Consistency of various ϕ-divergence statistics

Peter Harremoës and Igor Vajda

1 Divergences and divergence statistics

Let $M\,(k)$ be the simplex of all discrete probability distributions $P = (p_j : 1 \leq j \leq k)$ and

$$\hat{P}_n = \left(\hat{p}_{nj} \stackrel{\Delta}{=} \frac{X_{nj}}{n} : 1 \leq j \leq k\right) \tag{1}$$

a statistic based on the multinomially distributed observations

$$\boldsymbol{X}_n = (X_{nj} : 1 \leq j \leq k) \sim \mathit{Mult}_k(n, P_n)$$

for

$$P_n = (p_{nj} : 1 \leq j \leq k) \in M\,(k)\,, \quad n = 1, 2, \ldots \tag{2}$$

If the distributions P_n are unknown, then it is often important to decide whether the uniform hypothesis

$$\mathcal{H} : P_n = U \stackrel{\Delta}{=} (1/k : 1 \leq j \leq k) \in M\,(k)\,, \quad n = 1, 2, \ldots \tag{3}$$

holds or not. The decision is usually based on the value of one of the *ϕ-divergence statistics*

$$T_\phi = T_{\phi,n} = 2n\, D_\phi(\hat{P}_n, U) \tag{4}$$

where on the right is the ϕ-divergence of the empirical distribution $\hat{P}_n$ and the hypothetical distribution U corresponding to a convex function $\phi(t)$, $t > 0$ with $\phi\,(1) = 0$ and with $\phi\,(0)$ defined as the limit for $t \downarrow 0$. For arbitrary distributions $P = (p_j : 1 \leq j \leq k)$ and $Q = (q_j : 1 \leq j \leq k) \in M\,(k)$ the ϕ-divergence $D_\phi(P, Q) \geq 0$ is defined by the formula

$$D_\phi(P, Q) = \sum_{j=1}^{k} q_j\, \phi\left(\frac{p_j}{q_j}\right) \tag{5}$$

(for details about the definition (5) and properties of the ϕ-divergences, see [9] or [12]).

Next follow several simple but important examples from the class of f_α-divergences $D_{f_\alpha}(P, Q)$ defined for all real $\alpha \in \mathbb{R}$ in accordance with (5) by the convex functions $\phi = f_\alpha$ given in the domain $t > 0$ by the formula

$$f_\alpha(t) = \frac{|\,\alpha\,|}{\alpha(\alpha - 1)}\left(2^{\alpha-1}(t+1) - (t^{1/\alpha} + 1)^\alpha\right) \quad \text{when} \quad \alpha \neq 0, 1 \tag{6}$$

and by the corresponding limits

$$f_0(t) =\mid t-1 \mid /2, \tag{7}$$

$$f_1(t) = t\log t + (t+1)\log\frac{2}{t+1}, \quad \log = \log_e . \tag{8}$$

The subclass of these functions for nonnegative parameters $\alpha \geq 0$ was introduced in [10] where it was proved that all corresponding f_α-divergences define metrics $(D_{f_\alpha}(P,Q))^{\pi(\alpha)}$ on the space of probability distributions $M(k)$ for appropriate powers $\pi(\alpha) > 0$ (in fact, the f_α-divergences were introduced and their metricity was proved for the probability measures on arbitrary measurable space).

Example 1. By (6),

$$f_{-1}(t) = \frac{(t-1)^2}{2(t+1)} \tag{9}$$

The f_{-1}-divergence is called LeCam divergence because it first appeared in [8]. By (5),

$$LC(P,Q) = \frac{1}{2}\sum_{j=1}^{k} q_j \frac{\left(\frac{p_j}{q_j}-1\right)^2}{\frac{p_j}{q_j}+1} = \frac{1}{2}\sum_{j=1}^{k}\frac{(p_j-q_j)^2}{p_j+q_j}. \tag{10}$$

The metricity of the square root $(LC(P,Q))^{1/2}$ was proved in [7].

Example 2. The function $2f_0(t) = |t-1|$ (cf. (7)) defines the variational distance

$$V(P,Q) = \sum_{j=1}^{k} q_j \left|\frac{p_j}{q_j}-1\right| = \sum_{j=1}^{k}|p_j-q_j| \quad \text{(cf. (5))}. \tag{11}$$

which plays an important role in information theory and mathematical statistics (cf. [1] or [3]).

Example 3. The metricity of the square root $(D_{f_1}(P,Q))^{1/2}$ for the f_1-divergence given by the function (8) was established independently in [10] and [2]. In [4] this metric was further investigated and was shown to be of the negative type, which means that it admits an isometric embedding into a Hilbert space. Authors of [4] also coined the name Jensen Shannon divergence for $D_{f_1}(P,Q)$. By (5) and (8),

$$\begin{aligned} JS(P,Q) &= \sum_{j=1}^{k} q_j \left(\frac{p_j}{q_j}\log\frac{p_j}{q_j} + \left(\frac{p_j}{q_j}+1\right)\log\frac{2}{\frac{p_j}{q_j}+1}\right) \\ &= \sum_{j=1}^{k}\left(p_j\log\frac{2p_j}{p_j+q_j} + q_j\log\frac{2q_j}{p_j+q_j}\right). \end{aligned} \tag{12}$$

Example 4. By (6),

$$f_2(t) = (1-\sqrt{t})^2.$$

The function $2f_2(t)$ defines so-called Hellinger divergence taking in accordance with (5) the form

$$H(P,Q) = 2\sum_{j=1}^{k}\left(\sqrt{p_j}-\sqrt{q_j}\right)^2. \tag{13}$$

We shall refer to it later.

In (5) is often taken the convex function ϕ which is one of the power functions ϕ_α of order $\alpha \in \mathbb{R}$ given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \text{when} \;\; \alpha(\alpha-1) \neq 0 \tag{14}$$

and by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1 \quad \text{and} \quad \phi_1(t) = t\ln t - t + 1. \tag{15}$$

The ϕ-divergences

$$D_\alpha(P,Q) \triangleq D_{\phi_\alpha}(P,Q), \quad \alpha \in \mathbb{R} \tag{16}$$

based on (14) and (15) are usually referred to as power divergences of orders α. For details about the properties of power divergences, see [9] or [12]. Next we mention the best known members of the family of statistics (4), with a reference to the skew symmetry $D_\alpha(P,Q) = D_{1-\alpha}(Q,P)$ of the power divergences (16).

