Distribution Estimation Consistent in Total Variation and in Two Types of Information Divergence

Andrew R. Barron, *Member, IEEE*, László Györfi, *Member, IEEE*, and Edward C. van der Meulen, *Fellow, IEEE*

Abstract—**The problem of the nonparametric estimation of a probability distribution is considered from three viewpoints: the consistency in total variation, the consistency in information divergence, and consistency in reversed order information divergence. These types of consistencies are relatively strong criteria of convergence, and a probability distribution cannot be consistently estimated in either type of convergence without any restrictions on the class of probability distributions allowed. Histogram-based estimators of distribution are presented which, under certain conditions, converge in total variation, in information divergence, and in reversed order information divergence to the unknown probability distribution. Some *a priori* information about the true probability distribution is assumed in each case. As the concept of consistency in information divergence is stronger than that of convergence in total variation, additional assumptions are imposed in the cases of informational divergences.**

Index Terms—**Consistent distribution estimation, total variation, information divergence, reversed order information divergence, histogram-based estimate.**

I. Introduction

We consider the problem of estimating an unknown probability distribution μ, defined on an arbitrary measurable space $(\mathcal{X}, \mathcal{B})$, based on independent, identically distributed (i.i.d.) observations $X_1, \cdots, X_n$ from μ. Here, $\mathcal{X}$ could be the real line $\mathbb{R}$ or the Euclidean space $\mathbb{R}^d$, $d \geq 1$, in which case $\mathcal{B}$ is the collection of Borel sets. As generic notation for the distribution estimate of a set A we use

$$\mu_n^\star(A) = \mu_n^\star(A, X_1, X_2, \cdots, X_n), \tag{1.1}$$

where $\mu_n^\star$ is a measurable function of its arguments.

In this paper, we examine density estimation problems and related distribution estimation problems that are motivated by statistical applications. As density estimation implicitly results in the estimation of a distribution, a reasonable error criterion for density estimation should correspond to an error criterion for the estimation of distributions.

Such error criteria can be derived from dissimilarity measures of probability measures, like f-divergences introduced by Csiszár [9]. The f-divergences have several important properties, e.g., they are invariant under an invertible transformation of the sample space.

The two most important f-divergences in mathematical statistics and information theory are the total variation and the information divergence: if μ and ν are probability measures on $\mathcal{X}$, then the total variation and the information divergence are defined by

$$T(\mu, \nu) = \sup_A |\mu(A) - \nu(A)| = \frac{1}{2} \sup_{\{A_i\}} \sum |\mu(A_i) - \nu(A_i)|, \tag{1.2}$$

and

$$I(\mu, \nu) = \sup_{\{A_i\}} \sum_i \mu(A_i) \log \frac{\mu(A_i)}{\nu(A_i)}, \tag{1.3}$$

respectively, where for $\sup_A$ the supremum is taken over all measurable sets A, and where for $\sup_{\{A_i\}}$ the supremum is taken over all finite (measurable) partitions $\{A_1, A_2, \cdots, A_m\}$ of $\mathcal{X}$. In this paper, log means logarithm to the base 2.

It is well known (cf. Csiszár [9], Kemperman [23], and Kullback [25]) that

$$2(\log e)\{T(\mu, \nu)\}^2 \leq I(\mu, \nu). \tag{1.4}$$

If μ and ν are absolutely continuous with respect to a σ-finite measure λ (i.e., $\mu \ll \lambda$, and $\nu \ll \lambda$), with densities $f = d\mu/d\lambda$ and $g = d\nu/d\lambda$, respectively, then

$$T(\mu, \nu) = \frac{1}{2} \int |f(x) - g(x)| \lambda(dx), \tag{1.5}$$

which is one half the L_1 distance between f and g, and

$$I(\mu, \nu) = \int f(x) \log \frac{f(x)}{g(x)} \lambda(dx). \tag{1.6}$$

Manuscript received March 6, 1990. A. R. Barron was supported by an Office of Naval Research Grant No. N00014-89-J-1811. L. Györfi and E. C. van der Meulen were supported by the scientific exchange program between the Hungarian Academy of Sciences and the Royal Belgian Academy of Sciences.

A. R. Barron is with the Department of Statistics, the Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory at the University of Illinois, Urbana, Champaign, IL 61801.

L. Györfi is with the Department of Mathematics, Technical University of Budapest, Stoczek u.2, H-1521, Budapest, Hungary.

E. C. van der Meulen is with the Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200 B, B-3001 Heverlee, Belgium.

IEEE Log Number 9108036.

$I(\mu, \nu)$ is also called the I-divergence or Kullback–Leibler information number of μ with respect to ν, and will be denoted by $D(f, g)$ as well in the sequel if f and g exist. If $\mu \ll \nu$, and $f = d\mu/d\nu$, then

$$I(\mu, \nu) = \int f(x) \log f(x) \nu(dx) = \int \log f(x) \mu(dx). \tag{1.7}$$

In this paper, we shall be interested in finding conditions under which suitably defined distribution estimators $\mu_n^\star$ are consistent estimators of μ, either in total variation $T(\mu, \mu_n^\star)$, or in informational divergence $I(\mu, \mu_n^\star)$, or in reversed order informational divergence $I(\mu_n^\star, \mu)$.

We first recall the standard empirical measure μ_n which is often referred to next. For random variables $X_1, \cdots, X_n$, which are i.i.d. according to μ on $(\mathcal{X}, \mathcal{B})$, and any measurable set A we define the standard empirical measure μ_n of A by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \in A\}}, \tag{1.8}$$

where $1_B = 1$ if B occurs, and $= 0$ otherwise.

As we shall see, though, the standard empirical measure is not suited for estimating μ when we wish to have a.s. consistency in total variation or in information divergence.

We also recall the classical result by Glivenko–Cantelli, which states that if for a random variable X the distribution function F corresponding to μ on $(\mathbb{R}, \mathcal{B})$ is arbitrary, and F_n is the empirical distribution function, then

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{a.s.}, \tag{1.9}$$

as $n \to \infty$. This is actually a weak result in many respects, since the Glivenko–Cantelli distance (1.9) only involves measure of half lines, as $F_n(x) = \mu_n((-\infty, x])$ and $F(x) = \mu((-\infty, x])$, but not of any arbitrary (measurable) sets A. It becomes an entirely different matter when one considers the total variation $T(\mu, \mu_n)$ between μ and the standard empirical measure μ_n. In the definition of $T(\mu, \mu_n)$ (cf. (1.2)) the supremum is taken over all measurable sets A. Now it is not true that $T(\mu, \mu_n)$ will tend to zero a.s. for all measures μ. In fact, if μ is absolutely continuous and μ_n is the standard empirical measure, then $T(\mu, \mu_n) = 1$, as can be easily seen by considering the set A that is the finite support of μ_n. Then, $\mu_n(A) = 1$, $\mu(A) = 0$, and thus, $T(\mu, \mu_n) = 1$. Also, in this case $I(\mu, \mu_n) = \infty$, since μ is not absolutely continuous with respect to μ_n.

Hence, when we want to find distribution estimators which are a.s. consistent in either total variation or in information divergence, the standard empirical measure μ_n is not suitable if we allow all possible probability measures μ, including the case that μ is absolutely continuous. Yet more is true. Recently, Devroye and Györfi [15] established the following negative finding.

Lemma (Devroye and Györfi [15]): For any δ, $0 < \delta < 1/2$, and for any sequence of distribution estimators $\mu_n^\star$ on $(\mathbb{R}, \mathcal{B})$, there exists a probability measure μ such that

$$\inf_n T(\mu, \mu_n^\star) > 1/2 - \delta \quad \text{a.s.}, \tag{1.10}$$

and thus, (by (1.4)) both

$$\inf_n I(\mu, \mu_n^\star) > 2(\log e)(1/2 - \delta)^2 \quad \text{a.s.} \tag{1.11}$$

and

$$\inf_n I(\mu_n^\star, \mu) > 2(\log e)(1/2 - \delta)^2 \quad \text{a.s.} \tag{1.12}$$

This indicates that, in our search for a distribution estimator $\mu_n^\star$ which is a.s. consistent in either total variation or in information divergence, we should limit the class of probability measures for which we are estimating the unknown one. This can be achieved by assuming some *a priori* information about μ. In addition, this shows that the standard empirical measure μ_n should be modified if absolutely continuous probability measures are retained in the restricted class.

One of the main messages of the L_1 theory of density estimation is that there are estimators (histogram, kernel, etc.) which are a.s. consistent in L_1 without any condition on the unknown density f (see Devroye and Györfi [14]). Therefore, if the underlying distribution has a density, then such a.s. consistent estimators in L_1 provide distribution estimates a.s. consistent in total variation.

We now recall the standard histogram density estimator that is central in our work. For any given σ-finite measure η and any partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \cdots\}$ of $\mathcal{X}$ such that $\eta(A_{n,i}) > 0$ for each i, the corresponding standard histogram density estimator $\hat{f}_n$ with respect to η is defined by

$$\hat{f}_n(x) = \hat{f}_n(x, X_1, X_2, \cdots, X_n) = \frac{\mu_n(A_{n,i})}{\eta(A_{n,i})}, \quad \text{if } x \in A_{n,i}, \tag{1.13}$$

where μ_n is the standard empirical measure (1.8). The corresponding basic distribution estimator, denoted by $\hat{\mu}_n$, is defined by

$$\hat{\mu}_n(A) = \int_A \hat{f}_n(x) \eta(dx) \tag{1.14}$$

$$= \sum_{i=1}^{\infty} \mu_n(A_{n,i}) \frac{\eta(A \cap A_{n,i})}{\eta(A_{n,i})}. \tag{1.15}$$

In Section II, we present a certain histogram-based distribution estimate $\tilde{\mu}_n$ ((2.4)), which is a modification of $\hat{\mu}_n$, and which is a.s. consistent in total variation, i.e., $\lim_{n \to \infty} T(\mu, \tilde{\mu}_n) = 0$ a.s., and consistent in expected total variation, i.e., $\lim_{n \to \infty} E(T(\mu, \tilde{\mu}_n)) = 0$ (cf. Theorem 1 and (2.10)). In deriving Theorem 1 and its generalization Theorem 1′, we assume as *a priori* information about μ that the nonatomic part of μ is absolutely continuous with respect to a known σ-finite measure ν.

In Section III, we present another histogram-based distribution estimator μ_n^* ((3.2)), which is also a certain modification of $\hat{\mu}_n$, and which is a.s. consistent in information divergence, i.e., $\lim_{n\to\infty} I(\mu, \mu_n^*) = 0$ a.s., and consistent in expected information divergence, i.e., $\lim_{n\to\infty} E(I(\mu, \mu_n^*)) = 0$ (Theorem 2). In deriving Theorem 2, we assume that there exists a known probability measure ν such that $I(\mu, \nu) < \infty$.

At the end of Section III, we give implications of our results for the problem of universal source coding of finely quantized data. Similar implications of convergence in expected information divergence can be given for universal portfolio selection (cf. Barron and Cover [5]).

Earlier, Barron [3] considered the convergence in expected information divergence of probability density estimators. In particular, he considered a modified histogram estimator for estimating an unknown probability density function on $[0, 1]$. His estimator is given as a special case in Section III. In Barrow and Sheu [6], convergence in probability of the information divergence for ordinary histogram estimators is established for density functions having finite Fisher information. See also Remark 6.

Also earlier, Györfi and van der Meulen [20] showed a.s. consistency in information divergence of two histogram-based density estimators under certain conditions. These conditions included a convergent series condition. In Theorem 2, we are able to relax this condition and, moreover, focus on the general case of distribution estimation rather than density estimation. Techniques of the proof in Section III are related to entropy estimation based on histograms (see Györfi and van der Meulen [19]).

Finally, in Section IV (Theorem 5) we prove the consistency in expected reversed order information divergence, i.e., $\lim_{n\to\infty} E(I(\hat{\mu}_n, \mu)) = 0$, for the basic histogram-based distribution estimator $\hat{\mu}_n$ ((1.14)). Here, we assume not only that μ is absolutely continuous with respect to a known σ-finite measure ν, but also impose additional conditions on μ and ν. Theorem 5 is arrived at by decomposing $I(\hat{\mu}_n, \mu)$ into a variance component and a bias component. The convergence properties of the variance component are dealt with in Theorem 3 and those of the bias component are treated in Theorem 4.

II. Consistency in Total Variation

In this section, we consider the problem of estimating the unknown probability measure μ, defined on an arbitrary measurable space $\mathcal{X}$, from i.i.d. observations $X_1, \cdots, X_n$, such that the estimator used is a.s. consistent in total variation. In Section I, we saw that the standard empirical measure μ_n is not suitable for this purpose if all possible μ are allowed. When proving the negative result given in the Lemma of Section I, Devroye and Györfi [15] mentioned that a distribution might be estimated consistently in total variation, if the distribution is a mixture of just an atomic and an absolutely continuous part (and thus does not involve a continuous singular component). In the sequel, we slightly extend this program of thought, first for $\mathcal{X} = \mathbb{R}^d$ and after that for $\mathcal{X}$ being an arbitrary measurable space. We will present a histogram-based estimator of the distribution, which is a modification of $\hat{\mu}_n$ (defined in (1.14)), and which is indeed a.s. consistent in total variation under some conditions. It will be denoted by $\tilde{\mu}_n$. One of these conditions assumes some prior information about μ, viz. that there is a known (nonatomic) σ-finite measure ν which dominates the nonatomic part of μ.

