DIVERGENCE BETWEEN VARIOUS ESTIMATES OF QUANTIZED INFORMATION SOURCES

Domingo Morales[1], Leandro Pardo[1] and Igor Vajda[2]

This paper investigates the asymptotics of information-theoretic divergences between theoretical and empirical estimates and/or between nonparametric and parametric estimates of probabilistic source models. The results are applied to the source compression based on statistical estimation of unknown parameters, and to the testing hypotheses about information sources.

1. INTRODUCTION

Coding of information sources for transmission in discrete communication systems is based on quantization, i.e., decomposition of messages into disjoint sets $D_1, \dots, D_M$. The number M is required to satisfy a transmission rate condition $\log_2 M \le R$ and the sets themselves are required to minimize certain distortion function (see Chap. 13 in Cover and Thomas [5] or Berger [2]).

In this paper we consider parametrized sources models $(F_\theta : \theta \in \Theta)$ where $F_\theta(x)$ is a k-variate probability distribution function and θ varies over an open set $\Theta \subset R^m$. As shown e.g. in Berger [2] or Gersho and Gray [10], for most of practically interesting source models there exist ordered decompositions $\mathcal{D}_M = (D_1(\theta), \dots, D_M(\theta))$ of R^k optimal in the sense of rate-distortion theory.

The source probability distribution, in particular the true parameter value $\theta_0 \in \Theta$, are typically not known a priori. They have to be determined a posteriori, on the basis of observation of source messages $X_1, \dots, X_n$. In this paper the messages are assumed to be independent random variables defined on a basic probability space $(\Omega, \mathcal{A}, \boldsymbol{P})$, with a common sample distribution function F_{θ_0}.

The posterior source specification may be either parametric, $F_{\hat{\theta}_n}$, where $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ is a given point estimator $\hat{\theta}_n : R^k \to \Theta$, or nonparametric, $\hat{F}_n$, where $\hat{F}_n$ is a sample-dependent k-variate distribution function (see Barron, Györfi and van der Meulen [1]). The best known example is the standard empirical distribution

[1] Supported by grants DGICYT PB 94-0308 and PR 219/94-5307.

[2] Supported by GA ČR grant 102/94/0320.

function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}((-\infty, x]), \tag{1}$$

where δ_ξ is the Dirac distribution concentrated at $\xi \in R^k$ and $(-\infty, \xi]$ is the product of similar intervals for coordinates of ξ.

The main disadvantage of nonparametric estimates is that they do not allow to employ the well-established parametric methods of quantization $\mathcal{D}_M(\theta)$. It is therefore convenient to combine the advantages of both these estimation methods by projecting a nonparametric estimate $\hat{F}_n$ on a parametrized family $(F_\theta : \theta \in \Theta)$, which leads to parametric estimates $F_{\hat{\theta}_n}$ and, in particular, to point estimators $\hat{\theta}_n$ (see Györfi, Vajda and van der Meulen [11]).

In this paper we consider parametric estimates allowing to employ the parameter-depending quantizations under consideration. Unless otherwise explicitly stated, we consider arbitrary fixed: family $(F_\theta : \theta \in \Theta)$, estimator $\hat{\theta}_n$, quantization method $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ and quantization size $M > 1$. For brevity we put for any distribution function

$$F\{D\} = \int_D \mathrm{d}F, \quad D \subset R^k.$$

We define probability M-vectors $P, \hat{P}_n, \tilde{P}_n, \hat{Q}_n$ and $\tilde{Q}_n$ by their respective coordinates

$$\begin{aligned} p_j = F_{\theta_0}\{D_j(\theta_0)\}, \qquad &\hat{p}_{n,j} = \hat{F}_n\{D_j(\hat{\theta}_n)\}, \quad &\tilde{p}_{n,j} = \hat{F}_n\{D_j(\theta_0)\}, \\ &\hat{q}_{n,j} = F_{\hat{\theta}_n}\{D_j(\hat{\theta}_n)\}, \quad &\tilde{q}_{n,j} = F_{\theta_0}\{D_j(\hat{\theta}_n)\}, \end{aligned} \tag{2}$$

where $\hat{F}_n$ is defined in (1).

We see that P is the true probability distribution on the source quantized by using the true θ_0. On the other hand, $\tilde{Q}_n$ is the true distribution on the source quantized by using an estimate $\hat{\theta}_n$.$\hat{P}_n$ and $\tilde{P}_n$ are relative frequencies of cells when the quantization is based on $\hat{\theta}_n$ and θ_0 respectively. The distributions $P, \tilde{P}_n$ and $\tilde{Q}_n$ are not practically available without the knowledge of θ_0, while $\hat{P}_n$ and $\hat{Q}_n$ are. The distribution $\hat{Q}_n$ is a family-based alternative to $\hat{P}_n$, and all four sample-depending distributions are approximations to P. In this paper we evaluate the asymptotics of ϕ-divergences $D_\phi(\mu_n, \nu_n)$ as $n \to \infty$ for any pair μ_n, ν_n from the set of probability vector sequences $\{\hat{P}_n, \tilde{P}_n, \hat{Q}_n, \tilde{Q}_n\}$. Motivations for this result arising from information theory, statistics and neural networks are presented in Examples 1–3 below.