Example 5. The quadratic divergence (also called χ^2-divergence)

$$D_2(P,Q) = D_{-1}(Q,P) = \frac{1}{2}\sum_{j=1}^{k}\frac{(p_j-q_j)^2}{q_j} \tag{17}$$

leads to the well known Pearson and Neyman statistics

$$T_2 = T_{2,n} = \sum_{j=1}^{k}\frac{(X_{nj}-nq_j)^2}{nq_j} \;\; \text{and} \;\; T_{-1} = T_{-1,n} = \sum_{j=1}^{k}\frac{(X_{nj}-nq_j)^2}{X_{nj}}\,.$$

The logarithmic divergence

$$D_1(P,Q) = D_0(Q,P) = \sum_{j=1}^{k} p_j \ln\frac{p_j}{q_j} \tag{18}$$

leads to the log-likelihood ratio and reversed log-likelihood ratio statistics

$$T_1 = T_{1,n} = 2\sum_{j=1}^{k} X_{nj}\ln\frac{X_{nj}}{nq_j} \;\; \text{and} \;\; T_0 = T_{0,n} = 2nq_j\sum_{j=1}^{k}\ln\frac{nq_j}{X_{nj}}. \tag{19}$$

The symmetric Hellinger divergence $D_{1/2}(P,Q) = D_{1/2}(Q,P) = H(P,Q)$ given in (13) leads to the Freeman–Tukey statistic

$$T_{1/2} = T_{1/2,n} = 4\sum_{j=1}^{k}\left(X_{nj}^{1/2} - (nq_j)^{1/2}\right)^2. \tag{20}$$

Remark 6. Metric divergences $D_\phi(P,Q)$ must be symmetric in P,Q. The symmetry condition is

$$t\phi(1/t)=\phi(t),\quad t>0\quad \text{(cf. [13] or [9])}. \tag{21}$$

The metric divergences $D_{f_\alpha}(P,Q)$ from Examples 1 - 4 can be obtained by the symmetrization of some ϕ-divergences $D_\phi(P,Q)$ based on the formulas

$$D_\phi(P,Q)=D_{f^{(1)}}(P,(P+Q)/2)\quad\text{and}\quad D_\phi(P,Q)=D_{f^{(2)}}(Q,(P+Q)/2) \tag{22}$$

for the convex functions

$$f^{(1)}(u)=(2-u)\phi\left(\frac{u}{2-u}\right)\quad\text{and}\quad f^{(2)}(u)=u\phi\left(\frac{2-u}{u}\right),\quad 0<u<2 \tag{23}$$

(cf. (9) in [13]). This leads for every convex $\phi(t)$, $t>0$ to the inverse formulas

$$D_\phi(P,(P+Q)/2)=D_{\phi^{(1)}}(P,Q)\quad\text{and}\quad D_\phi(Q,(P+Q)/2)=D_{\phi^{(2)}}(P,Q)$$

where

$$\phi^{(1)}(t)=\frac{1+t}{2}\phi\left(\frac{2t}{1+t}\right)\quad\text{and}\quad \phi^{(2)}(t)=\frac{1+t}{2}\phi\left(\frac{2}{1+t}\right),\quad t>0 \tag{24}$$

are the convex functions functions studied previously in [13] and [14]. As a result we get the symmetrized version of arbitrary ϕ-divergence

$$D_\phi(P,(P+Q)/2)+D_\phi(Q,(P+Q)/2)=D_{\phi^{(1+2)}}(P,Q) \tag{25}$$

for the convex function

$$\phi^{(1+2)}(t)=\phi^{(1)}(t)+\phi^{(2)}(t),\quad t>0. \tag{26}$$

Since it holds

$$t\phi^{(1)}(1/t)=\phi^{(2)}(t)\quad\text{and}\quad t\phi^{(2)}(1/t)=\phi^{(1)}(t), \tag{27}$$

the symmetry condition (21) holds for $\phi^{(1+2)}(t)$ as it is expected.

Example 7. By definition, for the total variation $f_0=f_0^{(1)}=f_0^{(2)}$ so that the symmetrized total variation is the total variation itself. For the symmetric Hellinger divergence the corresponding power function $\phi_{1/2}$ leads to new symmetrized function $\phi_{1/2}^{(1+2)}$ with the corresponding $\phi_{1/2}^{(1+2)}$-divergence different from the Hellinger divergence. Therefore the symmetrized Hellinger divergence is not the Hellinger divergence itself. For the quadratic power function ϕ_2 of (14) it holds $\phi_2^{(1+2)}(t)=f_{-1}(t)$ where $f_{-1}(t)$ was defined by (9). Therefore the LeCam divergence is nothing but the symmetrized Pearson divergence.

Remark 8. If the original ϕ-divergence is symmetric then its symmetrized version may be identical (e.g. the total variation) or not identical (e.g. the Hellinger divergence). Similarly, the symmetrization may preserve an already symmetrized divergence (see again the total variation) or it may change it (see e.g. the symmetrization of the symmetrized Pearson divergence).

2 Consistency of ϕ-divergence statistics

Let us consider testing of the hypothesis $\mathcal{H}$ of (3) by means of some power divergence statistic $D_\alpha(\hat{P}_n, U)$. This testing is based on the assumption that the alternative to $\mathcal{H}$ is detectable by $D_\alpha(\hat{P}_n, U)$ in the sense that the values of the statistic significantly differ when $\mathcal{H}$ is true from the case where the alternative characterized by (2) is true. For brevity we denote this alternative by the symbol $\mathcal{A}$ (we say "*under* $\mathcal{H}$" when $\mathcal{H}$ is true and "*under* $\mathcal{A}$" in the opposite case). If the alternative $\mathcal{A}$ and the hypothesis $\mathcal{H}$ are to be logically exclusive then at least one of the distributions P_n in (2) must be nonuniform.

If $D_\alpha(\hat{P}_n, U)$ approximates $D_\alpha(P_n, U)$ well for large n in the sense that the difference $D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)$ tends stochastically to zero, then the above mentioned detectability is achieved if under $\mathcal{A}$ the nonnegative sequence $D_\alpha(P_n, U)$ has a positive limit since under $\mathcal{H}$ this sequence is identically 0. This motivates the following definition. In this definition, and in the rest of the paper, we admit that $k = k_n$ depends on n in a nondecreasing manner with $k_n \longrightarrow \infty$ but the subscript n is dropped in the sequel.

Note that unless otherwise explicitly stated, the convergences are in this paper considered for $n \longrightarrow \infty$.