For a probability measure μ on $\mathcal{X}$, we denote the atomic part of μ by μ_a and the nonatomic part of μ by μ_b, so that

$$\mu = p\mu_a + (1-p)\mu_b. \tag{2.1}$$

Let ν be a known σ-finite measure on $\mathcal{X}$ such that μ_b is absolutely continuous with respect to ν, i.e., $\mu_b \ll \nu$.

Moreover, given the sample $X_1, \cdots, X_n$, define

$$B_n = \left\{x\colon \mu_n(\{x\}) \geq \frac{2}{n}\right\}, \tag{2.2}$$

where μ_n is the standard empirical measure (1.8). Finally, let $N\{\cdots\}$ denote the cardinality of a set $\{\cdots\}$.

We first discuss the case $\mathcal{X} = \mathbb{R}^d$ and state Theorem 1. We next observe that the results can be formulated in a general setting and state Theorem 1′, which generalizes Theorem 1 to the case that $\mathcal{X}$ is an arbitrary measurable space. We then prove Theorem 1′, which *a fortiori* provides a proof of Theorem 1.

Now turning to the case $\mathcal{X} = \mathbb{R}^d$, let

$$\mathcal{P}_n = \{A_{n,1}, A_{n2}, \cdots\} \tag{2.3}$$

be a partition of $\mathbb{R}^d$ with $0 < \nu(A_{n,i}) < \infty$ for all $n \geq 1$, $i \geq 1$. Next define, for measurable subsets A of $\mathbb{R}^d$, the distribution estimate

$$\tilde{\mu}_n(A) = \sum_{i=1}^{\infty} \mu_n(A_{n,i} \cap B_n^c) \frac{\nu(A \cap A_{n,i})}{\nu(A_{n,i})} + \mu_n(A \cap B_n). \tag{2.4}$$

We then have the following theorem.

Theorem 1: Let μ be an unknown probability measure on $\mathbb{R}^d$. Suppose there exists a known σ-finite measure ν such that $\mu_b \ll \nu$. Let $\mathcal{P}_n$ be a sequence of partitions, as defined in (2.3), such that for each finite sphere S centered at the origin

$$\text{a)} \qquad \lim_{n\to\infty} \frac{N\{i\colon A_{n,i} \cap S \neq \varnothing\}}{n} = 0 \tag{2.5}$$

and

$$\text{b)} \qquad \lim_{n\to\infty} \max_{\{i\colon A_{n,i}\cap S \neq \varnothing\}} \left(\sup_{x,\, y \in A_{n,i}} \|x - y\| \right) = 0. \tag{2.6}$$

(Here, $\|\cdot\|$ denotes Euclidean norm.) Let $\tilde{\mu}_n$ be defined as in (2.4). Then,

$$\lim_{n\to\infty} T(\mu, \tilde{\mu}_n) = 0 \quad \text{a.s.} \tag{2.7}$$

Remark 1: We note that if the nonatomic part μ_p is absolutely continuous (with respect to Lebesgue measure λ) then the obvious choice for ν is λ. We also note that if the $A_{n,i}$'s in (2.3) are cubes of common size h_n, then conditions (2.5) and (2.6) can be replaced by the conditions

$$\lim_{n\to\infty} nh_n^d = \infty \tag{2.8}$$

and

$$\lim_{n\to\infty} h_n = 0, \tag{2.9}$$

respectively. Furthermore, since $T(\mu, \tilde{\mu}_n)$ is bounded between 0 and 1, the a.s. consistency of $T(\mu, \tilde{\mu}_n)$ given by (2.7) implies the consistency in expected total variation, i.e., that

$$\lim_{n\to\infty} E(T(\mu, \tilde{\mu}_n)) = 0 \tag{2.10}$$

under the conditions of Theorem 1.

Remark 2: The novelty of our result is the treatment of distributions with both discrete and absolutely continuous components. In what follows, we show how the result of Theorem 1 can be deduced by relating our situation to the case that the measure μ is dominated by a given σ-finite measure η. Indeed, the estimator $\tilde{\mu}_n$ ((2.4)) is shown to be close to a basic histogram estimator $\hat{\mu}_n$ with respect to a σ-finite measure η, which is a mixture of ν and an atomic measure κ. To make the arguments work, we will need less restrictive assumptions on the sequences of partitions. Furthermore, we assume $\mathcal{X}$ to be arbitrary. The necessary conditions will now be defined.

We recall the definition of the standard histogram-based density estimator $\hat{f}_n$ ((1.13)) with respect to a given σ-finite measure η and a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \cdots\}$ of an arbitrary measurable space $\mathcal{X}$ such that $\eta(A_{n,i}) > 0$. We also recall the definition (1.14) of the corresponding basic distribution estimate $\hat{\mu}_n$ for measurable subsets A of $\mathcal{X}$.

In a similar manner, we define

$$f_n(x) = \frac{\mu(A_{n,i})}{\eta(A_{n,i})}, \qquad \text{if } x \in A_{n,i}, \tag{2.13}$$

which is the expected value of the histogram estimator (i.e., $f_n(x) = E\hat{f}_n(x)$), and denote

$$\bar{\mu}_n(A) = \int_A f_n(x)\eta(dx). \tag{2.14}$$

Next, we recall the following definitions from Barron [4].

Definition 1: The ϵ-effective cardinality of a partition $\mathcal{P}$ of $\mathcal{X}$ with respect to a measure η restricted to a subset S of $\mathcal{X}$, denoted $m(\mathcal{P}, \eta, S, \epsilon)$, is the minimum number of sets in the partition such that the union of the remaining sets intersected with S has η-measure less than ϵ.

Definition 2 (Condition A): The effective cardinality of a sequence of partitions $\mathcal{P}_n$ with respect to a σ-finite measure η is said to be of order $o(n)$ if

$$\lim_{n\to\infty} \frac{m(\mathcal{P}_n, \eta, S, \epsilon)}{n} = 0, \tag{2.15}$$

for all $\epsilon > 0$ and all sets S with $\eta(S) < \infty$.

Condition A on the effective cardinality of a sequence of partitions of $\mathcal{X}$ can be shown (see Appendix A) to be equivalent to condition (D) in Abou-Jaoude [2], which he shows to be necessary and sufficient for the convergence to zero in probability of the variation term $T(\hat{\mu}_n, \bar{\mu}_n)$ for all probability measures μ dominated by η. This condition is also sufficient for the almost sure convergence of the variation term as shown in Abou-Jaoude [1].

In the case that the measurable space $\mathcal{X}$ is a metric space, a sufficient condition for the effective cardinality of a sequence of partitions to be of order $o(n)$ is that

$$\lim_{n\to\infty} \frac{N\{i:\ A_{n,i} \cap S \neq \varnothing\}}{n} = 0, \tag{2.16}$$

for each ball S centered at some point x_0. (This is the same as condition (2.5) in the case that $\mathcal{X} = \mathbb{R}^d$.) In particular, a partition of $\mathbb{R}^d$ into cubes of common width h_n is of effective cardinality $o(n)$ provided (cf. (2.8))

$$\lim_{n\to\infty} nh_n^d = \infty. \tag{2.17}$$

The following concept is due to Csiszár [10].

Definition 3 (Condition B): For a measure η on a measurable space $\mathcal{X}$, a sequence of partitions $\mathcal{P}_n$ of $\mathcal{X}$ is said to be η-approximating if, for every measurable set A with $\eta(A) < \infty$ and for every $\epsilon > 0$, there is for all n sufficiently large a set A_n equal to a union of sets in $\mathcal{P}_n$ such that

$$\eta(A_n \triangle A) < \epsilon, \tag{2.18}$$

where $A_n \triangle A$ denotes the symmetric difference of A_n and A.

Abou-Jaoude [2] showed that the condition that the sequence $\mathcal{P}_n$ is η-approximating (condition B) is necessary and sufficient for the convergence to zero of the bias term $T(\bar{\mu}_n, \mu)$, for all measures μ dominated by η.

For separable metric spaces, Csiszár [10, p. 168] showed that a sufficient condition, for a sequence of partitions $\mathcal{P}_n$ to be η-approximating for any σ-finite measure η, is that

$$\lim_{n\to\infty} \max_{\{i:\ A_{n,i}\cap S\neq\varnothing\}} \operatorname{diam}(A_{n,i}) = 0, \tag{2.19}$$

for each ball S centered at some point x_0. Here $\operatorname{diam}(A) = \sup_{x,y\in A} d(x, y)$, and $d(x, y)$ denotes the distance between points in the metric space. For partitions of $\mathbb{R}^d$ into cubes of width h_n this sufficient condition becomes (cf. (2.9))

$$\lim_{n\to\infty} h_n = 0. \tag{2.20}$$

This leads to the following generalization of Theorem 1.

Theorem 1': Let $\mathcal{X}$ be a measurable space with the probability distribution μ defined on it. Suppose there is a

known (nonatomic) σ-finite measure ν that dominates the nonatomic part of μ. Let $\mathscr{P}_n = \{A_{n,1}, A_{n,2}, \cdots\}$ be a sequence of partitions of $\mathcal{X}$ such that $0 < \nu(A_{n,i}) < \infty$ for all $n \geq 1$, $i \geq 1$. Assume that the sequence of partitions $\mathscr{P}_n$ of $\mathcal{X}$ is ν-approximating and has effective cardinality of order $o(n)$ with respect to ν. Let $\tilde{\mu}_n$ be defined on $\mathcal{X}$ as in (2.4), for subsets A of $\mathcal{X}$ and based on this $\mathscr{P}_n$. Then,

$$\lim_{n\to\infty} T(\mu, \tilde{\mu}_n) = 0 \quad \text{a.s.} \tag{2.21}$$

Proof of Theorem 1': If we introduce the notation

$$B = \{x: \mu(\{x\}) > 0\}, \tag{2.22}$$

then, clearly

$$B_n \subset B \quad \text{a.s.} \tag{2.23}$$

Let b_i, $i = 1, 2, \ldots$ denote the points in B, and let the partition generated by $\mathscr{P}_n$ and the point-sets in B be denoted by

$$\tilde{\mathscr{P}}_n = \{A_{n,1} \cap B^c, A_{n,2} \cap B^c, \cdots, \{b_1\}, \{b_2\}, \cdots\}. \tag{2.24}$$

Similar to (1.14) and (1.15), let $\hat{\tilde{\mu}}_n$ be the (auxiliary) histogram-based distribution "estimator" with respect to the partition $\tilde{\mathscr{P}}_n$ and the σ-finite measure $\eta = \nu + \kappa$, where κ is the counting measure restricted to B. Then,

$$\hat{\tilde{\mu}}_n(A) = \sum_{i=1}^{\infty} \mu_n(A_{n,i} \cap B^c) \frac{\nu(A \cap A_{n,i})}{\nu(A_{n,i})} + \mu_n(A \cap B). \tag{2.25}$$

The distribution $\hat{\tilde{\mu}}_n$ has density with respect to η given by

$$\hat{\tilde{f}}_n(x) = \begin{cases} \mu_n(A_{n,i} \cap B^c)/\nu(A_{n,i}), & \text{if } x \in A_{n,i} \cap B^c, \\ \mu_n(\{x\}), & \text{if } x \in B. \end{cases} \tag{2.26}$$

Note that, unlike the estimator $\tilde{\mu}_n$, $\hat{\tilde{\mu}}_n$ requires knowledge of B. Therefore, $\hat{\tilde{\mu}}_n$ is not an estimator in the true sense, but an auxiliary, artificial estimator introduced in the course of the proof. Note that $\hat{\tilde{\mu}}_n(A) = \tilde{\mu}_n(A)$ for $A \subset B_n$, and $\hat{\tilde{\mu}}_n(A) \leq \tilde{\mu}_n(A)$ for $A \subset B^c$, while $\tilde{\mu}_n(A) = 0$ for $A \subset B - B_n$. Thus, the set $A_0 \triangleq B - B_n$ has the property that $\hat{\tilde{\mu}}_n$ is greater than or equal to $\tilde{\mu}_n$ when restricted to A_0, while the reverse inequality holds when restricted to A_0^c, i.e.,

$$\tilde{\mu}_n(A \cap A_0) \leq \hat{\tilde{\mu}}_n(A \cap A_0)$$

and

$$\tilde{\mu}_n(A \cap A_0^c) \geq \hat{\tilde{\mu}}_n(A \cap A_0^c).$$

It is seen that the distribution $\tilde{\mu}_n$ is also absolutely continuous with respect to η and a version of the corresponding density is

$$\tilde{f}_n(x) = \begin{cases} \mu_n(A_{n,i} \cap B_n^c)/\nu(A_{n,i}), & \text{if } x \in A_{n,i} \cap B^c, \\ \mu_n(\{x\}) 1_{\{x \in B_n\}}, & \text{if } x \in B. \end{cases} \tag{2.27}$$

