About the Asymptotic Accuracy of Barron Density Estimates

Alain Berlinet, Igor Vajda, *Senior Member, IEEE*, and Edward C. van der Meulen, *Fellow, IEEE*

***Abstract*—By extending the information-theoretic arguments of previous papers dealing with the Barron-type density estimates, and their consistency in information divergence and chi-square divergence, the problem of consistency in Csiszár's ϕ-divergence is motivated for general convex functions ϕ. The problem of consistency in ϕ-divergence is solved for all ϕ with $\phi(0) < \infty$ and $\phi(t) = O(t \ln t)$ when $t \to \infty$. The problem of consistency in the expected ϕ-divergence is solved for all ϕ with $t\,\phi(1/t) + \phi(t) = O(t^2)$ when $t \to \infty$. Various stronger versions of these asymptotic restrictions are considered too. Assumptions about the model needed for the consistency are shown to depend on how strong these restrictions are.**

***Index Terms*— Barron density estimate, consistency in divergences and expected divergences, divergences of Csiszár, nonparametric density estimates, Rényi distances.**

I. INTRODUCTION

THERE is an extensive literature dealing with nonparametric density estimators consistent in the L_1-distance or L_2-distance. Relatively few papers have considered density estimators consistent in distances topologically stronger than these two. Bickel and Rosenblatt [7] proved consistency of the well-known kernel estimators in the reversed (Neyman-modified) χ^2-divergence. Barron [2], [4] introduced an estimator combining in some sense the advantages of histogram and kernel estimators. Later Barron *et al.* [4] introduced a whole class of such estimators and proved their consistency in the information divergence. Recently, Györfi *et al.* [18] established the consistency of the Barron estimator in the χ^2-divergence.

The Barron estimator proved to be a practically useful tool for nonparametric density estimation as part of the nonparametric density estimation package reported in [38]. It combines the computational simplicity of classical histograms with the accuracy of computationally much more complicated estimators. From this point of view it thus seems to be suitable for applications in many areas of communication where decisions depend on density estimates and various functionals of such estimates.

According to [4], [18], the information divergence and the χ^2-divergence between a true density f and any estimate f_n are appropriate measures of inaccuracy of f_n. Estimators f_n consistent in these two divergences are thus asymptotically accurate. In this paper, we consider the general class of ϕ-divergences $D_\phi(f, f_n)$ introduced by Csiszár [10], [11] and specified by convex functions ϕ: $(0, \infty) \to R$. We extend information-theoretic and statistical arguments in [2], [4], [18] to this class and we demonstrate the significance of the inaccuracy measure $D_\phi(f,\, f_n)$ for a quite wide family of convex functions ϕ. We prove the asymptotic accuracy of Barron estimator in the sense of consistency in ϕ-divergence or expected ϕ-divergence for most common types of convex functions ϕ. Conditions imposed on the statistical model by our theorems depend on how fast the convex function $\phi(t)$ increases in the neighborhoods of $t = 0$ and $t = \infty$. The faster ϕ increases, the stronger is the ϕ-divergence topology on the model and, consequently, the stricter restrictions on the model are needed.

The two types of information divergence considered in [4] and the two types of χ^2-divergence considered in [7] and [18] satisfy the aforementioned technical conditions. The same holds also for the total variation norm considered by Devroye and Györfi [16] (c.f. also Barron *et al.* [4] and Berlinet *et al.* [5]), the distances of Matusita [23], and the most important of the distances of Rényi [30] (c.f. their variant considered by Liese and Vajda [21] and Read and Cressie [29]). Our paper extends these results. In particular, our result concerning the reversed χ^2-divergence is complementary to the consistency of Barron estimator in the expected χ^2-divergence established in [18]. The consistency in the expected reversed information divergence is proved here under essentially weaker conditions on the model than in [4].

Consistency in ϕ-divergence and expected ϕ-divergence leads to the problem of asymptotic distribution of appropriately normalized differences $D_\phi(f,\, f_n) - E\,D_\phi(f,\, f_n)$. This problem has already been solved for some ϕ-divergences and estimators f_n. Namely, an asymptotic normality has been established for the reversed χ^2-divergence and kernel estimators in [7], for the total variation and histogram estimators in [5], and for the information divergence and Barron-type estimators in [6].

II. THE ESTIMATORS

We restrict ourselves to measures defined on the Borel line $(R, \mathcal{B})$, and to probability distributions defined by densities on R. Let ν be a probability measure or, more generally, a σ-finite measure (*Lebesgue* is a typical continuous example, and *counting* a typical discrete example). By $L_p(\nu)$, $1 \le p \le \infty$,

Manuscript received January 31, 1997; revised October 24, 1997. This work was supported under the Copernicus Grant 579 and by the University of Montpellier II.

A. Berlinet is with the Department of Statistics, University of Montpellier II, 34095 Montpellier Cedex, France.

I. Vajda is with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague, Czech Republic.

E. C. van der Meulen is with the Department of Mathematics, Catholic University of Leuven, 3001 Heverlee, Belgium.

Publisher Item Identifier S 0018-9448(98)02353-0.

we denote the normed space of measurable functions $g: R \to R$ with finite norms

$$\|g\|_p = \left(\int |g|^p \, \mathrm{d}\nu \right)^{1/p}$$

(equal to the limit $\nu - \operatorname{essup} |g|$ for $p = \infty$). By $\mathbb{F}_\nu$ we denote the set of all probability densities with respect to ν, i.e.,

$$\mathbb{F}_\nu = \{f \in L_1(\nu): f \geq 0, \|f\|_1 = 1\}.$$

In practice, many decisions depend on a probability distribution which is not precisely known. What is usually known are random observations $X_1, \cdots, X_n$ with density f in a given class $\mathbb{F} \subset \mathbb{F}_\nu$. The class $\mathbb{F}$ specifies the *statistical model*. If $\mathbb{F}$ is not simply parameterizable, one cannot simply use a parametric point estimate. In such situations one has to consider a nonparametric density estimate $f_n(x) = f_n(x; X_1, \cdots, X_n)$, i.e., a measurable mapping $f_n: R^{n+1} \to [0, \infty)$ such that a.s. (almost surely with respect to the distribution of observations $X_1, \cdots, X_n$ figuring in f_n)

$$\int f_n(x) \, \mathrm{d}\nu(x) = 1, \qquad \text{and } f_n \in \mathbb{F}.$$

A sequence of such mappings $(f_n, n \in N)$ is an *estimator* of the unknown density f.

If a distance or divergence $D(f, f_*)$ is defined for all $f, f_* \in \mathbb{F}$ and

$$\lim_{n \to \infty} D(f, f_n) = 0 \quad \text{a.s. for all } f \in \mathbb{F} \tag{1}$$

then the estimator is said to be *consistent in* D. If in this definition $D(f, f_n)$ is replaced by the expectation $E\, D(f, \hat{f}_n)$ taken with respect to the density $f(x_1) \cdots f(x_n)$ of observations $X_1, \cdots, X_n$ then we obtain the *consistency in the expected distance* D.

The most popular density estimator is the *histogram estimator* f_n^H defined by means of sequences of partitions $\mathcal{P}_n$ of the real line into intervals. We consider the variant defined for models $\mathbb{F} = \mathbb{F}_\nu$ specified by a nonatomic probability measure ν, with partitions $\mathcal{P}_n = \{A_{n,1}, \cdots, A_{n,m_n}\}$ defined by

$$A_{n,i} = G^{-1}\left(\frac{i-1}{m_n}, \frac{i}{m_n} \right), \qquad 1 \leq i \leq m_n$$

where $G(x) = \nu(-\infty, x)$ is the distribution function of ν, and where the sequence m_n increases slowly to infinity in the sense

$$\lim_{n \to \infty} m_n = \lim_{n \to \infty} \frac{n}{m_n} = \infty.$$

The nonatomic assumption means that G is continuous on R, but not necessarily strictly monotone. Therefore, $G^{-1}(t)$ is defined as $\inf\{x: G(x) \geq t\}$. The ν-probability of all sets $A \in \mathcal{P}_n$ is then the constant $h_n = 1/m_n$, and the histogram estimator is defined by the formula

$$f_n^H(x) = \frac{\mu_n^E(A)}{\nu(A)} = \frac{\mu_n^E(A)}{h_n}, \qquad \text{for all } x \in A, A \in \mathcal{P}_n \tag{2}$$

where

$$\mu_n^E(B) = \frac{1}{n} \sum_{i=1}^n 1_B(X_i), \qquad \text{for all } B \in \mathcal{B} \tag{3}$$

is the *empirical probability distribution*, with all mass uniformly concentrated at the observation points $X_1, \cdots, X_n$. References to the extensive literature dealing with histogram estimators can be found in Devroye and Györfi [16] and Scott [34].

