Efficiency of Entropy Testing

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

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

Abstract— **Recently it was shown that Shannon entropy is more Bahadur efficient than any Rényi entropy of order $\alpha > 1$. In this paper we shall show that relative Bahadur efficiency between any two Rényi entropies of orders $\alpha \in \,]0;1]$ is 1 when the relative Bahadur efficiency is defined according to [1]. Despite the fact that the relative Bahadur efficiency is 1 it is shown that in a certain sense Shannon entropy is more efficient than Rényi entropy for $\alpha \in \,]0;1[$. This indicates that the definition of relative efficiency given in [1] does not fully capture the notion of efficiency.**

I. POWER DIVERGENCE STATISTICS

Let $M(k)$ denote the set of all discrete probability distributions of the form $P = (p_1, ..., p_k)$ 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 bins $1, ..., k$ independently according to an unknown probability distribution $P_n \in M(k)$, which may depend on the number of balls n. This results in frequency counts $X_{n1}, ..., X_{nk}$ the vector of which $\boldsymbol{X}_n = (X_{n1}, ..., X_{nk}) \in \{0, 1, ...\}^k$ is multinomially distributed with parameters k, n and P_n. The problem is to decide on the basis of observations $\boldsymbol{X}_n$ whether the unknown law P_n is equal to a given $Q = (q_1, ..., q_k) \in M(k)$ or not.

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

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

and a 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 power divergence of order α of the 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

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

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 [2] and [3]. 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 $\mathcal{H} : P_n = U \triangleq (1/k, ..., 1/k) \in M(k|n)$. Then

$$\boldsymbol{X}_n \sim Multinomial_k(n, U) \quad \text{under } \mathcal{H}. \quad (6)$$

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

$$\boldsymbol{X}_n \sim Multinomial_k(n, P_n) \quad \text{under } \mathcal{A}_n \quad (7)$$

for $P_n \in M(k)$.

Example 1: Let μ, ν be probability measures on the Borel line $(\mathbb{R}, \mathcal{B})$ with continuous distribution functions F, G and $Y_1, ..., Y_n$ 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 $(Y_1, ..., Y_n)$ nor this sample itself. Nevertheless, he observes the frequencies $\boldsymbol{X}_n = (X_{n1}, ..., X_{nk})$ of the samples $Y_1, ..., Y_n$ in an interval partition $\mathcal{P}_n = \{A_{n1}, ..., A_{nk}\}$ of $\mathbb{R}$ chosen by him. Using $\boldsymbol{X}_n$ 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 $\mathcal{P}_n = \{A_{n1}, ..., A_{nk}\}$ under his control he obtains the observations generated by $P_n = (\mu(A_{n1}), ..., \mu(A_{nk}))$ 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

$\mathcal{P}_n = \{A_{n1}, \ldots, A_{nk}\}$ with the property $\nu(A_{nj}) = 1/k$ for $1 \le j \le k$. Hence for this special partition we get

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

and

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

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

Note that the sequence P_n is contiguous to the sequence of uniform distributions because μ is absolutely continuous to ν.

It is often convenient to replace the power divergences $D_\alpha(P, Q)$ of the orders $\alpha > 0$ by the one-one related Rényi divergences [4] (see also [2]) given by

$$D_\alpha(P\|Q) = \frac{\log\left(\sum p_i^\alpha q_i^{1-\alpha}\right)}{\alpha - 1}$$

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

$$D_\alpha(P\|U) = \ln k - H_\alpha(P) \quad \text{for } P \in M(k) \qquad (10)$$

where $H_\alpha(P) = (\alpha - 1)^{-1} \sum p_i^\alpha$ denotes the Rényi entropy of order $\alpha > 0$. Here $D_1(P\|Q)$ and $H_1(P)$ are supposed to be the limits of $D_\alpha(P\|Q)$ and $H_\alpha(P)$ for $\alpha \nearrow 1$ which happen to coincide with the classical information divergence $D_1(P, Q)$ and Shannon entropy $H(P)$.

II. BAHADUR EFFICIENCY

In this short report we focus on the typical situation where $k = k_n$ satisfies $k \to \infty$ and the average number of observations per bin tends to infinity, i.e.

$$\frac{n}{k} \to \infty \quad \text{for } n \to \infty. \qquad (11)$$

This condition implies that T_α 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 significance level.

