# CONSISTENCY OF $\phi$-DIVERGENCE ERRORS OF BARRON DENSITY ESTIMATES[1]

**Tomáš Hobza**

*Institute of Information Theory and Automation, Prague, Czech Republic*

*Department of Mathematics, Czech Technical University in Prague, Czech Republic, hobza@km1.fjfi.cvut.cz*

Abstract: Our research is motivated by the need of using nonparametric density estimates for optimization of the switching rules in the nods of telecommunication and computer networks. One of the applications in this area is to estimate tail probabilities of an unknown distribution. Thus, we need estimates as much precise as possible particularly in the tail areas. The Barron estimator, defined as a convex mixture of histogram estimate and some dominating probability distribution, showed up to be a convenient tool for this type of problems.

In this contribution we are interested in the asymptotic properties of the mentioned estimator if its error is measured by $\phi$-divergences with the true density for a general class of functions $\phi$. In quantized models leading to reduced $\phi$-divergences we prove the consistency of this estimator in the reduced $\phi$-divergence and in the expected reduced $\phi$-divergence. Further, we formulate conditions under which it is possible to extend the obtained results also to the original non-quantized models. Our conditions and results are then compared with those of Berlinet et al. (1998).

Keywords: Histogram estimate, Barron density estimate, $\phi$-divergence errors, Divergences of Csiszár, Consistency.

## 1. DEFINITIONS AND BASIC CONCEPTS

We consider on the Borel line $(I\!R, \mathcal{B})$ a probability distribution $Q$ with a Lebesgue density $g(x)$ (the Lebesgue measure will be denoted by $\lambda$). Further, let $P$ be an unknown distribution dominated by $Q$ with a Lebesgue density $f(x)$ and let $\mathcal{P}_n = \{A_{n1}, \ldots, A_{nm_n}\}$ be for $n = 1, \ldots$ interval partitions of $I\!R$ with the property $Q(A_{nj}) = 1/m_n$ for some increasing sequence of positive integers $m_n < n$. This means that we suppose from the beginning that the partitions $\mathcal{P}_n$ are equiprobable with respect to the dominating distribution $Q$. Let $\boldsymbol{p}_n = (p_{nj} \triangleq P(A_{nj}))_{j=1}^{m_n}$ be unknown discrete distributions, $\boldsymbol{q}_n = (q_{nj} \triangleq Q(A_{nj}) = 1/m_n)_{j=1}^{m_n}$, $\boldsymbol{X}_n = (X_{nj})_{j=1}^{m_n} \sim$ Multinomial$(n, \boldsymbol{p}_n)$ sequence of multinomially distributed random vectors with parameters $n, \boldsymbol{p}_n$, and $\widehat{\boldsymbol{p}}_n = (\widehat{p}_{nj} = X_{nj}/n)_{j=1}^{m_n}$ estimates

[1]Supported by the grant MSMTV 1M0572.

of the distributions $\boldsymbol{p}_n$. Further we consider on $\mathbb{R}$ the density $g_n(x) = (p_{nj}/q_{nj})\, g(x)$ if $x \in A_{nj}$ and the corresponding distribution is denoted by $Q_n$.

Now we define two estimates of the density $f(x)$. First of them is the *g-shaped histogram*

$$\widehat{g}_n(x) = \frac{\widehat{p}_{nj}}{q_{nj}}\, g(x) = \frac{m_n}{n}\, X_{nj}\, g(x) \quad \text{if } x \in A_{nj}\,. \tag{1}$$

The *Barron g-shaped histogram* or, simply, the *Barron estimator* is then defined as the convex mixture

$$\widetilde{g}_n(x) = \frac{n}{n+m_n}\, \widehat{g}_n(x) + \frac{m_n}{n+m_n}\, g(x) = \frac{m_n}{n+m_n}\, (X_{nj}+1)\, g(x) \quad \text{if } x \in A_{nj} \tag{2}$$

of the $g$-shaped histogram and the dominating density $g$. The corresponding distributions we denote by $\widehat{Q}_n$ and $\widetilde{Q}_n$. The Barron estimator was defined firstly in (Barron, 1988). It is easy to see that both estimates $\widetilde{g}_n(x)$ and $\widehat{g}_n(x)$ are probability densities, i.e. that they are *bona fide* estimates of the density $f(x)$.