Definition 9. *We say that the statistic $D_\phi(\hat{P}_n, U)$ is consistent if the alternative is detectable in the sense that there exists $0 < \Delta < \infty$ such that*

$$D_\phi(P_n, U) \longrightarrow \Delta \quad \textit{under } \mathcal{A} \tag{28}$$

and

$$D_\phi(\hat{P}_n, U) \xrightarrow{p} 0 \quad \textit{under } \mathcal{H}, \tag{29}$$

$$D_\phi(\hat{P}_n, U) \xrightarrow{p} \Delta \quad \textit{under } \mathcal{A}. \tag{30}$$

Note that the test rejecting $\mathcal{H}$ when $T_n = 2nD_\phi(\hat{P}_n, U)$ exceeds a critical value $x_{n\alpha} > 0$ is consistent if the statistic $D_\phi(\hat{P}_n, U)$ is consistent in the sense of the present definition. Indeed, (29) implies that the significance level (probability of the wrong decision under $\mathcal{H}$)

$$s = \mathsf{P}(T_n > x_n \mid \mathcal{H}) = \mathsf{P}\left(D_\phi(\hat{P}_n, U) > \frac{x_n}{2n} \,\middle|\, \mathcal{H}\right) \tag{31}$$

preserves a fixed level between 0 and 1 only if $x_{n\alpha}/n \longrightarrow 0$. However, then (30) implies that the test power (probability of the correct decision under $\mathcal{A}$)

$$\pi_n = \mathsf{P}(T_n > x_n \mid \mathcal{A}) = \mathsf{P}\left(D_\phi(\hat{P}_n, U) > \frac{x_n}{2n} \,\middle|\, \mathcal{A}\right) \tag{32}$$

tends to 1 which means the consistency of the test.

The importance of consistency of the power divergence statistics for conclusions about their relative Bahadur efficiencies was investigated in [6]. The present consistency definition is strictly weaker than the one considered in [6] and [11] (dealing only with $\alpha = 1$ and $\alpha = 2$) where (29) was replaced by $\mathsf{E}\left(D_\alpha\left(\hat{P}_n, U\right) \mid \mathcal{H}\right) \to 0$. To this end we show in the next example that this stronger consistency holds for $\alpha \geq 3$ only if

$$\frac{n}{k^2} \to \infty. \tag{33}$$

while later we prove that the present weaker consistency holds for all $\alpha > 2$ already if $n/(k \log k) \to \infty$.

Example 10. For $\alpha = 3$ we get

$$\mathsf{E}D_3\left(\hat{P}_n, U\right) = \frac{k^2 \mathsf{E}\left[\sum_{j=1}^k \hat{p}_j^3\right] - 1}{6} \tag{34}$$

where

$$\hat{p}_j^3 = p_{nj}^3 + 3p_{nj}^2\left(\hat{p}_j - p_{nj}\right) - 3p_j\left(\hat{p}_j - p_{nj}\right)^2 + \left(\hat{p}_j - p_{nj}\right)^3. \tag{35}$$

Therefore

$$\mathsf{E}D_3\left(\hat{P}_n, U\right) = \frac{k^2 p_{nj}^3 - 1}{6} + \frac{k^2}{6}\sum_{j=1}^k \left(\frac{3\mathsf{E}\left[\left(\hat{p}_j - p_{nj}\right)^2\right]}{k} + \mathsf{E}\left[\left(\hat{p}_j - p_{nj}\right)^3\right]\right).$$

Since $p_{nj} = 1/k$ under $\mathcal{H}$, we get

$$\begin{aligned}\mathsf{E}\left(D_3\left(\hat{P}_n, U\right) \mid \mathcal{H}\right) &= \frac{k^2}{6}\sum_{j=1}^k \left(3\frac{(1-1/k)/k}{nk} + \frac{\frac{1}{k}(1-1/k)(1-2/k)}{n}\right) \\ &= \frac{k^2}{6n}\sum_{j=1}^k \left(\frac{1}{k} - \frac{1}{k^3}\right) = \frac{k^2-1}{6n}.\end{aligned} \tag{36}$$

Hence $E\left(D_3(\hat{P}_n, U) \mid \mathcal{H}\right)$ tends to zero only if (33) holds.

In Section 3 we need the following auxiliary result.

Lemma 11. *For $0 \le x < 1$, $0 \le y \le 1$ and $1 < \alpha < 2$ it holds*

$$|y^\alpha - x^\alpha| \le \alpha x^{\alpha-1}|y - x| + (\alpha - 1)x^{\alpha-2}(y-x)^2. \tag{37}$$

Proof. Since $(\alpha-1)x^{\alpha-2}(y-x)^2$ is nonnegative, it suffices to prove

$$y^\alpha \ge x^\alpha + \alpha x^{\alpha-1}(y - x) \tag{38}$$

and

$$y^\alpha \le x^\alpha + \alpha x^{\alpha-1}(y-x) + (\alpha-1)x^{\alpha-2}(y-x)^2. \tag{39}$$

But (38) is evident since the function $y \to y^\alpha$ is convex. We shall prove that the function

$$f(y) = y^\alpha - \left(x^\alpha + \alpha x^{\alpha-1}(y-x) + (\alpha-1)x^{\alpha-2}(y-x)^2\right) \tag{40}$$

is non-positive on $[0,1]$. First we observe that $f(0) = f(1) = 0$. By differentiating $f(y)$ we get

$$f'(y) = \alpha y^{\alpha-1} - \left(\alpha x^{\alpha-1} + (\alpha-1)x^{\alpha-2}2(y-x)\right) = \alpha y^{\alpha-1} + (\alpha-2)x^{\alpha-1} + (2-2\alpha)x^{\alpha-2}y \tag{41}$$

so that $f'(x) = 0$. Differentiating once more we get

$$f''(y) = \alpha(\alpha-1)y^{\alpha-2} + (2-2\alpha)x^{\alpha-2} = (\alpha-1)\left(\alpha y^{\alpha-2} - 2x^{\alpha-2}\right). \tag{42}$$

Thus $f''(y) > 0$ for $y < x_\alpha \stackrel{\Delta}{=} (\alpha/2)^{\frac{1}{2-\alpha}} x$ and $f''(y) < 0$ for $y > x_\alpha$. Since $x_\alpha < x$ and $f(y)$ is concave on $[x_\alpha, 1]$, it is maximized on this interval at $y = x$ where $f(x) = 0$. Thus $f(y) \le 0$ on this interval and in particular $f(x_\alpha) \le 0$. This together with $f(0) = 0$ and the convexity of $f(y)$ on the interval $[0,\, x_\alpha]$ implies $f(y) \le 0$ on this interval. This completes the proof of the non-positivity, i.e. the proof of (39). ■

The main results of this paper are two general consistency theorems. One of them, formulated for ϕ-divergences, is given in this section. The other one, formulated for power divergences, is given in the next section. We start with some simple particular results useful in the proofs of general results, which however might be also of independent interest.

Theorem 12. *If the detectability condition (28) holds and*

$$\frac{n}{k} \longrightarrow \infty \tag{43}$$

then the Pearson divergence $D_2(\hat{P}_n, U)$ is consistent.