We are interested in the relative asymptotic efficiencies of the power divergence statistics T_{α_1} and T_{α_2} for $0 < \alpha_1 < \alpha_2 < \infty$. The condition (11) implies that the *Pitman asymptotic relative efficiencies* of all statistics T_α, $\alpha \in \mathbb{R}$ coincide (cf. e.g. [3]). In this situation preferences between these statistics must be based on the Bahadur efficiencies $BE(T_{\alpha_1} \mid T_{\alpha_2})$. Quine and Robinson [1] demonstrated that the log-likelihood ratio statistic T_1 is infinitely more Bahadur efficient than the Pearson statistic T_2. In [6] it was proved that T_1 is more Bahadur efficient than T_α for $\alpha > 1$.

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 $\alpha \in]0; 1[$. Our solution is based on the results on indices of coincidence derived in

[7]. 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 $\mathcal{A}_n$

$$\lim_{n \to \infty} D_\alpha(P_n, U) = \Delta_\alpha. \qquad (12)$$

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

$$\lim_{n \to \infty} D_\alpha(\hat{P}_n, U) \xrightarrow{p} 0 \quad \text{under } \mathcal{H} \qquad (13)$$

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

Note that the consistency condition (13) is slightly weaker than the one used in the literature [1], [6], and will allow us to get better consistency results. For calculation of Bahadur efficiency this weaker condition makes no difference. For $\alpha \ge 1$ the Bahadur condition implies that the sequence of alternatives P_n is contiguous to the sequence U. This is not the case for $\alpha \in]0; 1[$ where the Bahadur condition is fulfilled a sequence of alternative distributions concentrated on single points that is entirely separated from the sequence of uniform distributions.

The Bahadur condition (12) 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 [8] 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 $n \to \infty$ while under $\mathcal{A}_n$ we have $D_\alpha(\hat{P}_n, U) \xrightarrow{p} \Delta_\alpha > 0$.

The above considered Bahadur efficiency $BE(T_{\alpha_1} \mid T_{\alpha_2})$ 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 $\mathsf{P}(B_n)$ shall denote the probability of events B_n depending on the random observations $\boldsymbol{X}_n$ (cf. (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 $g_\alpha :]0; \infty[\to]0; \infty[$ such that under $\mathcal{H}$

$$\lim_{n \to \infty} -\frac{c_{\alpha,n}}{n} \ln \mathsf{P}(D_\alpha(\hat{P}_n, U) \ge \Delta) = g_\alpha(\Delta). \qquad (14)$$

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_{α_1}, g_{α_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* $BE(T_{\alpha_1} \mid T_{\alpha_2})$ of the corresponding power divergence T_{α_1} with respect to T_{α_2} is defined by

$$BE(T_{\alpha_1} \mid T_{\alpha_2}) = \frac{g_{\alpha_1}(\Delta_{\alpha_1})}{g_{\alpha_2}(\Delta_{\alpha_2})} \lim_{n\to\infty} \frac{c_{\alpha_2,n}}{c_{\alpha_1,n}} \qquad (15)$$

provided the limit exists in $[0, \infty]$.[1]

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 (14) 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

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

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

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

for $i = 1, 2$ under $\mathcal{H}$ where $s_{n,i} \to 0$ as $n \to \infty$ for $i = 1, 2$ under $\mathcal{H}$. By (14), the T_{α_i}-tests need different sample sizes

$$n_i = \frac{c_{\alpha_i,n}}{g_{\alpha_i}(\Delta_{\alpha_i})} \ln \frac{1}{s_n}, \quad i \in \{1, 2\} \qquad (16)$$

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 main theorem in this section presents consistency conditions for all statistics $D_{\alpha}(\hat{P}_n, U)$, $\alpha > 0$. It extends and refines a theorem in [6]. See [9] for more details on consistency of various ϕ-divergences. We shall use the following result about consistency of variational the distance V.

Lemma 5: If $\lim_{n\to\infty} \frac{n}{k} = \infty$ then $\mathsf{E}\left[V(\hat{P}_n, P_n)\right] \to 0$ for $n \to \infty$.