Some properties of the histogram estimator are undesirable. For instance, if ν is continuous, then f_n^H is a.s. discontinuous. Additionally, the support $S_n = \{f_n^H > 0\}$ of f_n^H may not contain the support $S_f = \{f > 0\}$ of f, i.e.,

$$f \not\ll f_n^H. \tag{4}$$

Some distances $D(f, f_*)$ defined on $\mathbb{F}_\nu$ are hypersensitive to discontinuities of densities f and f_*, or to situations when the domination $f \ll f_*$ fails. To circumvent such drawbacks various modifications of f_n^H have been proposed in the literature. The modifications include the *frequency polygon* (see Scott [34]) which is always continuous, and the *Barron estimator* f_n^B (see Barron [2]) which always dominates f.

The Barron estimator is defined for models $\mathbb{F}_\nu$ and partitions $\mathcal{P}_n$ considered in the definition of histogram estimator above, by the formula

$$f_n^B(x) = \frac{n\, h_n\, f_n^H(x) + 1}{n\, h_n + 1} = \frac{n\, \mu_n^E(A) + 1}{n\, h_n + 1}, \qquad \text{for all } x \in A, A \in \mathcal{P}_n \tag{5}$$

where $h_n = 1/m_n = \nu(A)$ for all $A \in \mathcal{P}_n$, $h_n \to 0$ and $n\, h_n \to \infty$ as $n \to \infty$. Each estimate f_n^B belongs a.s. to $\mathbb{F}_\nu$. Indeed, it is nothing but a convex mixture in the convex set $\mathbb{F}_\nu$ of probability densities

$$f_n^B = (1 - a_n)\, f_n^H + a_n, \qquad \text{for } a_n = \frac{1}{n\, h_n + 1}. \tag{6}$$

The second mixture component is the density $f_*(x) \equiv 1 \in \mathbb{F}_\nu$ of the dominating probability measure ν itself. Therefore, f_n^B is a probability density whose support $S_n = \{f_n^B > 0\}$ coincides with that of ν. Consequently, $f \ll f_n^B$ for all $f \in \mathbb{F}_\nu$.

Other examples of density estimators are the *kernel estimator* f_n^K defined as a convolution $K * \mu_n^E$ of a stochastic kernel K with μ_n^E (see Rosenblatt [31] and Parzen [28], or the monographs [16] and [34]), or the *minimum Kolmogorov distance estimator* f_n^{KOL} which minimizes the Kolmogorov distance between μ_n^E and $f_* \in \mathbb{F}_\nu$ (see Györfi *et al.* [17]).

III. The Distances

The most natural and popular candidate for the distance $D(f, f_*)$ considered in Section II is the L_1-norm $\|f - f_*\|_1$. It is defined for all $f, f_* \in \mathbb{F}_\nu$, it is further convex and bounded by 2 on $\mathbb{F}_\nu$, and possesses other nice properties. For this distance it is relatively easy to prove the consistency and other asymptotic properties of the above considered estimators. Some of these properties can be extended to the L_p-norms $\|f - f_*\|_p$, $1 < p \leq \infty$, for models $\mathbb{F} = \mathbb{F}_\nu \cap L_p(\nu) \subset \mathbb{F}_\nu$.

In fact, the original asymptotic results of Rosenblatt and Parzen concerning the kernel estimators were formulated for the L_2-norm. Asymptotic properties of histogram and kernel estimators under the L_1-norm can be found in Devroye and Györfi [16]. Beirlant *et al.* [5] is the most recent contribution in this area. The L_1-norm consistency of the Barron estimator was established in Barron *et al.* [4, Sec. II]. The L_2-norm properties of the histogram and kernel estimators were recently summarized in Scott [34].

Sometimes it is necessary to consider distances $D(f, f_*)$ different from the L_p-norms. These distances are often not topologically dominated by even the strongest of the L_p-norms, the L_∞-norm. Conversely, they topologically dominate weaker norms, such as the L_1-norm.

Definition: The *ϕ-divergence* of distributions represented by densities $f, f_* \in \mathbb{F}_\nu$ is defined as

$$D_\phi(f, f_*) = \int f_* \, \phi\left(\frac{f}{f_*}\right) \mathrm{d}\nu \tag{7}$$

where ϕ is assumed to be convex on $(0, \infty)$ and not linear in an open neighborhood of 1, with $\phi(1) = 0$. (The expression behind the integral is assumed to be $\phi(0) = \lim_{t\downarrow 0} \phi(t)$ on $S_{f_*} - S_f$ and $\phi(\infty)/\infty = \lim_{t\to\infty} \phi(t)/t$ on $S_f - S_{f_*}$. Outside $S_f \cup S_{f^*}$ this expression is assumed to be 0.)

The concept of ϕ-divergence was first introduced by Csiszár [10]. As proved by Csiszár [10], [11] and Vajda [35], [37], under the present assumptions on ϕ, the sum $\phi(0) + \phi(\infty)/\infty$ is strictly positive and $D_\phi(f, f_*)$ takes on values in the interval $[0, \phi(0) + \phi(\infty)/\infty]$. The equality $D_\phi(f, f_*) = 0$ takes place if and only if $f = f_*$ while

$$D_\phi(f, f_*) = \phi(0) + \phi(\infty)/\infty$$

if the orthogonality $f \perp f_*$ holds. For $\phi(0) + \phi(\infty)/\infty < \infty$ the last equality is equivalent to orthogonality of f and f_*. Finally, for $\phi(0) = \infty$, the finiteness of $D_\phi(f, f_*)$ implies $f_* \ll f$. For $\phi(\infty)/\infty = \infty$ the same finiteness implies the reversed domination $f \ll f_*$. Thus if $f \not\ll \hat{f}_n$ with a positive ν-probability then $D_\phi(f, \hat{f}_n) = \infty$ for all ϕ with $\phi(\infty)/\infty = \infty$.

According to Österreicher and Vajda [26], $D_\phi(f, f_*)$ is an average risk in the Bayesian dichotomy with conditional densities f, f_* and the 0–1 loss, provided the prior probabilities are randomly selected with a distribution depending on ϕ. For more details about the role of ϕ-divergences in Bayesian decision, we refer to Clarke and Barron [8].

Obviously, the ϕ-divergence remains unaltered when ϕ is replaced by the nonnegative function $\phi(t) - \phi'_+(1)\,(t-1)$ where $\phi'_+(1)$ is the right-hand derivative of $\phi(t)$ at $t = 1$. Consequently, we can restrict ourselves to nonnegative ϕ with the properties assumed in (7).

Example 1: χ^p-Divergences: The class of functions

$$\phi_p(t) = |t - 1|^p, \qquad 1 \le p < \infty$$

defines a particular class of ϕ-divergences on $\mathbb{F}_\nu$. These *χ^p-divergences* are cumulants of the likelihood ratio f/f_*, as defined by the formula

$$\chi^p(f, f_*) = \int \left|1 - \frac{f}{f_*}\right|^p f_* \, \mathrm{d}\nu.$$

As a special case, we have the L_1-norm

$$\chi^1(f, f_*) = \|f - f_*\|_1$$

and the well-known *χ^2-divergence*

$$\chi^2(f, f_*) = \int \frac{(f - f_*)^2}{f} \, \mathrm{d}\nu.$$

The upper bound is

$$\phi_p(0) + \phi_p(\infty)/\infty = \begin{cases} 2, & \text{if } p = 1 \\ \infty, & \text{otherwise} \end{cases}$$

and $\chi^p(f, f_*)^{1/p}$ is nondecreasing in the variable $p \ge 1$. This class of divergences has been systematically studied in [21] and [36].

Example 2: I_a-Divergences and Rényi Distances: The class of nonnegative convex functions

$$\phi_a(t) = \frac{t^a - a\,t + a - 1}{a(a-1)}, \qquad \text{for } a \ne 0 \text{ and } a \ne 1$$

with the corresponding limits

$$\phi_1(t) = t \log t - t + 1 \qquad \phi_0(t) = -\log t + t - 1$$

was introduced by Liese and Vajda [21] (unless otherwise explicitely stated, the natural logarithms are used throughout the paper). This class leads to the divergences

$$I_a(f, f_*) = D_{\phi_a}(f, f_*), \qquad \text{for all } a \in R. \tag{8}$$

called *information divergences of order a* (briefly, *I_a-divergences*). These divergences are skew-symmetric in the sense

$$I_a(f, f_*) = I_{1-a}(f_*, f), \qquad \text{for all } a \in R.$$

They are upper-bounded by $\phi_a(0) + \phi_a(\infty)/\infty$ where $\phi_a(0)$ is nonincreasing in $a \in R$ with $\phi_a(0) = 1/a$ for $a > 0$, and $\phi_a(\infty)/\infty$ is nondecreasing with $\phi_a(\infty)/\infty = 1/(1-a)$ for $a < 1$. Notice that

$$I_{1/2}(f, f_*) = 2 \int \left(\sqrt{f} - \sqrt{f_*}\right)^2 \mathrm{d}\nu$$

is twice the squared *Hellinger distance*. Other special cases include the classical *information divergence* (*I-divergence*)

$$I_1(f, f_*) = \int f \log \frac{f}{f_*} \, \mathrm{d}\nu \triangleq I(f, f_*),$$

the *reversed information divergence*

$$I_0(f, f_*) = I(f_*, f),$$

the *χ^2-divergence*

$$I_2(f, f_*) = \int \frac{f^2}{f_*} \, \mathrm{d}\nu - 1 = \int \frac{(f - f_*)^2}{f_*} \, \mathrm{d}\nu = \chi^2(f, f_*),$$

and the *reversed χ^2-divergence*

$$I_{-1}(f, f_*) = \chi^2(f_*, f).$$

ϕ-divergences and I_a-divergences of arbitrary distributions have been systematically studied in [21] and [37]. Properties of I_a-divergences of discrete distributions have been systematically studied by Read and Cressie [29].

