ON THE CONSISTENCY IN DIVERGENCE FOR A CLASS OF NONPARAMETRIC DISTRIBUTION ESTIMATES*

Tomáš Hobza

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic, hobza@km1.fjfi.cvut.cz

In this paper we consider the Barron estimates $\widetilde{P}_n$ of probability distributions P which are dominated by a given probability measure Q on a measurable space $(\mathcal{X}, \mathcal{A})$. These estimates are defined by the means of the histograms corresponding to n independent P-distributed observations from $\mathcal{X}$ and partitions of $\mathcal{X}$ into m_n measurable and Q-equiprobable sets. The consistency of $\widetilde{P}_n$ is studied for $m_n \to \infty$, $m_n/n \to 0$ when $n \to \infty$ and for the errors evaluated by means of the Csiszár ϕ-divergences of $\widetilde{P}_n$ and P.

Keywords: Histogram estimates, Barron estimates, Divergences of Csiszár, Divergence errors, Expected divergence errors, Consistency in divergence, Consistency in expected divergence

AMS Classification: 62B10, 62G07, 62G20

1 Introduction, basic concepts and auxiliary results

This paper deals with estimates $\widehat{P}_n$ of probability measures P on abstract measurable spaces $(\mathcal{X}, \mathcal{A})$ based on independent observations

$$X_i \sim (\mathcal{X}, \mathcal{A}, P), \quad 1 \leq i \leq n \,. \tag{1}$$

The error criteria are the ϕ-divergences $D_\phi(\widehat{P}_n, P)$ for convex functions $\phi : (0, \infty) \mapsto I\!R$ strictly convex at 1 with $\phi(1) = 0$ where $D_\phi(\widetilde{P}, P)$ denotes the divergence of Csiszár (1963) defined by the formula

$$D_\phi(\widetilde{P}, P) = \int p\, \phi\left(\frac{\widetilde{p}}{p}\right) dQ \,. \tag{2}$$

In this formula $\widetilde{P}, P, Q$ are arbitrary probability measures on $(\mathcal{X}, \mathcal{A})$ such that Q dominates $\{\widetilde{P}, P\}$, in symbols $Q \gg \{\widetilde{P}, P\}$, $\widetilde{p}, p$ are the Radon-Nikodym derivatives $d\widetilde{P}/dQ$, dP/dQ and

$$\phi\left(\frac{\widetilde{p}}{p}\right) = \phi(0) \triangleq \lim_{t \downarrow 0} \phi(t) \quad \text{when} \quad \widetilde{p} = 0,\ p > 0 \tag{3}$$

while

$$p\, \phi\left(\frac{\widetilde{p}}{p}\right) = \widetilde{p} \cdot \frac{\phi(\infty)}{\infty} \triangleq \widetilde{p} \cdot \lim_{t \to \infty} \frac{\phi(t)}{t} \quad \text{when} \quad \widetilde{p} \geq 0,\ p = 0 \tag{4}$$

with the convention $0 \cdot \infty = 0$ in (4). Table 1 presents the best known ϕ-divergences.

*Supported by the grant 1M0572.

$\boldsymbol{\phi(t),\ t>0}$	$\boldsymbol{D_\phi(\widetilde{P},P)}$	**Name**
$\lvert t-1\rvert$	$V(\widetilde{P},P)=\int\lvert\widetilde{p}-p\rvert\,d\mu$	Total variation
$t\ln t$	$I(\widetilde{P},P)=\int\widetilde{p}\ln\dfrac{\widetilde{p}}{p}\,d\mu$	Information divergence
$-\ln t$	$I(P,\widetilde{P})=\int p\ln\dfrac{p}{\widetilde{p}}\,d\mu$	Reversed information div.
$(t-1)\ln t$	$J(\widetilde{P},P)=\int(\widetilde{p}-p)\ln\dfrac{\widetilde{p}}{p}\,d\mu$	J-divergence
$(t-1)^2$	$\chi^2(\widetilde{P},P)=\int\dfrac{(\widetilde{p}-p)^2}{p}\,d\mu$	χ^2-divergence
$\dfrac{(t-1)^2}{t}$	$\chi^2(P,\widetilde{P})=\int\dfrac{(\widetilde{p}-p)^2}{\widetilde{p}}\,d\mu$	Reversed χ^2-divergence
$\lvert t-1\rvert^a,\ a\geq 1$	$\chi^a(\widetilde{P},P)=\int p^{1-a}\lvert\widetilde{p}-p\rvert^a\,d\mu$	χ^a-divergence
$\lvert t^a-1\rvert^{\frac{1}{a}},\ 0<a<1$	$M_a(\widetilde{P},P)=\int\lvert\widetilde{p}^a-p^a\rvert^{\frac{1}{a}}\,d\mu$	Matusita distance of order a
$\dfrac{t^a-1}{a(a-1)},\ a\neq 0,1$	$I_a(\widetilde{P},P)=\dfrac{1}{a(a-1)}\left(\int\widetilde{p}^a g^{1-a}\,d\mu-1\right)$	Power divergence of order a
$\dfrac{(t-1)^2}{t+1}$	$LC^2(\widetilde{P},P)=\int\dfrac{(\widetilde{p}-p)^2}{\widetilde{p}+p}\,d\mu$	Squared Le Cam distance
$(1-\sqrt{t})^2$	$H^2(\widetilde{P},P)=\int(\sqrt{\widetilde{p}}-\sqrt{p})^2\,d\mu$	Squared Hellinger distance

Table 1: Examples of classical ϕ-divergences $D_\phi(P,Q)$: Generating functions ϕ, symbols and formulas for $D_\phi(P,Q)$ and names of the divergences. Notice that $\chi^1(\widetilde{P},P)=V(\widetilde{P},P)$, $J(\widetilde{P},P)=I(\widetilde{P},P)+I(P,\widetilde{P})$, $\chi^2(\widetilde{P},P)=2I_2(\widetilde{P},P)$ and $I_{1/2}(\widetilde{P},P)=2H^2(\widetilde{P},P)=2M_{1/2}(\widetilde{P},P)$.

The divergence error criteria $D_\phi(\widehat{P}_n,P)$ are justified by the basic properties of ϕ-divergences. Namely, according to Csiszár (1963, 1967) and Liese, Vajda (1987),

$$0\leq D_\phi(\widetilde{P},P)\leq\phi(0)+\phi(\infty)/\infty\qquad\text{(cf. (3), (4))}$$

where $D_\phi(\widetilde{P},P)=0$ iff $\widetilde{P}=P$ and $D_\phi(\widetilde{P},P)=\phi(0)+\phi(\infty)/\infty$ if $\widetilde{P}\perp P$ (singularity, i.e. the supports of $\widetilde{P},P$ are disjoint). Moreover, the ϕ-divergences with the upper bound $\phi(0)+\phi(\infty)/\infty$ finite attain this bound only if $\widetilde{P}\perp P$. Thus $D_\phi(\widetilde{P},P)$ detects the position of the pair $\{\widetilde{P},P\}$ on the scale between the full similarity $\widetilde{P}=P$ and the full dissimilarity $\widetilde{P}\perp P$.

By the *consistency of $\widehat{P}_n$ in the ϕ-divergence* or *in the expected ϕ-divergence* we mean the convergence

$$D_\phi(\widehat{P}_n,P)\xrightarrow{P}0\quad\text{as}\quad n\to\infty\tag{5}$$

or

$$\mathsf{E}\,D_\phi(\widehat{P}_n,P)\longrightarrow 0\quad\text{as}\quad n\to\infty\tag{6}$$

respectively.

This definition is meaningless if $D_\phi(\widehat{P}_n,P)$ is not a random variable, i.e. a measurable function of the observations $X_1,\ldots,X_n$. Here a helpful tool is Theorem 6 in Vajda (1972) by which there exists a sequence of partitions

$$\mathcal{P}_n = \{A_{n1}, \ldots, A_{nm_n}\} \subset \mathcal{A}, \quad m_1 \le m_2 \ldots, \tag{7}$$

such that the restrictions $\widetilde{P}^{(n)}, P^{(n)}$ of $\widetilde{P}, P$ on the subalgebra $\mathcal{A}_n \subset \mathcal{A}$ generated by $\mathcal{P}_n$ satisfy the relation

$$D_\phi(\widetilde{P}, P) = \sup_n D_\phi(\widetilde{P}^{(n)}, P^{(n)})\,. \tag{8}$$

By (2),

$$D_\phi(\widetilde{P}^{(n)}, P^{(n)}) = \sum_{A \in \mathcal{P}_n} P(A)\, \phi\left(\frac{\widetilde{P}(A)}{P(A)}\right) \tag{9}$$

with the conventions (3), (4) behind the sum. Thus if all $\widehat{P}(A)$, $A \in \mathcal{A}$ are random variables then $D_\phi(\widehat{P}_n, P)$ in (5), (6) is a random variable too. Moreover, since the latter random variable is nonnegative, the expectation considered in (6) exists and takes on values in the closed interval $[0, \phi(0) + \phi(\infty)/\infty]$.

The consistency in the divergence sense (5) or (6) is usually stronger than the classical statistical consistency of point estimates. For example, let P be exponential on $I\!\!R$ with the Lebesgue density $f(x) = \boldsymbol{I}(x > 0)\,\theta \exp\{-\theta x\}$ and $\widehat{P}_n$ an estimate of P with the Lebesgue density $\widehat{f}_n(x) = \boldsymbol{I}(x > 0)\,\widehat{\theta}_n \exp\{-\widehat{\theta}_n x\}$ where

$$\widehat{\theta}_n = \frac{n}{X_1 + \ldots + X_n}$$

is the maximum likelihood point estimate of θ. Then $\widehat{\theta}_n$ is consistent in the classical statistical sense while $\mathsf{E}\, I(P, \widehat{P}_n) = \infty$ for all $n \ge 1$, i.e. $\widehat{P}_n$ is not consistent in the expected reversed information divergence (cf. Example 1 in Vajda and van der Meulen (2001)).