Let $A^* = \{x: \hat{\tilde{f}}_n(x) > \tilde{f}_n(x)\}$. Then, $A^* = A_0$. Now examining the total variation distance between $\hat{\tilde{\mu}}_n$ and $\tilde{\mu}_n$ (cf. definition (1.2)) we have that

$$\begin{aligned} T(\hat{\tilde{\mu}}_n, \tilde{\mu}_n) &= \sup_A |\hat{\tilde{\mu}}_n(A) - \tilde{\mu}_n(A)| \\ &= \frac{1}{2}\int |\hat{\tilde{f}}_n(x) - \tilde{f}_n(x)| \eta(dx) \\ &= \int_{A^*} (\hat{\tilde{f}}_n(x) - \tilde{f}_n(x)) \eta(dx) \\ &= \hat{\tilde{\mu}}_n(A^*) - \tilde{\mu}_n(A^*) \\ &= \mu_n(B - B_n) - 0 \\ &= \mu_n(B) - \mu_n(B_n). \end{aligned} \tag{2.28}$$

Now put, for $b_i \in B$, $p_i = \mu(\{b_i\})$ and $p_{ni} = \mu_n(\{b_i\})$. By the strong law of large numbers,

$$\lim_{n\to\infty} p_{ni} = p_i \quad \text{a.s.} \tag{2.29}$$

Next, let $M > 0$ be a fixed positive integer. Then,

$$\begin{aligned} &\mu_n(B) - \mu_n(B_n) \\ &\quad = \sum_{i=1}^{\infty} \mu_n(\{b_i\}) - \sum_{i=1}^{\infty} \mu_n(\{b_i\}) 1_{\{\mu_n(\{b_i\}) \geq 2/n\}} \\ &\quad = \sum_{i=1}^{\infty} \mu_n(\{b_i\}) 1_{\{\mu_n(\{b_i\}) \leq 1/n\}} \\ &\quad = \sum_{i=1}^{M} p_{ni} 1_{\{p_{ni} \leq 1/n\}} + \sum_{i=M+1}^{\infty} p_{ni} 1_{\{p_{ni} \leq 1/n\}}. \end{aligned} \tag{2.30}$$

Now, (2.29) implies that, for each fixed $M > 0$,

$$\lim_{n\to\infty} \sum_{i=1}^{M} p_{ni} 1_{\{p_{ni} \leq 1/n\}} = 0 \quad \text{a.s.} \tag{2.31}$$

Moreover,

$$\begin{aligned} &\limsup_{n\to\infty} \sum_{i=M+1}^{\infty} p_{ni} 1_{\{p_{ni} \leq 1/n\}} \\ &\qquad \leq \limsup_{n\to\infty} \sum_{i=M+1}^{\infty} p_{ni} = \sum_{i=M+1}^{\infty} p_i \text{a.s.} \end{aligned} \tag{2.32}$$

From (2.30), (2.31), and (2.32), we deduce that

$$0 \leq \limsup_{n\to\infty} (\mu_n(B) - \mu_n(B_n)) \leq \sum_{i=M+1}^{\infty} p_i \quad \text{a.s.}, \tag{2.33}$$

and since M was arbitrary it follows that

$$\lim_{n\to\infty} (\mu_n(B) - \mu_n(B_n)) = 0 \quad \text{a.s.} \tag{2.34}$$

Thus, by (2.28) and (2.34),

$$\lim_{n\to\infty} T\left(\hat{\tilde{\mu}}_n, \tilde{\mu}_n\right) = 0 \quad \text{a.s.} \tag{2.35}$$

By applying the triangle inequality, it now follows that $\tilde{\mu}_n$ converges a.s. to μ in total variation, if and only if $\hat{\tilde{\mu}}_n$ converges a.s. to μ in total variation. Now, since the sequence of partitions $\mathscr{P}_n$ is assumed to be ν-approximating and has effective cardinality of order $o(n)$ with respect to ν, it is easily checked that the sequence of partitions $\tilde{\mathscr{P}}_n$ is η-approximating and has effective cardinality of order $o(n)$ with respect to η. Therefore, by the results of Abou-Jaoude [1], [2] just discussed,

$$\lim_{n\to\infty} T\left(\hat{\tilde{\mu}}_n, \mu\right) = 0 \quad \text{a.s.} \tag{2.36}$$

and hence,

$$\lim_{n\to\infty} T\left(\tilde{\mu}_n, \mu\right) = 0 \quad \text{a.s.} \tag{2.37}$$

Thus, the proof of Theorem 1′ (and hence, of Theorem 1) is completed. □

III. Consistency in Information Divergence

Consider again the problem of estimating the unknown probability measure μ from i.i.d. observations $X_1, \cdots, X_n$ taking values in an arbitrary measurable space $\mathscr{X}$. We will now present another histogram-based estimator of the distribution, denoted by μ_n^*, which is also a certain modification of $\hat{\mu}_n$, and which is a.s. consistent in information divergence and consistent in expected information divergence. In order to derive these consistency results, we need to impose some additional assumptions. In particular, we assume that there exists a known probability measure ν on $\mathscr{X}$ such that $I(\mu, \nu) < \infty$. As is well-known, the condition that $I(\mu, \nu) < \infty$ implies that μ is absolutely continuous with respect to ν. Apart from the fact that ν is known, ν is otherwise arbitrary and thus may be discrete or have a discrete part.

Now define a sequence of integers m_n, $0 < m_n < n$, $n = 1, 2, \cdots$, and a sequence of real numbers $h_n > 0$. Furthermore, introduce a sequence of partitions $\mathscr{P}_n = \{A_{n,1}, A_{n,2}, \cdots, A_{n,m_n}\}$, $n = 1, 2, \cdots$, of $\mathscr{X}$ such that $\nu(A_{n,i}) \geq h_n$. Notice that this implies that $m_n \leq 1/h_n$. We require that $\nu(A_{n,i}) \geq h_n$ rather than $\nu(A_{n,i}) = h_n$, to allow for the possibility that ν is discrete since in that case $\nu(A_{n,i})$ may exceed the prescribed number h_n.

For a given sequence a_n, $0 < a_n < 1$, with

$$\lim_{n\to\infty} a_n = 0, \tag{3.1}$$

we now define the distribution estimator

$$\mu_n^*(A) = (1 - a_n) \sum_{i=1}^{m_n} \mu_n(A_{n,i}) \frac{\nu(A \cap A_{n,i})}{\nu(A_{n,i})} + a_n \nu(A), \tag{3.2}$$

where μ_n is again the standard empirical measure for the sample $X_1, X_2, \cdots, X_n$, as defined in (1.8).

Defining the standard histogram density estimator (with respect to $\mathscr{P}_n$ and ν) by (cf. (1.13))

$$\hat{f}_n(x) = \mu_n(A_{n,i})/\nu(A_{n,i}), \qquad \text{if } x \in A_{n,i}, \tag{3.3}$$

and the corresponding distribution estimator (cf. (1.14)–(1.15)) by

$$\hat{\mu}_n(A) = \int_A \hat{f}_n(x)\nu(dx),$$

we note that

$$\mu_n^*(A) = (1 - a_n)\hat{\mu}_n(A) + a_n \nu(A). \tag{3.4}$$

Observe that $\mu_n^* \ll \nu$ and that the density f_n^* of μ_n^* with respect to ν has the form

$$f_n^*(x) = (1 - a_n)\mu_n(A_{n,i})/\nu(A_{n,i}) + a_n, \qquad \text{if } x \in A_{n,i}. \tag{3.5}$$

We recall (cf. Section I) that if $\mu \ll \lambda$ and $\nu \ll \lambda$, with $f = d\mu/d\lambda$, $g = d\nu/d\lambda$, then,

$$I(\mu, \nu) = D(f, g) \triangleq \int f(x) \log \frac{f(x)}{g(x)} \lambda(dx). \tag{3.6}$$

Now, if in the above setup $\mathscr{X} = \mathbb{R}$, the $A_{n,i}$'s are intervals, and it is assumed that ν is absolutely continuous with respect to the Lebesgue measure λ, with $g = d\nu/d\lambda$, then the distribution estimate (3.2) can be derived from the density estimate

$$\tilde{f}_n^*(x) = ((1 - a_n)\mu_n(A_{n,i})/\nu(A_{n,i}) + a_n)g(x), \qquad \text{if } x \in A_{n,i}, \tag{3.7}$$

which can be regarded as a modified histogram (with respect to Lebesgue measure). It is easy to see that $\tilde{f}_n^*$ is itself a density and in this case

$$\mu_n^*(A) = \int_A \tilde{f}_n^*(x)\lambda(dx).$$

Furthermore, if in this case one chooses $\nu(A_{n,i}) = 1/m_n = h_n$, $1 \leq i \leq m_n$, and

$$a_n = \frac{m_n}{n + m_n}, \tag{3.8}$$

then $\tilde{f}_n^*$ becomes a density estimator introduced by Barron [3]. It takes the particular form

$$\tilde{p}_n^*(x) = \frac{m_n}{n + m_n}(n\mu_n(A_{n,i}) + 1)g(x), \qquad \text{if } x \in A_{n,i}. \tag{3.9}$$

Barron [3] showed that if

$$\lim_{n\to\infty} m_n = \infty \tag{3.10}$$

and

$$\lim_{n\to\infty} m_n/n = 0 \tag{3.11}$$

and μ has a density f with respect to Lebesgue measure such that

$$D(f,g) < \infty, \tag{3.12}$$

then

$$\lim_{n\to\infty} E(D(f,\tilde{p}_n^*)) = 0. \tag{3.13}$$

Whereas Barron [3] took the expected information divergence as a measure of accuracy (convergence criterion), we consider here the a.s. convergence in information divergence as our main objective. In the first part of Theorem 2, we extend his result. In the second part of Theorem 2, we show that the same conditions which guarantee convergence in expected information divergence are also sufficient conditions for the a.s. convergence in information divergence of our distribution estimator μ_n^* to the unknown probability measure μ. In doing so, we relax the sufficient conditions for a.s. convergence in information divergence reported in Györfi and van der Meulen [20]. When stating Theorem 2 we assume the general framework introduced at the beginning of this section.

Theorem 2: Let μ be an unknown probability measure on $\mathcal{X}$. Assume that there exists a known probability measure ν such that

$$I(\mu,\nu) < \infty. \tag{3.14}$$

Moreover, assume that (3.1) is satisfied, $\mathcal{P}_n$ is ν-approximating,

$$\lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{3.15}$$

and

$$\limsup_{n\to\infty} \frac{1}{na_nh_n} \le 1. \tag{3.16}$$

Let μ_n^* be defined as in (3.2). Then,

a) $$\lim_{n\to\infty} E(I(\mu,\mu_n^*)) = 0 \tag{3.17}$$

and

b) $$\lim_{n\to\infty} I(\mu,\mu_n^*) = 0 \quad \text{a.s.} \tag{3.18}$$

Remark 3: There is an obvious question whether it is really necessary to impose more conditions when estimating a distribution consistently in information divergence instead of consistently in total variation. For example, according to Theorem 1 each discrete distribution can be estimated consistently in total variation without any additional assumption on the underlying distribution. This is not the case for information divergence. Namely, if for an infinite discrete distribution, we fail to detect at least one point of the unknown infinite support, then, using any basic distribution estimator $\hat{\mu}_n$, the divergence $I(\mu,\hat{\mu}_n)$ is infinite, and it is impossible to discover from a finite sample with probability 1 the whole infinite support of a countably discrete distribution. On the other hand, our distribution estimator $\mu_n^*(A) = (1-a_n)\hat{\mu}_n(A) + a_n\nu(A)$ $= 0$ iff $\nu(A) = 0$, and thus $\mu \ll \mu_n^*$ a.s. Hence, by adding the term $a_n\nu(A)$ the effect of empty cells is circumvented and $I(\mu,\mu_n^*)$ is not necessarily infinite to start with.

If we now assume that $\mathcal{X} = \mathbb{R}$, that μ has a density f and ν has a density g, both with respect to Lebesgue measure, and we use $f_n^*(x)$ defined in (3.7) to construct μ_n^*, then Theorem 2 yields the following corollary.