A negative property of the $g$-shaped histogram is that it may not dominate the estimated density $f$. I.e. the set $\{x \in \mathbb{R} : \widehat{g}_n(x) > 0\}$ may not contain the set $\{x \in \mathbb{R} : f(x) > 0\}$. This may happen particularly in the tail areas since there can appear bins with no observations. Some divergence measures are sensitive to situation when the domination $f \ll \widehat{g}_n$ fails. This is not the case of the Barron estimator which dominates $f$ provided $f$ is dominated by the auxiliary density $g$. The difference between the $g$-shaped histogram and the Baron $g$-shaped histogram can be easily seen from formulas (1) and (2). In fact, in the case of Barron estimator one observation was added *a priori* to each cell $A_{nj}$ and thus the effect of empty cell was circumvented.

We are interested in the asymptotic properties of $\phi$-divergence errors of the above defined estimators. The $\phi$-divergence between two probability distributions $P$ and $Q$ with densities $f$ and $g$ was firstly used in (Csiszár, 1963) and is defined by the formula

$$D_\phi(P,Q) = D_\phi(f,g) = \int g(x)\, \phi\left(\frac{f(x)}{g(x)}\right) \mathrm{d}\lambda(x)\,,$$

where the function $\phi(t) : [0,\infty) \to (-\infty,\infty]$ is convex, strictly convex at $t = 1$, with $\phi(1) = 0$ and where some additional rules are used to specify the undefined expressions like $0\,\phi(0/0)$ behind the integral. Typical examples of $\phi$-divergences are the *information divergence* $I(f,g) = \int f \ln(f/g)\,\mathrm{d}\lambda$ defined by $\phi(t) = t \ln t$ and the $\chi^2$*-divergence* $\chi^2(f,g) = \int (f-g)^2/g\,\mathrm{d}\lambda = \int f^2/g\,\mathrm{d}\lambda - 1$ defined by $\phi(t) = (t-1)^2$. The importance of $\phi$-divergence errors in information theory and telecommunications has been explained in detail in Applications 1-6 of (Berlinet, *et al.*, 1998).

## 2. KNOWN RELATED ASYMPTOTIC RESULTS

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 the additional condition $I(f,g) < \infty$ they proved also the consistency in the information divergence and the expected information divergence. The asymptotic normality of the error $I(f,\widetilde{g}_n)$ was proved in (Berlinet, *et al.*, 1997). Györfi, *et al.* (1998) presented arguments in favor of the $\chi^2$-divergence

error $\chi^2(f,\widetilde{g}_n)$ and proved the consistency in the expected $\chi^2$-divergence under some additional regularity assumptions on the density $f$. They obtained also the optimal rate of convergence $E\,\chi^2(f,\widetilde{g}_n) = O\left(n^{-\frac{2}{3}}\right)$ .

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

Our research is complementary to that of Berlinet, *et al.* (1998) and our results are more general in some respects.

## 3. NEW ASYMPTOTIC RESULTS

We suppose the restrictions $P^{(n)}$ and $Q^{(n)}$ of a probability distributions $P$ and $Q$ given on $(I\!R,\mathcal{B})$ on the subfield $\mathcal{B}^{(n)}$ of $\mathcal{B}$ generated by $\mathcal{P}_n$ and by $\widehat{Q}_n^{(n)}$ and $\widetilde{Q}_n^{(n)}$ we denote the same restrictions of the estimated probability distributions $\widehat{Q}_n$ and $\widetilde{Q}_n$. The discrete distribution $P^{(n)}$ is characterized by the vector $\boldsymbol{p}_n$ of probabilities of atoms of $\mathcal{B}^{(n)}$, $\widehat{Q}_n^{(n)}$ by the vector $\widehat{\boldsymbol{p}}_n$ and $\widetilde{Q}_n^{(n)}$ is characterized by the vector $\widehat{\boldsymbol{b}}_n = n/(n+m_n)\,\widehat{\boldsymbol{p}}_n + m_n/(n+m_n)\,\boldsymbol{q}_n$. We are interested in the asymptotics of the *reduced $\phi$ - divergence errors*

$$D_\phi\left(\widetilde{Q}_n^{(n)}, P^{(n)}\right) = \sum_{j=1}^{m_n} p_{nj}\,\phi\left(\frac{\widehat{b}_{nj}}{p_{nj}}\right)$$

of the the Barron $g$-shaped histogram $\widetilde{Q}_n$.