The classical application of the consistency principle (5) or (6) in the nonparametric statistics is the application to the total variation $V(\widehat{P}_n, P) = V(P, \widehat{P}_n)$ which is nothing but the $L_1(Q)$-distance of the densities $\widehat{p}_n = d\widehat{P}_n/dQ$ and $p = dP/dQ$ for $Q \gg \{\widehat{P}_n, P\}$. The density estimates consistent in this distance were systematically studied in the monograph Devroy, Györfi (1985)) and some references cited there.

Barron (1988) and Barron et al. (1992) studied distribution estimates consistent in the sense (5) and (6) for the total variation $V(\widehat{P}_n, P) = V(P, \widehat{P}_n)$ and also for the information divergence $I(\widehat{P}_n, P)$ and the reversed information divergence $I(P, \widehat{P}_n)$. Later Györfi et al. (1998) and Vajda, van der Meulen (1998, 2001) studied the consistency in the sense of (5) and (6) for the reversed χ^2-divergence $\chi^2(P, \widehat{P}_n)$. Berlinet et al. (1998) presented arguments for evaluation of the estimation errors by means of a wider class of ϕ-divergences $D_\phi(\widehat{P}_n, P)$. These authors studied the consistency for such a wider class including the power divergences of orders $-1 \le a \le 2$.

The special role of the consistency in the total variation $V(\widehat{P}_n, P)$ follows from the fact that every ϕ-divergence $D_\phi(\widetilde{P}, P)$ satisfies the relation

$$D_\phi(\widetilde{P}, P) \ge L_\phi\left(V(\widetilde{P}, P)\right) \tag{10}$$

where $L_\phi(t)$ is strictly increasing and convex in the domain $0 \le t \le 2$ with $L_\phi(0) = 0$. Here $L_\phi(t)$ denotes the lower bound of the ϕ-divergences $D_\phi(\widetilde{P}, P)$ over all pairs $\widetilde{P}, P$ with

$V(\widetilde{P}, P) \geq t$. This characterization of the lower bound L_ϕ was proved in Vajda (1995) but for typical functions ϕ which are convex at all $t \in (0, \infty)$ it was proved already in Theorem 3 of Vajda (1972). The inequality (10) implies that the consistency in the total variation $V(\widehat{P}_n, P)$ follows from the consistency in any ϕ-divergence $D_\phi(\widehat{P}_n, P)$. The converse is not true. For example if $\mathcal{X} = \{0, 1\}$ and

$$P = (1-p, p), \qquad \widehat{P}_n = \left(1 - \frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i\right)$$

then

$$\mathsf{E}\, V(P, \widehat{P}_n) = 2\,\mathsf{E}\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| \leq 2\left(\mathsf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i - p\right)^2\right)^{1/2} = 2\sqrt{\frac{p(1-p)}{n}} \to 0$$

as $n \to \infty$ while $\mathsf{E}\, I(P, \widehat{P}_n) = \infty$ for all $n \geq 1$. The consistency in the total variation (expected total variation) is thus the weakest of the consistencies considered in (5) or (6).

In this paper we consider sequences of subalgebras $\mathcal{A}_n \subset \mathcal{A}$ generated by similar partitions $\mathcal{P}_n$ as considered in (7) and the error criteria $D_\phi(\widehat{P}_n^{(n)}, P^{(n)})$ for the restrictions $\widehat{P}_n^{(n)}, P^{(n)}$ of $\widehat{P}_n, P$ on $\mathcal{A}_n$. By the monotonicity of ϕ-divergences (see e.g. Theorem 1.24 in Liese and Vajda (1987)),

$$D_\phi(\widehat{P}_n^{(n)}, P^{(n)}) \leq D_\phi(\widehat{P}_n, P) \tag{11}$$

so that

$$D_\phi(\widehat{P}_n^{(n)}, P^{(n)}) = \sum_{a \in \mathcal{P}_n} P(A)\,\phi\left(\frac{\widehat{P}_n(A)}{P(A)}\right) \qquad \text{(cf. (9))} \tag{12}$$

is a restricted version of the ϕ-divergence $D_\phi(\widehat{P}_n, P)$.

We are interested in the estimates $\widehat{P}_n$ *consistent in the restricted ϕ-divergence*, i.e. satisfying the condition

$$D_\phi(\widehat{P}_n^{(n)}, P^{(n)}) \xrightarrow{P} 0 \qquad \text{as } n \to \infty\,, \tag{13}$$

and those *consistent in the expected restricted ϕ-divergence* , i.e. satisfying the condition

$$\mathsf{E}\, D_\phi(\widehat{P}_n^{(n)}, P^{(n)}) \longrightarrow 0 \qquad \text{as } n \to \infty\,. \tag{14}$$

It is clear from (12) that if the probabilities $\widehat{P}_n(A)$, $A \in \mathcal{A}_n$, are random variables then $D_\phi(\widehat{P}_n^{(n)}, P^{(n)})$ in (13), (14) is a random variable and, similarly as $\mathsf{E}\, D_\phi(\widehat{P}_n, P)$ in (6), the expectations figuring in (14) exist and take on values in the interval $[0, \phi(0) + \phi(\infty)/\infty]$.

We see from (11) that the convergences (13), (14) follow from those in (5), (6) so that the concepts of consistency considered in the present paper are not stronger than those considered in the above cited papers. In some situations these concepts are equally strong (they coincide). Next we list two such situations.

Situation 1. The ϕ-divergence is total variation (i.e. $\phi(t) = |t-1|$), the partitions $\mathcal{P}_n$ are nested in the sense that $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \ldots$, the union $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots$ generates the σ-algebra $\mathcal{A}$ and the densities $\widehat{p}_n = d\widehat{P}_n/dQ$ are for each $A \in \mathcal{P}_n$ constant on A. Denote by P_n the probability measure with a density $dP_n/dQ = p_n$ where for each $A \in \mathcal{P}_n$

$$p_n(x) = \frac{P(A)}{Q(A)} \qquad \text{for all} \quad x \in A\,. \tag{15}$$

Then by the martingale convergence theorem (see e.g. section VII.4 in Doob (1990))

$$p_n \longrightarrow p \qquad Q-\text{a.s.} \quad \text{for} \quad n \to \infty \tag{16}$$

and, by the Lebesgue dominated convergence theorem,

$$V(P_n, P) = \int |p_n - p|\, dQ \longrightarrow 0 \quad \text{for} \quad n \to \infty\,.$$

On the other hand, by the triangle inequality,

$$V(\widehat{P}_n, P) \le V(\widehat{P}_n, P_n) + V(P_n, P)$$

where $V(\widehat{P}_n, P_n) = V(\widehat{P}_n^{(n)}, P^{(n)})$ because both densities $\widehat{p}_n$ and p_n are constant on the sets $A \in \mathcal{P}_n$. Thus (13) or (14) for $V(\widehat{P}_n^{(n)}, P^{(n)})$ implies (5) or (6) for $V(\widehat{P}_n, P)$, respectively.

The following situation is more general. It uses the fact that if $\widehat{p}_n = d\widehat{P}_n/dQ$ is constant on the sets $A \in \mathcal{P}_n$ then

$$\widehat{p}_n(x) = \frac{\widehat{P}_n(A)}{Q(A)} \qquad \text{for all} \quad x \in A\,.$$

Hence the density p_n of Situation 1 satisfies the relation

$$\int_A p_n\, \phi\left(\frac{\widehat{p}_n}{p_n}\right) dQ = P(A)\, \phi\left(\frac{\widehat{P}_n(A)}{P(A)}\right)\,.$$

By (2) and (12), this implies for every ϕ-divergence

$$D_\phi(\widehat{P}_n, P_n) = D_\phi(\widehat{P}_n^{(n)}, P^{(n)})\,. \tag{17}$$

The special case for $\phi(t) = |t-1|$ was used in Situation 1.

Situation 2. For some $\alpha > 0$, the α-th power $D_\phi^\alpha(\widetilde{P}, P)$ of a ϕ-divergence $D_\phi(\widetilde{P}, P)$ is metric, the densities $\widehat{p}_n = d\widehat{P}_n/dQ$ are for each $A \in \mathcal{P}_n$ constant on A, and the distributions P_n introduced in Situation 1 satisfy the condition

$$D_\phi(P, P_n) \longrightarrow 0 \qquad \text{for} \quad n \to \infty\,. \tag{18}$$

Then

$$D_\phi(\widehat{P}_n, P) \le \left(D_\phi^\alpha(\widehat{P}_n, P_n) + D_\phi^\alpha(P, P_n)\right)^{1/\alpha}$$

where $D_\phi^\alpha(\widehat{P}_n, P_n) = (D_\phi(\widehat{P}_n^{(n)}, P^{(n)}))^\alpha$ by (17). Hence under (18) the conditions (13) or (14) imply (5) or (6) respectively. This means that in Situation 2 the consistencies studied in the present paper are equally strong as those defined by the conditions (5) or (6) respectively.

Some conditions for the convergence (18) can be found in the previous literature. For example, $\phi(t) = t \ln t$ generates the information divergence for which

$$\begin{aligned} I(P, P_n) &= \int p \ln \frac{p}{p_n}\, dQ = \int p \ln p \, dQ - \int p_n \ln p_n \, dQ \\ &= I(P, Q) - I(P^{(n)}, Q^{(n)}) \end{aligned}$$

provided $I(P, Q) < \infty$ where $P^{(n)}, Q^{(n)}$ are restrictions of the distributions P, Q on the algebra $\mathcal{A}_n$ generated by $\mathcal{P}_n$. If $(\mathcal{X}, \mathcal{A})$ is an Euclidean space $\mathbb{R}^d$ with the Borel σ-field and $\mathcal{P}_n$ is a rectangular partition of $\mathbb{R}^d$ then, by Vajda (2002),

$$I(P^{(n)}, Q^{(n)}) \longrightarrow I(P, Q) \qquad \text{for} \quad n \to \infty$$

provided

$$\max_{A \in \mathcal{P}_n} Q(A) \longrightarrow 0 \qquad \text{for} \quad n \to \infty\,.$$

Österreicher and Vajda (2003) proved that

$$\phi_\beta(t) = \frac{1}{1-\beta}\, \left[(1 + t^{1/\beta})^\beta - 2^{\beta-1}(1+t)\right] \quad \beta > 0, \beta \neq 1$$

with the corresponding limit

$$\phi_1(t) = \ln \frac{2}{1+t} + t \, \ln \frac{2t}{1+t}$$

are functions strictly convex on $(0, \infty)$ and that for $\alpha = \min\{1/2, 1/\beta\}$ the α-th powers of the ϕ_β-divergences are metrics on the space of probability measures. In particular,

$$D_{\phi_1}(\widetilde{P}, P) = I(\widetilde{P}, (P + \widetilde{P})/2) + I(P, (P + \widetilde{P})/2)$$

is the ϕ_1-divergence and the square root of this divergence is metric.