Liese and Vajda [21] also introduced the Rényi distances of order $a \in R$

$$R_a(f, f_*) = \frac{1}{a(a-1)} \log \int f^a \, f_*^{1-a} \, \mathrm{d}\nu, \qquad \text{for } a \ne 0 \text{ and } a \ne 1$$

with the corresponding limits

$$R_0(f, f_*) = \lim_{a\downarrow 0} R_a(f, f_*) = I_0(f, f_*)$$

and

$$R_1(f, f_*) = \lim_{a\uparrow 1} R_a(f, f_*) = I_1(f, f_*).$$

These distances were originally proposed by Rényi [30] for $a > 0$ in a slightly different form. Rényi distances and I_a-divergences are one-to-one related, with mutual coincidence at $a = 0$ and $a = 1$. In this sense, the I_a-divergences can be considered as versions of the Rényi distances.

We now present several applications that motivate the present study.

Application 1: Consider a partition $\mathcal{P} = (A_1, \cdots, A_m)$ not depending on n, and the theoretical and empirical probability vectors

$$\boldsymbol{\mu} = \left(\mu(A_i) = \int_{A_i} f\, \mathrm{d}\nu \colon 1 \le i \le m\right), \qquad f \in \mathbb{F}_\nu$$

and

$$\boldsymbol{\mu}_n^E = \left(\mu_n^E(A_i)\colon 1 \le i \le m\right) \quad \text{(c.f. (3))}.$$

By the strong law of large numbers, the estimate $\boldsymbol{\mu}_n^E$ of $\boldsymbol{\mu}$ is consistent in the L_p-norms $\|\boldsymbol{\mu} - \boldsymbol{\mu}_n^E\|_p$, $1 \le p \le \infty$. It is thus reasonable to reject the hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ when the distances $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^E\|_p$ exceed certain critical values $C(p, n) > 0$. It is however too hard to calculate critical values leading to given asymptotic test sizes $0 < \alpha < 1$, i.e., to specify scaling factors $c(p, n)$ leading to known asymptotic distributions for $c(p, n)\, \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_n^E\|_p$. This negative conclusion remains valid even if the L_p-norm is replaced by the power $\|\boldsymbol{\mu} - \boldsymbol{\mu}_n^E\|_p^p$ or any other one-to-one mapping. On the other hand, if the L_p-norms are replaced by the ϕ-divergences

$$D_\phi(\boldsymbol{\mu}_n^E, \boldsymbol{\mu}_0) = \sum_{i=0}^m \mu_0(A_i)\, \phi\left(\frac{\mu_n^E(A_i)}{\mu(A_i)}\right)$$

with $\phi(t)$ twice continuously differentiable at $t = 1$ and $\phi''(1) > 0$, then the scaling factors

$$c(\phi, n) = \frac{2n}{\phi''(1)}$$

lead to the chi-squared asymptotic distribution of statistics $c(\phi, n) D_\phi(\boldsymbol{\mu}_n^E, \boldsymbol{\mu}_0)$, with $m - 1$ degrees of freedom (c.f. Section II in Menéndez *et al.* [24]). The particular choice $\phi = \phi_2$ (c.f. Example 2) leads to the Pearson statistic

$$c(\phi_2, n)\, I_2(\boldsymbol{\mu}_n^E, \boldsymbol{\mu}) = n \sum_{i=1}^n \frac{(\mu_n^E(A_i) - \mu(A_i))^2}{\mu(A_i)}$$

for which the mentioned asymptotic distribution is a classical statistical result. The choice $\phi = \phi_{-1}$ leads to the Neyman statistic

$$c(\phi_{-1}, n)\, I_{-1}(\boldsymbol{\mu}_n^E, \boldsymbol{\mu}) = n \sum_{i=1}^n \frac{(\mu_n^E(A_i) - \mu(A_i))^2}{\mu_n^E(A_i)}$$

(reversed χ^2-divergence statistic of Pearson, see Example 2), with the same asymptotic distribution. For some alternatives and fixed sample sizes n the power of the Neyman test exceeds that of the Pearson test, while for other alternatives the converse was established to be true (see, e.g., [29, Table 5.2]). Extensions of some of these testing results to the case where the partition $\mathcal{P}$ varies with the sample size n can be found in [24].

Application 2: Consider now data compression, where the redundancy of a code based on a density $f_* \in \mathbb{F}_\nu$ equals the I-divergence $I(f, f_*)$ (see Davisson [13]). A simple binary example easily demonstrates that $I(f, f_*)$ may be arbitrarily large while f_* is arbitrarily close in the L_∞-norm to f. Therefore, even the estimates f_n consistent in all L_p-norms may provide asymptotically redundant codes, while estimates consistent in the I-divergence achieve asymptotic nonredundancy. On the other hand, by the Pinsker inequality $\|f - f_*\|_1^2 \le 2I(f, f_*)$, the I-divergence topologically dominates the L_1-norm. The consistency of f_n in the I-divergence guarantees the a.s. convergence of $\|f - f_n\|_1$ to zero. Therefore, e.g., the estimate $\int T f_n \, \mathrm{d}\nu$ of the expectation $\int T f \, \mathrm{d}\nu$ of any bounded statistics $T\colon R \to R$ is consistent too. The consistency of f_n in the expected I-divergence implies that the estimate $\int T f_n \mathrm{d}\nu$ is asymptotically unbiased. The consistencies of Barron estimator f_n^B in the I-divergence and expected I-divergence were proved in [4] for models with $I(f, 1)$ finite (c.f. [4, Lemma 2, Appendix]).

Application 3: Consider the χ^p-divergences. Similarly as in the previous situation, all χ^p-divergences with $1 < p < \infty$ dominate the L_1-norm, and none of them is dominated by the L_∞-norm. Consistency of an estimate f_n in the χ^p-divergence for $1/p + 1/q = 1$ implies consistency of estimates $\int T f_n \, \mathrm{d}\nu$ of the expectations $\int T f \, \mathrm{d}\nu$ for all statistics $T\colon R \to R$ satisfying the condition

$$\sup_{f \in \mathbb{F}} \int |T|^q f \, \mathrm{d}\nu < \infty.$$

Similarly, consistency in the expected χ^2-divergence implies asymptotic unbiasedness of estimates $\int T f_n \, \mathrm{d}\nu$. These conclusions follow from the fact that, by the Hölder inequality, the normalized absolute deviation

$$\frac{|\int T f \, \mathrm{d}\nu - \int T f_* \, \mathrm{d}\nu|}{\left(\int |T|^q f_* \, \mathrm{d}\nu\right)^{1/q}}$$

is bounded above by $\chi^p(f, f_*)^{1/p}$ (c.f. [36]). Thus if in a model $\mathbb{F} \subset \mathbb{F}_\nu$ the moment $\int |x|^{2k} f(x) \, \mathrm{d}\nu(x)$ is bounded on $\mathbb{F}$ for some $k \ge 1$, then for every estimate f_n consistent in the χ^2-divergence and every $m \ge 1$, the statistic

$$\phi_n(\theta_1, \cdots, \theta_n) = \sum_{k=1}^m \theta_k \int \varphi_k f_n \, \mathrm{d}\nu$$

consistently estimates the linear function

$$\phi(\theta_1, \cdots, \theta_n) = \sum_{k=1}^m \theta_k \int \varphi_k f \, \mathrm{d}\nu$$

with polynomials $\varphi_1, \cdots, \varphi_m$ of degree at most k. Estimates of such functions were considered by Barron and Sheu [3]. Note that all previous conclusions also hold with the absolute moments of order q or $2k$ replaced by the corresponding centralized alternatives.

Since

$$I(f, f_*) \le \chi^2(f, f_*) \le \chi^p(f, f_*)^{2/p}, \qquad \text{for all } p > 2$$

consistency in the (expected) χ^p-divergence for any $p \ge 2$ is stronger than the consistency in the (expected) I-divergence. Consistency of the Barron estimator f_n^B in the χ^2-divergence and expected χ^2-divergence was proved in [18] for models satisfying stronger conditions than $\chi^2(f, 1) < \infty$ (c.f. [18, Lemma 4, Appendix]).