Before we present the main theoretical results let us list the assumptions imposed on the probability model, the partitions of the real line and the divergence generating function $\phi$. We suppose that the partitions $\mathcal{P}_n$ and the probability distributions $\boldsymbol{p}_n$ satisfy the following $P_\infty$-assumptions.

***P*$_\infty$-assumptions:** It holds $m_n \to \infty$ and $m_n/n \to 0$ for $n \to \infty$ and there exists $\beta \geq 1$ such that

$$\liminf_{n\to\infty} m_n^\beta \min_{1\leq j\leq m_n} p_{nj} > 0 \quad \text{and} \quad \frac{m_n^{1+\beta}}{n} = o(1)\,.$$

The divergence generating function $\phi$ is supposed to satisfy the following conditions which will be referred to as $F$-assumptions.

***F*-assumptions:** $\phi$ is finite and convex on $(0,\infty)$ extended in $[0,\infty)$ by the rule $\phi(0) = \lim_{t\to 0_+}\phi(t)$. Further, $\phi$ is twice continuously differentiable in a neighborhood of 1, satisfying the conditions $\phi(1) = \phi'(1) = 0$, $\phi''(1) > 0$ with the second derivative $\phi''(t)$ Lipschitz in a neighborhood of $t = 1$.

Let us note that both these sets of conditions are standard and can be found in the literature dealing with similar problems of nonparametric statistics.

*3.1 Consistency for the Barron g-shaped histogram*

Now we are ready to present the main theoretical results concerning the consistency of the Barron $g$-shaped histogram in the reduced $\phi$-divergence and the expected reduced $\phi$-divergence.

**Theorem 1** *If the partitions $\mathcal{P}_n$ and the probability distributions $\boldsymbol{p}_n$ satisfy the $P_\infty$-assumptions, then the Barron estimator $\widetilde{Q}_n$ of the probability distribution $P$ is consistent in the reduced $\phi$-divergence for all $\phi$ satisfying the F-assumptions, i.e.*

$$D_\phi\left(\widetilde{Q}_n^{(n)}, P^{(n)}\right) = o_P(1)\,. \tag{3}$$

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

$$\sup_{n>n_1} E\,\left(D_\phi\left(\widetilde{Q}_n^{(n)}, P^{(n)}\right)\right)^2 \;<\; +\infty\,, \tag{4}$$

*then the Barron estimator $\widetilde{Q}_n$ of the probability distribution $P$ is consistent in the expected reduced $\phi$-divergence for all $\phi$ satisfying the F-assumptions, i.e.*

$$E\,D_\phi\left(\widetilde{Q}_n^{(n)}, P^{(n)}\right) = o(1)\,. \tag{5}$$

The basic idea of the proofs of Theorems 1 and 2 is the following. First we show the desired asymptotic relation for the particular case of $\chi^2$-divergence and then, with a help of the following inequality

$$\left|D_\phi(\boldsymbol{p}_n, \boldsymbol{q}_n) - \frac{\phi''(1)}{2}\,\chi^2(\boldsymbol{p}_n, \boldsymbol{q}_n)\right| \le \frac{L_\phi}{2}\,\sum_{j=1}^{m_n}\frac{|p_{nj}-q_{nj}|^3}{q_{nj}^2}\,,$$

we argue the same also for the general $\phi$-divergence.

Since it can be quite difficult to check the validity of the condition (4) and it is not obvious what restriction it inquires, we have found simple conditions on the divergence generating function $\phi$ which are sufficient for (4) in the model satisfying the $P_\infty$-assumptions.