Notice that in the class of ϕ_β-divergences all upper bounds

$$\phi_\beta(0) + \phi_\beta(\infty)/\infty = \begin{cases} 2(2^{\beta-1} - 1)/(\beta - 1) & \text{if} \quad \beta \neq 1 \\ 2 \ln 2 & \text{if} \quad \beta = 1 \end{cases} \tag{19}$$

are finite. For all bounded divergences we can give the following simple condition for the convergence (18).

Proposition 1 *All bounded ϕ-divergences satisfy (18) if the partitions $\mathcal{P}_n$ are nested in the sense that the corresponding algebras $\mathcal{A}_n$ monotonically increase and their union generates $\mathcal{A}$.*

Proof. Each ϕ under consideration defines the same divergence as the nonnegative convex function

$$\widetilde{\phi}(t) = \phi(t) - \phi'_+(1)(t-1), \quad 0 < t < \infty \tag{20}$$

where $\phi'_+(1)$ denotes the right hand derivative at 1. Obviously, $\phi(0) + \phi(\infty)/\infty$ is finite iff $\widetilde{\phi}(0) + \widetilde{\phi}(\infty)/\infty$ is finite. Thus we can assume that ϕ is nonnegative with $\phi(0) + \phi(\infty)/\infty$ finite. By Jensen's inequality,

$$\phi(t) \leq \left(1 - \frac{t}{s}\right) \phi(0) + \frac{t}{s}\, \phi(s) = \phi(0) + t \cdot \frac{\phi(s) - \phi(0)}{s}$$

so that

$$0 \le \phi(t) \le \phi(0) + t\,\phi(\infty)/\infty\,. \tag{21}$$

By (2),

$$D_\phi(P, P_n) = \int p_n\,\phi\left(\frac{p}{p_n}\right) dQ\,.$$

By (16), the integrand tends Q-a.s. to $p\,\phi(1) = 0$. By (21), it is bounded above by the Q-integrable function $p_n\phi(0) + p\,\phi(\infty)/\infty$. Thus the desired convergence (18) follows from the Lebesgue bounded convergence theorem for integrals. □

In the important case where $(\mathcal{X}, \mathcal{A})$ is the Borel line, Györfi et al. (1998) found a relatively simple condition guaranteeing (18) for the χ^2-divergence $D_\phi(P, P_n) = \chi^2(P, P_n)$ with infinite $\phi(0) + \phi(\infty)/\infty$. The problem of convergence (18) is of its own interest with applications beyond the scope of the above considered Situation 2. In the Appendix we show that a certain modification of the condition of Györfi et al. guarantees (18) for a wide class of ϕ-divergences, e.g. for all those considered in Table 1 except the total variation and some of the χ^a-divergences.

Let us finish this introductory section by specifying the class of estimates $\widehat{P}_n$ studied in this paper. As shown by Devroy and Györfi (1990), no estimate $\widehat{P}_n$ is consistent in total variation for all distributions P on nontrivial observation spaces $(\mathcal{X}, \mathcal{A})$. Using the inequality (10) we can extend this negative conclusion to the consistency in an arbitrary ϕ-divergence. However, as shown in the papers cited above, consistent estimates $\widehat{P}_n$ may exist for all distributions P dominated by a σ-finite measure (which is in fact equivalent to the domination by a probability measure Q).

For $P \ll Q$ one can consider the *Q-shaped histogram estimates* $\widehat{P}_n$ dominated by Q with the densities

$$\widehat{p}_n = \frac{d\widehat{P}_n}{dQ} \triangleq \sum_{j=1}^{m_n} \boldsymbol{I}_{A_{nj}} \frac{Y_{nj}}{n\,Q(A_{nj})} \tag{22}$$

where $\boldsymbol{I}$ stands for the indicator function and

$$Y_{nj} = \sum_{i=1}^{n} \boldsymbol{I}_{A_{nj}}(X_i)$$

are the random counts of observations in the bins A_{nj} of the finite partitions $\mathcal{P}_n$ (see (7)). If Q has a density q with respect to a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$ then

$$\frac{d\widehat{P}_n}{d\lambda} = \sum_{j=1}^{m_n} \boldsymbol{I}_{A_{nj}}\, q\, \frac{Y_{nj}}{n\,Q(A_{nj})}\,, \tag{23}$$

i.e. the density of $\widehat{P}_n$ is shaped by q inside the partition sets A_{nj}.

The histogram estimates (22), (23) are computationally simpler than the kernel estimates, and in some sense they are also simpler than the histogram estimates based on infinite partitions. However, similarly as the histograms based on the infinite partitions, they encounter problems arising from empty bins which may appear with nonzero probabilities.

To avoid such problems, Barron (1988) proposed the modified Q-shaped histogram $\widetilde{P}_n$ where (23) is replaced by the density

$$\frac{d\widetilde{P}_n}{d\lambda} = (1 - \varepsilon_n)\,\frac{d\widehat{P}_n}{d\lambda} + \varepsilon_n\, q$$

for $d\widehat{P}_n/d\lambda$ given by (23) and $0 < \varepsilon_n < 1$ decreasing to 0 for $n \to \infty$. Formally simpler is the density

$$\widetilde{p}_n = \frac{d\widetilde{P}_n}{dQ} = (1 - \varepsilon_n)\widehat{p}_n + \varepsilon_n$$

where $\widehat{p}_n$ is given by (22). We shall consider the version with $\varepsilon = m_n/(n + m_n)$, i.e.

$$\widetilde{p}_n = \frac{d\widetilde{P}_n}{dQ} = \frac{n}{n + m_n}\,\widehat{p}_n + \frac{m_n}{n + m_n} \tag{24}$$

for $m_n \to \infty$ and $m_n/n \to 0$ as $n \to \infty$. This estimate is called *Barron estimate* in the sequel.

Barron et al. (1992) proved the consistency of the Barron estimator in the L_1-error and the expected L_1-error for all densities $f \in \mathbb{F}_Q$, where $\mathbb{F}_Q$ is the class of all probability densities with respect to Q, and for partitions $\mathcal{P}_n$ defined by Q and m_n satisfying $m_n \to \infty$ and $m_n/n \to 0$ for n increasing to infinity. Under $I(P, Q) < \infty$ and some additional conditions they proved also the consistency in the information divergence and reversed information divergence and in the expected versions of these divergences. Györfi et al. (1998) presented arguments in favor of the reversed χ^2-divergence error $\chi^2(P, \widetilde{P}_n)$ and proved the consistency in the expected χ^2-divergence under some additional regularity assumptions on the density f. They obtained also the optimal rate of convergence $\mathsf{E}\,\chi^2(P, \widetilde{P}_n) = O(n^{-2/3})$.

Berlinet et al. (1998) was the first paper dealing with the asymptotic properties of the Barron estimator when the errors are expressed by more general ϕ-divergences. Under some regularity assumptions about the distribution P, the interval partitions $\mathcal{P}_n$ and the divergence generating function ϕ, they proved that the Barron estimator is consistent in the ϕ-divergence and the expected ϕ-divergence, i.e. that $\lim_{n\to\infty} D_\phi(P, \widetilde{P}_n) = 0$ a.s. and $\lim_{n\to\infty} \mathsf{E}\, D_\phi(P, \widetilde{P}_n) = 0$.

Our paper goes in some respect deeper than Berlinet et al. (1998). For the Barron estimate we prove results about the consistency in the reduced ϕ-divergences and expected reduced ϕ-divergences which are considerably stronger than similar results of Berlinet et al. These results are summarized in Theorems 1 and 2 and Proposition 2 in the next section. Note that the asymptotic normality of the error $I(P, \widetilde{P}_n)$ was proved in Berlinet et al. (1997) and the asymptotic normality of the error $\chi^2(P, \widetilde{P}_n)$ was proved in Vajda and van der Meulen (1998). Results about the asymptotic normality of general ϕ-divergence errors $D_\phi(\widetilde{P}_n, P)$ established in Hobza (2003) will be presented in a forthcoming paper.

2 Main results

Restrict ourselves for simplicity to the partitions $\mathcal{P}_n = \{A_{n1}, \dots, A_{nm_n}\}$ introduced in (7) which are uniform in the sense that $Q(A_{nj}) = 1/m_n$ and put for $n = 1, 2, \dots$

$$\boldsymbol{q}_n = \left(q_{nj} \triangleq Q(A_{nj}) = \frac{1}{m_n} : \ 1 \le j \le m_n \right) . \tag{25}$$

Further, put

$$\boldsymbol{p}_n = \left(p_{nj} \triangleq P(A_{nj}) : \ 1 \le j \le m_n \right) \tag{26}$$

and for

$$\boldsymbol{Y}_n = (Y_{nj}:\ 1 \le j \le m_n) \sim \text{Multinomial}(n, \boldsymbol{p}_n) \tag{27}$$

define the empirical distributions

$$\widehat{\boldsymbol{p}}_n = \left(\widehat{p}_{nj} = \frac{Y_{nj}}{n}:\ 1 \le j \le m_n\right) \tag{28}$$

which estimate the true $\boldsymbol{p}_n$. Finally, put

$$\widetilde{\boldsymbol{p}}_n = (1-\varepsilon_n)\,\widehat{\boldsymbol{p}}_n + \varepsilon_n \boldsymbol{q}_n, \qquad \varepsilon_n = \frac{m_n}{n+m_n}\,. \tag{29}$$

By (12),

$$D_\phi\left(\widetilde{P}_n^{(n)}, P^{(n)}\right) = D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) = \sum_{j=1}^{m_n} p_{nj}\,\phi\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right) \tag{30}$$

where

$$D_\phi(\widetilde{\boldsymbol{p}}, \boldsymbol{p}) = \sum_{j=1}^{m} p_j\,\phi\left(\frac{\widetilde{p}_j}{p_j}\right) \tag{31}$$

stands in accordance with (2) for the ϕ-divergence of distributions $\widetilde{\boldsymbol{p}} = (\widetilde{p}_1, \ldots, \widetilde{p}_m)$ and $\boldsymbol{p} = (p_1, \ldots, p_m)$ on the finite observation space $\mathcal{X} = \{1, \ldots, m\}$.