Proof. Since

$$\mathsf{E}(\hat{p}_j - p_{nj})^2 = \frac{p_{nj}(1-p_{nj})}{n} \le \frac{p_{nj}}{n}, \quad 1 \le j \le k \tag{44}$$

it is obvious that

$$\mathsf{E}D_2\left(\hat{P}_n, P_n\right) = \sum_{j=1}^{k} \frac{\mathsf{E}\left(\hat{p}_j - p_{nj}\right)^2}{p_{nj}} \le \sum_{j=1}^{k} \frac{1}{n} = \frac{k}{n}\,. \tag{45}$$

■

Theorem 13. *If the detectability conditions (28) and (43) hold then the variational distance $V(\hat{P}_n, U)$ is consistent.*

Proof. Variational distance is a metric so that

$$\left|V(\hat{P}_n, U) - V(P_n, U)\right| \le V(\hat{P}_n, P_n). \tag{46}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} V(\hat{P}_n, P_n) &= \sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right| p_j = \sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right| p_j^{1/2} p_j^{1/2} \\ &\le \left(\sum_{j=1}^{k} \left|\frac{\hat{p}_j}{p_j} - 1\right|^2 p_j\right)^{1/2} \cdot \left(\sum_{j=1}^{k} p_j\right)^{1/2} \\ &= \left(D_2\left(\hat{P}_n, P_n\right)\right)^{1/2}. \end{aligned} \tag{47}$$

Hence

$$\mathsf{E}V(\hat{P}_n, P_n) \leq \left(\mathsf{E}D_2\left(\hat{P}_n, P_n\right)\right)^{1/2} \leq \left(\frac{k}{n}\right)^{1/2}. \tag{48}$$

■

As well known, the right derivative ϕ'_+ of the convex function ϕ exists and is non-decreasing. Define

$$\phi'_+(\infty) = \lim_{t\to\infty} \phi'_+(t). \tag{49}$$

Remark 14. It is easy to verify that a continuous convex function $\phi : [0;\infty[\to R$ is uniformly continuous if and only if $\phi'_+(\infty) < \infty$. Notice that the condition $\phi(0) + \phi'_+(\infty) < \infty$ is weaker than $\phi(0) + \phi^*(0) < \infty$ where $\phi^*(0)$ is continuous extension of $\phi^*(t) = t\phi(1/t)$ to $t = 0$ because for every $0 < t < 1$

$$\frac{\phi^*(t) - t}{1-t} \leq \phi'_+(1/t) \leq \phi'_+(\infty). \tag{50}$$

As proved in [13] (see also [9]), ϕ-divergences take on values between 0 and $\phi(0) + \phi^*(0)$. Hence the ϕ-divergences with uniformly continuous functions ϕ are bounded but in the reversed direction this statement is not in general true. It is true e.g. if $\phi(0) < 0$ and the symmetry (21) takes place.

Theorem 15. *Let $\phi : [0;\infty[\to \mathbb{R}$ be uniformly continuous. If the conditions (28) and 43) hold then $D_\phi(\hat{P}_n, U)$ is consistent.*

Proof. First assume that $\left|\phi'_+(0)\right| < \infty$. Then by convexity of ϕ we have $\phi'_+(0) \leq \phi'(t) \leq \phi'_+(\infty)$ for all x. Define $\lambda = \max\left\{\left|\phi'_+(0)\right|, \left|\phi'_+(\infty)\right|\right\}$. Then ϕ is Lipschitz with the Lipschitz constant λ , i.e. $|\phi(t) - \phi(s)| \leq \delta|t - s|$ for all $t, s > 0$. Then

$$\begin{aligned}
\left|D_\phi(\hat{P}_n, U) - D_\phi(P_n, U)\right| &= \left|\sum_{j=1}^k \frac{1}{k}\phi\left(\frac{\hat{p}_j}{1/k}\right) - \sum_{j=1}^k \frac{1}{k}\phi\left(\frac{p_j}{1/k}\right)\right| \\
&= \sum_{j=1}^k \frac{1}{k}\left|\phi\left(\frac{\hat{p}_j}{1/k}\right) - \phi\left(\frac{p_j}{1/k}\right)\right| \\
&\leq \sum_{j=1}^k \frac{\lambda}{k}\left|\frac{\hat{p}_j}{1/k} - \frac{p_j}{1/k}\right| \\
&= \delta V(\hat{P}_n, P_n).
\end{aligned} \tag{51}$$

Therefore

$$\mathsf{E}\left|D_\phi(\hat{P}_n, U) - D_\phi(P_n, U)\right| \leq \delta\mathsf{E}V(\hat{P}_n, P_n) \leq \delta\left(\frac{k}{n}\right)^{1/2}. \tag{52}$$

If $\phi'_+(0) = -\infty$ choose some $t_0 > 0$ and define

$$\phi^*(t) = \begin{cases} \phi(t) & \text{for } t \geq t_0, \\ \phi(t_0) + \phi'_+(t_0)(t - t_0) & \text{for } t < t_0. \end{cases} \tag{53}$$

Then

$$0 \leq \phi(t) - \phi^*(t) \leq \phi(0) - \phi^*(0) \tag{54}$$

and

$$0 \leq D_{\phi}\left(P,Q\right) - D_{\phi^*}\left(P,Q\right) \leq \phi\left(0\right) - \phi^*\left(0\right). \tag{55}$$

The function ϕ^* is Lipschitz with the Lipschitz constant $\max\left\{\left|\phi'_+\left(x_0\right)\right|, \phi'_+\left(\infty\right)\right\}$. This implies that

$$\mathsf{E}\left|D_{\phi^*}(\hat{P}_n,U) - D_{\phi^*}(P_n,U)\right| \leq \max\left\{\left|\phi'_+\left(t_0\right)\right|, \phi'_+\left(\infty\right)\right\}\left(\frac{k}{n}\right)^{1/2}. \tag{56}$$

Therefore

$$\mathsf{E}\left|D_{\phi}(\hat{P}_n,U) - D_{\phi}(P_n,U)\right| \leq 2\left(\phi\left(0\right) - \phi^*\left(0\right)\right) + \max\left\{\left|\phi'_+\left(t_0\right)\right|, \phi'_+\left(\infty\right)\right\}\left(\frac{k}{n}\right)^{1/2} \tag{57}$$

and

$$\limsup_{n\to\infty}\ \mathsf{E}\left|D_{\phi}(\hat{P}_n,U) - D_{\phi}(P_n,U)\right| \leq 2\left(\phi\left(0\right) - \phi^*\left(0\right)\right). \tag{58}$$

This holds for all $t_0 > 0$. The result follows because $\phi^*\left(0\right) \to \phi\left(0\right)$ for $t_0 \to 0$. ■

The functions $f_\alpha(t)$ defined by (6) - (8) are continuous in $0 \leq t < \infty$ and differentiable at $t > 0$ different from 1 with the derivatives

$$f'_\alpha(t) = \begin{cases} \frac{t-1}{2|t-1|}, & \text{when } \ \alpha = 0\ ; \\ \frac{|\alpha|}{\alpha(\alpha-1)}\left(2^{\alpha-1} - \frac{(t^{1/\alpha}+1)^{\alpha-1}}{t^{1-1/\alpha}}\right), & \text{when } \ \alpha \neq 0,1\ ; \\ \log 2 - \log\frac{t+1}{t}, & \text{when } \ \alpha = 1. \end{cases} \tag{59}$$

