$

on the sequence $k = k_n$. This observation leads to the following lemma.

Lemma 6: Assume that for some $\alpha \in \mathbb{R}$ and $k = k_n$ there exist $c_{\alpha,n} > 0$ satisfying (24) such that the sequence of functions

$$G_{\alpha,\Delta}(k|n) = c_{\alpha,n} \left(\inf_{P \in A_{\alpha,\Delta}(k|n)} D_1(P, U) \right), \quad \Delta > 0 \quad (25)$$

converges to a positive limit

$$g_\alpha(\Delta) = \lim_{n \rightarrow \infty} G_{\alpha,\Delta}(k|n), \quad \Delta > 0. \quad (26)$$

Then g_α is the Bahadur function for the power divergence statistic T_α .

In the following assertion we consider for arbitrary $\alpha \in \mathbb{R}$, $k = k_n$ and $c_{\alpha,n} > 0$ the sequence of functions

$$G_{\alpha,\Delta}(k) = c_{\alpha,n} \left(\inf_{P \in A_{\alpha,\Delta}(k)} D_1(P, U) \right), \quad \Delta > 0. \quad (27)$$

Obviously, $G_{\alpha,\Delta}(k) \leq G_{\alpha,\Delta}(k|n)$.

Lemma 7: Let for some $\alpha \in \mathbb{R}$ the Bahadur condition hold and $c_{\alpha,n} > 0$ satisfy (24). If the corresponding sequences of functions (25) and (27) asymptotically coincide in the sense

$$\lim_{n \rightarrow \infty} [G_{\alpha,\Delta}(k|n) - G_{\alpha,\Delta}(k)] = 0 \quad (28)$$

and at the same time $G_{\alpha,\Delta}(k)$ converges to a positive limit

$$g_\alpha(\Delta) = \lim_{n \rightarrow \infty} G_{\alpha,\Delta}(k), \quad \Delta > 0. \quad (29)$$

Then g_α is the Bahadur function for the power divergence statistic T_α .

According to Lemma 7 the problem of determining the Bahadur efficiency is essentially equivalent to minimizing $D_1(P, U)$ under the condition $P \in A_{\alpha,\Delta}(k)$ or equivalently maximizing the Shannon entropy for a fixed value of the index of coincidence. In [3] it was proved that for every $x \in [k^{1-\alpha}, 1]$ and for the Dirac distribution $\mathbf{1} = (1, 0, \dots, 0) \in M(k)$, the equation

$$IC_\alpha(s\mathbf{1} + (1-s)U) = x \quad (30)$$

has a unique solution $s \in [0, 1]$ and that this solution satisfies the relation

$$\sup_{IC_\alpha(P) \geq x} H(P) = H(s\mathbf{1} + (1-s)U). \quad (31)$$

This leads to the following lemma using the weight sequences

$$a_k = 1/k \quad \text{and} \quad b_k = 1 - a_k \quad (32)$$

and the sequence of functions

$$\psi_k(s) = a_k(ks + 1 - s)^\alpha + b_k(1 - s)^\alpha, \quad s \in [0, 1]. \quad (33)$$

The equation

$$\psi_k(s) = 1 + \alpha(\alpha - 1)\Delta \quad (34)$$

has a unique solution $s = s_k \in [0, 1]$ when $\alpha > 1$, $k = k_n$ and

$$\frac{1}{\alpha(\alpha - 1)k} < \Delta \leq \frac{k^{\alpha-1} - 1}{\alpha(\alpha - 1)}, \quad (35)$$

This solution satisfies the relation

$$\lim_{n \rightarrow \infty} s_k = 0 \quad (36)$$

and the equality

$$\begin{aligned} \inf_{P \in A_{\alpha,\Delta}(k)} D_1(P, U) &= a_k(ks_k + 1 - s_k) \ln(ks_k + 1 - s_k) \\ &+ b_k(1 - s_k) \ln(1 - s_k). \end{aligned} \quad (37)$$

Using the explicit result (37) one can prove that for every $\alpha > 1$ and $\Delta > 0$,

$$\begin{aligned} \inf_{P \in A_{\alpha,\Delta}(k)} D_1(P, U) &= \\ &= \left([\alpha(\alpha - 1)\Delta]^{1/\alpha} + o(1) \right) \frac{k^{1/\alpha} \ln k^{1/\alpha}}{k} \quad \text{as } n \rightarrow \infty. \end{aligned}$$

If we define

$$c_{\alpha,n} = \frac{k}{k^{1/\alpha} \ln k^{1/\alpha}} \quad (38)$$

for $\alpha > 1$ and $k = k_n$ satisfying (13), we can give explicit formulas for the Bahadur functions of the statistics T_α , $\alpha > 1$.

Theorem 8: Assume that $k = k_n$ increases so slowly that (13) holds. Then (17) holds for the sequence $c_{\alpha,n}$ given by (38) and for the function

$$g_\alpha(\Delta) = [\alpha(\alpha - 1)\Delta]^{1/\alpha}, \quad \Delta > 0 \quad (39)$$

i.e., (39) is the Bahadur function of the statistic T_α . For $\alpha = 1$ we have $g_1(\Delta) = \Delta$ for the sequence $c_{1,n} = 1$.

Proof: Let $\alpha > 1$ be arbitrary and fixed. If $c_{\alpha,n}$ is given by (38) then (13) as well as (24) hold. Hence it follows from Lemma 7 that (16) holds for the $c_{\alpha,n}$ under consideration and for g_α given by (39). We find that (29) reduces to (39) and the last part follows directly from (23) and (24). ■

V. MAIN RESULTS

The functions g_α as well as the normalizing sequences $c_{\alpha,n}$ have been explicitly evaluated in Theorem 8 for all $\alpha \geq 1$. Therefore (17) provides explicit Bahadur efficiencies $BE(T_{\alpha_1} | T_{\alpha_2})$ on the whole domain $\alpha_1, \alpha_2 \geq 1$. These efficiencies are given in the following main result of this paper.