We restrict ourselves to the ϕ-divergences for the restricted class of the above considered functions ϕ satisfying the following assumptions.

ϕ - assumptions

ϕ is finite and convex on $(0,\infty)$ extended on $[0,\infty)$ by the rule

$$\phi(0) = \lim_{t\to 0_+} \phi(t)\,,$$

twice continuously differentiable in a neighborhood of 1, satisfying the condition

$$\phi(1) = \phi'(1) = 0, \quad \phi''(1) > 0 \tag{32}$$

with the second derivative $\phi''(t)$ Lipschitz in this neighborhood.

Here $\phi''(1) > 0$ and the continuity of ϕ'' implies the strict convexity in the neighborhood of 1 and $\phi'(1) = 0$ implies that ϕ is positive on $[0,\infty) - 1$ (cf. (20)).

We suppose that the partitions $\mathcal{P}_n$ satisfy the following assumptions.

$\boldsymbol{\mathcal{P}_n}$-assumptions

It holds

$$\lim_{n\to\infty} m_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{m_n}{n} = 0 \tag{33}$$

and there exists $\beta \geq 1$ such that

$$\liminf_{n\to\infty} m_n^{\beta} \min_{1\leq j\leq m_n} p_{nj} > 0 \tag{34}$$

and

$$\frac{m_n^{1+\beta}}{n} = o(1)\,. \tag{35}$$

Remark 1 Since for every $\boldsymbol{p}_n$ it holds $\min_{1\leq j\leq m_n} p_{nj} \leq 1/m_n$, (34) cannot hold for $\beta < 1$.

Our main results are the following two theorems and the following proposition.

Theorem 1 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions, then for all ϕ satisfying the ϕ-assumptions the Barron estimator $\widetilde{P}_n$ is consistent in the reduced ϕ-divergence, i.e.*

$$D_\phi\left(\widetilde{P}_n^{(n)}, P^{(n)}\right) = o_P(1) \qquad \textit{as} \quad n \to \infty\,. \tag{36}$$

Theorem 2 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions, and if there exists $n_1 \in \mathbb{N}$ such that*

$$\sup_{n>n_1} \mathsf{E}\left(D_\phi\left(\widetilde{P}_n^{(n)}, P^{(n)}\right)\right)^2 \;<\; +\infty\,, \tag{37}$$

then for all ϕ satisfying the ϕ-assumptions the Barron estimator $\widetilde{P}_n$ is consistent in the expected reduced ϕ-divergence, i.e.

$$\mathsf{E}\, D_\phi\left(\widetilde{P}_n^{(n)}, P^{(n)}\right) = o(1) \qquad \textit{as} \quad n \to \infty\,. \tag{38}$$

Since it is difficult to check the condition (37) directly we have found a simple condition on the divergence function ϕ sufficient for (37).

Proposition 2 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions then the condition*

$$\phi\left(\frac{1}{t}\right) + \phi(t) = O(t^k) \quad \textit{for} \quad t \to \infty \quad \textit{and some} \quad k \in \mathbb{N} \tag{39}$$

is sufficient for (37).

Remark 2 The condition (39) on the convex function ϕ is weaker than the condition $t\phi(1/t) + \phi(t) = O(t^2)$ for $t \to \infty$ assumed in Berlinet et al. (1998). We allow $\phi(t)$ to increase polynomially and not only quadratically in the neighborhood of $t = 1$ and $t = \infty$. Here it is to be mentioned that we treat consistency in the reduced ϕ-divergences only. However, as argued in the previous section, for some ϕ-divergences this is equivalent to the consistency in the same non-reduced sense as in Berlinet et al. (1998).

Remark 3 Let us note that the same results as Theorems 1, 2 and Proposition 2 can be stated also in the case of Q-shaped histogram $\widehat{P}_n$ (see Hobza (2003)). The only difference is that instead of the condition (39) we need to suppose $\phi(0) < \infty$ and $\phi(t) = O(t^k)$ for $t \to \infty$. The reason is that the Q-shaped histogram can attain zero value with positive probability (due to empty cells) and thus $\phi(0) = \infty$ would imply infinite value of $\mathsf{E}\, D_\phi(\widehat{P}_n^{(n)}, P^{(n)})$. This does not affect the Barron estimator which is positive with Q-probability 1 and thus $\phi(0) = \infty$ is admitted.

Remark 4 Define ϕ^* conjugated to ϕ in the sense that

$$\phi^*(t) = t\,\phi\left(\frac{1}{t}\right) \quad \text{for all } t > 0\,. \tag{40}$$

It can be proved (cf. Lemma 1 on page 41 in Hobza (2003)) that if the function ϕ satisfy the ϕ-assumptions then also the conjugated function ϕ^* satisfies the ϕ-assumptions and

$$D_{\phi^*}(P, Q) = D_\phi(Q, P)$$

for all distributions P, Q. If, moreover, ϕ satisfies (39) then it holds

$$\phi^*\left(\frac{1}{t}\right) + \phi^*(t) = \frac{1}{t}\,\phi(t) + t\,\phi\left(\frac{1}{t}\right) = O(t^{k+1})\,,$$

i.e. ϕ^* satisfies the condition (39) too. Hence if ϕ satisfies the ϕ-assumptions and (39) then the conditions of Theorem 2 are satisfied also for ϕ^* and thus this theorem implies also

$$\mathsf{E}\, D_\phi\left(P^{(n)}, \widetilde{P}_n^{(n)}\right) = o(1)\,.$$

However, similar conclusion does not apply to the Q-shaped histogram $\widehat{P}_n$, since the conditions $\phi(0) < \infty$ and $\phi(t) = O(t^k)$ do not imply $\phi^*(0) < \infty$.

The proofs of the above stated theorems will be divided into several steps. The basic idea is to show the desired asymptotic relations for the particular χ^2-divergence and then, with the help of an inequality from Lemma 9 below, to extend these relations to the remaining ϕ-divergences. Therefore the first two lemmas that follow state the consistency of the Barron estimate in the reduced and expected reduced χ^2-divergence.

Lemma 1 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions then the Barron estimator $\widetilde{P}_n$ is consistent in the reduced chi-square divergence, i.e.*

$$\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \xrightarrow{P} 0, \quad n \to \infty.$$

Proof. Since the distribution of Y_{nj} is $Bi(n, p_{nj})$, for all $j = 1, \ldots, m_n$, it holds

$$\mathsf{E}\,(\widehat{p}_{nj} - p_{nj})^2 = \frac{1}{n^2}\,\mathsf{E}\,(Y_{nj} - np_{nj})^2 = \frac{p_{nj}(1 - p_{nj})}{n}\,. \tag{41}$$

Using the relations $\mathsf{E}\,(\widehat{p}_{nj} - p_{nj}) = 0$, $|p_{nj} - q_{nj}| \leq 1$ and $0 < \varepsilon_n < 1$ we conclude that for all $j = 1, \ldots, m_n$

$$\mathsf{E}\,(\widetilde{p}_{nj} - p_{nj})^2 \;=\; \mathsf{E}\,((1 - \varepsilon_n)(\widehat{p}_{nj} - p_{nj}) - \varepsilon_n(p_{nj} - q_{nj}))^2 \leq \frac{p_{nj}}{n} + \varepsilon_n^2(p_{nj} - q_{nj})^2\,. \tag{42}$$

Thus, using the Markov inequality and (34), (35), we obtain for all $\varepsilon > 0$

$$P\ \left(\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) > \varepsilon\right) = P\left(\sum_{j=1}^{m_n} \frac{(\widetilde{p}_{nj} - p_{nj})^2}{p_{nj}} > \varepsilon\right) \le \frac{1}{\varepsilon}\sum_{j=1}^{m_n} \frac{\mathsf{E}\,(\widetilde{p}_{nj} - p_{nj})^2}{p_{nj}}$$

$$\le\ \frac{1}{\varepsilon}\left(\frac{m_n}{n} + \varepsilon_n^2\left(1 + \frac{1}{m_n \min\limits_{1\le j\le m_n} p_{nj}}\right)\right) \longrightarrow 0, \quad n \to \infty\,.$$

□

Lemma 2 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions then the Barron estimator $\widetilde{P}_n$ is consistent in the expected reduced chi-square divergence, i.e.*

$$\mathsf{E}\ \chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \longrightarrow 0, \quad n \to \infty.$$

Proof. Since

$$\mathsf{E}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) = \quad \mathsf{E}\sum_{j=1}^{m_n} \frac{(\widetilde{p}_{nj} - p_{nj})^2}{p_{nj}} \quad = \quad \sum_{j=1}^{m_n} \frac{\mathsf{E}\,(\widetilde{p}_{nj} - p_{nj})^2}{p_{nj}}\,,$$

the proof can be finished in the same way as the proof of Lemma 1. □

We will also need the following result.

Lemma 3 *Under the $\mathcal{P}_n$-assumptions, for all but finitely many n,*

$$\mathsf{E}\ |\widetilde{p}_{nj} - p_{nj}|^3\ \le\ 2\sqrt{3}\left(\frac{p_{nj}}{n}\right)^{\frac{3}{2}}, \qquad 1 \le j \le m_n\,.$$

Proof. The proof follows similar steps as the proof of Lemma 2 in Györfi and Vajda (2002). A detailed version can be found in Hobza (2003, Lemma 17 on page 70). □

The next step is to prove (37) for the χ^2-divergence.