By the symmetry condition (21),

$$f_\alpha(t) = t f_\alpha(1/t), \quad t > 0. \tag{60}$$

Thus at the differentiability points

$$f'_\alpha(t) = f_\alpha(1/t) - \frac{f'_\alpha(1/t)}{t}. \tag{61}$$

Together with the above given formulas for $f'_\alpha(t)$, this implies

$$f_\alpha(0) = \begin{cases} f'_\alpha(\infty) = \frac{1}{2}, & \text{when } \ \alpha = 0\ ; \\ \frac{|\alpha|}{\alpha}\cdot\frac{2^{\alpha-1}-1}{\alpha-1}, & \text{when } \ \alpha \neq 0,1\ ; \\ \log 2, & \text{when } \ \alpha = 1. \end{cases} \tag{62}$$

Hence we see from Remark 14 that all functions $f_\alpha(t)$ defined by (6) – (8) are uniformly continuous. Hence all f_α-divergence statistics are consistent if $n/k \to \infty$. The above directly studied total variation statistic as well as the Le Cam, the Jensen Shannon and the Hellinger statistics are consistent if $n/k \to \infty$. Similarly one can prove that all ϕ-divergence statistics defined by the symmetrized ϕ-divergences are consistent if $n/k \to \infty$. However, the above studied Pearson statistic as well as the important log-likelihood statistics are not in this class, and thus their consistency is not covered by the previous general theorem. The general theorem presented in Section 3 is thus an important complement of what is established in this section.

3 Consistency of power divergence statistics

In this section we study the consistency of the class of power divergence statistics $D_\alpha(\hat{P}_n, U)$, $\alpha \in \mathbb{R}, \alpha \neq 0$. We consider the corresponding versions of detectability and consistency,

$$D_\alpha(P_n, U) \longrightarrow \Delta_\alpha \quad \text{under } \mathcal{A} \tag{63}$$

and

$$D_\alpha(\hat{P}_n, U) \xrightarrow{p} 0 \qquad \text{under } \mathcal{H}, \tag{64}$$

$$D_\alpha(\hat{P}_n, U) \xrightarrow{p} \Delta_\alpha \quad \text{under } \mathcal{A}. \tag{65}$$

The main result of the paper is the following theorem.

Theorem 16. *If detectability condition (63) holds then $D_\alpha(\hat{P}_n, U)$ is consistent provided*

$$\alpha < 0 \qquad \textit{and} \qquad \frac{n}{k^{2-1/\alpha} \log k} \longrightarrow \infty, \tag{66}$$

or

$$0 < \alpha \leq 2 \qquad \textit{and} \qquad \frac{n}{k} \longrightarrow \infty, \tag{67}$$

or

$$\alpha > 2 \qquad \textit{and} \qquad \frac{n}{k \log k} \longrightarrow \infty. \tag{68}$$

Proof. Let $\alpha \in \mathbb{R}$ be arbitrary fixed. Under $\mathcal{H}$ we have $D_\alpha(P_n, U) = D_\alpha(U, U) = 0$. Hence it suffices to prove

$$|\Lambda_{\alpha,n}| \xrightarrow{p} 0 \quad \text{under both } \mathcal{H} \text{ and } \mathcal{A} \tag{69}$$

for $\Lambda_{\alpha,n} = D_\alpha(\hat{P}_n, U) - D_\alpha(P_n, U)$. For simplicity we skip the subscript n in the symbols $\hat{P}_n, P_n$, i.e. we substitute

$$\hat{P}_n = \hat{P} = (\hat{p}_j : 1 \leq j \leq k), \quad P_n = P = (p_j : 1 \leq j \leq k). \tag{70}$$

This leads to the simplified formula $\Lambda_{\alpha,n} = D_\alpha(\hat{P}, U) - D_\alpha(P, U)$. We can without loss of generality assume that $D_\alpha(P, U)$ is constant not only under $\mathcal{H}$ (where the constant is automatically 0) but also under $\mathcal{A}$ (where the assumed detectability implies only the convergence $D_\alpha(P, U) \longrightarrow \Delta_\alpha$). In other words, this assumption says that for all $n = 1, 2, ...$

$$\Delta_\alpha = D_\alpha(P, U) = \frac{\sum \frac{1}{k} \left(\frac{p_j}{1/k} \right)^\alpha - 1}{\alpha\,(\alpha - 1)} = \frac{1}{\alpha(\alpha - 1)} \left(k^{\alpha - 1} \sum_{j=1}^{k} p_j^\alpha - 1 \right) \tag{71}$$

under both $\mathcal{H}$ and $\mathcal{A}$. Obviously, $\Delta_\alpha = 0$ under $\mathcal{H}$ because then $P = U \stackrel{\Delta}{=} (1/k, \ldots, 1/k)$ and $0 < \Delta_\alpha < \infty$ under $\mathcal{A}$. Therefore it suffices to prove (69) for

$$\Lambda_{\alpha,n} = D_\alpha(\hat{P}, U) - \Delta_\alpha \quad \text{with } \Delta_\alpha \text{ given by (71).} \tag{72}$$

If $\alpha \neq 0, 1$ then (71) leads to the useful formula

$$\sum_{j=1}^{k} p_j^{\alpha} = [\Delta_\alpha \alpha (\alpha - 1) + 1] k^{1-\alpha}. \tag{73}$$

In the proof we treat the following cases separately :

$$\mathbf{i} : \alpha < 0, \quad \mathbf{ii} : 0 < \alpha < 1, \quad \mathbf{iii} : \alpha = 1, \quad \mathbf{iv} : 1 < \alpha \leq 2, \quad \text{and} \quad \mathbf{v} : \ \alpha > 2.$$

Case i: $\alpha < 0$. The distribution of the random variable $X_j = X_{nj}$ appearing in (1) is approximately Poisson, $Po(np_j)$, so that

$$\mathsf{P}(\hat{p}_j \leq bp_j) = \mathsf{P}(X_j \leq bnp_j) \leq \exp\{-D_1(Po(bnp_j), Po(np_j))\}.$$

But

$$D_1(Po(bnp_j), Po(np_j)) = bnp_j \log \frac{bnp_j}{np_j} + np_j - bnp_j = n\phi_1(b) p_j$$

for ϕ_1 defined in (15). Therefore the probability

$$\boldsymbol{\pi}_n \stackrel{\Delta}{=} \mathsf{P}(\cup_j E_{nj}) \tag{74}$$

of the union of events $E_{nj} = \{\hat{p}_j \leq bp_j\}$ is upperbounded by

$$\sum_{j=1}^{k} \exp\{-n\phi_1(b) p_j\} \leq k \exp\{-n\phi_1(b) p_{\min}\}$$

where, by (73),

$$p_{\min} \geq \left((\Delta_\alpha \alpha (\alpha - 1) + 1) k^{1-\alpha}\right)^{1/\alpha} = [\Delta_\alpha \alpha (\alpha - 1) + 1]^{1/\alpha} k^{(1-\alpha)/\alpha}. \tag{75}$$