Corollary: Assume $\mathcal{X} = \mathbb{R}$, that the $A_{n,i}$'s are intervals of equal ν-measure $\nu(A_{n,i}) = h_n = 1/m_n$, that (3.1), (3.10), (3.15), and (3.16) are satisfied, that μ has a density f, and that there exists a known absolutely continuous probability measure ν with density g such that

$$D(f,g) < \infty. \tag{3.19}$$

Let $\tilde{f}_n^*$ be defined as in (3.7). Then,

a) $$\lim_{n\to\infty} E\left(D\left(f,\tilde{f}_n^*\right)\right) = 0 \tag{3.20}$$

and

b) $$\lim_{n\to\infty} D\left(f,\tilde{f}_n^*\right) = 0 \quad \text{a.s.} \tag{3.21}$$

Proof of Theorem 2: Before giving the specific proofs of parts a) and b), we start with some general facts which are used in both proofs. Given the partition $\mathcal{P}_n$, the sequence a_n, and the probability measure ν, let $f_n^*(x)$ be defined as in (3.5). Recall the definition (3.3) of $\hat{f}_n(x)$. Since $\mu \ll \nu$, let $f = d\mu/d\nu$. We also define, for $x \in A_{n,i}$,

$$f_n(x) = E\hat{f}_n(x) = \frac{\mu(A_{n,i})}{\nu(A_{n,i})} \tag{3.22}$$

and

$$\bar{\mu}_n(A) = \int_A f_n(x)\nu(dx). \tag{3.23}$$

We first decompose $I(\mu,\mu_n^*)$. We then have

$$\begin{aligned} I(\mu,\mu_n^*) = D(f,f_n^*) &= \int f(x)\log\frac{f(x)}{f_n^*(x)}\nu(dx) \\ &= \int f(x)\log\frac{f(x)}{E\hat{f}_n(x)}\nu(dx) \\ &\quad + \int f(x)\log\frac{E\hat{f}_n(x)}{f_n^*(x)}\nu(dx) \\ &= U_n + V_n. \end{aligned} \tag{3.24}$$

Now,

$$U_n = \int f(x)\log f(x)\nu(dx) - \sum_{i=1}^{m_n} \mu(A_{n,i})\log\frac{\mu(A_{n,i})}{\nu(A_{n,i})}, \tag{3.25}$$

and

$$V_n = \int f(x) \log \frac{f_n(x)}{f_n^*(x)} \nu(dx)$$
$$= \sum_{i=1}^{m_n} \mu(A_{n,i}) \log \frac{\mu(A_{n,i})}{(1-a_n)\mu_n(A_{n,i}) + a_n \nu(A_{n,i})}$$
$$= I(\bar{\mu}_n, \mu_n^*). \tag{3.26}$$

Since $\mathscr{P}_n$ is ν-approximating, and ν is a probability measure, we have by Theorem 1 of Csiszár [10] that

$$\lim_{n\to\infty} U_n = 0. \tag{3.27}$$

a) We first prove (3.17). Since V_n is an I-divergence, $V_n \geq 0$ a.s. Therefore, because of (3.27) it suffices to show that

$$\limsup_{n\to\infty} E(V_n) \leq 0. \tag{3.28}$$

By Jensen's inequality

$$E(V_n) = E\left(\int f(x) \log \frac{f_n(x)}{f_n^*(x)} \nu(dx)\right)$$
$$\leq \int f(x) \log \left\{ f_n(x) E\left(\frac{1}{f_n^*(x)}\right)\right\} \nu(dx). \tag{3.29}$$

Moreover, for $x \in A_{n,i}$,

$$E\left(\frac{1}{f_n^*(x)}\right)$$
$$= E\left(\frac{1}{(1-a_n)\mu_n(A_{n,i})/\nu(A_{n,i}) + a_n}\right)$$
$$\leq E\left(\frac{1}{\min\{(1-a_n)/[n\nu(A_{n,i})], a_n\}[n\mu_n(A_{n,i})+1]}\right)$$
$$= \frac{1}{\min\{(1-a_n)/[n\nu(A_{n,i})], a_n\}} E\left(\frac{1}{n\mu_n(A_{n,i})+1}\right)$$
$$\leq \frac{1}{\min\{(1-a_n)/[n\nu(A_{n,i})], a_n\}} \frac{1}{n\mu(A_{n,i})}$$
$$= \frac{1}{\min\{(1-a_n)/n, a_n\nu(A_{n,i})\}} \frac{1}{n\hat{Ef}_n(x)},$$
$$\leq \frac{1}{\min\{(1-a_n)/n, a_n h_n\}} \frac{1}{nf_n(x)}, \tag{3.30}$$

where the second inequality follows from an inequality for the binomial distribution (see Lemma 1 in Appendix B). From (3.29) and (3.30), we get that

$$E(V_n) \leq \log\left(\frac{1}{n\min\{(1-a_n)/n, a_n h_n\}}\right). \tag{3.31}$$

Now, conditions (3.1) and (3.16) imply that

$$\limsup_{n\to\infty} E(V_n) \leq \limsup_{n\to\infty} \log \cdot \left(\frac{1}{n\min\{(1-a_n)/n, a_n h_n\}}\right) \leq 0, \tag{3.32}$$

and thus the proof of (3.17) is complete. □

b) We next prove (3.18). Because of (3.24) and (3.27), it remains to prove that

$$\lim_{n\to\infty} V_n = 0 \quad \text{a.s.} \tag{3.33}$$

Choose

$$c_n = \frac{(\log n)^2}{n}, \tag{3.34}$$

and decompose V_n as

$$V_n = V_n^{(1)} + V_n^{(2)} + V_n^{(3)}, \tag{3.35}$$

where, for $k = 1, 2, 3$,

$$V_n^{(k)} = \sum_{\{i \in I_n^{(k)}\}} \mu(A_{n,i}) \cdot \log \frac{\mu(A_{n,i})}{(1-a_n)\mu_n(A_{n,i}) + a_n \nu(A_{n,i})}, \tag{3.36}$$

with

$$I_n^{(1)} = \{i: c_n < \mu(A_{n,i})\}, \tag{3.37}$$
$$I_n^{(2)} = \{i: b_{n,i} < \mu(A_{n,i}) \leq c_n\}, \tag{3.38}$$
$$I_n^{(3)} = \{i: \mu(A_{n,i}) \leq b_{n,i}\}, \tag{3.39}$$

and

$$b_{n,i} = \min\{a_n \nu(A_{n,i}), c_n\}. \tag{3.40}$$

Now $V_n \geq 0$, whereas $V_n^{(3)} \leq 0$ because $\mu(A_{n,i}) \leq a_n \nu(A_{n,i})$, for all $i \in I_n^{(3)}$. Therefore, (3.33) is proved if for all $\epsilon > 0$

$$\sum_{n=1}^{\infty} P(V_n^{(1)} > \epsilon) < \infty \tag{3.41}$$

and

$$\sum_{n=1}^{\infty} P(V_n^{(2)} > \epsilon) < \infty. \tag{3.42}$$

We first prove (3.41). Choose n sufficiently large so that $-\epsilon/2 < \log(1-a_n) < 0$. Then,

$$P\left(\log \frac{\mu(A)}{(1-a_n)\mu_n(A) + a_n\nu(A)} > \epsilon\right)$$
$$\leq P\left(\log \frac{\mu(A)}{(1-a_n)\mu_n(A)} > \epsilon\right)$$
$$= P\left(\log \frac{\mu(A)}{\mu_n(A)} > \epsilon + \log(1-a_n)\right)$$
$$\leq P\left(\log \frac{\mu(A)}{\mu_n(A)} > \epsilon/2\right). \tag{3.43}$$

From Lemma 4 of Györfi and van der Meulen [19], it follows that

$$P\left(\left|\log \frac{\mu(A)}{\mu_n(A)}\right| > \frac{\epsilon}{2}\right)$$
$$\leq 2\exp\left(-n\mu(A)(1-2^{-\epsilon/2})^2/4\right),$$

and thus, if $\mu(A) > c_n$, that

$$P\left(\left|\log\frac{\mu(A)}{\mu_n(A)}\right| > \frac{\epsilon}{2}\right) < 2\exp\left(-nc_n(1-2^{-\epsilon/2})^2/4\right). \tag{3.44}$$

Hence, observing that $m_n \le 1/h_n$, we have from (3.34), (3.43), and (3.44) that for all $\epsilon > 0$

$$\begin{aligned}
&\sum_{n=1}^{\infty} P(V_n^{(1)} > \epsilon) \\
&\quad \le 2\sum_{n=1}^{\infty} m_n \\
&\qquad \cdot \sup_{\substack{A: \\ \mu(A) > c_n}} P\left(\log\frac{\mu(A)}{(1-a_n)\mu_n(A) + a_n\nu(A)} > \epsilon\right) \\
&\quad \le 2\sum_{n=1}^{\infty} m_n \exp\left(-nc_n(1-2^{-\epsilon/2})^2/4\right) \\
&\quad \le 2\left(\max_i \frac{m_i}{i}\right)\sum_{n=1}^{\infty} \\
&\qquad \cdot \exp\left(-(\log n)^2(1-2^{-\epsilon/2})^2/4 + \ln n\right) < \infty,
\end{aligned} \tag{3.45}$$

and thus, (3.41) is proved. We next prove (3.42), and use thereby the technique of Poissonization (cf. Devroye and Györfi [14, p. 13] and Hoeffding's inequality (Hoeffding [21]). Let the partition $\mathscr{P}_n$ be fixed. Let $X_1, X_2, \cdots$ be an infinite sequence of independent random variables identically distributed according to μ, and let $\hat{N}$ be a Poisson (n) random variable independent of the sequence $\{X_i\}$. Define, for $i = 1, \cdots, m_n$,

$$\check{\mu}_n(A_{n,i}) = \frac{N\left\{j: X_j \in A_{n,i}, 1 \le j \le \hat{N}\right\}}{n}. \tag{3.46}$$

Then, $n\check{\mu}_n(A_{n,1})$, $n\check{\mu}_n(A_{n,2}), \cdots, n\check{\mu}_n(A_{n,m_n})$ are indepen-dent Poisson random variables with means $n\mu(A_{n,1})$, $n\mu(A_{n,2}), \cdots, n\mu(A_{n,m_n})$. Now, put

$$\begin{aligned}
\hat{V}_n^{(2)} = &\sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \\
&\cdot \log\frac{\mu(A_{n,i})}{(1-a_n)\min\{\check{\mu}_n(A_{n,i}), 1\} + a_n\nu(A_{n,i})}.
\end{aligned} \tag{3.47}$$

On the event $\{\hat{N} = n\}$, $\check{\mu}_n(A_{n,i}) = \mu_n(A_{n,i})$, and therefore, on the event $\{\hat{N} = n\}$,

$$V_n^{(2)} \le \hat{V}_n^{(2)}. \tag{3.48}$$

Hence,

$$\begin{aligned}
P(V_n^{(2)} > \epsilon) &= P\left(V_n^{(2)} > \epsilon | \hat{N} = n\right) \\
&\le P\left(\hat{V}_n^{(2)} > \epsilon | \hat{N} = n\right) \\
&\le P\left(\hat{V}_n^{(2)} > \epsilon\right)/P(\hat{N} = n) \\
&\le (2\pi n)^{1/2} e^{1/(12n)} P\left(\hat{V}_n^{(2)} > \epsilon\right).
\end{aligned} \tag{3.49}$$

For $i \in I_n^{(2)}$, we have $\min\{a_n\nu(A_{n,i}), c_n\} = a_n\nu(A_{n,i})$ and

$$\begin{aligned}
&\mu(A_{n,i})\log\frac{\mu(A_{n,i})}{(1-a_n)\min\{\check{\mu}_n(A_{n,i}), 1\} + a_n\nu(A_{n,i})} \\
&\quad \le \mu(A_{n,i})\log\frac{c_n}{a_n h_n} \\
&\quad \le \mu(A_{n,i})O(\log\log n),
\end{aligned} \tag{3.50}$$

and

$$\begin{aligned}
&\mu(A_{n,i})\log\frac{\mu(A_{n,i})}{(1-a_n)\min\{\check{\mu}_n(A_{n,i}), 1\} + a_n\nu(A_{n,i})} \\
&\quad \ge \mu(A_{n,i})\log\frac{a_n\nu(A_{n,i})}{1 + a_n\nu(A_{n,i})} \\
&\quad = -\mu(A_{n,i})\log\left(1 + \frac{1}{a_n\nu(A_{n,i})}\right) \\
&\quad \ge -\mu(A_{n,i})\log\left(1 + \frac{1}{a_n h_n}\right) \\
&\quad = -\mu(A_{n,i})O(\log n).
\end{aligned} \tag{3.51}$$

Therefore, $\hat{V}_n^{(2)}$ is the sum of bounded independent random variables, and $\mu(A_{n,i})O(\log n)$ is a bound on the absolute value of the ith term. Using Hoeffding's inequality (Hoeffding [21, Theorem 2]), we have that

$$\begin{aligned}
P\left(\hat{V}_n^{(2)} > \epsilon\right) &= P\left(\hat{V}_n^{(2)} - E\hat{V}_n^{(2)} > \epsilon - E\hat{V}_n^{(2)}\right) \\
&\le \exp\left(-\frac{2\left(\left[\epsilon - E\hat{V}_n^{(2)}\right]^+\right)^2}{\sum_{i \in I_n^{(2)}} (\mu(A_{n,i}))^2 O(\log n)^2}\right) \\
&\le \exp\left(-\frac{2\left(\left[\epsilon - E\hat{V}_n^{(2)}\right]^+\right)^2}{c_n O(\log n)^2}\right).
\end{aligned} \tag{3.52}$$

Therefore, by (3.34), (3.49), and (3.52), we get

$$P(V_n^{(2)} > \epsilon) \le (2\pi e n)^{1/2}\exp\left(-\frac{2n\left(\left[\epsilon - E\hat{V}_n^{(2)}\right]^+\right)^2}{O(\log n)^4}\right), \tag{3.53}$$

which is summable, i.e., (3.42) holds, if

$$\limsup_{n\to\infty} E\hat{V}_n^{(2)} \le 0. \tag{3.54}$$

It remains to prove (3.54). This can be done in much the same way as in deriving (3.32). The only difference is that in the case of (3.30) we applied an inequality for the binomial distribution, whereas here we use an inequality for the Poisson distribution, i.e.,

$$E\left(\frac{1}{n\check{\mu}_n(A_{n,i}) + 1}\right) \le \frac{1}{n\mu(A_{n,i})}. \tag{3.55}$$