Lemma 4 *Under the $\mathcal{P}_n$-assumptions there exists $n_0 \in I\!N$ such that*

$$\sup_{n>n_0} \mathsf{E}\ \left(\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2\ <\ +\infty\,.$$

Proof. The proof can be carried out in a similar straightforward way as the previous proofs but it is technically more complicated and therefore too long. However, since the chi-square divergence $\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)$ is the ϕ-divergence for $\phi(t) = (t-1)^2$ and for this ϕ

$$\phi(0) = 1 \quad \text{and} \quad \phi(t) = O(t^2), \quad t \to \infty\,,$$

the desired result follows from Proposition 2 proved below without reference to the results of this section. □

In the next lemma, as well as in the sequel, we consider for probability distributions $\boldsymbol{p} = (p_1, \ldots, p_m)$, $\boldsymbol{q} = (q_1, \ldots, q_m)$ the relative deviation

$$\Delta(\boldsymbol{p}, \boldsymbol{q}) = \max_{1\le j\le m}\left|\frac{p_j}{q_j} - 1\right|. \tag{43}$$

Remark 5 In the sequel we use the inequality

$$\left| D_\phi(\boldsymbol{p},\boldsymbol{q}) - \frac{\phi''(1)}{2}\,\chi^2(\boldsymbol{p},\boldsymbol{q}) \right| \le \frac{L_\phi}{2} \sum_{j=1}^{m} \frac{|p_j - q_j|^3}{q_j^2}\,, \tag{44}$$

valid for all ϕ satisfying the ϕ-assumptions and for all discrete probability distributions $\boldsymbol{p},\boldsymbol{q}$ with sufficiently small $\Delta(\boldsymbol{p},\boldsymbol{q})$. This is nothing but a discrete version of Lemma 9 in the next section. This version follows by multiplying both sides of (62) by q_j, substituting $t = p_j/q_j$ and summing up both sides over $1 \le j \le m$.

Lemma 5 *Under the $\mathcal{P}_n$-assumptions it holds*

$$\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \xrightarrow{P} 0 \quad as \quad n \to \infty\,,$$

for $\Delta(\boldsymbol{p},\boldsymbol{q})$ defined by (43).

Proof. Applying the Chebyshev inequality, we obtain for arbitrary n and $\varepsilon > 0$

$$\begin{aligned} P\left(\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) > \varepsilon\right) \quad &= \quad P\left(\max_{1\le j\le m_n} \left| \frac{\widetilde{p}_{nj}}{p_{nj}} - 1 \right| > \varepsilon \right) \\ \le \quad P\left(\bigcup_{j=1}^{m_n} \left(\left| \frac{\widetilde{p}_{nj}}{p_{nj}} - 1 \right| > \varepsilon \right) \right) \quad &\le \quad \sum_{j=1}^{m_n} P\left(\left| \frac{\widetilde{p}_{nj}}{p_{nj}} - 1 \right| > \varepsilon \right) \\ \le \quad \sum_{j=1}^{m_n} \frac{1}{\varepsilon^2}\, \mathsf{E}\left(\frac{\widetilde{p}_{nj}}{p_{nj}} - 1 \right)^2 \quad &= \quad \frac{1}{\varepsilon^2} \sum_{j=1}^{m_n} \frac{\mathsf{E}\left(\widetilde{p}_{nj} - p_{nj}\right)^2}{p_{nj}^2}\,. \end{aligned}$$

Thus, using the upper bound (42) and the $\mathcal{P}_n$-assumptions, we find that

$$P\left(\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) > \varepsilon\right) \le \frac{1}{\varepsilon^2} \left(\frac{m_n}{n \min\limits_{1\le j\le m_n} p_{nj}} + \varepsilon_n^2 \left(m_n + \frac{1}{m_n (\min\limits_{1\le j\le m_n} p_{nj})^2} \right) \right) = o(1) \tag{45}$$

as $n \to \infty$. □

Next follows the last lemma needed for proving Theorems 1 and 2.

Lemma 6 *If the partitions $\mathcal{P}_n$ satisfy the $\mathcal{P}_n$-assumptions and (37) is fulfilled then*

$$\mathsf{E}\left| D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \;-\; \frac{\phi''(1)}{2}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right| \longrightarrow 0, \quad n \to \infty$$

for all ϕ satisfying the ϕ-assumptions.

Proof. Let us denote

$$Z_n \triangleq \left| D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \;-\; \frac{\phi''(1)}{2}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right|\,.$$

For all n, $\varepsilon > 0$ and sufficiently small $\delta > 0$

$$P\left(Z_n > \varepsilon\right) \le P\left(Z_n > \varepsilon,\, \Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \le \delta\right) + P\left(\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) > \delta\right)\,.$$

From Lemma 5 it follows $P\left(\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) > \delta\right) = o(1)$ and under the condition $\Delta\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \leq \delta$ we get from (44) that

$$Z_n \leq \frac{L_\phi}{2} \sum_{j=1}^{m_n} \frac{|\widetilde{p}_{nj} - p_{nj}|^3}{p_{nj}^2} \,. \tag{46}$$

Thus, by the Markov inequality,

$$P\left(Z_n > \varepsilon\right) \leq \; P\left(\frac{L_\phi}{2} \sum_{j=1}^{m_n} \frac{|\widetilde{p}_{nj} - p_{nj}|^3}{p_{nj}^2} > \varepsilon \right) + o(1) \;\leq\; \frac{L_\phi}{2\varepsilon} \sum_{j=1}^{m_n} \frac{\mathsf{E}\,|\widetilde{p}_{nj} - p_{nj}|^3}{p_{nj}^2} + o(1)\,.$$

Using Lemma 3 we obtain

$$P\left(Z_n > \varepsilon\right) \leq \frac{\sqrt{3}L_\phi}{\varepsilon} \sum_{j=1}^{m_n} \frac{1}{n^{\frac{3}{2}} p_{nj}^{\frac{1}{2}}} + o(1) \;\leq\; \frac{\sqrt{3}L_\phi}{\varepsilon} \cdot \frac{m_n}{n^{\frac{3}{2}} \big(\min\limits_{1\leq j \leq m_n} p_{nj} \big)^{\frac{1}{2}}} + o(1)\,.$$

The $\mathcal{P}_n$-assumptions (34) and (35) then imply

$$Z_n \xrightarrow{P} 0, \quad n \to \infty\,. \tag{47}$$

Finally, to prove $\mathsf{E}\,Z_n \to 0$ it suffices to show that $\{Z_n\}$ is uniformly integrable (see e.g. Theorem A, p. 14 in Serfling (1980)). We shall check for $\gamma = 1$ the sufficient condition

$$\sup_n \mathsf{E}\,|Z_n|^{1+\gamma} < \infty$$

for uniform integrability of the sequence $\{Z_n\}$. It holds

$$\mathsf{E}\,Z_n^2 \leq \mathsf{E}\,\left(D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2 + \left(\frac{\phi''(1)}{2}\right)^2 \mathsf{E}\,\left(\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2 \,. \tag{48}$$

Thus (37) and Lemma 4 imply that there exists $n_2 \in I\!N$ ($n_2 = \max\{n_0, n_1\}$) such that $\sup_{n>n_2} \mathsf{E}\,Z_n^2 < +\infty$, which finishes the proof. □

Proof of Theorem 1. By (47),

$$\left| D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \; - \; \frac{\phi''(1)}{2}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right| \xrightarrow{P} 0, \quad n \to \infty\,.$$

Thus the desired statement follows from Lemma 1 and the Slutsky theorem. □

Proof of Theorem 2. Lemma 4 and (37) justify to use Jensen's inequality and to write for all $n > \max\{n_0, n_1\}$

$$\left| \mathsf{E}\,D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \; - \; \frac{\phi''(1)}{2}\,\mathsf{E}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right| \;\leq\; \mathsf{E}\,\left| D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \; - \; \frac{\phi''(1)}{2}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right| \,.$$

Therefore Lemma 6 implies

$$\left| \mathsf{E}\,D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \; - \; \frac{\phi''(1)}{2}\,\mathsf{E}\,\chi^2\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \right| \quad \longrightarrow \quad 0\,, \quad n \to \infty.$$

Combining this with Lemma 2 we obtain the desired result $\mathsf{E}\,D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right) \to 0\,,\; n \to \infty.$ □

To prove Proposition 2 we need an upper bound on the function ϕ and a new upper bound on the expectation of powers of a simple function of binomially distributed random variable X.

Lemma 7 *Let the function ϕ be convex on $(0,\infty)$ with $\phi(1)=0$. If ϕ satisfies the condition*

$$\phi\left(\frac{1}{t}\right)+\phi(t)=O(t^k) \quad \text{for} \quad t\to\infty \quad \text{and some} \quad k\in I\!N$$

then there exist $C_1>0$, $C_2>0$ such that

$$\phi(t)\le C_1\frac{1}{t^k}+C_2t^k \qquad \text{for all} \quad t\in(0,\infty)\,. \tag{49}$$

Proof. The proof can be carried out in a similar way as the proof of Lemma 7 in Berlinet et al. (1998). For more details see the proof of Lemma 21 on page 76 in Hobza (2003). □

Lemma 8 *If a random variable X has binomial distribution with parameters n and p then for all $r\in I\!N$,*

$$\mathsf{E}\left(\frac{1}{X+1}\right)^r\le\frac{r!}{(n+1)^r\,p^r}\,.$$

Proof. By the definition of binomial distribution,

$$\mathsf{E}\left(\frac{1}{X+1}\right)^r=\sum_{i=0}^{n}\left(\frac{1}{i+1}\right)^r\left(\begin{array}{c} n \\ i \end{array}\right)p^i(1-p)^{n-i}$$

For all $i=0,\ldots,n$ it holds

$$\begin{aligned}\left(\frac{1}{i+1}\right)^r &= \left(\frac{1}{i+1}\cdot\frac{1}{i+2}\cdots\frac{1}{i+r}\right)\cdot\left(1+\frac{1}{i+1}\right)\cdot\left(1+\frac{2}{i+1}\right)\cdots\left(1+\frac{r-1}{i+1}\right)\\ &\le \left(\frac{1}{i+1}\cdot\frac{1}{i+2}\cdots\frac{1}{i+r}\right)\cdot r!\,,\end{aligned}$$

since all the functions $(1+j/(i+1))$, $j=1,\ldots,r-1$, are decreasing as functions of i and thus attain their maximum at $i=0$. Using this result we obtain

$$\begin{aligned}\mathsf{E}\left(\frac{1}{X+1}\right)^r &\le r!\sum_{i=0}^{n}\frac{n!}{(n-i)!(i+r)!}\,p^i(1-p)^{n-i}\\ &= r!\frac{1}{(n+1)\cdots(n+r)}\sum_{i=0}^{n}\frac{(n+r)!}{(n-i)!(i+r)!}\,p^i(1-p)^{n-i}\\ &\le \frac{r!}{(n+1)^r}\sum_{j=r}^{n+r}\left(\begin{array}{c} n+r \\ j \end{array}\right)p^{j-r}(1-p)^{n+r-j}\\ &\le \frac{r!}{(n+1)^rp^r}\sum_{j=0}^{n+r}\left(\begin{array}{c} n+r \\ j \end{array}\right)p^{j}(1-p)^{n+r-j}=\frac{r!}{(n+1)^rp^r}\,.\end{aligned}$$

□

Remark 6 This lemma is a generalization of Lemma 1 playing a fundamental role in the paper of Barron et al. (1992). It was presented there for $r=1$.