Proof: By the Cauchy–Schwarz inequality,

$$V(\hat{P}_n, P_n) = \sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right| p_j = \sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right| p_j^{1/2} p_j^{1/2}$$

$$\leq \left(\sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right|^2 p_j\right)^{1/2} \cdot \left(\sum_{j=1}^{k} p_j\right)^{1/2}$$

$$= \left(D_2\left(\hat{P}_n, P_n\right)\right)^{1/2}.$$

Hence

$$\mathsf{E}V(\hat{P}_n, P_n) \leq \left(\mathsf{E}D_2\left(\hat{P}_n, P_n\right)\right)^{1/2} = \left(\frac{k-1}{n}\right)^{1/2}.$$

∎

[1] In [6] due to a misprint, α_1 and α_2 were interchanged behind the limit there, but the formula was used in the correct form.

Theorem 6: For all $\alpha \geq 1$ let the Bahadur condition (12) hold. Then $D_{\alpha}(\hat{P}_n, U)$ is consistent if

$$\alpha \in]0; 2] \quad \text{and} \quad \lim_{n\to\infty} \frac{n}{k} = \infty, \qquad (17)$$

or

$$\alpha > 2 \quad \text{and} \quad \lim_{n\to\infty} \frac{n}{k \log k} = \infty. \qquad (18)$$

Proof: First assume that $\alpha < 1$. The function ϕ_α is uniformly continuous on $[0; \infty[$ so for all $\varepsilon > 0$ there exists $\delta > 0$ such that $|\phi_\alpha(t) - \phi_\alpha(s)| \leq \delta |t - s| + \varepsilon$ for all $t, s > 0$.

$$\left|D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)\right|$$

$$\leq \sum_{j=1}^{k} \frac{1}{k} \left|\phi_\alpha\left(\frac{\hat{p}_j}{1/k}\right) - \left(\frac{p_j}{1/k}\right)\right|$$

$$\leq \sum_{j=1}^{k} \frac{1}{k} \left(\delta \left|\frac{\hat{p}_j}{1/k} - \frac{p_j}{1/k}\right| + \varepsilon\right)$$

$$= \delta V(\hat{P}_n, P_n) + \varepsilon.$$

Therefore

$$\mathsf{E}\left|D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)\right| \leq \delta \mathsf{E}\left[V(\hat{P}_n, P_n)\right] + \varepsilon.$$

Hence

$$\lim_{n\to\infty} \sup \mathsf{E}\left|D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)\right| \leq \varepsilon.$$

This holds for all $\varepsilon > 0$ so

$$\mathsf{E}\left|D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)\right| \to 0 \text{ for } n \to \infty.$$

For $\alpha \in [1; 2]$ the results was proved in [6].
For $\alpha > 2$ we use the Taylor expansion

$$\hat{p}_j^\alpha = p_j^\alpha + \alpha \, p_j^{\alpha-1}(\hat{p}_j - p_j) + \frac{\alpha(\alpha - 1)}{2} \xi_j^{\alpha-2}(\hat{p}_j - p_j)^2 \quad (19)$$

where ξ_j is between p_j and $\hat{p}_j$. We need a highly probable upper bound on $\hat{p}_j$. For this choose some number $b > 1$ and consider the random event

$$E_{nj}(b) = \{\hat{p}_j \geq b \max\{p_j, 1/k\}\}.$$

We shall prove that under assumptions (18)

$$\pi_n(b) \overset{\Delta}{=} \mathsf{P}(\cup_j E_{nj}(b)) \longrightarrow 0. \qquad (20)$$

Obviously,

$$\pi_n(b) \leq \sum_j \mathsf{P}(\hat{p}_j \geq b \max\{p_j, 1/k\})$$

$$\leq \sum_j \exp\{-D_1(Po(b \max\{np_j, n/k\}), Po(np_j))\}$$

$$= \sum_j \exp\{-D_1(Po(bn/k), Po(n/k))\}$$

$$= k \exp\left\{-\left(bn/k \log \frac{bn/k}{n/k} + n/k - bn/k\right)\right\}$$

$$= k \exp\left\{-\frac{n\phi_1(b)}{k}\right\} = k^{1 - n\phi_1(b)/(k \log k)} \qquad (21)$$

for the function $\phi_1(b) > 0$ introduced in (5). Assumption (18) implies that the exponent in (21) tends to $-\infty$ so that (20) holds. Therefore it suffices to prove (12) under the condition that the random events $\cup_j E_{n,j}(b)$ fail to take place, i.e. that

$$\hat{p}_j > b \max\{p_j, 1/k\} \quad \text{for all } \ 1 \leq j \leq k. \quad (22)$$