**Theorem 3** *Let the partitions $\mathcal{P}_n$ and the probability distributions $\boldsymbol{p}_n$ satisfy the $P_\infty$-assumptions. Then the condition*

$$\phi\left(\frac{1}{t}\right) + \phi(t) = O(t^k), \quad \textit{for} \quad t \to \infty \quad \textit{and some} \quad k \in \mathbb{N} \tag{6}$$

*is sufficient for (4).*

The condition (6) on the convex function $\phi$ is more general than conditions in (Berlinet, *et al.*, 1998) where the most general condition on the function $\phi$ was $t\phi(1/t)+\phi(t) = O(t^2)$ for $t \to \infty$. In our case the function $\phi(t)$ is allowed to increase with arbitrary polynomial rate for $t \to \infty$ or $t \to 0_+$. But, we have to say that we treat consistency only in the reduced $\phi$-divergences. However, if we suppose similar conditions on the model and partitions as in (Berlinet, *et al.*, 1998) we should be able to deduce from the consistency in the reduced $\phi$-divergence also something about the consistency in the non-reduced $\phi$-divergence for some convex functions $\phi$. This problem is treated more in detail in the next section.

*3.2 Quantization*

Let us denote by $p$ and $r_n$ the density functions of the probability distributions $P$ and $Q_n$ with respect to the dominating measure $Q$ and by $\mathbb{F}_Q$ the set of all probability densities with respect to $Q$.

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

$$\lim_{n\to\infty} E_Q\left(f|\mathcal{P}_n\right) = f \quad Q-a.s.\ , \tag{7}$$

*where $E_Q\left(\cdot|\mathcal{P}_n\right)$ is the corresponding conditional expectation $E_Q$ with respect to the $\sigma$-field generated by $\mathcal{P}_n$.*

We suppose $p \in L_3(Q)$ and define on $(I\!R, \mathcal{B}, Q)$ the conditional expectations

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

Further we consider the convex function $\psi(t) = t^3 - 1,\ t \in (0,\infty)$ and the corresponding $\phi$-divergence $D_\psi(P,Q)$.

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

Similar properties 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.

The following assertion can be proved.

*Let $\phi$ satisfy the F-assumptions and let the corresponding $\phi$-divergence be metric or, more generally, let the $\phi$-divergence satisfy for some $0 < \alpha < \infty$ the inequality*

$$\left(D_\phi\left(P,Q\right)\right)^\alpha \le \left(D_\phi\left(P,Q'\right)\right)^\alpha + \left(D_\phi\left(Q,Q'\right)\right)^\alpha \tag{8}$$

*for arbitrary probability distributions $P, Q, Q'$ on $(I\!R, \mathcal{B})$. If further $D_\psi(P,Q) < +\infty$ and the sequence of partitions $\mathcal{P}_n$ is $(P,Q)$-approximating, then from the consistency of the Barron estimator in the reduced $\phi$-divergence follows also the consistency of the Barron estimator in the non-reduced $\phi$-divergence, i.e. $D_\phi(\widetilde{Q}_n^{(n)}, P^{(n)}) = o_P(1) \Rightarrow D_\phi(\widetilde{Q}_n, P) = o_P(1)$ and in the case that the $\phi$-divergence is metric the same holds also for the expected $\phi$-divergence, i.e. $E\, D_\phi(\widetilde{Q}_n^{(n)}, P^{(n)}) = o(1) \Rightarrow E\, D_\phi(\widetilde{Q}_n, P) = o(1)$.*

We note that the inequality (8) is fulfilled e.g. for a large class of divergences introduced in (Österreicher and Vajda, 2003) and defined by convex functions

$$\phi(t) = \frac{\beta}{\beta - 1}\left[(1+t^\beta)^{1/\beta} - 2^{1/\beta - 1}\,(1+t)\right]$$

for $\beta \neq 1, \beta \neq 0$ and for the corresponding limits at $\beta = 1$ and $\beta = \infty$.

Thus, we found conditions on the partitions of the observation space and the statistical model under which we can deduce from the consistency of assumed estimates in the reduced $\phi$-divergences also the consistency in the stronger non-reduced $\phi$-divergences. These conditions are similar to those considered in the literature dealing with the consistency of Barron estimator directly in the non-reduced $\phi$-divergences. Some restriction is that

the above stated conclusion can be done only for the $\phi$-divergences which are metrics or whose powers are metrics.

Nevertheless even with this restriction our consistency results for the non-reduced $\phi$-divergences considerably extend similar results established for the non-reduced $\phi$-divergences in the previous literature, particularly in (Berlinet, *et al.*, 1998).

REFERENCES


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

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.

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

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*, vol. **55**, no. 3, pp. 639-653.