Now we can deal with the proof of Proposition 2.

Proof of Proposition 2. By the Schwarz inequality, for all $n \in \mathbb{N}$

$$\mathsf{E}\,\left(D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2 = \mathsf{E}\,\left(\sum_{j=1}^{m_n} p_{nj}\,\phi\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)\right)^2$$

$$\leq \sum_{j=1}^{m_n} p_{nj}^2\,\mathsf{E}\,\phi^2\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right) + 2\sum_{j=1}^{m_n}\sum_{\ell=j+1}^{m_n} p_{nj}p_{n\ell}\left(\mathsf{E}\,\phi^2\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)\mathsf{E}\,\phi^2\left(\frac{\widetilde{p}_{n\ell}}{p_{n\ell}}\right)\right)^{\frac{1}{2}}. \tag{50}$$

From the definition of the Barron estimate $\widetilde{P}_n$, particularly from the definition of the probability vector $\widetilde{\boldsymbol{p}}_n$ corresponding to the distribution $\widetilde{P}_n^{(n)}$ (cf. (24)), it follows that $\widetilde{p}_{nj} \neq 0$ for all $1 \leq j \leq m_n$. Thus we can use Lemma 7 to obtain the following upper bound

$$\mathsf{E}\,\phi^2\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right) \leq C_1^2\,\mathsf{E}\left(\frac{p_{nj}}{\widetilde{p}_{nj}}\right)^{2k} + 2\,C_1C_2 + C_2^2\,\mathsf{E}\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)^{2k}. \tag{51}$$

We shall upperbound the expectations

$$\mathsf{E}\left(\frac{p_{nj}}{\widetilde{p}_{nj}}\right)^{2k} \quad \text{and} \quad \mathsf{E}\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)^{2k}. \tag{52}$$

Let us start with the first of them. Since

$$1 - \varepsilon_n = \frac{n\varepsilon_n}{m_n} = \frac{n}{n+m_n},$$

we get for all n and $1 \leq j \leq m_n$

$$\mathsf{E}\left(\frac{p_{nj}}{\widetilde{p}_{nj}}\right)^{2k} = p_{nj}^{2k}\,\mathsf{E}\left(\frac{1}{(1-\varepsilon_n)\frac{Y_{nj}}{n} + \varepsilon_n q_{nj}}\right)^{2k} = (np_{nj})^{2k}\left(\frac{n+m_n}{n}\right)^{2k}\mathsf{E}\left(\frac{1}{Y_{nj}+1}\right)^{2k}.$$

Further, Y_{nj} are binomially distributed with parameters n, p_{nj}. Thus using Lemma 8 we obtain for all n and $1 \leq j \leq m_n$

$$\mathsf{E}\left(\frac{p_{nj}}{\widetilde{p}_{nj}}\right)^{2k} \leq (2k)!\left(\frac{n+m_n}{n}\right)^{2k} \quad \text{where} \quad (2k)!\left(\frac{n+m_n}{n}\right)^{2k} \longrightarrow (2k)! \qquad \text{(cf. (33))}.$$

Hence, there exists such $n_A \in \mathbb{N}$ and a finite real constant $K_1 > 0$ such that

$$\mathsf{E}\left(\frac{p_{nj}}{\widetilde{p}_{nj}}\right)^{2k} \leq K_1 \tag{53}$$

for all $n > n_A$ and for all $1 \leq j \leq m_n$. The second expectation of (52) satisfies for all n and $1 \leq j \leq m_n$

$$\begin{aligned}
E\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)^{2k} &= \mathsf{E}\left((1-\varepsilon_n)\frac{\widehat{p}_{nj}}{p_{nj}} + \varepsilon_n\frac{q_{nj}}{p_{nj}}\right)^{2k} \\
&= \sum_{i=0}^{2k}\binom{2k}{i}(1-\varepsilon_n)^{2k-i}\left(\varepsilon_n\frac{q_{nj}}{p_{nj}}\right)^i\mathsf{E}\left(\frac{\widehat{p}_{nj}}{p_{nj}}\right)^{2k-i} \\
&\leq \sum_{i=0}^{2k}\binom{2k}{i}\left(\varepsilon_n\frac{q_{nj}}{p_{nj}}\right)^i\mathsf{E}\left(\frac{\widehat{p}_{nj}}{p_{nj}}\right)^{2k-i}.
\end{aligned} \tag{54}$$

From formula (3) on p. 51 in Johnson and Kotz (1969) we know that a binomially distributed X with parameters (n,p) satisfies the relation

$$\mathsf{E}\, X^r = \sum_{\ell=1}^{r} S_{r,\ell}\ \mu_{(\ell)}(X), \qquad r \in I\!N,$$

where $\mu_{(\ell)}(X) = n(n-1)\cdots(n-\ell+1)\cdot p^\ell$ and $S_{r,\ell}$ are so called *Stirling numbers of the second kind* defined for all natural numbers r and $1 \le \ell \le r$. Since

$$\mu_{(\ell)}(X) \le n^\ell p^\ell,$$

we have an upper bound

$$\mathsf{E}\, X^r \le \sum_{\ell=1}^{r} S_{r,\ell}\ n^\ell p^\ell\,. \tag{55}$$

Now, we can use formula (55) to get for all $1 \le r \le 2k$ and all $1 \le j \le m_n$

$$\begin{aligned} \mathsf{E}\left(\frac{\widehat{p}_{nj}}{p_{nj}}\right)^r &= \frac{1}{(np_{nj})^r}\ \mathsf{E}\,(Y_{nj})^r \ \le\ \frac{1}{(np_{nj})^r}\sum_{\ell=1}^{r} S_{r\ell} n^\ell p_{nj}^\ell \\ &= \sum_{\ell=1}^{r} S_{r\ell}\frac{1}{(np_{nj})^{r-\ell}} \ \le\ \sum_{\ell=1}^{r} S_{r\ell}\left(\frac{m_n^\beta}{n}\cdot\frac{1}{m_n^\beta \min\limits_{1\le j\le m_n} p_{nj}}\right)^{r-\ell} \quad \to S_{rr} = 1\,, \end{aligned}$$

since the remaining terms in the sum tend by the $\mathcal{P}_n$-assumptions to zero. This means that there exist a finite real constant $K > 0$ and for all $1 \le r \le 2k$ natural numbers n_r such that for all $n > n_r$

$$\mathsf{E}\left(\frac{\widehat{p}_{nj}}{p_{nj}}\right)^r < K\,. \tag{56}$$

If we choose $n_K = \max\{n_1,\ldots,n_{2k}\}$ then for every $n > n_K$ (56) holds for all $1 \le r \le 2k$ and $1 \le j \le m_n$. At the same time, (34) and (35) imply

$$\varepsilon_n \frac{q_{nj}}{p_{nj}} \le \frac{m_n^\beta}{n+m_n}\cdot\frac{1}{m_n^\beta \min\limits_{1\le j\le m_n} p_{nj}} \longrightarrow 0$$

which means that there exist $J > 0$ and $n_J \in I\!N$ such that for all $n > n_J$, $\varepsilon_n(q_{nj}/p_{nj}) < J$. Combining this with (54) and (56), we get for all $n > n_B = \max\{n_J, n_K\}$ and for all $1 \le j \le m_n$

$$E\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right)^{2k} \le \sum_{i=0}^{2k}\begin{pmatrix}2k\\ i\end{pmatrix} J^i K = K(1+J)^{2k} \triangleq K_2 < +\infty\,.$$

Substituting this and (53) into (51), we get for all $n > n_C = \max\{n_A, n_B\}$ and for all $1 \le j \le m_n$

$$\mathsf{E}\,\phi^2\left(\frac{\widetilde{p}_{nj}}{p_{nj}}\right) \le C_1^2 K_1 + 2C_1C_2 + C_2^2 K_2 \triangleq K_3 < +\infty\,.$$

Applying this bound in the formula (50), we finally get

$$\mathsf{E}\ \left(D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2 \le K_3 \sum_{j=1}^{m_n} p_{nj}^2 + 2K_3 \sum_{j=1}^{m_n}\sum_{\ell=j+1}^{m_n} p_{nj}p_{n\ell} \le 3K_3$$

and thus

$$\sup_{n>n_1} \mathsf{E}\ \left(D_\phi\left(\widetilde{\boldsymbol{p}}_n, \boldsymbol{p}_n\right)\right)^2 \ <\ +\infty$$

for $n_1 = n_C$. □

Appendix

In this section we state conditions on the supposed model guaranteeing relation (18) for a wide class of ϕ-divergences.