Application 4: Another application of density estimators consistent in χ^p-divergence is the call admission control and call admission policing in asynchronous transmission mode (ATM) networks, such as broadband integrated service digital networks (ISDN's) (see De Prycker [15] or Hui [19]). Data rates achieved by the individual user during the transmission period followed a distribution $f \in \mathbb{F} \subset \mathbb{F}_\nu$ (here $f(x) = 0$ for $x \le 0$). Relevant user characteristics include the tail probabilities

$$\int_t^\infty f(x)\, \mathrm{d}\nu(x), \qquad t > 0$$

and the moments

$$\int x^k f(x)\, \mathrm{d}\nu(x), \qquad 1 \le k \le m.$$

(Usually it suffices to consider $m = 2$.) By sampling a user's data rates one can obtain various estimates f_n of his density f. In models $\mathbb{F}$ with centralized or noncentralized moments of order $mp/(p-1)$ bounded for some $p > 1$, each density estimator f_n consistent in χ^p-divergence guarantees the a.s. convergence of both

$$\sup_{t>0} \left| \int_t^\infty f\, \mathrm{d}\nu - \int_t^\infty f_n\, \mathrm{d}\nu \right|$$

and

$$\max_{1 \le k \le m} \left| \int x^k f(x)\, \mathrm{d}\nu(x) - \int x^k f_n(x)\, \mathrm{d}\nu(x) \right|$$

to zero. Estimators consistent in a distance topologically weaker than the χ^p-divergence cannot guarantee this. For instance, the Kolmogorov distance guarantees the convergence of tail probabilities but neither the convergence of expectations nor the convergence of variances.

Application 5: Consider a sequence $\tau = (\tau_n\colon n \in N)$ of tests $\tau_n\colon R^n \to \{0, 1\}$ of the null hypothesis $\mathcal{H}_0{:}f$ against the alternative $H_*{:}f_*$ based on n independent observations, and the corresponding *first-* and *second-kind errors*

$$\alpha_n(\tau) = \int (1 - \tau_n) f^n\, \mathrm{d}\nu^n \quad \text{and} \quad \beta_n(\tau) = \int \tau_n f_*^n\, \mathrm{d}\nu^n.$$

As follows from Csiszár [12, Theorem 1], the Rényi distance $R_a(f, f_*)$ with $0 < a < 1$ is nothing but the supremum of real numbers R_a for which there exists a sequence of tests τ such that for all $R > 0$

$$\alpha_n(\tau) \le e^{-naR} \quad \text{and} \quad \beta_n(\tau) \le e^{-n(1-a)(R_a - R) + o(n)} \qquad \text{as } n \to \infty.$$

Therefore, the Rényi distances represent the cutoff rates in testing one data source against another.

This result of Csiszár means that the data sources f and f_* are exponentially separable (in an obvious statistical sense), and that the Rényi distance $R_a(f, f_*)$ characterizes the rate of this separation.

Let f_n be an estimator of the density f, $c_n \uparrow \infty$ a sequence of positive integers, $\tau = (\tau_n\colon R^{c_n} \to \{0, 1\})$ a sequence of tests of $\mathcal{H}_0{:}f$ against the alternatives $\mathcal{H}_n{:}f_n$ based on c_n independent observations, and $\alpha_n(\tau)$, $\beta_n(\tau)$ the corresponding sequences of errors. We shall consider two extremal situations: entire separability of pairs of the c_n-dimensional product sources $f_n^{c_n}$, f^{c_n} and contiguity of these pairs (see Roussas [32] and Liptser *et al.* [20]).

The sequences $f_n^{c_n}$ and f^{c_n} of product densities are said to be *entirely separable* if there exists a sequence τ of tests under consideration such that

$$\lim_{n\to\infty} (\alpha_n(\tau) + \beta_n(\tau)) = 0 \quad \text{a.s.}$$

The sequences $f_n^{c_n}$ and f^{c_n} are said to be *contiguous* if for any sequence τ of tests under consideration

$$\lim_{n\to\infty} \alpha_n(\tau) = 0 \quad \text{implies} \quad \lim_{n\to\infty} \beta_n(\tau) = 1 \text{ a.s.}$$

and

$$\lim_{n\to\infty} \beta_n(\tau) = 0 \quad \text{implies} \quad \lim_{n\to\infty} \alpha_n(\tau) = 1.$$

We see that in the case of entire separation, the data sources f and f_n can asymptotically be distinguished on the basis of c_n independent observations with the error tending a.s. to zero. In the case of contiguity such a distinction is impossible. It is preferable to use density estimates f_n such that $f_n^{c_n}$ and f^{c_n} are contiguous for c_n increasing as fast as possible. By Liese and Vajda [21, Proposition 7.8], the contiguity takes place if and only if the I_a-divergences satisfy the relation

$$\begin{aligned} \lim_{a\uparrow 1} \limsup_{n\to\infty} c_n (1-a) I_a(f, f_n) &= \lim_{a\downarrow 1} \limsup_{n\to\infty} c_n\, a\, I_a(f, f_n) \\ &= 0 \text{ a.s.} \end{aligned}$$

It is seen from here that the desired contiguity is hardly possible without the consistency of f_n in the I_a-divergence for all $0 < a < 1$. On the other hand, the contiguity takes place if f_n is order c_n consistent in the I_a-divergence (i.e., $\lim_{n\to\infty} c_n I_a(f, f_n) = 0$ a.s.) for each $0 < a < 1$. Thus the new approach to optimality of density estimates formulated in this application leads to consistency in a subfamily of generalized information divergences.

Another application where accuracy of density estimates f_n is measured by some divergence $D_\phi(f, f_n)$ can be found in the extensive literature on classification, pattern recognition, and neural networks, see the first two chapters in Devroye *et al.* [14] and references therein.

An additional argument in favor of ϕ-divergences follows from Pardo and Vajda [27]. Namely, ϕ-divergences are the only distances that satisfy the information processing theorem of information theory (c.f., Cover and Thomas [9]), in the sense that they are invariant with respect to all information-preserving transformations of data. Distances $D(f, f_*)$ that lack this invariance property (e.g., the L_p-norms for $p > 1$, and the Kolmogorov distance) can be dramatically changed by 1–1 recodings of data spaces.

IV. The Results

In this section we consider models $\mathbb{F} \subset \mathbb{F}_\nu$ where the dominating ν is a probability measure. We also consider special models $\mathbb{F} = \mathbb{F}_{(k)}$ defined by the formulas

$$\mathbb{F}_{(1)} = \{f \in \mathbb{F}_\nu\colon f \log f \in L_1(\nu)\} \tag{9}$$

$$\mathbb{F}_{(2)} = \{f \in \mathbb{F}_\nu\colon \log f \in L_\beta(\nu)\} \cap L_\alpha(\nu) \tag{10}$$

$$\mathbb{F}_{(3)} = \{f \in \mathbb{F}_\nu\colon 1/f \in L_\beta(\nu)\} \cap L_{2\alpha}(\nu) \tag{11}$$

where $1 < \alpha, \beta \le \infty$ are constants satisfying the relation

$$\frac{1}{\alpha} + \frac{1}{\beta} < 1 \tag{12}$$

where we put $1/\infty = 0$. Moreover, the definition of $\mathbb{F}_{(2)}$ is extended also to $\alpha = \infty$, $\beta = 1$, and $\alpha = 1$, $\beta = \infty$. These added pairs satisfy (12) with $<$ replaced by equality.

The inequalities $t \log t \ge -1/e$ (valid for $0 < t \le 1$) and $\log t \le t - 1$ (valid for $t > 0$) imply

$$|f \log f| \le \frac{2}{e} + f \log f \quad \text{and} \quad |\log f| \le -\log f + 2f$$

for all $f \in \mathbb{F}_\nu$. Therefore, the conditions $f \log f \in L_1(\nu)$ and $\log f \in L_1(\nu)$ are, respectively, equivalent to

$$I_1(f, 1) \equiv \int f \log f \, \mathrm{d}\nu < \infty$$

and

$$I_0(f, 1) \equiv -\int \log f \, \mathrm{d}\nu < \infty.$$

Similarly, for all $a > 1$ and $t > 0$

$$\log t \le \frac{t^{a-1} - 1}{a - 1} < \frac{t^{a-1}}{a-1}$$

and for all $t > 0$

$$\log t \ge 1 - \frac{1}{t}.$$

Using the first inequality with $a = \alpha$, we obtain $f \in L_\alpha(\nu) \implies f \log f \in L_1(\nu)$, i.e., $\mathbb{F}_{(1)} \supset L_\alpha(\nu) \supset \mathbb{F}_{(2)}$ for any $1 < \alpha < \infty$ and $1 \le \beta \le \infty$. Further, taking the first inequality with $a = 1 + \alpha/\beta$ and combining it with the second inequality, we obtain

$$|\log f| \le \frac{1_{(0,1)}(f)}{f} + \frac{1_{[1,\infty)}(f)\, f^{\alpha/\beta}}{\alpha/\beta}.$$

Hence, for any $1 \le \alpha, \beta < \infty$, the assumptions $f \in L_\alpha(\nu)$, $1/f \in L_\beta(\nu)$ imply $\log f \in L_\beta(\nu)$. Therefore, $\mathbb{F}_{(2)} \supset \mathbb{F}_{(3)}$. It is easy to see that if $\beta = \infty$ this inclusion remains valid for all $1 \le \alpha \le \infty$ and if $\alpha = \infty$, for all $1 \le \beta \le \infty$. Thus we have proved the following result.