This inequality is also proved in Lemma 1 of Appendix B. Thus proceeding, we have

$$
\begin{aligned}
&E\hat{V}_n^{(2)} \\
&= \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \\
&\quad \cdot E\left(\log \frac{\mu(A_{n,i})}{(1-a_n)\min\{\breve{\mu}_n(A_{n,i}),1\} + a_n \nu(A_{n,i})}\right) \\
&\le \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \\
&\quad \cdot \log E\left(\frac{\mu(A_{n,i})}{(1-a_n)\min\{\breve{\mu}_n(A_{n,i}),1\} + a_n \nu(A_{n,i})}\right) \\
&\le \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \log \left\{\frac{\mu(A_{n,i})}{\min\{(1-a_n)/n, a_n \nu(A_{n,i})\}}\right. \\
&\quad \left. \cdot E\left(\frac{1}{\min\{n\breve{\mu}_n(A_{n,i}), n\} + 1}\right)\right\} \\
&\le \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \log \left\{\frac{\mu(A_{n,i})}{\min\{(1-a_n)/n, a_n h_n\}}\right. \\
&\quad \left. \cdot \left(\frac{1}{n\mu(A_{n,i})} + \frac{1}{n} P(n\breve{\mu}_n(A_{n,i}) > n)\right)\right\} \\
&\le \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \log \left\{\frac{1}{\min\{(1-a_n), na_n h_n\}}\right. \\
&\quad \left. \cdot (1 + P(\breve{\mu}_n(A_{n,i}) > 1))\right\} \\
&\le \sum_{i \in I_n^{(2)}} \mu(A_{n,i}) \\
&\quad \cdot \log \left\{\frac{1}{\min\{(1-a_n), na_n h_n\}}(1+c_n)\right\}, \qquad (3.56)
\end{aligned}
$$

where the last inequality is by Markov's inequality applied to $P(\breve{\mu}_n(A_{n,i}) > 1)$ for cells with $\mu(A_{n,i}) \le c_n$. Hence,

$$
E\hat{V}_n^{(2)} \le \log \left\{\frac{1}{\min\{(1-a_n), na_n h_n\}}(1+c_n)\right\}. \qquad (3.57)
$$

Since $c_n = o(1)$ and (3.16) holds, (3.54) follows from (3.57). □

Remark 4: The immediate question arises whether the standard histogram estimate $\hat{\mu}_n$ (which corresponds to the case $a_n = 0$ in (3.2)) is consistent in divergence. If we apply the decomposition of the divergence $I(\mu, \mu_n^*)$ into $U_n + V_n$ as in the proof of Theorem 2 (cf. (3.24)) for the case $a_n = 0$, then U_n, defined in (3.25), tends to zero (cf. (3.27)). Therefore, it is the behavior of V_n (defined in (3.26)) which will determine the answer to our question. Clearly, if $a_n = 0$, $V_n = \infty$ with positive probability. Therefore, for the standard histogram estimate, $E(I(\mu, \hat{\mu}_n)) = \infty$, and thus obviously we cannot have consistency in expected information divergence when using this density estimate. We now show by example that the standard, unmodified, histogram estimate cannot be a.s. consistent in information divergence either. Let ν be the uniform distribution on $[0,1]$. Assume $\gamma > 1$ such that

$$
\lim_{n \to \infty} nh_n^{\gamma} = 0. \qquad (3.58)
$$

Such $\gamma > 1$ exists if $h_n = O(n^{-\alpha})$, $0 < \alpha < 1$. Define a density f of μ with respect to ν on $[0,1]$ by

$$
f(x) = \gamma x^{\gamma - 1},
$$

and let

$$
A_{n,i} = [(i-1)h_n, ih_n).
$$

Then,

$$
\mu(A_{n,1}) = h_n^{\gamma}. \qquad (3.59)
$$

Now, setting $a_n = 0$ in the definition of V_n, we have for any $\epsilon > 0$ that

$$
\begin{aligned}
P(V_n > \epsilon) &= P\left(\sum_i \mu(A_{n,i}) \log \frac{\mu(A_{n,i})}{\mu_n(A_{n,i})} > \epsilon\right) \\
&\ge P\left(\sum_i \mu(A_{n,i}) \log \frac{\mu(A_{n,i})}{\mu_n(A_{n,i})} = \infty\right) \\
&\ge P(\mu_n(A_{n,1}) = 0) = (1 - \mu(A_{n,1}))^n \\
&= (1 - h_n^{\gamma})^n \ge 1 - nh_n^{\gamma} \to 1, \qquad \text{as } n \to \infty. \qquad (3.60)
\end{aligned}
$$

Hence V_n does not converge to zero in probability, and a fortiori does not converge to zero almost surely. This proves that the standard histogram-based distribution estimate $\hat{\mu}_n$ cannot be a.s. consistent in information divergence, thus providing a negative answer to the question previously raised.

Remark 5: An interpretation of our modified histogram-based distribution estimator μ_n^* may be given in terms of Bayes rule. Consider the restriction of the distribution to the partition $\mathcal{P}_n$. Suppose a prior distribution is assigned to the simplex of cell probabilities $\mu(A_{n,1}), \cdots, \mu(A_{n,m_n})$ that takes the form of a Dirichlet distribution with parameters $\lambda_{n,1}, \cdots, \lambda_{n,m_n}$. Then, the posterior distribution of these cell probabilities given the data is Dirichlet with parameters $n\mu(A_{n,i}) + \lambda_{n,i}$ for $i = 1, 2, \cdots, m_n$ (see, e.g., Ferguson [16]). In particular, the Bayes estimator of the cell probabilities (given by the mean of the posterior distribution) is

$$
\mu_n^{**}(A_{n,i}) = \frac{n\mu_n(A_{n,i}) + \lambda_{n,i}}{n + b_n}, \qquad (3.61)
$$

for $i = 1, 2, \cdots, m_n$, where $b_n = \sum_{i=1}^{m_n} \lambda_{n,i}$. Note that when $\lambda_{n,i}$ is an integer, the Bayes estimator may be interpreted as the relative frequency of $A_{n,i}$ with $\lambda_{n,i}$ additional (fictitious) observations in each cell. The modified histogram estimator given by

$$
f_n^{**}(x) = \frac{\mu_n^{**}(A_{n,i})}{\nu(A_{n,i})}, \qquad (3.62)
$$

if $x \in A_{n,i}$, defines a density with respect to ν, which is the Bayes estimator for a Bayesian who takes the prior distribution of the probabilities of cells in $\mathcal{P}_n$ to be Dirichlet and who takes the conditional distribution within each cell to be fixed and equal to that given by ν.

A choice of parameters $\lambda_{n,i}$ proportional to $\nu(A_{n,i})$, i.e.,

$$\lambda_{n,i} = b_n \nu(A_{n,i}), \tag{3.63}$$

makes the Bayesian estimator f_n^{**} the same as the modified histogram estimator f_n^* (cf. (3.5)). The relationship between the constants $0 < a_n < 1$ and $b_n > 0$ is

$$a_n = \frac{b_n}{n + b_n}. \tag{3.64}$$

For the choices $\nu(A_{n,i}) = 1/m_n$ and $b_n = m_n$, we obtain $\lambda_{n,i} = 1$, the prior becomes the uniform distribution over the simplex of cell probabilities, and the density estimator becomes the one previously considered, with a_n given as in (3.8). This choice of prior distribution leads to some appealing universal data compression properties, first considered by Gilbert [18], see also Cover [8]. (Advantageous properties of the Dirichlet prior with parameters $\lambda_{n,i} = 1/2$ instead of 1 are given in Krichevsky and Trofimov [24]. Unfortunately, condition (3.16) effectively requires that $\lambda_{n,i} \geq 1$ for the application of our theorem.)

Remark 6: Some observations concerning the rate of convergence in expected information divergence (made by Barron [3]) follow from the proof of part a) of Theorem 2 in the case that $\nu(A_{n,i}) = h_n = 1/m_n$ and $a_n = m_n/(n + m_n)$ as in (3.8). In this case, $f_n^*(x)$ takes the form

$$p_n^*(x) = \frac{m_n}{n + m_n}(n\mu_n(A_{n,i}) + 1), \qquad x \in A_{n,i}.$$

Now, setting again $f = d\mu/d\nu$, the reasoning which led to the derivation of inequality (3.31) can be simplified as follows:

$$\begin{aligned} E(D(f, p_n^*)) &= E\left(\int f \log \frac{f}{p_n^*}\, d\nu\right) \\ &\leq \int f \log \left(fE\left(\frac{1}{p_n^*}\right)\right) d\nu \\ &= \sum_i \int_{A_{n,i}} f(x) \\ &\quad \cdot \log \left(f(x)E\left(\frac{n + m_n}{m_n(n\mu_n(A_{n,i}) + 1)}\right)\right)\nu(dx) \\ &\leq \int f(x) \log \left(f(x)\frac{n + m_n}{nf_n(x)}\right)\nu(dx) \\ &= \log\left(1 + \frac{m_n}{n}\right) + D(f, f_n), \end{aligned} \tag{3.65}$$

where $f_n(x)$ is the histogram-shaped density (3.22) based on the true probabilities μ. By (3.11), the first term on the right-hand side of (3.65) tends to zero.

The approximation error term $D(f, f_n)$ is examined in Barron and Sheu [6] in the case of a partition of $[0, 1]$ into equal-spaced intervals of width $\nu(A_{n,i}) = h_n = 1/m_n$, where ν is taken there to be Lebesgue measure. (Barron and Sheu examine exponential family models based on splines, polynomials, and trigonometric series. Histograms arise in the case of splines of order 0.) It is shown there that $D(f, f_n)$ is of order $O(1/m_n)^2$ for density functions on $[0, 1]$ for which the derivative of $\log f(x)$ is square integrable with respect to Lebesgue measure. For such densities, we have that the modified histogram estimator p_n^* satisfies

$$\begin{aligned} E(D(f, p_n^*)) &\leq \log\left(1 + \frac{m_n}{n}\right) + D(f, f_n) \\ &\leq \frac{m_n}{n} + O\left(\frac{1}{m_n}\right)^2. \end{aligned} \tag{3.66}$$

In particular, choosing m_n of order $n^{1/3}$ to optimize the order of the bound, we get for log-densities with square-integrable derivatives that

$$E(D(f, p_n^*)) = O\left(\frac{1}{n}\right)^{2/3}. \tag{3.67}$$

Using this bound on the expected informational divergence, order $O(1/n)^{2/3}$ bounds may also be given for the redundancy of universal codes (which turns out to have a representation as a Cesàro average of expected informational divergences as discussed below). A somewhat different proof of the same bounds on expected informational divergence and redundancy for the modified histogram, also based on inequalities from Barron and Sheu [6], are given in Rissanen, Speed, and Yu [26].

A. Implications for Universal Source Coding of Finely Partitioned Data

We point out here how consistent estimation of a probability distribution in expected informational divergence $E(I(\mu, \mu_n^*))$ (cf. (3.17)) leads naturally to a universal source code for arbitrarily fine quantizations of the data. Similar applications were noted in Clarke and Barron [7], but the examples there were limited to finite-dimensional parametric families.

Let $\{0, 1\}^*$ be the set of finite length binary strings and let $|s|$ denote the length of a string $s \in \{0, 1\}^*$. Recall (cf. Gallager [17]) that for any countable alphabet $\mathcal{A}$, there exists a unique decodable code $\phi: \mathcal{A} \to \{0, 1\}^*$ with lengths $|\phi(a)|$, $a \in \mathcal{A}$, if and only if the Kraft–McMillan inequality is satisfied, i.e.,

$$\sum_{a \in \mathcal{A}} 2^{-|\phi(a)|} \leq 1.$$

If $\mathcal{X}$ is not a discrete space, data sequences $X_1, \cdots, X_n$ cannot be represented exactly by a noiseless source code. Nevertheless, for any given partition of $\mathcal{X}^n$ (no matter how fine), we can code the element of the partition that includes the data in a uniquely decodable way.

For any given probability distribution η_n on $\mathcal{X}^n$, the Shannon code for elements of a partition $\{A_{n,i}\}$ assigns a codeword of length

$$|\phi(A_{n,i})| = \lceil \log 1/\eta_n(A_{n,i}) \rceil. \tag{3.68}$$

Suppose $X_1, \cdots, X_n$ are independently drawn from an unknown probability distribution μ on the space $\mathcal{X}$. The lack of knowledge of the distribution leads in some cases to a nonnegligible redundancy of the code. The redundancy of a code is defined as the difference between the expected length and the entropy, divided by the sample size n. In the present case of description of elements of a partition, the redundancy is given by

$$\frac{1}{n}\sum_{i=1}^{m_n} \mu^n(A_{n,i})|\phi(A_{n,i})| - \frac{1}{n}\sum_{i=1}^{m_n} \mu^n(A_{n,i}) \log (1/\mu^n(A_{n,i})). \tag{3.69}$$

The effect of rounding up to the nearest integer in (3.68) on the redundancy is bounded by $1/n$. Ignoring the rounding effect, the redundancy of the Shannon code based on η_n for elements of a partition $\mathcal{P}_n = \{A_{n,i}\}$ of $\mathcal{X}^n$ is given by

$$\begin{aligned} R_n(\mathcal{P}_n) &= \frac{1}{n}\sum_{i=1}^{m_n} \mu^n(A_{n,i}) \log \frac{1}{\eta_n(A_{n,i})} \\ &\quad - \frac{1}{n}\sum_{i=1}^{m_n} \mu^n(A_{n,i}) \log \frac{1}{\mu^n(A_{n,i})} \\ &= \frac{1}{n}\sum_{i=1}^{m_n} \mu^n(A_{n,i}) \log \frac{\mu^n(A_{n,i})}{\eta_n(A_{n,i})}, \end{aligned} \tag{3.70}$$

which we recognize as $1/n$ times the information divergence between the distribution μ^n and η_n restricted to the partition $\mathcal{P}_n = \{A_{n,i}\}$ of $\mathcal{X}^n$. Taking the supremum over partitions of $\mathcal{X}^n$ yields the least upper bound on the redundancy denoted by

$$R_n = \sup_{\mathcal{P}_n} R_n(\mathcal{P}_n). \tag{3.71}$$

Thus, using definition (1.3), the redundancy is given in terms of the informational divergence by

$$R_n = \frac{1}{n} I(\mu^n, \eta_n). \tag{3.72}$$

As in Davisson [12], a code based on a given sequence η_n is said to be universal for a class of distributions if $\lim_{n\to\infty} R_n = 0$ for all distributions μ in this class. Because of (1.11), no universal code exists for the class of all distributions.