Consequently,

$$\begin{aligned}
\boldsymbol{\pi}_n &\leq k \exp\left\{-n\phi_1(b) (\Delta_\alpha \alpha (\alpha - 1) + 1)^{1/\alpha} k^{(1-\alpha)/\alpha}\right\} \\
&= k \exp\left\{-\frac{n}{k^{1-1/\alpha}} \phi_1(b) [\Delta_\alpha \alpha (\alpha - 1) + 1]^{1/\alpha}\right\} \\
&= \exp\left\{\log k \left(1 - \frac{n}{k^{1-1/\alpha} \log k} \phi_1(b) [\Delta_\alpha \alpha (\alpha - 1) + 1]^{1/\alpha}\right)\right\}.
\end{aligned}$$

From here we see that assumption (66) implies the convergence $\boldsymbol{\pi}_n \longrightarrow 0$. This means that it suffices to prove (69) under the condition that the random events $\cup_j E_{nj}$ do not take place, i.e. that

$$\hat{p}_j > bp_j \quad \text{for all} \ \ 1 \leq j \leq k. \tag{76}$$

This is done in the next paragraph.

The second order Taylor expansion of y^α gives

$$y^\alpha = x^\alpha + \alpha x^{\alpha-1} (y - x) + \alpha (\alpha - 1) \zeta^{\alpha-2} (y - x)^2 / 2 \tag{77}$$

for ζ between x and y. Therefore

$$\begin{aligned} |y^\alpha - x^\alpha| &\le \alpha x^{\alpha-1}|y-x| + \alpha(\alpha-1)\zeta^{\alpha-2}(y-x)^2/2 \\ &\le \alpha x^{\alpha-1}|y-x| + \alpha(\alpha-1)\max\left\{x^{\alpha-2}, y^{\alpha-2}\right\}(y-x)^2/2. \end{aligned} \tag{78}$$

First we note that $x \to x^\alpha$ is convex so that $y^\alpha \ge x^\alpha + \alpha x^{\alpha-1}(y-x)$. Thus we get

$$\begin{aligned} D_\alpha\left(\hat{P}, U\right) &= \frac{\sum \frac{1}{k}\left(\left(\frac{\hat{p}_j}{1/k}\right)^\alpha - 1\right)}{\alpha(\alpha-1)} = \frac{1}{\alpha(\alpha-1)}\left(\sum \frac{\hat{p}_j^\alpha}{k^{1-\alpha}} - 1\right) \\ &\ge \frac{1}{\alpha(\alpha-1)}\left(\sum \frac{p_j^\alpha + \alpha p_j^{\alpha-1}(\hat{p}_j - p_j)}{k^{1-\alpha}} - 1\right) \\ &= \Delta_\alpha + \frac{\sum p_j^{\alpha-1}(\hat{p}_j - p_j)}{(\alpha-1)k^{1-\alpha}} \qquad \text{(cf. (73))}. \end{aligned} \tag{79}$$

Under (76),

$$\begin{aligned} \left|\hat{p}_j^\alpha - p_j^\alpha\right| &\le \alpha p_j^{\alpha-1}|\hat{p}_j - p_j| + \alpha(\alpha-1)\max\left\{p_j^{\alpha-2}, \hat{p}_j^{\alpha-2}\right\}(\hat{p}_j - p_j)^2 \\ &\le \alpha p_j^{\alpha-1}|\hat{p}_j - p_j| + \alpha(\alpha-1)b^{\alpha-2}p_j^{\alpha-2}(\hat{p}_j - p_j)^2. \end{aligned} \tag{80}$$

By inserting this in (72), using the Schwarz inequality and applying (73) we obtain

$$\begin{aligned} |\Lambda_{\alpha,n}| &\le \frac{\sum p_j^{\alpha-1}|\hat{p}_j - p|}{(\alpha-1)k^{1-\alpha}} + \frac{b^{\alpha-2}\sum p_j^{\alpha-2}(\hat{p}_j - p_j)^2}{\alpha(\alpha-1)k^{1-\alpha}} \\ &\le \frac{\left(\sum\left(p_j^{\alpha/2}\right)^2\right)^{1/2}\left(\sum p_j^{\alpha-2}(\hat{p}_j - p_j)^2\right)^{1/2}}{(\alpha-1)k^{1-\alpha}} + \frac{b^{\alpha-2}\sum p_j^{\alpha-2}(\hat{p}_j - p_j)^2}{\alpha(\alpha-1)k^{1-\alpha}} \\ &= \frac{[\alpha(\alpha-1)\Delta_\alpha + 1]^{1/2}}{(\alpha-1)}\left(\frac{\sum p_j^{\alpha-2}(\hat{p}_j - p)^2}{k^{1-\alpha}}\right)^{1/2} + \frac{b^{\alpha-2}}{\alpha(\alpha-1)}\frac{\sum p_j^{\alpha-2}(\hat{p}_j - p_j)^2}{k^{1-\alpha}}. \end{aligned} \tag{81}$$

Since

$$\mathsf{E}\sum_{j=1}^{k} p_j^{\alpha-2}(\hat{p}_j - p_j)^2 = \frac{\sum p_j^{\alpha-1}(1-p_j)}{n} \le \frac{\sum p_j^{\alpha-1}}{n}, \tag{82}$$

we see that under (76) the desired relation (69) holds if

$$\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha}n} \longrightarrow 0 \quad \text{under both } \mathcal{H} \text{ and } \mathcal{A} \tag{83}$$

follows from assumption (66). Here

$$\begin{aligned} \sum p_j^{\alpha-1} &\le k p_{\min}^{\alpha-1} \\ &\le k\left([\Delta\alpha(\alpha-1)+1]^{1/\alpha}k^{(1-\alpha)/\alpha}\right)^{\alpha-1} \qquad \text{(cf. (75))} \\ &= [\Delta\alpha(\alpha-1)+1]^{(\alpha-1)/\alpha}k^{1-(\alpha-1)^2/\alpha}. \end{aligned} \tag{84}$$

Therefore

$$\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha}n} \le \frac{\left[\Delta\alpha\left(\alpha-1\right)+1\right]^{(\alpha-1)/\alpha} k^{1-(\alpha-1)^2/\alpha}}{k^{1-\alpha}n} = \left[\Delta\alpha\left(\alpha-1\right)+1\right]^{(\alpha-1)/\alpha} \frac{k^{2-1/\alpha}}{n}. \quad (85)$$

Since the right hand side tends to zero under (66), we see that (83) is valid.

Case ii: $0 < \alpha < 1$. It is easy to check that in this case the function $\phi_\alpha(t)$ given in (14) is uniformly continuous so that the desired consistency follows from Theorem 15.