Proposition 1: For any pair $1 < \alpha, \beta \le \infty$ with the property (12), the models (9)–(11) satisfy the relations

$$\mathbb{F}_\nu \supset \mathbb{F}_{(1)} \supset \mathbb{F}_{(2)} \supset \mathbb{F}_{(3)}.$$

The restricted relations

$$\mathbb{F}_\nu \supset \mathbb{F}_{(1)} \supset \mathbb{F}_{(2)}$$

remain valid also for $\mathbb{F}_{(2)}$ with $(\alpha, \beta) = (1, \infty)$ and $(\alpha, \beta) = (\infty, 1)$.

In the following theorems we consider the Barron estimator f_n^B defined by (5) for probability measures ν and partitions $\mathcal{P}_n$ satisfying the condition

$$\lim_{n\to\infty} E_\nu(f|\mathcal{P}_n) = f \quad \nu\text{–a.s. for all } f \in \mathbb{F}_\nu. \tag{13}$$

Here $E_\nu(f|\mathcal{P}_n)$ denotes the conditional ν-expectation of the density f on the $\mathcal{P}_n$-generated subfield of $\mathcal{B}$, i.e., a function constant on the sets of $\mathcal{P}_n$, with values given by the formula

$$E_\nu(f|\mathcal{P}_n) = \frac{\int_A f \, \mathrm{d}\nu}{\nu(A)} = \frac{1}{h_n} \int_A f \, \mathrm{d}\nu, \qquad \text{for all } x \in A \in \mathcal{P}_n. \tag{14}$$

(Notice that the expectation E_ν with respect to the distribution ν differs from the expectation E with respect to the distribution defined by the product density $f(x_1) \cdots f(x_n)$ which is also used in the paper.) Since $E_\nu(f|\mathcal{P}_n)$ is a piecewise-constant approximation of f by mean values taken with respect to ν, arbitrary partitions of R satisfying (13) have been called *ν-approximating* in [4] and [18]. Applying (13) to the positive and negative part of any $f \in L_1(\nu)$ one easily obtains that (13) is equivalent to

$$\lim_{n\to\infty} E_\nu(f|\mathcal{P}_n) = f \quad \nu\text{–a.s. for all } f \in L_1(\nu).$$

The partitions $\mathcal{P}_n$ considered in (5) are for a given ν specified by a slowly increasing sequence of integers m_n. According to the next Proposition, if $m_n = 2^{c_n}$ for a sequence of positive integers c_n, (13) holds for every probability measure ν. Replacing m_n by

$$m_n^* = 2^{[\log_2 m_n]}$$

we obtain modified partitions $\mathcal{P}_n^*$ differing from $\mathcal{P}_n$ just by slightly reduced cardinality, but with the guaranteed ν-approximating property for any ν.

Note that $m_n = 2^{c_n}$ satisfies the relation $n/m_n \to \infty$ if and only if

$$\limsup_{n\to\infty} \frac{c_n}{\log_2 n} < 1.$$

If $n/m_n \to \infty$ then this condition holds for $c_n = [\log_2 m_n]$ so that also $n/m_n^* \to 0$.

The binary exponential specification of m_n also simplifies the numerical calculations involving f_n^B. Indeed, this estimate can then be evaluated iteratively for all $1 \le n \le n_{\max}$ by storing in memory the frequency counts $n\,\mu_n^E(A)$ for all intervals $A \in \mathcal{P}_{n_{\max}}$. Then $(n+1)\,\mu_{n+1}^E$ differs from $n\,\mu_n^E$ on just one interval $A \in \mathcal{P}_{n_{\max}}$ covering the new observation X_{n+1}. Addressing the memory cells by an appropriate c_n-bit binary code, one can easily calculate the values $f_n^B(x)$ for all

$1 \le n \le n_{\max}$ and all $x \in R$, as well as the values of linear functionals

$$\int w(x)\, f_n^B(x)\, \mathrm{d}\nu(x) = \sum_{A\in\mathcal{P}_n} \frac{n\,\mu_n^E(A)+1}{n\,h_n+1} \int_A w(x)\, \mathrm{d}G(x)$$

of the corresponding distribution estimate μ_n^B. Practically, the only task is to recover the values $\mu_n^E(A)$ for $A \in \mathcal{P}$ by summing up the contents of all memory cells corresponding to the intervals of $\mathcal{P}_{n_{\max}}$ contained in A.

Proposition 2: If $m_n = b^{c_n}$ for some integers $b > 1$ and $c_n \uparrow \infty$ then every probability measure ν and the corresponding partitions $\mathcal{P}_n$ satisfy the relation (13).

Proof of this statement and the main results below are presented in Section V.

Theorem 1: Let ϕ be a function satisfying the assumptions of the Definition with $\phi(0)$ finite. Then f_n^B is consistent ϕ-divergence and expected ϕ-divergence

i) in the model $\mathbb{F} = \mathbb{F}_\nu$ if $\phi(\infty)/\infty < \infty$, and

ii) in the model $\mathbb{F} = \mathbb{F}_{(1)}$ if $\phi(t) = O(t\log t)$ for $t \to \infty$.

All I_a-divergences of order $0 < a < 1$ satisfy the condition in i). This condition is satisfied also by $\phi(t) = |t-1|$ (leading to the L_1-norm $\|f - f_*\|$), by all $\phi_a(t)$ considered in (8) with $0 < a < 1$, by

$$\phi_a(t) = |t^a + 1|^{1/a}\, 2^{(1-a)/a}(t+1), \qquad a > 1$$

(leading to the metrics $D_{\phi_a}(f, f^*)^{1/2}$ on $\mathbb{F}_\nu$, see Österreicher [25]), by

$$\phi_a(t) = \frac{1}{a + (1-a)\,t}, \qquad 0 < a < 1$$

(leading to divergences with an interesting statistical application in Rukhin [33]), and by

$$\phi_p(t) = |t^{1/p} - 1|^p, \qquad p \ge 1$$

(leading to the L_p-norms $\|f^{1/p} - f_*^{1/p}\|_p$ known as Matusita distances, see Matusita [23]).

The conditions in ii) and $\phi(0) < \infty$ hold, e.g., for the divergences of Lin [22] defined by the convex functions

$$\phi_a(t) = a\,t \log \frac{t}{a\,t + 1 - a} - (1-a)\log(a\,t + 1 - a), \qquad 0 < a < 1$$

and for I_a-divergences of order $0 < a \le 1$.

Theorem 2: Let $\mathbb{F} = \mathbb{F}_{(2)}$. The estimator f_n^B is consistent in the expected ϕ-divergence for all ϕ considered in the Definition with $t\,\phi(1/t) + \phi(t) = O(t\log t)$ when $t \to \infty$.

The condition of Theorem 2 holds for all I_a-divergences of order $0 \le a \le 1$.

Theorem 3: Let $\mathbb{F} = \mathbb{F}_{(3)}$. The estimator f_n^B is consistent in the expected ϕ-divergence for all ϕ considered in the Definition with $t\,\phi(1/t) + \phi(t) = O(t^2)$ when $t \to \infty$.

The condition of this theorem holds for all I_a-divergences of order $-1 \le a \le 2$, and for χ^p-divergences with $1 \le p \le 2$. These functions can be linearly combined (using positive coefficients) with the ϕ-functions satisfying the condition of i) or ii) in Theorem 1. Indeed, the conditions i) and ii) are stronger than that of Theorem 3, and each positive linear combination of functions ϕ satisfying the assumptions of Definition satisfies these assumptions too. A similar remark applies, of course, also to the examples satisfying Theorem 2.

The conditions imposed on ϕ-divergences in Theorems 1–3 obviously do not hold for χ^p-divergences with $p > 2$ and I_a-divergences with $a \notin [-1, 2]$.

V. Proofs of the Results

Proof of Proposition 2: Let G be the continuous distribution function of ν

$$G^{-1}(t) = \inf\{x \in R\colon G(x) \ge t\}$$

and S the set of all $x \in R$ such that

$$G(y) < G(x), \qquad \text{for all } y < x.$$

Obviously, the complement R–S is a union of disjoint intervals (open on the left) on which the function G takes on different constant values. The set S is thus measurable and $\nu(S) = 1$.