Now, we construct a universal code for a large class of distributions. The construction is based on our distribution estimate (3.2). If, as in (1.1), we are given a sequence of estimators of a distribution on $\mathcal{X}$, which we denote here as

$$\mu_n^*(A) = \mu_n^*(A|X_1, \cdots, X_n), \tag{3.73}$$

then in order to construct a distribution η_n on $\mathcal{X}^n$, we define η_n by its conditional distributions $\eta_{k+1}(X_{k+1} \in A|X_1, \cdots, X_k) = \mu_k^*(A)$, for $k = 1, 2, \cdots, n-1$, and by the distribution of X_1 given by $\eta_1 = \mu_0^* = \nu$.

This distribution on $\mathcal{X}^n$ that we have constructed from a sequence of estimators of distributions on $\mathcal{X}$, may be used to provide the Shannon codes for partitions of $\mathcal{X}^n$. By the chain rule for the information divergence (cf. [11, p. 50]), the redundancy R_n is equal to the Cesàro average of expected information divergences:

$$R_n = \frac{1}{n} I(\mu^n, \eta_n) = \frac{1}{n}\sum_{k=0}^{n-1} E(I(\mu, \mu_k^*)). \tag{3.74}$$

If $\lim_{n\to\infty} E(I(\mu, \mu_n^*)) = 0$, then the Cesàro average in (3.74) must also converge to zero. Thus, we have the desired conclusion that if a sequence of distribution estimators is consistent in expected information divergence, then the redundancy of the code tends to zero as $n \to \infty$.

Theorem 2 gives conditions such that a sequence of estimators μ_n^* converges to μ, in the sense that $\lim_{n\to\infty} E(I(\mu, \mu_n^*)) = 0$, for all μ for which $I(\mu, \nu) < \infty$, for a given probability measure ν which serves the role of the dominating measure. Consequently, the redundancy of the code based on this sequence of estimators converges to zero for all such μ. Thus, the estimator μ_n^* defined in (3.2) provides a universal code for the class of all measures μ with $I(\mu, \nu) < \infty$.

IV. Consistency in Reversed Order Information Divergence

In this section, we consider the problem of consistency of the histogram estimator with respect to an informational divergence criterion in which we reverse the order of the estimating and the hypothetical distribution. Suppose again that $X_1, \cdots, X_n$ are i.i.d. observations from an unknown distribution with probability measure μ taking values in an arbitrary measurable space $\mathcal{X}$. Assume also that there exists a known σ-finite measure ν such that μ is absolutely continuous with respect to ν. Let $f(x)$ denote the density of μ with respect to ν. Let $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \cdots, A_{n,m_n}\}$, $n = 1, 2, \cdots$, be a sequence of partitions of $\mathcal{X}$, with m_n either finite or infinite, such that $0 < \nu(A_{n,i}) < \infty$ for each i. Let μ_n be the standard empirical measure for $X_1, \cdots, X_n$, as defined in (1.8). We also define, as in (3.3), the standard histogram density estimator with respect to ν and $\mathcal{P}_n$ by

$$\hat{f}_n(x) = \frac{\mu_n(A_{n,i})}{\nu(A_{n,i})}, \qquad \text{if } x \in A_{n,i}, \tag{4.1}$$

and the corresponding estimator $\hat{\mu}_n$ of the distribution μ by

$$\hat{\mu}_n(A) = \int_A \hat{f}_n(x)\nu(dx). \tag{4.2}$$

We are interested in studying the consistency of $\hat{\mu}_n$ as the estimator of μ with $I(\hat{\mu}_n, \mu)$ as the informational divergence criterion. If for an arbitrary estimator $\mu_n^\star$ we want

$I(\mu_n^\star, \mu) < \infty$, then $\mu_n^\star$ must be absolutely continuous with respect to μ, which is not only a condition on $\mu_n^\star$ but a strict condition on μ. However under this strict condition (to be specified later) it is enough to consider the most simple estimator, the histogram $\hat{\mu}_n$, rather than involved modifications of it. This contrary to the cases of estimating μ in total variation or in I-divergence $I(\mu, \hat{\mu}_n)$, considered in Sections II and III, where $\hat{\mu}_n$ needed to be modified to $\bar{\mu}_n$ and μ_n^*, respectively.

We furthermore define, as in (3.22) and (3.23), $f_n(x) = E\hat{f}_n(x)$, and

$$\bar{\mu}_n(A) = \int_A f_n(x)\nu(dx). \tag{4.3}$$

We then have

$$\begin{aligned} I(\hat{\mu}_n, \mu) &= D(\hat{f}_n, f) = \int \hat{f}_n(x) \log \frac{\hat{f}_n(x)}{f(x)} \nu(dx) \\ &= \int \hat{f}_n(x) \log \frac{\hat{f}_n(x)}{f_n(x)} \nu(dx) \\ &\quad + \int \hat{f}_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx) \\ &= D(\hat{f}_n, f_n) + \int \hat{f}_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx) \\ &= I(\hat{\mu}_n, \bar{\mu}_n) + \int \hat{f}_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx) \\ &= I(\hat{\mu}_n, \bar{\mu}_n) + W_n. \end{aligned} \tag{4.4}$$

The first term in (4.4) is called the estimation error (or "variance") component of the reversed order informational divergence. Convergence properties of it are given in Theorem 3. The expected value of the second term W_n is called the approximation error (or "bias") component of $I(\hat{\mu}_n, \mu)$. The convergence properties of this bias component are studied after Theorem 3. We first examine the convergence properties of $D(\hat{f}_n, f_n)$.

Given a partition $\mathscr{P}_n = \{A_{n,1}, A_{n,2}, \cdots, A_{n,m_n}\}$, we define the informational divergence between the empirical distribution μ_n and the true distribution μ, restricted to $\mathscr{P}_n$, by

$$I_n(\mu_n, \mu) = \sum_{i=1}^{m_n} \mu_n(A_{n,i}) \log \frac{\mu_n(A_{n,i})}{\mu(A_{n,i})}. \tag{4.5}$$

Thus, if $\mu \ll \nu$, and $\hat{f}_n$, f_n, $\hat{\mu}_n$, and $\bar{\mu}_n$ are as defined in (4.1)–(4.3), then $I_n(\mu_n, \mu) = I(\hat{\mu}_n, \bar{\mu}_n) = D(\hat{f}_n, f_n)$. Let m_n be the total number of sets in $\mathscr{P}_n$, either finite or infinite, and denote the number of sets in $\mathscr{P}_n$ that are contained in a given set A by

$$m_n(A) = N\{B \in \mathscr{P}_n : B \subset A\}. \tag{4.6}$$

Furthermore, given a certain partition $\mathscr{P}$, we define the entropy of μ restricted to $\mathscr{P}$ by

$$H^{\mathscr{P}}(\mu) = -\sum_{A \in \mathscr{P}} \mu(A) \log \mu(A). \tag{4.7}$$

The average number of sets in $\mathscr{P}_n$ that are contained in a set A of $\mathscr{P}$ is defined by

$$m_n(\mu, \mathscr{P}) = \sum_{A \in \mathscr{P}} m_n(A)\mu(A). \tag{4.8}$$

The first claim in the next theorem requires the total number of sets m_n in each partition of a sequence of partitions $\mathscr{P}_n$ to be finite, and obtains almost sure convergence of $I_n(\mu_n, \mu)$, as well as of the expected value $E(I_n(\mu_n, \mu))$. The second claim allows for infinitely many sets in each partition $\mathscr{P}_n$, provided the average number $m_n(\mu, \mathscr{P})$ is finite, but only convergence of $E(I_n(\mu_n, \mu))$ is established in that case.

Theorem 3: Let μ be a probability measure on $\mathfrak{X}$, suppose $X_1, \cdots, X_n$ are i.i.d. observations from μ, and let μ_n be the standard empirical measure for $X_1, \cdots, X_n$. Let $\{\mathscr{P}_n\}$ be a sequence of partitions of $\mathfrak{X}$.

a) If the cardinality m_n of $\mathscr{P}_n$ satisfies

$$\lim_{n \to \infty} \frac{m_n}{n} = 0, \tag{4.9}$$

then

$$\lim_{n \to \infty} I_n(\mu_n, \mu) = 0 \quad \text{a.s.} \tag{4.10}$$

and

$$\lim_{n \to \infty} E(I_n(\mu_n, \mu)) = 0. \tag{4.11}$$

b) More generally, if each partition $\mathscr{P}_n$ is a refinement of a certain partition $\mathscr{P}$, if

$$H^{\mathscr{P}}(\mu) < \infty, \tag{4.12}$$

and if the average number of sets $m_n(\mu, \mathscr{P})$ satisfies

$$\lim_{n \to \infty} \frac{m_n(\mu, \mathscr{P})}{n} = 0, \tag{4.13}$$

then,

$$\lim_{n \to \infty} E(I_n(\mu_n, \mu)) = 0. \tag{4.14}$$

Remark 7: For $\mathfrak{X} = \mathbb{R}$, suppose $\mathscr{P}_n = \{A_{n,i}, i = \cdots, -1, 0, 1, \cdots\}$ is a partition of $\mathbb{R}$ into intervals $A_{n,i} = [ih_n, (i+1)h_n)$ of common length $h_n = 1/k_n$, where k_n is a sequence of integers. Let $\mathscr{P} = \{A_i, i = \cdots, -1, 0, 1, \cdots\}$ be the partition of $\mathbb{R}$ consisting of unit intervals $A_i = [i, i+1)$. Then $\mathscr{P}_n$ is a refinement of $\mathscr{P}$, condition (4.13) becomes $k_n = o(n)$, and $H^{\mathscr{P}}(\mu)$ can be interpreted as the entropy of a probability mass function on the integers. Note that the second set of conditions in Theorem 3 is more general than the first one. This can be seen by taking $\mathscr{P}$ to be the trivial partition $\{\mathfrak{X}, \varnothing\}$. In this case $H^{\mathscr{P}}(\mu) = 0$ and $m_n(\mu, \mathscr{P}) = m_n$ for all μ. Hence, for this choice of $\mathscr{P}$, (4.12) is always satisfied and (4.13) implies (4.9).

Proof of Theorem 3

a) We first take the case of a finite partition of cardinality $m = m_n$. The first step in the proof follows Barron [4, pp. 112–113] and is based on the method of

types (Csiszár and Körner [11, Section 1.2]) and an inequality due to Hoeffding [22, (2.8)].

If $\mathscr{P}_n = \{A_{n,1}, \cdots, A_{n,m}\}$, let $\mu_n^{\mathscr{P}_n} = \{\mu_n(A_{n,i}): A_{n,i} \in \mathscr{P}_n\}$ denote the empirical distribution of the sample $\underset{\sim}{X} = (X_1, \cdots, X_n)$ restricted to $\mathscr{P}_n$. Then $\mu_n^{\mathscr{P}_n}$ is also the type Q of a sequence in $\mathscr{P}_n^n$. Clearly, the number of types Q of sequences in $\mathscr{P}_n^n$ equals $\binom{n+m-1}{m-1}$. By Hoeffding's inequality, it holds for any type Q of a sequence in $\mathscr{P}_n^n$ that

$$P\{\mu_n^{\mathscr{P}_n} = Q\} \le 2^{-n I_n(Q,\mu)}. \tag{4.15}$$

Therefore, given $\delta > 0$,

$$\begin{aligned} P\{I_n(\mu_n, \mu) \ge \delta\} &= \sum_{\{Q: I_n(Q,\mu) \ge \delta\}} P\{\mu_n^{\mathscr{P}_n} = Q\} \\ &\le \sum_{\{Q: I_n(Q,\mu) \ge \delta\}} 2^{-n I_n(Q,\mu)} \\ &\le \binom{n+m}{m} 2^{-n\delta} \\ &\le 2^{-n(\delta - \epsilon_n)}, \end{aligned} \tag{4.16}$$

where $\epsilon_n = (1 + m/n)h(m/(n+m))$ with $h(p) = -p \log p - (1-p)\log(1-p)$. Clearly $\epsilon_n \to 0$ as $m/n \to 0$.