Under (22) it holds $\xi_j \leq \{bp_j, b/k\}$ and, consequently,

$$\xi_j^{\alpha-2} \leq (\max\{bp_j, b/k\})^{\alpha-2} \leq b^{\alpha-2} p_j^{\alpha-2} + \frac{b^{\alpha-2}}{k^{\alpha-2}}. \quad (23)$$

However, (19) together with (23) implies

$$\left| \hat{p}_j^\alpha - p_j^\alpha \right| \leq \alpha\, p_j^{\alpha-1} |\hat{p}_j - p_j| +$$
$$\frac{\alpha(\alpha-1)b^{\alpha-2}}{2} \left(p_j^{\alpha-2} + \frac{1}{k^{\alpha-2}} \right) (\hat{p}_j - p_j)^2.$$

The rest of the proof is the same as the proof of [6, Thm. 1, Eq. 40]. ∎

IV. BAHADUR FUNCTION

We can now calculate the Bahadur function for the statistic $\mathcal{T}_\alpha = 2nD_\alpha(\hat{P}_n\|U)$ for $\alpha \in]0,1]$. To this end it suffices to evaluate $\inf D_1(P\|U)$ under the constraint $D_\alpha(P\|U) \geq \Delta$ as a function of $\Delta > 0$ [10]. The inequality $D_\alpha(P\|U) \leq D_1(P\|U)$ for $\alpha \in]0;1]$ implies that

$$\inf_{D_\alpha(P\|U) \geq \Delta} D_1(P\|U) \geq \inf_{D_\alpha(P\|U) \geq \Delta} D_\alpha(P\|U)$$
$$= \Delta.$$

Next we note that given $D_\alpha(P\|U) \geq \Delta$ the minimum of $D_1(P\|U)$ is a mixture of two uniform distributions U_l and U_{l+1} on two different subsets where one is a subset of the other and contains one more element. The subscript l indicated the number of elements in the support of U_l. The number l is determined by the condition

$$\log \frac{k}{l+1} \leq \Delta \leq \log \frac{k}{l}.$$

Note that

$$\log \frac{k}{l} - \Delta \leq \log \frac{k}{l} - \log \frac{k}{l+1}$$
$$= \log \left(1 + \frac{1}{l} \right) \leq \frac{1}{l}.$$

Assume that $P = sU_l + (1-s)U_{l+1}$. Then

$$D_1(P\|U) \leq sD_1(U_l\|U) + (1-s)D_1(U_{l+1}\|U)$$
$$= s \log \frac{k}{l} + (1-s) \log \frac{k}{l+1}$$
$$\leq \log \frac{k}{l} \leq \Delta + \frac{1}{l}.$$

For a fixed value of Δ the number l will tend to infinity for k tending to infinity. Hence

$$\inf_{D_\alpha(P\|U) \geq \Delta} D_1(P\|U) \to \Delta \text{ for } k \to \infty.$$

Thus the Bahadur function of the statistic $\mathcal{T}_\alpha$ is $g_\alpha^\mathcal{T}(\Delta) = \Delta$. In particular the relative Bahadur efficiency of two of the statistics $D_\alpha(P\|U)$, $\alpha \in]0;1]$ is 1.