Let us now consider the subfields $\mathcal{B}_1, \mathcal{B}_2, \cdots$ of the Borel field $\mathcal{B}$ generated by $\mathcal{P}_n = G^{-1}(\Pi_n)$, where Π_n is the partition of (0, 1) into m_n disjoint intervals defined by the equidistant points $0, 1/m_n, \cdots, (m_n - 1)/m_n, 1$. The assumed exponential form of m_n guarantees that these subfields are nested in the sense that $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots$. If $\mathcal{B}_* \subset B$ is the subfield generated by the union $\mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots$ then the well-known Lévy martingale convergence theorem implies that the conditional expectations $f_* = E_\nu(f|\mathcal{B}_*)$ and $E_\nu(f|\mathcal{P}_n) = E_\nu(f|\mathcal{B}_n)$ satisfy the asymptotic relation

$$\lim_{n\to\infty} E_\nu(f|\mathcal{P}_n) = f_* \quad \nu\text{–a.s. for all } f \in \mathbb{F}_\nu.$$

We shall prove that $f_* = f$ ν–a.s.

It follows from the exponential form of m_n that the subsets

$$C_n = \{1/m, 2/m_n, \cdots, (m_n - 1)/m_n\} \subset (0, 1)$$

and

$$D_n = G^{-1}(C_n) \subset R$$

are nested in the sense that $C_1 \subset C_2 \subset \cdots$, and $D_1 \subset D_2 \subset \cdots$. The union $D_1 \cup D_2 \cup \cdots$ is dense in S. Indeed, for any $x \in S$ and $\varepsilon > 0$, there exists $j/m_n \in C_n$ with the property

$$G(x - \varepsilon) < \frac{j}{m_n} < G(x)$$

so that, by the assumed strict monotonicity of G, the point $a_{nj} = G^{-1}(j/m_n) \in D_n$ satisfies the relation

$$x - \varepsilon \le a_{nj} < x.$$

By the theorem proved in Abou–Jaude [1, pp. 216–219] (c.f. also [16, Theorem 5]), it follows that

$$\lim_{n\to\infty} \|E_\nu(f|\mathcal{P}_n) - f\|_1 = 0.$$

As is well known, this implies ν–a.s. convergence of subsequences of $E_\nu(f|\mathcal{P}_n)$ to f. But in view of the martingale convergence established above, this means that $f = f_*$ ν–a.s. which completes the proof.

Proofs of Theorems 1–3 are based on a chain of lemmas. The first lemma is proved in Theorem 1 and [4, eq. (2.10)].

Lemma 1: Let $\mathbb{F} = \mathbb{F}_\nu$. The estimator f_n^B is consistent in the L_1-norm and expected L_1-norm.

The following result follows from [4, Theorem 2].

Lemma 2: Let $\mathbb{F} = \mathbb{F}_{(1)}$. The estimator f_n^B is consistent in the I_1-divergence and expected I_1-divergence.

The consistency of f_n^H established in the next lemma was proved for the particular model $\mathbb{F}_{(2)}$ with $\alpha = \infty$ and $\beta = 1$ in Barron *et al.* [4, Theorem 5]. The present stronger and wider result is thus of its own interest in information theory, with similar applications as the mentioned Theorem 5.

Lemma 3: Let $\mathbb{F} = \mathbb{F}_{(2)}$. The estimators f_n^B and f_n^H are consistent in the expected I_0-divergence.

Proof: By (6) and by the Jensen inequality applied to the convex function $\phi(x) = x \log x$

$$\begin{aligned}I_0(f, f_n^B) &= \int ((1-a_n) f_n^H + a_n) \log \frac{((1-a_n) f_n^H + a_n)}{f} \,\mathrm{d}\nu \\ &\le \int \big[(1-a_n) f_n^H \log f_n^H + a_n 1 \cdot \log 1 \\ &\quad + ((1-a_n) f_n^H + a_n) \log 1/f\big] \,\mathrm{d}\nu \\ &= (1-a_n) I_0(f, f_n^H) + a_n I_0(f, 1)\end{aligned}$$

where, by the assumption $f \in \mathbb{F}_{(1)}$, $I_0(f, 1)$ is finite.

Denoting for simplicity the conditional expectation (14) by z_n we obtain

$$\begin{aligned}I_0(f, f_n^H) &= \int f_n^H \log \frac{f_n^H}{z_n} \,\mathrm{d}\nu + \int f_n^H \log \frac{z_n}{f} \,\mathrm{d}\nu \\ &\triangleq I_n + Z_n\end{aligned}$$

where, by [4, Theorem 3], $E I_n$ tends to zero. Thus it remains to prove that $E Z_n$ tends to zero. By (2)

$$\begin{aligned}Z_n &= \sum_{A \in \mathcal{P}_n} \int_A \frac{\mu_n(A)}{\nu(A)} \log \frac{z_n}{f} \,\mathrm{d}\nu \\ &= \sum_{A \in \mathcal{P}_n} \int_A \frac{\mu_n(A)}{\mu(A)} z_n \log \frac{z_n}{f} \,\mathrm{d}\nu\end{aligned}$$

so the obvious relation $E \mu_n(A) = \mu(A)$ implies

$$E Z_n = \int z_n \log \frac{z_n}{f} \,\mathrm{d}\nu.$$

Since $z_n \in \mathbb{F}_\nu$, we have

$$E Z_n = I_1(z_n, f) = I_1(z_n, 1) - E_\nu(z_n \log f) \ge 0.$$

By the monotonicity of I_1-divergence (c.f. [21] or [37])

$$I_1(z_n, 1) = I_1(E_\nu(f|\mathcal{P}_n), E_\nu(1|\mathcal{P}_n)) \le I_1(f, 1) < \infty.$$

Therefore, if we prove

$$\lim_{n\to\infty} E_\nu(z_n \log f) = I_1(f, 1) \tag{15}$$

we have established the desired consistency. We see from (13) that (15) holds when the limit and the expectation can be interchanged. If $\alpha = \infty$ or $\beta = \infty$ in (10), this interchange is justified for $f \in \mathbb{F}_{(2)}$ by the Lebesgue bounded convergence theorem. Indeed, if $f \in L_\infty(\nu)$ and $\log f \in L_\beta(\nu)$ for some $\beta \ge 1$, then $|z_n \log f| \le \|f\|_\infty \,|\log f|$ ν–a.s., where $|\log f|$ is ν-integrable. If $\log f \in L_\infty(\nu)$ then f is ν–a.s. bounded below and above by positive constants. Therefore, $z_n \log f$ is absolutely ν–a.s. bounded. It remains to investigate the case $\alpha + \beta < \infty$. Here the Hölder inequality implies that

$$E_\nu |z_n \log f|^c \le (E_\nu\, z_n^{cp})^{1/p} (E_\nu |\log f|^{cq})^{1/q}$$

for all $c > 1$ and $1 < p,\ q < \infty$ conjugated in the usual sense. Applying Jensen's inequality in the convex function $\phi(t) = t^{cp}$ and the conditional ν-expectation, we obtain $z_n^{cp} \le E_\nu(f^{cp}|\mathcal{P}_n)$. Therefore,

$$E_\nu\, z_n^{cp} \le E_\nu(E_\nu(f^{cp}|\mathcal{P}_n)) = E_\nu\, f^{cp}.$$

Choosing

$$p = \frac{\alpha+\beta}{\beta}, \quad q = \frac{\alpha+\beta}{\alpha}, \quad \text{and} \quad c = \frac{\alpha\beta}{\alpha+\beta}$$

we obtain

$$\sup_n E_\nu |z_n \log f|^c < \infty$$

for any $f \in \mathbb{F}_{(2)}$. Since $1/c = 1/\alpha + 1/\beta$, the strict inequality in (12) implies that $c > 1$. This implies the uniform ν-integrability of the sequence $(z_n \log f)$, and therefore the commutativity of $\lim$ and E_ν in (15). This proves (15) and thus completes the proof of consistency in the expected I_0-divergence.

Since $I_2(f, f_*) = \chi^2(f, f_*)$ is the χ^2-divergence, the following result follows from Proposition 2 and [18, Theorem 1].

Lemma 4: Let $\mathbb{F} = \mathbb{F}_{(3)}$. The estimator f_n^B is consistent in the expected I_2-divergence.

Since $I_{-1}(f, f_*) = \chi^2(f_*, f)$ is the reversed χ^2-divergence, the following lemma is complementary to the previously mentioned result of [18].

Lemma 5: Let $\mathbb{F} = \mathbb{F}_{(3)}$. The estimator f_n^B is consistent in the expected I_{-1}-divergence.