It follows that $P\{I_n(\mu_n, \mu) \ge \delta\}$ converges to zero exponentially as $n \to \infty$ under condition (4.9). Conclusion (4.10) follows by the Borel–Cantelli lemma, and conclusion (4.11) results after carrying out the following integration:

$$\begin{aligned} E(I_n(\mu_n, \mu)) &= \int_0^\infty P\{I_n(\mu_n, \mu) \ge \delta\}\, d\delta \\ &\le \epsilon_n + \int_{\epsilon_n}^\infty 2^{-n(\delta - \epsilon_n)}\, d\delta \\ &\le \epsilon_n + \frac{1}{n \ln 2}. \end{aligned} \tag{4.17}$$

b) Next consider the case of a sequence of partitions $\mathscr{P}_n$ satisfying (4.12) and (4.13). We obtain the following sequence of bounds. First,

$$E(I_n(\mu_n, \mu)) = E\left(\sum_{i=1}^{m_n} \mu_n(A_{n,i}) \log \frac{\mu_n(A_{n,i})}{\mu(A_{n,i})}\right) \tag{4.18}$$

$$= \sum_{i=1}^{m_n} E\left(\mu_n(A_{n,i}) \log \frac{\mu_n(A_{n,i})}{\mu(A_{n,i})}\right)$$

$$\le \sum_{i=1}^{m_n} \mu(A_{n,i}) \log\left(1 + \frac{1}{n\mu(A_{n,i})}\right). \tag{4.19}$$

Here, the exchange of expectation and summation is by the Fubini–Tonelli theorem applied to $Z_{n,i} = \mu_n(A_{n,i}) \log \mu_n(A_{n,i})/\mu(A_{n,i}) + \mu_n(A_{n,i}) - \mu(A_{n,i})$, which is nonnegative and has the same sum as in (4.18). The inequality (4.19) is by an inequality for the binomial distribution (see Lemma 2 in Appendix B). Furthermore,

$$\begin{aligned} &\sum_{i=1}^{m_n} \mu(A_{n,i}) \log\left(1 + \frac{1}{n\mu(A_{n,i})}\right) \\ &= \sum_{A \in \mathscr{P}} \sum_{A_{n,i} \subset A} \mu(A_{n,i}) \log\left(1 + \frac{1}{n\mu(A_{n,i})}\right) \\ &= \sum_{A \in \mathscr{P}} \mu(A) \sum_{A_{n,i} \subset A} \mu(A_{n,i}|A) \log\left(1 + \frac{1}{n\mu(A_{n,i})}\right) \\ &\le \sum_{A \in \mathscr{P}} \mu(A) \log\left(1 + \frac{m_n(A)}{n\mu(A)}\right). \end{aligned} \tag{4.20}$$

The inequality (4.20) follows form the concavity of the logarithm and Jensen's inequality, since $\sum_{A_{n,i} \subset A} \mu(A_{n,i} \mid A)(1 + 1/(n\mu(A_{n,i}))) = 1 + m_n(A)/(n\mu(A))$. Next, we use the inequality

$$\left(1 + \frac{x}{y}\right) \le (1 + x) \max\left(\frac{1}{y}, 1\right) \tag{4.21}$$

which is valid for $x > 0$, $y > 0$.

Taking into account (4.18)–(4.20) and applying (4.21), we obtain that for any $c > 1$,

$$\begin{aligned} &E(I_n(\mu_n, \mu)) \\ &\le \sum_{A \in \mathscr{P}} \mu(A) \log\left(1 + \frac{m_n(A)}{n\mu(A)}\right) \\ &\le \sum_{A \in \mathscr{P}} \mu(A) \log\left[\left(1 + \frac{cm_n(A)}{n}\right)\right. \\ &\quad \left.\cdot \max\left(\frac{1}{c\mu(A)}, 1\right)\right] \\ &\le \sum_{A \in \mathscr{P}} \mu(A) \log\left(1 + \frac{cm_n(A)}{n}\right) \\ &\quad + \sum_{A \in \mathscr{P}} \mu(A) \left(\log \frac{1}{\mu(A)}\right) 1_{\{c\mu(A) < 1\}} \\ &\le \log\left(1 + \frac{cm_n(\mu, \mathscr{P})}{n}\right) \\ &\quad + \sum_{A \in \mathscr{P}} \mu(A) \left(\log \frac{1}{\mu(A)}\right) 1_{\{c\mu(A) < 1\}}, \end{aligned} \tag{4.22}$$

where in the last inequality we have used again the concavity of the logarithm and Jensen's inequality.

Since $H^{\mathscr{P}}(\mu) < \infty$ is assumed, it follows by the dominated convergence theorem that the second term in the bound in (4.22) tends to zero for any sequence $c = c_n$ such that $c_n \to \infty$. Finally, since $m_n(\mu, \mathscr{P})/n \to 0$ by assumption (4.13), the bound in (4.22) tends to zero for any sequence $c = c_n$ chosen to satisfy $c_n \to \infty$ and $c_n m_n(\mu, \mathscr{P})/n \to 0$ as $n \to \infty$. Hence, conclusion (4.14) follows and the proof of Theorem 3 is completed.

We next turn to the investigation of the convergence properties of the bias component (cf. (4.4))

$$E(W_n) = E\left(\int \hat{f}_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx)\right). \quad (4.23)$$

In Theorem 4, we will derive sufficient conditions under which $I(\bar{\mu}_n, \mu)$ tends to zero, where

$$0 \le I(\bar{\mu}_n, \mu) = \int f_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx). \quad (4.24)$$

It then follows by the Fubini theorem that under the same conditions

$$\lim_{n\to\infty} E(W_n) = \lim_{n\to\infty} I(\bar{\mu}_n, \mu) = 0. \quad (4.25)$$

Remark 8: Unlike the case of distribution estimation with the criterion $I(\mu, \mu_n^*)$, as dealt with in Section III, for which a.s. convergence to zero obtains (for a proper choice of μ_n^*) for all μ with $I(\mu, \nu) < \infty$, more restrictive conditions on μ are required for the convergence to zero of $I(\hat{\mu}_n, \mu)$. Indeed, from inspection of the integral in (4.24) it is seen that, in order for (4.25) to hold, it is clearly necessary that the function $\log(1/f(x))$ be integrable with respect to ν on each set $A_{n,i}$ of $\mathcal{P}_n$ that has positive μ-probability (and this for all large n). Thus, a restriction that is necessarily imposed is that boundaries of the support set of μ must coincide with boundaries of the cells of the partition, for all large n. For, if there is a set $A_{n,i}$ in $\mathcal{P}_n$ for which $f(x)$ is zero on a nontrivial portion of this set and $f(x) > 0$ on the remaining portion (so that $f_n(x) > 0$ on the entire set), then $I(\bar{\mu}_n, \mu) = D(f_n, f) = \infty$.

Remark 9: The necessary condition of Remark 8 rules out many simple examples. For instance, if μ is the uniform distribution on $[0, a]$ with $a < 1$ unknown irrational, ν is the Lebesgue measure on $[0, 1]$, and $\mathcal{P}_n = \{[ih_n, (i+1)h_n)\colon i = \cdots, -1, 0, 1, \cdots\}$ with h_n rational, then $D(f_n, f) = \infty$ for all $h_n < a$. Before formulating Theorem 4, we note that $I(\bar{\mu}_n, \mu) = D(f_n, f)$ can be decomposed as

$$D(f_n, f) = \int f_n(x) \log f_n(x) \nu(dx) - \int f_n(x) \log f(x) \nu(dx). \quad (4.26)$$

Now suppose that $\mathcal{X} = \mathbb{R}^d$, and ν is Lebesgue measure. Also suppose that the density $f(x) = d\mu/d\nu$ is continuous, positive, and of known rectangular compact support. Denoting this support set by C, let $\mathcal{P} = \{C, \mathcal{X} - C\}$ be an initial partition of $\mathcal{X}$. Let $\mathcal{P}_n$ be a sequence of partitions of $\mathcal{X}$ into cells, such that each $\mathcal{P}_n$ is a refinement of $\mathcal{P}$ and the maximum width of the cells tends to zero as $n \to \infty$. Furthermore suppose the Lebesgue (and Lebesgue-Stieltjes) integral

$$H_\nu(\mu) = -\int_{\mathcal{X}} f(x) \log f(x) \nu(dx) = -\int_{\mathcal{X}} (\log f(x)) \mu(dx) \quad (4.27)$$

is finite. It can then be shown by the theory of Riemann–Stieltjes integration that both terms in (4.26) tend to $-H_\nu(\mu)$ as $n \to \infty$. Therefore, in the case that $\mathcal{X} = \mathbb{R}^d$, and $f(x)$ is continuous, positive, and of known bounded rectangular support, and $|H_\nu(\mu)| < \infty$, one has that

$$\lim_{n\to\infty} D(f_n, f) = 0. \quad (4.28)$$

The reason that this result does not apply to the above example (μ uniform on $[0, a]$, $a < 1$ unknown irrational) is that a and thus the support are unknown in this case.

We now present a result on the convergence of $D(f_n, f)$ to zero, if the underlying space $\mathcal{X}$ is not necessarily $\mathbb{R}^d$, and $f(x)$ is not necessarily continuous.

Theorem 4: Let $\mathcal{X}$ be an arbitrary measurable space. Let μ be an unknown probability measure and ν a given σ-finite measure on $\mathcal{X}$, such that $\mu \ll \nu$. Let $\mathcal{P}$ be a given partition of $\mathcal{X}$ and define $\bar{f}(x) = \mu(A_i)/\nu(A_i)$ if $x \in A_i \in \mathcal{P}$ and $\bar{\mu}(A) = \int_A \bar{f}(x)\nu(dx)$. Suppose that each partition $\mathcal{P}_n$ is a refinement of $\mathcal{P}$, that $f(x)/\bar{f}(x) = d\mu/d\bar{\mu}$ is bounded and that

$$I(\bar{\mu}, \mu) < \infty. \quad (4.29)$$

Moreover, suppose that the sequence of partitions $\mathcal{P}_n$ is ν-approximating. Then,

$$\lim_{n\to\infty} I(\bar{\mu}_n, \mu) = 0, \quad (4.30)$$

with $I(\bar{\mu}_n, \mu)$ as defined in (4.24).

Remark 10: Note that, since

$$I(\bar{\mu}, \mu) = \int \bar{f}(x) \log(\bar{f}(x)/f(x)) \nu(dx),$$

the condition $I(\bar{\mu}, \mu) < \infty$ necessitates that $f(x) > 0$ ν-almost everywhere within each set A_i in $\mathcal{P}$ that has positive μ-probability, in accordance with the point made in Remark 8.

Theorem 4 yields the following corollary when ν is assumed to be a probability measure and we take for $\mathcal{P}$ the trivial partition $\{\mathcal{X}, \varnothing\}$. In this case it is assumed that $f(x)$ is bounded and $I(\nu, \mu) < \infty$.

Corollary: Let $\mathcal{X}$ be an arbitrary measurable space. Let μ be an unknown probability measure and ν a given probability measure on $\mathcal{X}$, such that $\mu \ll \nu$ and the density $f(x) = d\mu/d\nu$ is bounded. Suppose that

$$I(\nu, \mu) < \infty \quad (4.31)$$

and that the sequence of partitions $\mathcal{P}_n$ is ν-approximating. Then,

$$\lim_{n\to\infty} I(\bar{\mu}_n, \mu) = 0. \quad (4.32)$$

Proof of Theorem 4: Suppose $f(x)/\bar{f}(x) \le c$. Then, for each measurable set A

$$\mu(A) = \int_A \frac{f(x)}{\bar{f}(x)} \bar{\mu}(dx) \le c\bar{\mu}(A). \tag{4.33}$$

The informational divergence between $\bar{\mu}_n$ and μ can be written as

$$\begin{aligned} I(\bar{\mu}_n, \mu) &= \int f_n(x) \log \frac{f_n(x)}{f(x)} \nu(dx) \\ &= \int \frac{f_n(x)}{\bar{f}(x)} \log \frac{f_n(x)/\bar{f}(x)}{f(x)/\bar{f}(x)} \bar{\mu}(dx) \\ &= \sum_{i=1}^{m_n} \frac{\mu(A_{n,i})}{\bar{\mu}(A_{n,i})} \int_{A_{n,i}} \\ &\quad \cdot \log \frac{\mu(A_{n,i})/\bar{\mu}(A_{n,i})}{f(x)/\bar{f}(x)} \bar{\mu}(dx). \end{aligned} \tag{4.34}$$