If we use power divergence statistics instead of Rényi divergence we get

$$D_1(P,U) = \log \frac{k}{l}$$

and

$$D_\alpha(P,U) = \frac{\sum_{i=1}^l \left(\frac{1/l}{1/k} \right)^\alpha \frac{1}{k} - 1}{\alpha(\alpha-1)}$$
$$= \frac{1 - \left(\frac{k}{l} \right)^{\alpha-1}}{\alpha(1-\alpha)}$$

implying

$$\frac{k}{l} = (1 - \alpha(1-\alpha)D_\alpha)^{\frac{1}{\alpha-1}}$$

Thus

$$\inf_{D_\alpha \geq \Delta} D_1 \approx \log \frac{k}{l}$$
$$= \log \left((1 - \alpha(1-\alpha)\Delta)^{\frac{1}{\alpha-1}} \right)$$
$$= \frac{1}{\alpha-1} \log (1 + \alpha(\alpha-1)\Delta).$$

This does not depend on n so for $\alpha \in]0;1[$. Hence according to [6], the Bahadur function is

$$g_\alpha^\mathcal{T}(\Delta) = \frac{1}{\alpha-1} \log (1 + \alpha(\alpha-1)\Delta).$$

V. MAIN RESULT AND DISCUSSION

The functions g_α as well as the normalizing sequences $c_{\alpha,n}$ have been explicitly evaluated in [6] using the results of [7] for all $\alpha \geq 1$. The results of Section IV enable us to calculate the explicit Bahadur efficiencies $BE(\mathcal{T}_{\alpha_1} \mid \mathcal{T}_{\alpha_2})$ on the domain $\alpha_1, \alpha_2 > 0$. These efficiencies are given in the following main result.

Theorem 7: Let $0 < \alpha_1 < \alpha_2 < \infty$.
(i) If $\alpha_2 \leq 1$ and

$$\frac{n}{k \ln n} \to \infty \text{ for } n \to \infty \quad (24)$$

then $BE(\mathcal{T}_{\alpha_1} \mid \mathcal{T}_{\alpha_2}) = 1$.
(ii) If $\alpha_2 > 1$ and $k = k_n$ increases to infinity slowly in the sense that

$$\frac{n}{k^{2-1/\alpha_2} \ln n} \text{ for } n \to \infty \quad (25)$$

then

$$BE(\mathcal{T}_{\alpha_1} \mid \mathcal{T}_{\alpha_2}) = \infty. \quad (26)$$

Proof: **(i)** Condition 24 implies that both D_{α_1} and D_{α_2} are consistent.
(ii) According to [6] it is sufficient to prove that (25) implies that D_{α_1} and D_{α_2} are consistent. We have

$$\frac{n}{k \log k} = \frac{n}{k^{2-1/\alpha_2} \ln n} \frac{k^{1-1/\alpha_2}}{\log k} \ln n$$

and each of the factors tends to infinity for n tending to infinity. ∎

The formula for relative efficiency of power divergence instead of Rényi divergence is given by

$$BE\left(T_\alpha \mid T_1\right) = \frac{g_\alpha\left(\Delta\right)}{g_1\left(\Delta\right)} = \frac{\log\left(1 + \alpha\left(\alpha - 1\right)\Delta\right)}{\left(\alpha - 1\right)\Delta}.$$

The above theorem may be interpreted in two ways. One interpretation is that all statistics $D_\alpha\left(\hat{P}_n, U\right)$, $\alpha \in {]0;1]}$ are equally efficient and an other interpretation is that the definition of efficiency is not sufficiently refined to distinguish the different statistics. Here we shall not introduce a new general definition of efficiency but will discuss an example where D_1 is better than $D_\alpha, \alpha \in {]0;1[}$ in distinguishing U from certain alternatives.

Example 8: We consider alternatives $\mathcal{A}_n$ defined by

$$\boldsymbol{X}_n \sim Multinomial_k(n, P_n) \quad \text{under } \mathcal{A}_n$$

where $P_n \in M\left(k\right)$ is given by

$$P_n = s_n\delta + \left(1 - s_n\right)U.$$

We shall fix the power of the tests to some number $p \in {]0;1[}$. Then the acceptance region will be approximately

$$\left\{\hat{P}_n | D_\alpha\left(\hat{P}, U\right) \le D_1\left(P_n, U\right)\right\}.$$

Under the null hypothesis the significance level will satisfy

$$\frac{\log P\left\{\hat{P}_n | D_\alpha\left(\hat{P}, U\right) > D_1\left(P_n, U\right)\right\}}{n}$$
$$\approx \inf_{P \in \left\{\hat{P}_n | D_\alpha\left(\hat{P}, U\right) > D_1\left(P_n, U\right)\right\}} D_1\left(P, U\right).$$