Proof: The proof is similar to that of Lemma 3. Jensen's inequality implies that

$$\begin{aligned}I_{-1}(f, f_n^B) + 1 &= \int \frac{((1-a_n) f_n^H + a_n)^2}{f} \,\mathrm{d}\nu \\ &\le (1-a_n) \int \frac{(f_n^H)^2}{f} \,\mathrm{d}\nu + a_n (I_{-1}(f, 1) + 1)\end{aligned}$$

where $I_{-1}(f, 1) < \infty$ if $f \in \mathbb{F}_{(3)}$. Thus it suffices to consider the random variables

$$Z_n = \int \frac{(f_n^H)^2}{f} \,\mathrm{d}\nu.$$

By definitions (2) and (3)

$$Z_n = \frac{1}{(n h_n)^2} \sum_{A \in \mathcal{P}_n} \int_A \frac{1}{f} \sum_{i,k=1}^n 1_A(X_i)\, 1_A(X_k) \,\mathrm{d}\nu$$

so that

$$\begin{aligned} E\,Z_n &= \frac{1}{(n\,h_n)^2} \sum_{A\in\mathcal{P}_n} \int_A \frac{1}{f} \\ &\quad \cdot \left[n\int_A f\,\mathrm{d}\nu + n(n-1)\left(\int_A f\,\mathrm{d}\nu\right)^2\right] \mathrm{d}\nu \\ &= \sum_{A\in\mathcal{P}_n} \int_A \left(\frac{z_n}{n\,h_n\,f} + \frac{n(n-1)\,z_n^2}{n^2\,f}\right) \mathrm{d}\nu \\ &= \frac{1}{n\,h_n}\int \frac{z_n}{f}\,\mathrm{d}\nu + \left(1-\frac{1}{n}\right)\int \frac{z_n^2}{f}\,\mathrm{d}\nu. \end{aligned}$$

By the Cauchy–Schwarz inequality

$$\int \frac{z_n}{f}\,\mathrm{d}\nu \le \left(\int \frac{z_n^2}{f}\,\mathrm{d}\nu\,(I_{-1}(f,\,1)+1)\right)^{1/2}$$

so that the relation

$$\lim_{n\to\infty} \int \frac{z_n^2}{f}\,\mathrm{d}\nu = \int \lim_{n\to\infty} \frac{z_n^2}{f}\,\mathrm{d}\nu = 1 \qquad (16)$$

would imply the desired consistency. Obviously, (16) holds if the sequence (z_n^2/f) is uniformly integrable. For $\alpha+\beta<\infty$ we can consider the same p, q, and c as in the proof of Lemma 3. We obtain a similar result

$$\begin{aligned} E_\nu(z_n^2/f)^c &\le (E_\nu\,f^{2cp})^{1/p}\,(E_\nu(1/f)^{cq})^{1/q} \\ &= (E_\nu\,f^{2\alpha})^{\beta/(\alpha+\beta)}\,(E_\nu(1/f)^{\beta})^{\alpha/(\alpha+\beta)}. \end{aligned}$$

Hence by the definition of $\mathbb{F}_{(3)}$

$$\sup_n E_\nu(z_n^2/f)^c < \infty.$$

Since $c>1$, this implies the desired uniform integrability, and thus the validity of (16). If $\alpha=\infty$ then (12) implies $\beta>1$ and if $\beta=\infty$ then $\alpha>1$. The modification of the previous procedure in the case $\alpha+\beta=\infty$ is thus obvious.

Lemma 6: If ϕ satisfies the conditions of Theorem 1, part i), then there exists $c>0$ such that

$$c|t-1| \ge \phi(t), \qquad \text{for all } t>0.$$

Proof: Since ϕ is convex

$$\psi(t) = \frac{\phi(t)-\phi(1)}{|t-1|} = \frac{\phi(t)}{|t-1|}$$

is nonincreasing in the domain $0<t<1$ and nondecreasing in the domain $t>1$. Therefore, $\psi(t)$ is bounded above by

$$c \triangleq \max\left\{\lim_{t\downarrow 0}\psi(t),\ \lim_{t\to\infty}\psi(t)\right\} = \max\{\phi(0),\,\phi(\infty)/\infty\}$$

which is by assumption finite and positive. □

Lemma 7: If ϕ satisfies the conditions of Theorem 1, part ii), then there exist positive c_1 and c_2 such that

$$c_1|t-1| + c_2\,t\,\log\,t \ge \phi(t), \qquad \text{for all } t>0.$$

Proof: In view of Lemma 6 we may restrict ourselves to ϕ's with $\phi(\infty)/\infty=\infty$. Using the same argument as in the previous proof

$$\phi(t) \le \phi(0)\,(1-t), \qquad \text{for all } 0<t\le 1. \qquad (17)$$

Further, $\psi(t)=t\,\log\,t$ is convex on the interval $(0,\,1]$ with $\psi(1)=0$ and with the derivative $\psi'(1)=1$. This implies

$$\psi(x)+1-t \ge 0, \qquad \text{for all } 0<t\le 1. \qquad (18)$$

Consequently, if $c_2>0$ and $c_1=\phi(0)+c_2>0$ then

$$c_1(1-t)+c_2\,t\log\,t \ge \phi(t), \qquad \text{for all } 0<t\le 1.$$

Thus it suffices to prove that the assumptions $\phi(t) = O(t\,\log\,t)$ when $t\to\infty$ and $\phi(\infty)/\infty=\infty$ imply the existence of $c_2>0$ such that

$$c_2\,t\,\log\,t \ge \phi(t), \qquad \text{for all } t>1. \qquad (19)$$

By the assumption, there exist $a>0$ and $b>1$ such that $\phi(b)/(b-1)\ge 1$ and $\phi(t)\le at\,\log\,t$ for all $t\ge b$. We shall prove that (19) holds for

$$c_2 = \frac{\phi(b)}{b-1}\max\{a,\,1\}.$$

The convexity of functions ϕ and $\psi(t)=t\,\log\,t$ implies for all $1<t<b$

$$\phi(t) \le \frac{b-t}{b-1}\,\phi(1) + \frac{t-1}{b-1}\,\phi(b) = \frac{t-1}{b-1}\,\phi(b)$$

and

$$\frac{\psi(t)}{t-1} = \frac{\psi(t)-\psi(1)}{t-1} \ge \psi'(1) = 1.$$

Therefore,

$$\frac{\phi(b)}{b-1}\,\psi(t) = \frac{t-1}{b-1}\,\phi(b)\,\frac{\psi(t)}{t-1} \ge \phi(t)\cdot 1$$

i.e., (19) with the above considered c_2 holds for $1<t<b$. If $t\ge b$ then

$$a\,\frac{\phi(b)}{b-1}\,\psi(t) \ge \frac{\phi(b)}{b-1}\,\phi(t) \ge 1\cdot\phi(t)$$

and we see that the previous conclusion remains valid. □

Lemma 8: If ϕ satisfies the condition of Theorem 2 then there exist positive c_1, c_2, and c_3 such that

$$c_1|t-1| + c_2\log\frac{1}{t} + c_3\,t\,\log\,t \ge \phi(t), \qquad \text{for all } t>0.$$

Proof: Let ϕ satisfy the condition of Theorem 3. Replacing $\phi(t)$ in the domain $0<t\le 1$ by $\phi(1)-\phi'_+(1)\,(t-1)$ one obtains a modification of ϕ satisfying the conditions of Theorem 1, part ii). By (19) there exists $c_3>0$ such that

$$c_3t\,\log\,t \ge \phi(t), \qquad \text{for all } t\ge 1$$

and, by (18), all $c_1\ge c_3$ satisfy the relation

$$c_1(1-t)+c_3\,t\,\log\,t \ge 0, \qquad \text{for all } 0<t<1.$$

Thus is suffices to prove the existence of $c_2 > 0$ such that

$$c_2 \log \frac{1}{t} \geq \phi(t), \qquad \text{for all } 0 < t < 1. \tag{20}$$

By substituting $y = 1/t$ in (20) and multiplying both sides of the inequality by $y > 0$ one obtains that (20) is equivalent to

$$c_2 \, y \log y \geq y \, \phi\left(\frac{1}{y}\right), \qquad \text{for all } \; y > 1.$$

It is easy to see that $\tilde{\phi}(y) = y \, \phi(1/y)$ possesses the properties assumed in (7) over the whole domain $y > 0$. Moreover, the condition of Theorem 3 concerning ϕ implies that $\tilde{\phi}(y)$ satisfies the assumptions of Theorem 1, part ii) over the domain $y > 1$. Thus (20) follows from (19) which has already been proved above. □

Lemma 9: If ϕ satisfies the condition of Theorem 3, then there exist positive c_1, c_2, and c_3 such that

$$c_1 |t - 1| + c_2 \left(\frac{1}{t} - 1\right) + c_3 (t^2 - 1) \geq \phi(t), \qquad \text{for all } t > 0.$$

Proof: Let ϕ satisfy the condition of Theorem 3. It suffices to prove that there exist positive c_1, c_2, c_3 such that

$$c_1 |t - 1| + c_2 \frac{(t-1)^2}{t} + c_3 (t-1)^2 \geq \phi(t), \qquad \text{for all } t > 0.$$

Indeed, by taking c_3 large enough and putting $c_2 = 2c_3$, one obtains the relation of Lemma 9.