Notice that, by convexity, for each measurable set A,

$$\int_A \log(\bar{f}(x)/f(x)) \bar{\mu}(dx) \ge \int_A \left(\log \frac{\bar{\mu}(A)}{\mu(A)} \right) \bar{\mu}(dx),$$

so that the summands in (4.34) are nonnegative. Hence,

$$\begin{aligned} I(\bar{\mu}_n, \mu) &\le \sum_{i=1}^{m_n} c \int_{A_{n,i}} \log \frac{\mu(A_{n,i})/\bar{\mu}(A_{n,i})}{f(x)/\bar{f}(x)} \bar{\mu}(dx) \\ &= c \sum_{i=1}^{m_n} \left(\int_{A_{n,i}} \log \frac{\bar{f}(x)}{f(x)} \bar{\mu}(dx) \right. \\ &\quad \left. + \bar{\mu}(A_{n,i}) \log \frac{\mu(A_{n,i})}{\bar{\mu}(A_{n,i})} \right) \\ &= c \left(\int \log \frac{\bar{f}(x)}{f(x)} \bar{\mu}(dx) \right. \\ &\quad \left. - \sum_{i=1}^{m_n} \bar{\mu}(A_{n,i}) \log \frac{\bar{\mu}(A_{n,i})}{\mu(A_{n,i})} \right) \\ &= c(I(\bar{\mu}, \mu) - I_n(\bar{\mu}, \mu)). \end{aligned} \tag{4.35}$$

Now, by Theorem 1 of Csiszár [10], since $\bar{\mu} \ll \nu$ and $\mu \ll \nu$ and $\mathscr{P}_n$ is assumed to be ν-approximating, it follows that

$$\lim_{n \to \infty} I_n(\bar{\mu}, \mu) = I(\bar{\mu}, \mu). \tag{4.36}$$

Therefore, since $I(\bar{\mu}, \mu) < \infty$, it follows that

$$\lim_{n \to \infty} I(\bar{\mu}_n, \mu) = 0. \tag{4.37}$$

Thus, the proof of Theorem 4 is completed. □

Remark 11: As in Remark 7, for $\mathcal{X} = \mathbb{R}$, suppose $\mathscr{P}_n = \{A_{n,i}, i = \cdots, -1, 0, 1, \cdots\}$ is the partition into intervals $A_{n,i} = [ih_n, (i+1)h_n)$ of common length $h_n = 1/k_n$, where k_n is a sequence of integers. Then the partition of $\mathbb{R}$ consisting of unit intervals $A_i = [i, i+1)$ is a natural choice for $\mathscr{P}$ in Theorem 4. The conditions that $f(x)/\bar{f}(x)$ be bounded and that $D(\bar{f}, f) < \infty$, are automatically satisfied if the ratio of the maximum value to the minimum value of $f(x)$ on the cells of the partition is bounded. A sufficient condition in the case of densities on the real line is that $\log f(x)$ be uniformly continuous. For density functions that are continuous and positive on a compact support set C, we may take for $\mathscr{P}$ the partition of $\mathcal{X}$ into C and its complement. Then the conditions of Theorem 4 are satisfied provided $\mathscr{P}_n$ is a refinement of $\mathscr{P}$. Therefore, as a special case, Theorem 4 includes the case covered by the Riemann–Stieltjes integration theory mentioned in Remark 9.

Finally, by the decomposition in (4.4), and by (4.14), (4.25), and (4.37), the results in Theorem 3 on the one hand, and those in Theorem 4 on the other, can be combined to yield conditions under which the expected reversed order informational divergence between the histogram estimator $\hat{\mu}_n$ and the distribution μ converges to zero. Combining Theorem 3, part b), and Theorem 4, we obtain the following result.

Theorem 5: Let μ be an unknown probability measure on an arbitrary measurable space $\mathcal{X}$. Assume there exists a known σ-finite measure ν on $\mathcal{X}$ such that $\mu \ll \nu$. Let $\mathscr{P}$ be a given partition of $\mathcal{X}$ such that $H^{\mathscr{P}}(\mu) < \infty$. Define $\bar{f}(x) = \mu(A_i)/\nu(A_i)$ if $x \in A_i \in \mathscr{P}$ and $\bar{\mu}(A) = \int_A \bar{f}(x)\nu(dx)$. Assume that $f(x)/\bar{f}(x) = d\mu/d\bar{\mu}$ is bounded and $I(\bar{\mu}, \mu) < \infty$. Let $\mathscr{P}_n$ be a sequence of partitions such that each $\mathscr{P}_n$ is a refinement of $\mathscr{P}$ and $m_n(\mu, \mathscr{P}) = o(n)$. Suppose further that the sequence $\mathscr{P}_n$ is ν-approximating. Let $X_1, \cdots, X_n$ be i.i.d. observations from μ, and let $\hat{\mu}_n$ be defined as in (4.2). Then,

$$\lim_{n \to \infty} E(I(\hat{\mu}_n, \mu)) = 0. \tag{4.38}$$

ACKNOWLEDGMENT

The authors are grateful to an anonymous referee for making useful suggestions and raising relevant questions which led to a considerable improvement and extension of this paper.

APPENDIX A

Here, we show the equivalence of the condition of effective cardinality of order $o(n)$ (condition A) with condition (D) of Abou-Jaoude [2], for a sequence of partitions $\mathscr{P}_n$ and a σ-finite measure η. Abou-Jaoude's condition (D) is that for every set S with $\eta(S) < \infty$, and every c and $\epsilon > 0$, there is an $N(\epsilon, c, S)$ such that for all $n \ge N(\epsilon, c, S)$,

$$\sum_{i:\, \eta(A_{n,i} \cap S) \le \frac{c}{n}} \eta(A_{n,i} \cap S) \le \epsilon. \tag{A.1}$$

To show the equivalence, first suppose Abou-Jaoude's condition is satisfied. Then, the cardinality of sets $A_{n,i} \cap S$ with $\eta(A_{n,i} \cap S) > \frac{c}{n}$ satisfies

$$N\left\{i: \eta(A_{n,i} \cap S) > \frac{c}{n}\right\} = \sum_{i:\, \eta(A_{n,i}\cap S) > \frac{c}{n}} 1$$
$$\leq \sum_i \frac{n}{c}\eta(A_{n,i} \cap S)$$
$$\leq \frac{n}{c}\eta(S). \tag{A.2}$$

From this, it follows that the ϵ-effective cardinality of $\mathscr{P}_n$ with respect to η restricted to S satisfies

$$\frac{m(\mathscr{P}_n, \eta, S, \epsilon)}{n} \leq \frac{1}{c}\eta(S), \tag{A.3}$$

for all $n \geq N(\epsilon, c, S)$. Since for any S with $\eta(S) < \infty$ this bound holds with c arbitrarily large, it follows that

$$\lim_{n\to\infty} \frac{m(\mathscr{P}_n, \eta, S, \epsilon)}{n} = 0. \tag{A.4}$$

Thus the sequence $\mathscr{P}_n$ has effective cardinality with respect to η of order $o(n)$.

Conversely, suppose the effective cardinality of the sequence $\mathscr{P}_n$ with respect to η is of order $o(n)$. Let an arbitrary set S with $\eta(S) < \infty$ and $\epsilon > 0$ be given. Let $m_n = m(\mathscr{P}_n, \eta, S, \epsilon/2)$ be the $\epsilon/2$-effective cardinality of $\mathscr{P}_n$ restricted to S. Then,

$$\lim_{n\to\infty} \frac{m_n}{n} = 0.$$

Thus, for every c, we have that for all n sufficiently large,

$$\frac{m_n}{n} < \frac{\epsilon}{2c}. \tag{A.5}$$

Now, let $A_{n,i}$, $i \in G_n$, denote a collection of m_n sets with the largest values for $\eta(A_{n,i} \cap S)$. By the definition of $\epsilon/2$-effective cardinality we have

$$\sum_{i\in G_n^c} \eta(A_{n,i} \cap S) \leq \frac{\epsilon}{2}. \tag{A.6}$$

Now defining $F_n = \{i: \eta(A_{n,i} \cap S) \leq \epsilon/(2m_n)\}$, it follows that for all n sufficiently large

$$\sum_{i:\, \eta(A_{n,i}\cap S) \leq \frac{c}{n}} \eta(A_{n,i} \cap S) \leq \sum_{i:\, \eta(A_{n,i}\cap S) \leq \frac{\epsilon}{2m_n}} \eta(A_{n,i} \cap S)$$
$$\leq \sum_{i\in G_n^c} \eta(A_{n,i} \cap S) + \sum_{i\in G_n \cap F_n} \eta(A_{n,i} \cap S)$$
$$\leq \frac{\epsilon}{2} + \sum_{i\in G_n} \frac{\epsilon}{2m_n}$$
$$= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \tag{A.7}$$

This completes the demonstration of the equivalence of the condition of effective cardinality of order $o(n)$ (condition A) and Abou-Jaoude's condition (D).

Appendix B

Here, we state and prove two useful inequalities used in the proofs of Theorem 2 and Theorem 3, respectively.

Lemma 1: If X has a binomial distribution with parameters n and p, or if X has a Poisson distribution with parameter np, then,

$$E\left(\frac{1}{X+1}\right) \leq \frac{1}{np}. \tag{B.1}$$

Proof: If X is binomial (n, p), then,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{n} \left(\frac{1}{k+1}\right)\binom{n}{k} p^k (1-p)^{n-k}$$
$$= \frac{1}{(n+1)p} \sum_{j=1}^{n+1} \binom{n+1}{j} p^j (1-p)^{n+1-j}$$
$$= \frac{1}{(n+1)p}(1 - (1-p)^{n+1})$$
$$\leq \frac{1}{(n+1)p} \leq \frac{1}{np}. \tag{B.2}$$

If X has a Poisson (np) distribution, let $\lambda = np$. Then,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^{\infty} \frac{1}{k+1}\frac{\lambda^k}{k!} e^{-\lambda}$$
$$= \frac{1}{\lambda}(1 - e^{-\lambda})$$
$$\leq \frac{1}{\lambda} = \frac{1}{np}. \qquad \square \tag{B.3}$$

Lemma 2: If X has a binomial distribution with parameters n and p, and $p_n = X/n$, then,

$$E\left(p_n \log \frac{p_n}{p}\right) \leq p \log\left(1 + \frac{1}{np}\right). \tag{B.4}$$

Proof: For $0 < q < 1$,

$$E\left(p_n \log \frac{p_n}{p}\right) = E\left(p_n \log \frac{p_n}{q}\right) + E\left(p_n \log \frac{q}{p}\right)$$
$$\leq E\left(p_n\left(\frac{p_n}{q} - 1\right)\right) + p \log \frac{q}{p}$$
$$= \frac{p^2 + (1/n)p(1-p)}{q} - p + p \log \frac{q}{p}, \tag{B.5}$$

and setting $q = p + (1/n)(1-p)$, which can be shown to optimize this bound, we have

$$= 0 + p \log\left(1 + \frac{1-p}{np}\right)$$
$$\leq p \log\left(1 + \frac{1}{np}\right). \qquad \square \tag{B.6}$$

References

[1] S. Abou-Jaoude, "Sur une condition nécessaire et suffisante de L_1-convergence presque complète de l'estimateur de la partition fixe pour une densité," *Comptes Rendus de l'Académie des Sciences de Paris Série A*, t. 283, pp. 1107–1110, 1976.

[2] ——, "Conditions nécessaires et suffisantes de convergence L_1 enprobabilité de l'histogramme pour une densité", *Annales de l'Institut Henri Poincaré, vol. XII, pp.* 213–231, 1976.

[3] A. R. Barron, "The convergence in information of probability density estimators," presented at *IEEE Int. Symp Inform. Theory*, Kobe, Japan, June 19–24, 1988.
[4] ——, " Uniformly powerful goodness of fit tests," *Ann. Statist.*, vol. 17, pp. 107–124, Mar. 1989.
[5] A. R. Barron and T. M. Cover, "A bound on the financial value of information," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1097–1100, Sept. 1988.
[6] A. R. Barron and C.-H. Sheu, "Approximation of density functions by sequences of exponential families," *Ann. Statist.*, vol. 19, pp. 1347–1369, Sept. 1991.
[7] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.
[8] T. M. Cover, "The admissibility properties of Gilbert's encoding for unknown source probabilities," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 216–217, Jan. 1972.
[9] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.
[10] ——, "Generalized entropy and quantization problems," in *Trans. Sixth Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*. Prague: Academia, 1973, pp. 159–174.
[11] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest: Akadémiai Kiadó, 1981.
[12] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[13] L. Devroye, *A Course in Density Estimation*. Boston: Birkhäuser, 1987.
[14] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. New York: Wiley, 1985.
[15] ——, "No empirical probability measure can converge in total variation sense for all distributions," *Ann. Statist.*, vol. 18, pp. 1496–1499, May 1990.
[16] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, pp. 209–230, Jan. 1973.
[17] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[18] E. N. Gilbert, "Codes based on inaccurate source probabilities," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 304–314, May 1971.
[19] L. Györfi and E. C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425–436, 1987.
[20] ——, "Density estimation a.s. consistent in information divergence," in *Proc. Fourth Joint Swedish–Soviet Int. Workshop Inform. Theory*, Gotland, Sweden, Aug. 27–Sept. 1, 1989, pp. 112–116.
[21] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, Mar. 1963.
[22] ——, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–401, Apr. 1965.
[23] J. H. B. Kemperman, "On the optimum rate of transmitting information," *Ann. Math. Statist.*, vol. 40, pp. 2156–2177, Dec. 1969.
[24] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Mar. 1981.
[25] S. Kullback, "A lower bound for discrimination in terms of variation," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 126–127, Jan. 1967.
[26] J. Rissanen, T. P. Speed, and B. Yu, "Density estimation by stochastic complexity," in *IEEE Trans. Inform. Theory*, vol. 38, pp. 315–323, Mar. 1992.