Case iii: $\alpha = 1$. This case was treated in [6]. For the sake of completeness we repeat the argument here. By (72),

$$\Lambda_{1,n} = \sum_{j=1}^{k} \left(\hat{p}_j \log \hat{p}_j - p_j \log p_j\right) = \sum_{j=1}^{k} \hat{p}_j \log \frac{\hat{p}_j}{p_j} - \sum_{j=1}^{k} (\hat{p}_j - p_j) \log \frac{1}{p_j} \quad (86)$$

so that

$$|\Lambda_{1,n}| \le D_1(\hat{P}_n, P_n) + \left| \sum_{j=1}^{k} (\hat{p}_j - p_j) \log \frac{1}{p_j} \right|. \quad (87)$$

Since $D_1(\hat{P}_n, P_n) \le 2D_2(\hat{P}_n, P_n)$, it holds

$$\mathsf{E}\,|\Lambda_{1,n}| \le 2\mathsf{E}\, D_2(\hat{P}_n, P_n) + \mathsf{E} \left| \sum_{j=1}^{k} (\hat{p}_j - p_j) \log \frac{1}{p_j} \right|. \quad (88)$$

Let $X_j = X_{nj}$ be the observations introduced in (1) and $\mathsf{Cov}\,(X_i, X_j)$ and $\mathsf{Var}\,(X_i)$ their covariances and variances. Then, using Jensen's inequality, the last term in (88) can be bounded as follows:

$$\begin{aligned} \mathsf{E} \left| \sum_{j=1}^{k} (\hat{p}_j - p_j) \log \frac{1}{p_j} \right| &\le \left(\mathsf{E} \left[\sum_{j=1}^{k} (\hat{p}_j - p_j) \log \frac{1}{p_j} \right]^2 \right)^{1/2} \\ &= \left(\sum_{i,j=1}^{k} \log p_j \log p_i \mathsf{Cov}\,(\hat{p}_i, \hat{p}_j) \right)^{1/2} \\ &= \left(\sum_{i,j=1}^{k} \log p_j \log p_i \frac{\mathsf{Cov}\,(\hat{n}_i, \hat{n}_j)}{n^2} \right)^{1/2}. \end{aligned} \quad (89)$$

Further,

$$\begin{aligned} &\sum_{i,j=1}^{k} \log p_j \log p_i \frac{\mathsf{Cov}\,(\hat{n}_i, \hat{n}_j)}{n^2} = \sum_{i=1}^{k} (\log p_i)^2 \frac{\mathsf{Var}\,(\hat{n}_i)}{n^2} + \sum_{i \ne j} \log p_j \log p_i \frac{\mathsf{Cov}\,(\hat{n}_i, \hat{n}_j)}{n^2} \\ &\le \sum_{j=1}^{k} (\log p_j)^2 \frac{p_j}{n} + \sum_{i \ne j} \log p_j \log p_i \frac{n p_i p_j}{n^2} = \frac{1}{n} \sum_{j=1}^{k} p_i^2 \log p_j + \frac{1}{n} \left(\sum_{j=1}^{k} p_j \log p_j \right)^2. \end{aligned} \quad (90)$$

The function $x \to x \ln^2 x$ is concave in the interval $[0; e^{-1}]$ and convex in the interval $[e^{-1}; 1]$. Therefore we we can apply the method of [5, Theorem 3.1] to verify that $\sum_{i=1}^{k} p_i (\ln p_i)^2$ attains its maximum for a mixture of uniform distributions on l and $l-1$ points for some $l \le k$. For $l \ge 2$ we have

$$\sum_{i=1}^{k} p_i \log^2 p_j \le \sum_{i=1}^{l} \frac{1}{l-1} \ln^2 \left(\frac{1}{l}\right) = \frac{l \log^2 l}{l-1} \le 2 \log^2 k. \tag{91}$$

Inequality 91 trivially holds for $l = 1$. The sum $\sum_{i=1}^{k} p_i \log p_j$ equals minus the entropy, which has maximum $\log k$. By combining (88), (89) and (90) we get

$$\mathsf{E}\,|\Lambda_{1,n}| \le \frac{2k}{n} + \left(\frac{3 \log^2 k}{n}\right)^{1/2}. \tag{92}$$

Under assumptions (67) the right hand side tends to zero so that the desired relation (69) holds.

Case iv: $1 < \alpha \le 2$. Here we get from (72)

$$\Lambda_{\alpha,n} = \frac{k^{\alpha-1}}{\alpha\,(\alpha-1)} \sum_{j=1}^{k} (\hat{p}_j^{\alpha} - p_j^{\alpha}) \tag{93}$$

so that Lemma 11 implies

$$\begin{aligned} |\Lambda_{\alpha,n}| &\le \frac{k^{\alpha-1}}{\alpha\,(\alpha-1)} \sum_{j=1}^{k} \left(\alpha p_j^{\alpha-1} |\hat{p}_j - p_j| + (\alpha-1)\, p_j^{\alpha-2} (\hat{p}_j - p_j)^2\right) \qquad (94) \\ &\le \frac{\left(\sum \left(p_j^{\alpha/2}\right)^2\right)^{1/2} \left(\sum p_j^{\alpha-2} (\hat{p}_j - p)^2\right)^{1/2}}{(\alpha-1)\, k^{1-\alpha}} + \frac{k^{\alpha-1}}{\alpha} \sum_{j=1}^{k} p_j^{\alpha-2} (\hat{p}_j - p_j)^2. \end{aligned}$$

Employing the expectation formula (82) we see that under condition (76) the desired relation (69) holds if (83) follows from assumptions (67). To prove the latter take into account that the function $x \to x^{\alpha-1}$ is concave and thus the Jensen inequality implies

$$\sum_{j=1}^{k} p_j^{\alpha-1} \le k \left(\sum_{j=1}^{k} \frac{p_j}{k}\right)^{\alpha-1} = k^{2-\alpha}. \tag{95}$$

Therefore

$$\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha} n} \le \frac{k^{2-\alpha}}{k^{1-\alpha} n} = \frac{k}{n}. \tag{96}$$

so that (67) implies (83).