The infimum is achieved for a mixture of uniform distributions on l and $l+1$ points And for such a distribution we have $D_1\left(P, U\right) \approx D_\alpha\left(P, U\right)$ implying that

$$\inf_{P \in \left\{\hat{P}_n | D_\alpha\left(\hat{P}, U\right) > D_1\left(P_n, U\right)\right\}} D_1\left(P, U\right) \approx D_1\left(P_n, U\right)$$

Thus the order of the sample size is given by

$$n_\alpha \approx \frac{\log P\left\{\hat{P}_n | D_\alpha\left(\hat{P}, U\right) > D_1\left(P_n, U\right)\right\}}{D_\alpha\left(P_n, U\right)}$$

and for $\alpha_1, \alpha_2 \in {]0;1]}$ we have

$$\lim \frac{n_{\alpha_1}}{n_{\alpha_2}} = \lim \frac{\frac{\log P\left\{\hat{P}_n | D_{\alpha_2}\left(\hat{P}, U\right) > D\left(P_n, U\right)\right\}}{D_{\alpha_1}\left(P_n, U\right)}}{\frac{\log P\left\{\hat{P}_n | D_{\alpha_2}\left(\hat{P}, U\right) > D\left(P_n, U\right)\right\}}{D_{\alpha_2}\left(P_n, U\right)}}$$
$$= \lim \frac{D_{\alpha_2}\left(P_n, U\right)}{D_{\alpha_1}\left(P_n, U\right)}.$$

For simplicity choose $\alpha_1 = 1$ and s_n such that $D_1\left(P_n, U\right)$ is constant. Then

$$\lim \frac{n_{\alpha_1}}{n_{\alpha_2}} = \frac{\lim D_{\alpha_2}\left(P_n, U\right)}{D_1\left(P_n, U\right)}$$

and this limit is 0 which follows from calculations that were essentially carried out in [6, Eq. 66-75].

The conclusion of the example is that with the specified sequence of alternatives Shannon entropy is asymptotically infinitely more efficient in characterizing the uniform distribution than any Rényi entropy of order $\alpha \in {]0;1[}$. The essential difference between the treatment of the cases $\alpha > 1$ and $\alpha \in {]0;1[}$ is the choice of the sequence of alternatives. In the case $\alpha \in {]0;1[}$ one should choose a sequence of alternatives of the form $P_n = s_n\delta + \left(1 - s_n\right)U$ that is not contiguous to the sequence of uniform distributions. In the case $\alpha > 1$ the sequence of alternatives can be chosen of the form $P_n = U_{l_n}$ where U_{l_n} is uniform on l_n points. The number of points l_n can then be chosen so that $D_\alpha\left(U_{l_n}\|U\right) = \log\frac{n}{l_n}$ is approximately constant, i.e. n/l_n is approximately constant. We note that with these alternatives $D_\alpha\left(P_n\|U\right)$ is independent of α. With this sequence of alternatives the calculations of the relative Bahadur efficiency reduces to evaluating 15 with $\Delta_1 = \Delta_2$. Thus, the Shannon entropy is as efficient as any Rényi entropy of order $\alpha \in {]0;1[}$ for problems satisfying the Bahadur condition like in Example 1, but if the sequence of alternatives is not contiguous Shannon entropy may be more efficient than any Rényi entropy of order $\alpha \in \mathbb{R}_+ \backslash \{1\}$.

VI. Acknowledgement

References

[1] 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.

[2] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.

[3] T. R. C. Read and N. Cressie, *Goodness of Fit Statistics for Discrete Multivariate Data*. Berlin: Springer, 1988.

[4] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, vol. 1, (Berkely), pp. 547–561, Univ. Calif. Press, 1961.

[5] L. Györfi and I. Vajda, "Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models," *Stat. Probab. Letters*, vol. 56, no. 1, pp. 57–67, 2002.

[6] 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.

[7] 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.

[8] I. Vajda, "On convergence of information contained in quantized observations," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2163–2172, Aug. 2002.

[9] P. Harremoës and I. Vajda, "Consistence of various ϕ-divergence statistics," Research Report ÚTIA AV ČR 2218, Institute of Information Theory and Automation, Praha, March 2008.

[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.