Theorem 9: If $k = k_n$ increases so slowly that

$$\lim_{n \rightarrow \infty} \frac{k^\beta}{n} = 0$$

for some $\beta \geq 2$ then the Bahadur efficiency of the statistic T_{α_1} with respect to T_{α_2} satisfies the relation

$$BE(T_{\alpha_1} | T_{\alpha_2}) = \infty \quad (40)$$

for all $1 \leq \alpha_1 < \alpha_2 \leq \beta + 1$.

Proof: If $1 < \alpha_1 < \alpha_2 \leq \beta + 1$ then, with the condition on the growth rate of $k = k_n$, conditions (19) or (20) hold. By Theorem 8, the sequences $c_{\alpha_1,n}$ and $c_{\alpha_2,n}$ given by (38) lead to the corresponding Bahadur functions g_{α_1} and g_{α_2} given by (39) and to the limit

$$\lim_{n \rightarrow \infty} \frac{c_{\alpha_2,n}}{c_{\alpha_1,n}} = \lim_{n \rightarrow \infty} \frac{k^{1/\alpha_1} \ln k^{1/\alpha_1}}{k^{1/\alpha_2} \ln k^{1/\alpha_2}} = \infty. \quad (41)$$

Relation (40) thus follows from (17) in Definition 3. If the assumptions hold for $1 = \alpha_1 < \alpha_2 < \infty$ then instead of the above considered general Bahadur function g_{α_1} given by (16) we have the particular function $g_{\alpha_1}(\Delta) = \Delta$ given by Theorem 9, and instead of $c_{\alpha_1,n} = k_n/k_n^{1/\alpha_1} \ln k_n^{1/\alpha_1}$ we have $c_{\alpha_1,n} = 1$. Therefore the limit

$$\lim_{n \rightarrow \infty} \frac{c_{\alpha_2,n}}{c_{\alpha_1,n}} = \lim_{n \rightarrow \infty} \frac{k_n}{k_n^{1/\alpha_2} \ln k_n^{1/\alpha_2}}$$

remains to be infinite as in (41). ■

In [10] it was proved that the log-likelihood ratio statistic T_1 is most Bahadur efficient in the class of power divergence statistics $\{T_{-1}, T_0, T_1, T_2\}$. The last theorem shows that T_1 is most Bahadur efficient also in the class $\{T_\alpha : \alpha \geq 1\}$. Hence, by (10), the entropy $H(\hat{P}_n)$ defines the statistic that is most Bahadur efficient in the class of all statistics $\{T_\alpha : \alpha \geq -1\}$, with the possible exception of the subset $\{T_\alpha : \alpha \in]-1; 0[\cup]0; 1[\}$. For this subset the Bahadur efficiency is not known.

Obviously only the uniform distribution can be characterized as having maximal entropy, but one may ask the same question for more restricted classes of distributions. For instance the binomial distribution with parameters k and $1/2$ has maximal entropy among distributions of sums of k independent Bernoulli random variables. To answer such a question one would need results like the ones obtained in [3], but such results have not been derived yet.

The main motivation for studying the characterization of the uniform distribution was Example 1, where the sample space \mathbb{R} was quantized in bins each with the same probability under the null hypothesis. If the sample space is more general than the real line or if a distortion measure is given, it is often natural to consider other quantizations than a uniform quantization with respect to the null hypothesis. It is natural to ask which test is most efficient if the quantization is allowed to depend on both the null hypothesis and the sample, but this is much harder even to state this problem precisely. If the sample space is a compact group the rate distortion function is maximal at the uniform distribution so for a compact group one may ask to what extent the rate distortion function of the sample evaluated in a point would serve as an efficient statistic for testing uniformity.

Acknowledgement The authors want to thank Wouter Koolen for many comments that have improved the quality of this paper.

REFERENCES

- [1] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.
- [2] T. R. C. Read and N. Cressie, *Goodness of Fit Statistics for Discrete Multivariate Data*. Berlin: Springer, 1988.
- [3] P. Harremoës and F. Topsøe, "Inequalities between entropy and index of coincidence derived from information diagrams," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2944–2960, Nov. 2001.
- [4] P. Harremoës and I. Vajda, "On the Bahadur-efficient testing of uniformity by means of the entropy." Submitted for publication in *IEEE Trans. Inform. Theory*, March 2007.
- [5] L. Györfi and I. Vajda, "Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models," *Probability and Statistics Letters*, vol. 56, no. 1, pp. 57–67, 2002.
- [6] M. P. Quine and J. Robinson, "Efficiencies of chi-square and likelihood ratio goodness-of-fit tests," *Ann. Statist.*, vol. 13, pp. 727–742, 1985.
- [7] R. R. Bahadur, *Some Limit Theorems in Statistics*. Philadelphia: SIAM, 1981.
- [8] Y. Nikulin, *Asymptotic Efficiency of Nonparametric Tests*. Cambridge: Cambridge University Press, 1995.
- [9] J. Beirlant, L. Devroye, L. Györfi, and I. Vajda, "Large deviations of divergence measures on partitions," *J. Statist. Planning and Infer.*, vol. 93, pp. 1–16, 2001.
- [10] L. Györfi, G. Morvai, and I. Vajda, "Information-theoretic methods in testing the goodness-of-fit," in *Proc. International Symposium on Information Theory, Sorrento, Italy, June 25–30*, p. 28, 2000.
- [11] I. Vajda, "On convergence of information contained in quantized observations," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2163–2172, Aug. 2002.