Case v: $\alpha > 2$. Here $\Lambda_{\alpha,n}$ is given by (93) as in the Case iv. Similarly as in (77), we use the Taylor expansion

$$\hat{p}_j^{\alpha} = p_j^{\alpha} + \alpha\, p_j^{\alpha-1} (\hat{p}_j - p_j) + \frac{\alpha(\alpha-1)}{2}\, \xi_j^{\alpha-2} (\hat{p}_j - p_j)^2 \tag{97}$$

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 (68)

$$\boldsymbol{\pi}_n(b) \stackrel{\Delta}{=} \mathsf{P}\left(\cup_j E_{nj}(b)\right) \longrightarrow 0. \tag{98}$$

Obviously,

$$\begin{aligned}
\boldsymbol{\pi}_n(b) &\leq \sum_j \mathsf{P}\left(\hat{p}_j \geq b \max\{p_j, 1/k\}\right) \\
&\leq \sum_j \exp\{-D_1\left(Po\left(b \max\{np_j, n/k\}\right), Po\left(np_j\right)\right)\} \\
&= \sum_j \exp\left\{-D_1\left(Po\left(bn/k\right), Po\left(n/k\right)\right)\right\} \\
&= k \exp\left\{-\left(bn/k \log \frac{bn/k}{n/k} + n/k - bn/k\right)\right\} \\
&= k \exp\left\{-\frac{n\phi_1\left(b\right)}{k}\right\} = k^{1-n\phi_1(b)/(k \log k)}
\end{aligned} \tag{99}$$

for the function $\phi_1(b) > 0$ introduced in (15). Assumption (68) implies that the exponent in (99) tends to $-\infty$ so that (98) holds. Therefore it suffices to prove (69) under the condition that the random events $\cup_j E_{n,j}(b)$ fail to take place, i.e. that

$$\hat{p}_j > b \max\{p_j, 1/k\} \quad \text{for all} \;\; 1 \leq j \leq k. \tag{100}$$

This is done in the next paragraph.

Under (100) it holds $\xi_j \leq \{bp_j, b/k\}$ and, consequently,

$$\xi_j^{\alpha-2} \leq \left(\max\{bp_j, b/k\}\right)^{\alpha-2} \leq b^{\alpha-2} p_j^{\alpha-2} + \frac{b^{\alpha-2}}{k^{\alpha-2}}. \tag{101}$$

However, (97) together with (101) implies

$$\left|\hat{p}_j^{\alpha} - p_j^{\alpha}\right| \leq \alpha\, p_j^{\alpha-1} \left|\hat{p}_j - p_j\right| + \frac{\alpha(\alpha-1)b^{\alpha-2}}{2} \left(p_j^{\alpha-2} + \frac{1}{k^{\alpha-2}}\right) (\hat{p}_j - p_j)^2. \tag{102}$$

Hence, under (100) $\Lambda_{\alpha,n}$ is bounded above by

$$\frac{k^{\alpha-1}}{\alpha\left(\alpha-1\right)} \sum_{j=1}^{n} \left(\alpha\, p_j^{\alpha-1} \left|\hat{p}_j - p_j\right| + \frac{\alpha(\alpha-1)b^{\alpha-2}}{2} \left(p_j^{\alpha-2} + \frac{1}{k^{\alpha-2}}\right) (\hat{p}_j - p_j)^2\right). \tag{103}$$

Using Jensen's inequality and the expectation bound (82), the mean value of (103) can be upperbounded by

$$\frac{\left[\alpha\left(\alpha-1\right)\Delta+1\right]^{1/2}}{\alpha\left(\alpha-1\right)}\left(\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha}n}\right)^{1/2}+\frac{b^{\alpha-2}k^{\alpha-1}}{2}\sum_{j=1}^{k}\left(p_j^{\alpha-2}+\frac{1}{k^{\alpha-2}}\right)\mathsf{E}\left[(\hat{p}_j-p_j)^2\right]$$

$$\leq\frac{\left[\alpha\left(\alpha-1\right)\Delta+1\right]^{1/2}}{\alpha\left(\alpha-1\right)}\left(\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha}n}\right)^{1/2}+\frac{b^{\alpha-2}k^{\alpha-1}}{2}\sum_{j=1}^{k}\left(p_j^{\alpha-2}+\frac{1}{k^{\alpha-2}}\right)\frac{p_j}{n}$$

$$=\frac{\left[\alpha\left(\alpha-1\right)\Delta+1\right]^{1/2}}{\alpha\left(\alpha-1\right)}\left(\frac{\sum p_j^{\alpha-1}}{k^{1-\alpha}n}\right)^{1/2}+\frac{b^{\alpha-2}}{2}\frac{k^{\alpha-1}\sum_{j=1}^{k}p_j^{\alpha-1}}{n}+\frac{b^{\alpha-2}k}{2n}.$$

We see that under (76) the desired relation (69) holds if (83) holds under assumption (68). By Schwarz inequality and (75),

$$\sum_{j=1}^{k}p_j^{\alpha-1}=\sum_{j=1}^{k}p_j\ \left(p_j^{\alpha-1}\right)^{\frac{\alpha-2}{\alpha-1}}\leq\left(\sum_{j=1}^{k}p_jp_j^{\alpha-1}\right)^{\frac{\alpha-2}{\alpha-1}}\tag{104}$$

$$=\left(\sum_{j=1}^{k}p_j^{\alpha}\right)^{\frac{\alpha-2}{\alpha-1}}\leq\left(\frac{\alpha\left(\alpha-1\right)\Delta+1}{k^{\alpha-1}}\right)^{\frac{\alpha-2}{\alpha-1}}=\frac{\left(\alpha\left(\alpha-1\right)\Delta+1\right)^{\frac{\alpha-2}{\alpha-1}}}{k^{\alpha-2}}$$

so that the validity of (83) under (68) is obvious and the proof is complete. ■

References

[1] A. R. Barron, L. Györfi, and E. C. van der Meulen. Distribution estimates consistent in total variation and in two types of information divergence. *IEEE Trans. Inform. Theory*, 38(9):1437–1454, Sept. 1992.

[2] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49:1858–60, 2003.

[3] Alexei Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory*, 49(6):1491–1498, June 2003.

[4] B. Fuglede and F. Topsøe. Jensen-Shannon Divergence and Hilbert space embedding. In *Proceedings 2004 International Symposium on Information Theory*, page 31, 2004.

[5] P. Harremoës and F. Topsøe. Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Trans. Inform. Theory*, 47(7):2944–2960, Nov. 2001.

[6] P. Harremoës and I. Vajda. On the Bahadur-efficient testing of uniformity by means of the entropy. *IEEE Trans. Inform Theory*, 54(1):321–331, Jan. 2008.

[7] P. Kafka, F. Österreicher, and I. Vincze. On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, 26:415–422, 1991.

[8] L. Le Cam. *Asymptotic Methods in Statistical Theory.* Springer-Verlag, New York, 1986.

[9] F. Liese and I. Vajda. On divergence and informations in statistics and information theory. *IEEE Tranns. Inform. Theory*, 52(10):4394 – 4412, Oct. 2006.

[10] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55:639–653, 2003.

[11] M. P. Quine and J. Robinson. Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *Ann. Statist.*, 13:727–742, 1985.

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

[13] I. Vajda. On the f-divergence and singularity of probabiltiy measures. *Periodica Math. Hungar.*, 2:223–234, 1972.

[14] I. Vajda. *Theory of Statistical Inference and Information.* Kluwer, Dordrecht, 1989.