Let us consider on the Borel line $(I\!\!R, \mathcal{B})$ probability distributions P, Q where P is dominated by Q with a density $p(x)$ w.r.t. Q. Further, let $\mathcal{P}_n = \{A_{nj} : \ j \in J\}$ be an interval partition of $I\!\!R$. Under this setup the density function p_n of the distribution P_n (defined by (15)) with respect to Q can be written in the form

$$p_n(x) = \frac{P(A_n(x))}{Q(A_n(x))}, \quad \text{where} \quad A_n(x) = A_{nj} \in \mathcal{P}_n \ \text{ if } \ x \in A_{nj},$$

for $x \in I\!\!R$, $n = 1, 2, \ldots$. The functions $p_n(x)$ can be viewed as random variables on the probability space $(I\!\!R, \mathcal{B}, Q)$, namely $p_n = E_Q\,(p|\mathcal{P}_n)$, where $E_Q\,(\cdot|\mathcal{P}_n)$ is the corresponding conditional expectation E_Q with respect to the σ-field of $\mathcal{B}$ generated by $\mathcal{P}_n$, i.e.

$$E_Q\,(p|\mathcal{P}_n)_x = \frac{1}{Q(A_n(x))} \int_{A_n(x)} p\, dQ = \frac{P(A_n(x))}{Q(A_n(x))}\,.$$

Let $\mathbb{F}_Q$ be the set of all probability densities with respect to Q, i.e.

$$\mathbb{F}_Q = \{f \in L_1(Q) : f \geq 0, \int f\, dQ = 1\}\,.$$

Definition 1 *The sequence of partitions $\mathcal{P}_n$ is called Q-approximating if for every $f \in \mathbb{F}_Q$*

$$\lim_{n\to\infty} E_Q\,(f|\mathcal{P}_n) = f \quad Q - a.s. \tag{57}$$

Similar property of partitions figuring in the definition of Barron estimator was required in Barron et al. (1992), Györfi et al. (1998) and other papers dealing with this estimator.

Remark 7 Applying (57) to the positive and negative part of any $h \in L_1(Q)$, one easily obtains that (57) is equivalent to

$$\lim_{n\to\infty} E_Q\,(h|\mathcal{P}_n) = h \quad Q - \text{a.s.} \quad \text{for all} \quad h \in L_1(Q).$$

Now we will try to outline what this condition means and when it is satisfied.

Proposition 3 *Let the distribution Q on $(I\!\!R, \mathcal{B})$ be non-atomic in the sense that $Q(\{x\}) = 0$ for all $x \in I\!\!R$. If $\mathcal{P}_n$ is Q-approximating then*

$$\lim_{n\to\infty} Q(A_n(x)) = 0 \qquad \textit{for all} \quad x \in I\!\!R\,. \tag{58}$$

Proof. If (58) doesn't hold then there is $x \in I\!\!R$ such that

$$\lim_{n\to\infty} Q(A_n(x)) > 0\,.$$

Since Q is non-atomic, the intervals $A_n(x)$ cannot shrink to a singleton $\{x\}$ when $n \to \infty$. Thus, there exists a nonempty interval (a, c) and an increasing sequence n_i, $i = 1, 2, \ldots$, such that $0 < Q((a,c)) < 1$ and $(a,c) \subset A_{n_i}(x)$ for all i. Further, there exists $a < b < c$ such that $Q((a,b))$ belongs to $(0, Q((a,c)))$. Let us define

$$\begin{aligned} h(x) &= 1 \qquad x \in (a,b) \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Then $h \in L_1(Q)$ and for all $x \in (a,b)$ and the sequence A_{n_i} it holds

$$\lim_{i\to\infty} E_Q \left(h|\mathcal{P}_{n_i}\right)_x = \lim_{i\to\infty} \frac{1}{Q(A_{n_i})} \int_a^b dQ = \lim_{i\to\infty} \frac{Q(a,b)}{Q(A_{n_i})} \le \frac{Q(a,b)}{Q(a,c)} < h(x) = 1\,,$$

which contradicts (57). □

Remark 8 In practical applications probability distributions are usually defined on (bounded or unbounded) intervals where their densities are almost everywhere positive. Thus they dominate on their supports the Lebesgue measure. If we moreover consider only the Q-a.e. continuous densities f then (58) is also sufficient condition for the Q-approximating property (57) of the partition $\mathcal{P}_n$. This can be seen from the relations

$$\begin{aligned} &\left|\frac{1}{Q(A_n(x))} \int_{A_n(x)} f(y)\, dQ(y) - f(x)\right| = \left|\frac{1}{Q(A_n(x))} \int_{A_n(x)} f(y) - f(x)\, dQ(y)\right| \\ &\le \frac{1}{Q(A_n(x))} \int_{A_n(x)} |f(y) - f(x)|\, dQ(y) \le \sup_{y\in A_n(x)} |f(y) - f(x)|\,. \end{aligned}$$

Remark 9 The condition (57) is fulfilled also for all partitions $\mathcal{P}_n$ which are nested in the sense $\sigma(\mathcal{P}_1) \subset \sigma(\mathcal{P}_2) \ldots$ (where $\sigma(\mathcal{P}_i)$ denotes the σ-field generated by $\mathcal{P}_i$) and their union generates the σ-field $\mathcal{B}$ of Borel sets in $I\!\!R$. This follows for all Q from the Lévy martingale convergence theorem (see e.g. Section VII.4 in Doob (1990)).

In order to be able to prove property (18) we need another condition. Let us suppose $p \in L_3(Q)$ and define on $(I\!\!R, \mathcal{B}, Q)$ the conditional expectations

$$y_n = E_Q \left(p^2|\mathcal{P}_n\right) \qquad \text{and} \qquad z_n = E_Q \left(p^3|\mathcal{P}_n\right).$$

We consider also the convex function $\phi(t) = t^3 - 1,\ t \in (0,\infty)$ and the corresponding power divergence $I_3^*(P,Q) = 6\, I_3(P,Q)$ (cf. Table 1).

Definition 2 *The sequence of partitions $\mathcal{P}_n$ is called $\underline{(P,Q)\text{-approximating}}$ for P and Q with $I_3(P,Q) < \infty$ if it is Q-approximating and the corresponding random sequences y_n/p_n and z_n/p_n^2 are uniformly Q-integrable.*

Remark 10 Since by definition

$$I_3(P,Q) = \frac{1}{6} \left(\int p^3 \, dQ - 1 \right)$$

the conditions

$$I_3(P,Q) < \infty \quad \text{and} \quad p \in L_3(Q)$$

are equivalent.

Now we present two conditions sufficient for the uniform Q-integrability. The first one states that it suffices to suppose that the density $p = dP/dQ$ is Q-a.s. bounded away from zero. This takes place for some densities considered in the statistical theory as well as in the practise.

Proposition 4 *Let P be a probability measure dominated by probability measure Q with density $p = dP/dQ$ and $I_3(P,Q) < \infty$. If there exists $\gamma > 0$, such that*

$$p(x) \geq \gamma \quad Q - a.s.\,,$$

then every Q-approximating sequence of partitions $\mathcal{P}_n$ is also (P,Q)-approximating.

Proof. According to Remark 10, $p \in L_3(Q)$ and thus also $p \in L_2(Q)$. We have to prove that the sequences y_n/p_n and z_n/p_n^2 are uniformly Q-integrable. By definition,

$$\frac{y_n(x)}{p_n(x)} = \frac{y_n(x)}{\frac{1}{Q(A_n(x))}\int_{A_n(x)} p\, dQ} \leq \frac{1}{\gamma}\, y_n(x)$$

and

$$\frac{z_n(x)}{(p_n(x))^2} = \frac{z_n(x)}{\left(\frac{1}{Q(A_n(x))}\int_{A_n(x)} p\, dQ\right)^2} \leq \frac{1}{\gamma^2}\, z_n(x)\,.$$

Thus, it remains to show the uniform Q-integrability of the sequences y_n and z_n. We do this for z_n since for y_n the proof is similar. Since $p \in L_3(Q)$ and the sequence $\mathcal{P}_n$ is supposed to be Q-approximating we know that $z_n \to p^3$ Q-a.s.. Further,

$$E_Q\left(z_n\right) = E_Q\left(E_Q\left(p^3|\mathcal{P}_n\right)\right) = E_Q\left(p^3\right) < \infty$$

and the uniform Q-integrability of the random sequence z_n follows e.g. from Lemma B on page 15 in Serfling (1980). □

The next lemma provides another sufficient condition for the uniform integrability in Definition 2.

Proposition 5 *A Q-approximating sequence of partitions $\mathcal{P}_n$ is (P,Q)-approximating for a probability measure P dominated by Q if there exist positive α and β for which*

$$\frac{3}{\alpha} + \frac{2}{\beta} < 1 \tag{59}$$

and the density $p = dP/dQ$ satisfies the conditions

$$p \in L_\alpha(Q) \quad and \quad \frac{1}{p} \in L_\beta(Q)\,.$$

Proof. The proof of this assertion is a slightly adapted version of the proof of Proposition 2 in Györfi et al. (1998). For more details see the proof of Lemma 4 on page 49 in Hobza (2003). □

Now we state the main result of this section.

Theorem 3 *If $I_3(P,Q) < +\infty$ and the sequence of partitions $\mathcal{P}_n$ is (P,Q)-approximating then*

$$D_\phi(P, P_n) = o(1), \qquad n \to \infty\,,$$

for all ϕ satisfying the ϕ-assumptions.

Remark 11 By Lemma 1 of Györfi et al. (1998) in the particular case of the χ^2-divergence $D_\phi(\cdot,\cdot) = \chi^2(\cdot,\cdot)$ a weaker form of (P,Q)-approximating property is sufficient where the uniform Q-integrability of z_n/p_n^2 is not required. The uniform integrability of y_n/p_n cannot be omitted. Indeed, the cited paper presented an example demonstrating that the convergence

$$\chi^2(P, P_n) = o(1)$$

need not hold without the uniform Q-integrability of y_n/p_n even if the partitions $\mathcal{P}_n$ are nested and generate $\mathcal{B}$, i.e. if they are Q-approximating.

The proof of Theorem 3 will be divided into several steps.