1) If $\phi(0)$ is finite then (17) holds. If $c_2 > 0$ and $c_1 = \phi(0) + c_2 > 0$ then

$$c_1(1 - t) + c_2 \frac{(t-1)^2}{t} \geq \phi(t), \qquad \text{for all } \; 0 < t \leq 1.$$

Thus it suffices to prove the existence of $c_3 > 0$ such that

$$c_3 (t-1)^2 \geq \phi(t), \qquad \text{for all } t > 1. \tag{21}$$

Proof of this relation for ϕ considered in (7) with $\phi(t) = O(t^2)$ when $t \to \infty$ follows the lines of proof of (19) and is thus omitted here.

2) If $\phi(0) = \infty$ then the validity of (21) remains unaffected, and it suffices to prove the existence of $c_2 > 0$ such that

$$c_2 \frac{(t-1)^2}{t} \geq \phi(t), \qquad \text{for all } 0 < t < 1. \tag{22}$$

The proof of (22) is similar to that of (20) above. Namely, by substituting $y = 1/t$ in (22) and multiplying both sides of the inequality by $y > 0$, (22) is seen to be equivalent to

$$c_2 (y - 1)^2 \geq \tilde{\phi}(y), \qquad \text{for all } y > 1$$

where $\tilde{\phi}(y) = y \, \phi(1/y)$ possesses the properties assumed in (7) over the whole domain $y > 0$. Moreover, the condition of Theorem 3 concerning ϕ implies that $\tilde{\phi}(y) = O(y^2)$ when $y \to \infty$. Thus the existence of $c_2 > 0$ satisfying (22) is equivalent to the existence of $c_3 > 0$ satisfying (21). This completes the proof. □

Proofs of Theorems 1–3: Since the ϕ-divergence is nonnegative, part i) of Theorem 1 follows from Lemmas 1 and 6 part ii) from Lemmas 2 and 7. Similarly, Theorem 2 follows from Lemmas 2, 3, and 8, and Theorem 3 from Lemmas 4, 5, and 9.

VI. Conclusions

Asymptotic accuracy of the Barron nonparametric density estimator f_n^B introduced in [2] for models dominated by a probability measure ν on R has been established in the sense that the ϕ-divergence $D_\phi(f, f_n^B)$ and/or the expected ϕ-divergence $E D_\phi(f, f_n^B)$ between the true and estimated model tend to zero. The conditions for the consistency in expected ϕ-divergence are

$$\begin{aligned} \phi(t) &= O(t^{-1}), && \text{when } t \downarrow 0 \\ \phi(t) &= O(t^2), && \text{when } t \to \infty \end{aligned}$$

and

$$\int f^{2\alpha} \, d\nu < \infty \qquad \int f^{-\beta} \, d\nu < \infty$$

for some $1/\alpha + 1/\beta < 1$. In particular, this implies consistency in the reversed χ^2-divergence defined by $\phi(t) = (t-1)^2/t$.

For functions satisfying the stronger restrictions

$$\begin{aligned} \phi(t) &= O(\log t^{-1}), && \text{when } t \downarrow 0 \\ \phi(t) &= O(t \log t), && \text{when } t \to \infty \end{aligned}$$

the condition for the above considered consistency is weaker, namely,

$$\int f^{\alpha} \, d\nu < \infty \quad \text{and} \quad \int |\log f|^{\beta} \, d\nu < \infty.$$

This fact leads to a new result about consistency in the expected reversed I-divergence defined by $\phi(t) = \log t^{-1}$.

If $\phi(t) = O(1)$ when $t \downarrow 0$ and/or $\phi(t) = O(t)$ when $t \to \infty$ then the restrictions on f are even weaker, and f is consistent for all models dominated by ν.

For purely atomic probability measures ν there exist density estimators f_n consistent in all ϕ-divergences for any $f \in \mathbb{F}_\nu$. The Barron estimators f_n^B exist in the sense specified in this paper if and only if ν is nonatomic. By Theorem 1, if ν is a nonatomic probability measure on R, then f_n^B is consistent in infinitely many ϕ-divergences and expected ϕ-divergences for all $f \in \mathbb{F}_\nu$. It is however not clear, whether there exist nonatomic probability measures ν on R with f_n^B consistent in all ϕ-divergences or expected ϕ-divergences for at least one $f \in \mathbb{F}_\nu$. The existence of such pairs (f, ν), or characterization of convex function ϕ such that for a given nonatomic ν the corresponding estimators f_n^B are inconsistent in the expected ϕ-divergence for all $f \in \mathbb{F}_\nu$, are interesting open problems.

References

[1] S. Abou-Jaoude, “Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l’histogramme pour une densité,” *Ann. l’Inst. Henri Poincaré*, vol. 12, pp. 213–231.

[2] A. R. Barron, “The convergence in information of probability density estimators,” in *IEEE Int. Symp. Information Theory* (Kobe, Japan, June 19–24, 1988).

[3] A. R. Barron and C. H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, pp. 1347–1369, 1991.

[4] A. R. Barron, L. Györfi, and E. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1437–1454, 1992.

[5] A. Berlinet, L. Devroye, and L. Györfi, "Asymptotic normality of L_1-error in density estimation," *Statistics*, vol. 26, pp. 329–343, 1995.

[6] A. Berlinet, L. Györfi, and E. van der Meulen, "Asymptotic normality of relative entropy in multivariate density estimation," *Publ. l'Inst. Statist. l'Univ. Paris*, vol. 41, pp. 3–27, 1997.

[7] P. J. Bickel and M. Rosenblatt, "On some global measures of the deviations of density function estimates," *Ann. Statist.*, vol. 1, pp. 1071–1095, 1973.

[8] B. S. Clarke and A. R. Barron, "Information–theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, 1990.

[9] T. M. Cover and J. B. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[10] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitât on Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, ser. A, vol. 8, pp. 84–108, 1963.

[11] ——, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[12] ——, "Generalized cutoff rates and Rényi's information measures, *IEEE Trans. Inform. Theory*, vol. 41, pp. 26–34, Jan. 1995.

[13] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, 1973.

[14] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer-Verlag, 1996.

[15] M. de Prycker, *Asynchronous Transfer Mode Solution for Broadband ISDN.* London, U.K.: Ellis Harwood, 1991.

[16] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1-View.* New York: Wiley, 1985.

[17] L. Györfi, I. Vajda, and E. van der Meulen, "Minimum Kolmogorov distance estimates of parameters and parametrized distributions," *Metrika*, vol. 42, pp. 237–255, 1995.

[18] L. Györfi, F. Liese, I. Vajda, and E. van der Meulen, "Distribution estimates consistent in χ^2-divergence," *Statistics*, vol. 31, 1998, in print.

[19] J. Hui, *Switching and Traffic Theory for Integrated Broadband Networks in Telecommunications.* Boston, MA: Kluwer, 1990.

[20] R. S. Liptser, F. Pukelheim, and A. N. Shiryayev, "On necessary and sufficient conditions for contiguity and entire separation of probability measures," *Uspekhi Matemat. Nauk*, vol. 37, pp. 97–124, 1982.

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

[22] J. Lin, "Divergence measures based on Shannon entropy," *IEEE Trans. Inform. Theory*, vol. 37, pp. 145–151, 1991.

[23] K. Matusita, "Distances and decision rules," *Ann. Inst. Statist. Math.*, vol. 16, pp. 305–320, 1964.

[24] M. L. Menéndez, D. Morales, L. Pardo, and I. Vajda, "Asymptotic distributions of ϕ-divergences of hypothetical and observed frequencies on refined partitions," *Statist. Nederland.*, to be published.

[25] F. Österreicher, "On a class of perimeter-type distances of probability distributions," *Kybernetika*, vol. 32, pp. 389–393, 1996.

[26] F. Österreicher and I. Vajda, "Statistical information and discrimination," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1036–1039, 1993.

[27] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem on information theory," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1288–1293, 1997.

[28] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[29] R. C. Read and N. A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New York: Springer, 1988.

[30] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Probability Theory and Mathematical Statistics* vol. 1, pp. 547–561, 1961.

[31] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 47, pp. 832–837, 1956.

[32] G. G. Roussas, *Contiguity of Probability Measures.* Cambridge, U.K.: Cambridge Univ. Press, 1972.

[33] A. L. Rukhin, "Optimal estimator of the mixture parameter by the method of moments and information affinity," in *Trans. 12th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes* (Prague: Czech Acad. Sci. and Charles Univ., 1994), pp. 214–219.

[34] D. W. Scott, *Multivariate Density Estimation.* New York: Wiley, 1992.

[35] I. Vajda, "On the f-divergence and singularity of probability measures," *Period. Math. Hungar.*, vol. 2, pp. 223–234, 1972.

[36] ——, χ^α-divergence and generalized Fisher's information," in *Trans. 6th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes* Prague, Czechoslovakia: Academia, 1972, pp. 873–886.

[37] ——, *Theory of Statistical Inference and Information.* Boston, MA: Kluwer, 1989.

[38] I. Vajda and V. Kůs, "Adaptive density estimates for ATM networks," Prague, Czech Rep., Inst. Inform. Theory and Automation, Res. Rep. 1893, Dec. 1996.