Let us consider probability distributions $P, \tilde{P}$ dominated by Q with corresponding densities $p, \tilde{p}$ and the relative deviation

$$\Delta_{P,\tilde{P}}(x) = \left|\frac{p(x)}{\tilde{p}(x)} - 1\right|. \tag{60}$$

By Table 1,

$$\chi^3(P, \widetilde{P}) = \int \frac{|p - \tilde{p}|^3}{\tilde{p}^2}\, dQ.$$

Lemma 9 *For all distributions $P, \tilde{P}$ with $\Delta_{P,\tilde{P}}$ sufficiently small Q-a.s., and for all functions ϕ satisfying the ϕ-assumptions,*

$$\left|D_\phi(P,\tilde{P}) - \frac{\phi''(1)}{2}\,\chi^2(P,\tilde{P})\right| \le \frac{L_\phi}{2}\,\chi^3(P,\tilde{P}) \le 3\,L_\phi\,I_3(P,\tilde{P})$$

where L_ϕ is the Lipschitz constant from the ϕ-assumptions.

Proof. By the ϕ-assumptions, there exist $\delta_\phi > 0$ and $L_\phi > 0$ such that

$$\left|\phi''(t) - \phi''(1)\right| \le L_\phi |t-1| \qquad \text{if} \qquad |t-1| < \delta_\phi. \tag{61}$$

By the Taylor expansion of $\phi(t)$ in the neighborhood of $t = 1$ we get for all t with sufficiently small $|t-1|$

$$\phi(t) = \phi(1) + \phi'(1)(t-1) + \frac{\phi''(t^*)}{2}\,(t-1)^2,$$

where $|t^* - 1| \le |t-1|$. Using (32) we obtain

$$\phi(t) - \frac{\phi''(1)}{2}\,(t-1)^2 \;=\; \frac{\phi''(t^*) - \phi''(1)}{2}\,(t-1)^2$$

and applying (61) on the right-hand side we get

$$\left|\phi(t) - \frac{\phi''(1)}{2}\,(t-1)^2\right| \le \frac{L_\phi}{2}|t-1|^3. \tag{62}$$

If $\Delta_{P,\tilde{P}}$ is small enough Q-a.s. then, by multiplying the inequality by $\tilde{p}$, substituting $t = p/\tilde{p}$ and integrating on both sides, we get

$$\int \left| \tilde{p}\, \phi\left(\frac{p}{\tilde{p}}\right) - \frac{\phi''(1)}{2}\, \tilde{p}\, \left(\frac{p}{\tilde{p}} - 1\right)^2 \right| dQ \le \frac{L_\phi}{2} \int \tilde{p} \left| \frac{p}{\tilde{p}} - 1 \right|^3 dQ.$$

Thus also,

$$\left| \int \tilde{p}\, \phi\left(\frac{p}{\tilde{p}}\right) dQ - \frac{\phi''(1)}{2} \int \tilde{p}\, \left(\frac{p}{\tilde{p}} - 1\right)^2 dQ \right| \le \frac{L_\phi}{2} \int \tilde{p} \left| \frac{p}{\tilde{p}} - 1 \right|^3 dQ,$$

which means

$$\left| D_\phi(P, \tilde{P}) - \frac{\phi''(1)}{2}\, \chi^2(P, \tilde{P}) \right| \le \frac{L_\phi}{2}\, \chi^3(P, \tilde{P}).$$

The inequality $\chi^3(P, \tilde{P}) \le I_3^*(P, \tilde{P})$ can be easily seen from the facts that $I_3^*(P, \tilde{P}) = D_\phi(P, \tilde{P})$ for $\phi(t) = t^3 - 3(t-1) - 1$ where $|t-1|^3 \le \phi(t)$ for all $t \in (0, \infty)$. □

Lemma 10 *If $\chi^2(P, Q) < +\infty$, $\mathcal{P}_n$ is Q-approximating and the random sequence y_n/p_n is uniformly Q-integrable then*

$$\chi^2(P, P_n) = o(1), \qquad n \to \infty\,.$$

Proof. See Györfi et al. (1998). □

Lemma 11 *If $I_3(P, Q) < +\infty$, $\mathcal{P}_n$ is Q-approximating and the random sequence z_n/p_n^2 is uniformly Q-integrable then*

$$I_3(P, P_n) = o(1), \qquad n \to \infty\,.$$

Proof. By definition,

$$6\, I_3(P, P_n) = \int \frac{p^3}{p_n^2}\, dQ - 1$$

and $I_3(P, Q) < +\infty$ implies $p \in L_3(Q)$. Since $\mathcal{P}_n$ is Q-approximating, z_n tends Q-a.s. to p^3 and p_n to p when $n \to \infty$. Consequently,

$$\rho_n = \frac{z_n}{p_n^2}$$

tends Q-a.s. to p and from the uniform Q-integrability of ρ_n it follows (cf. e.g. Theorem A on page 14 in Serfling (1980))

$$\lim_{n \to \infty} E_Q\left(\rho_n\right) = E_Q\left(p\right) = 1\,.$$

The assertion of the lemma follows from here and from the equalities

$$\begin{aligned} E_Q\left(\rho_n\right) &= \int_{I\!R} \left(\frac{\frac{1}{Q(A_n(x))} \int_{A_n(x)} p^3(x)\, dQ(x)}{\left(\frac{P(A_n(x))}{Q(A_n(x))}\right)^2} \right) dQ(x) \\ &= \int_{I\!R} \frac{1}{Q(A_n(x))} \left(\int_{A_n(x)} \frac{p^3(x)}{p_n^2(x)}\, dQ(x) \right) dQ(x) \\ &= E_Q\left(E_Q\left(\frac{p^3}{p_n^2} \middle| \mathcal{P}_n \right) \right) = E_Q\left(\frac{p^3}{p_n^2} \right) = 6\, I_3(P, P_n) + 1\,, \end{aligned}$$

where we have used property of conditional expectation w.r.t. a σ-field. □

Proof of Theorem 3. Since $\mathcal{P}_n$ is Q-approximating, p_n tends Q-a.s. to p as $n \to \infty$. Consequently, p/p_n tends to 1 Q-a.s. and

$$\Delta_{P,P_n} = \left| \frac{p}{p_n} - 1 \right|$$

tends to zero Q-a.s. Thus, by Lemma 9

$$\left| D_\phi(P, P_n) - \frac{\phi''(1)}{2} \chi^2(P, P_n) \right| \leq 3\, L_\phi\, I_3(P, P_n)$$

for sufficiently large n. Since $I_3(P,Q) < +\infty$ implies $\chi^2(P,Q) < +\infty$ (this can be seen in the manner used at the end of proof of Lemma 9), the proof is completed by applying Lemmas 10 and 11. □

Acknowledgement

The author thanks Igor Vajda for his valuable comments and help with improvement of the paper. This paper is supported by the grant GA CR 201/02/1391.

References

Barron, A.R. (1988) *The convergence in information of probability density estimators*, presented at IEEE Int. Symp. Inform. Theory, Kobe, Japan, June 19-24.

Barron, A.R., Györfi, L. and van der Meulen, E. (1992) *Distribution estimation consistent in total variation and in two types of information divergence*, IEEE transactions on Information Theory, vol. 38, no. 5, pp. 1437-1454.

Berlinet A., Györfi, L. and van der Meulen, E. (1997) *The asymptotic normality of I-divergence in multivariate density estimation*, Publ. Inst. Stat. Univ. Paris, 41, pp. 3-27.

Berlinet, A., Vajda, I. and van der Meulen, E. C. (1998) *About the asymptotic accuracy of Barron density estimates*, IEEE transactions on Information Theory, vol. 44, no. 3, pp. 999-1009.

Csiszár, I. (1963) *Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität on Markhoffschen Ketten*, Publ. Math. Inst. Hungar. Acad. Sci. Ser. A, nr. 8, pp. 84-108.

Csiszár, I. (1967) *Information-type measures of difference of probability distributions and indirect observations* Studia Sci. Math. Hungar. 2, pp. 299-318.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L_1-view*, Wiley, New York.

Devroye, L. and Györfi, L. (1990) *No empirical measure can converge in total variation sense for all distributions*, Ann. Statist. vol. 24, pp. 508-539.

Doob, J. L. (1990) *Stochastic Processes*, Wiley Classic Library Edition, Wiley, New York.

Györfi, L., Liese, F., Vajda, I. and van der Meulen, E. C. (1998) *Distribution estimates consistent in χ^2-divergence*, Statistics 32, pp. 31-57.

Györfi, L. and Vajda, I. (2002) *Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models*, Statistics & Probability Letters, 56, pp. 57-67.

Hobza, T. (2003) *Asymptotics of some histogram-based density estimates*, Ph.D. dissertation, Czech Technical University, Prague. Available online at: http://tjn.fjfi.cvut.cz/~hobza/papers

Johnson, N. L. and Kotz, S. (1969) *Distributions in Statistics: Discrete Distributions*, Houghton Mifflin Comp., Boston.

Liese, F. and Vajda, I. (1987) *Convex Statistical Distances*, Teubner, Leipzig.

Österreicher, F. and Vajda, I. (2003) *A new class of metric divergences on probability spaces and its applicability in statistics*, Annals of the Institute of Statistical Mathematics (in print).

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Vajda, I. (1972) *On the f-divergence and singularity of probability measures*, Periodica Mathematica Hungarica 2, 223-234.

Vajda, I. (1989) *Theory of Statistical Inference and Information*, Kluwer, Boston.

Vajda, I. (1995) *Information-theoretic Methods in Statistics*, Research Report n. 1834, Inst. Inform. Theory and Automation, Prague, Czech Republic.

Vajda, I. and van der Meulen, E. C. (1998) *The chi-square error of Barron estimator of regular density is asymptotically normal*, Publ. Inst. Stat. Univ. Paris, 42, pp. 93-110.

Vajda, I. and van der Meulen, E. C. (2001) *Optimization of Barron density estimates*, IEEE transactions on Information Theory, vol. 47, no. 5, pp. 1867-1883.

Vajda, I. (2002) *On convergence of information contained in quantized observations*, IEEE transactions on Information Theory, vol. 48, no. 8, pp. 2163-2172.