Remind that the ϕ-divergence of arbitrary probability M-vectors $\mu = (\mu_1, \ldots, \mu_M)$ and $\nu = (\nu_1, \ldots, \nu_M)$ is defined for convex functions $\phi : [0, \infty) \to (-\infty, \infty]$ finite everywhere except possibly at 0, by the formula

$$D_\phi(\mu, \nu) = \sum_{j=1}^{M} \nu_j \phi\left(\frac{\mu_j}{\nu_j}\right). \tag{3}$$

The eventually undefined expressions behind the sum are assumed to be specified in the same way as in Csiszár [7] or Liese and Vajda [15].

The role of divergences obtained for $\phi(u) = -\log u$ and $\phi(u) = u \cdot \log u$ is well known in information theory. Applications of this divergence in neural networks were described e. g. in Vajda [21] and Veselý and Vajda [22]. The importance of other ϕ-divergences has been documented recently e. g. in [4, 8, 18, 20 and 21].

Next we illustrate the application of the above mentioned asymptotics in the case of

$$I(\mu, \nu) = \sum_{j=1}^{M} \nu_j \log \frac{\nu_j}{\mu_j}, \tag{4}$$

which is the ϕ-divergence for $\phi(u) = -\log u$, and

$$X^2(\mu, \nu) = \sum_{j=1}^{M} \frac{(\mu_j - \nu_j)^2}{\nu_j}, \tag{5}$$

which is the ϕ-divergence for $\phi(u) = (1-u)^2$.

Example 1. Let the source quantized by $\mathcal{D}_M(\hat{\theta}_n)$ be Shannon-coded either on the basis of the essentially nonparametric estimate $\hat{P}_n$ or on the basis of the parametric estimate $\hat{Q}_n$. Denote by $L(\hat{P}_n)$ and $L(\hat{Q}_n)$ the respective average codelengths. As well known, the Shannon entropy $H(\tilde{Q}_n)$ is the lower bound to both $L(\hat{P}_n)$ and $L(\hat{Q}_n)$. By Theorem 5.4.3 of Cover and Thomas [5], it holds

$$H(\tilde{Q}_n) + I(\hat{P}_n, \tilde{Q}_n) \le L(\hat{P}_n) \le H(\tilde{Q}_n) + I(\hat{P}_n, \tilde{Q}_n) + 1 \tag{6}$$

$$H(\tilde{Q}_n) + I(\hat{Q}_n, \tilde{Q}_n) \le L(\hat{Q}_n) \le H(\tilde{Q}_n) + I(\hat{Q}_n, \tilde{Q}_n) + 1. \tag{7}$$

We see that the divergences $I(\hat{P}_n, \tilde{Q}_n)$ and $I(\hat{Q}_n, \tilde{Q}_n)$ defined by (4) represent inefficiencies of source estimates $\hat{P}_n$ and $\hat{Q}_n$ in the given information-theoretic context. The asymptotics of divergences describes the rates of inefficiencies of these estimates. The estimate with a smaller inefficiency is obviously preferable. Our general result provides the asymptotics for $n\, I(\hat{P}_n, \tilde{Q}_n)$ and $n\, I(\hat{Q}_n, \tilde{Q}_n)$ for various quantizations $\mathcal{D}_M(\theta)$ and estimates $\hat{\theta}_n$.

Example 2. Our assumption that the source model belongs to a family $(F_\theta : \theta \in \Theta)$ is a composite statistical hypothesis. In some cases this hypothesis is a priori acceptable but, quite often, is has to be tested on the basis of data $X_1, \dots, X_n$ introduced above. The classical statistical testing procedure is specified for univariate observations and for quantizations $\mathcal{D}_M(\theta)$ not depending on $\theta \in \Theta$. The testing statistic $nX^2(\hat{P}_n, \hat{Q}_n)$ is according to (5) evaluated for $\hat{P}_n, \hat{Q}_n$ given by (2) with a suitable estimator $\hat{\theta}_n$, and the hypothesis is rejected if the statistic exceeds a positive critical value. Chernoff and Lehman [3] found a method for evaluating the critical values in case of the MLE $\hat{\theta}_n$. They proved for sufficiently regular family $(F_\theta : \theta \in \Theta)$ and $M \ge m+1$ that $nX^2(\hat{P}_n, \hat{Q}_n)$ tends under the hypothesis to

$$\sum_{j=1}^{M-1} \lambda_j\, Z_j^2 \tag{8}$$

in law, where $Z_1, \ldots, Z_{M-1}$ are independent standard normal variates and $\lambda_j = 1$ for $1 \le j \le M-m-1$. The remaining parameters $0 < \lambda_j < 1$ for $M-m \le j \le M-1$ are well-defined by means of the family and quantization. Later many authors extended this result to quantizations depending on $\theta \in \Theta$ and investigated optimality of various quantization methods $(\mathcal{D}_M(\theta) : \theta \in \Theta, M > 1)$, (see Dahiya and Gurland [9], Kallenberg, Oosterhoff and Schriever [13] and references therein). For extensions to multivariate observations see Moore [16].

By extending the results of previous authors to arbitrary test statistics $D_\phi(\hat{P}_n, \hat{Q}_n)$ we provide a wide class of statistical tests for the hypothesis under consideration. Empirical studies (e. g. Sec. 5 in [15]) indicate that the powers of tests in this class depend on ϕ, and for typical alternatives the power is maximized by a statistics $D_\phi(\hat{P}_n, \hat{Q}_n)$ different from $X^2(\hat{P}_n, \hat{Q}_n)$.

Example 3. The hypotheses testing result of Example 2 has an immediate application in classification by neural networks. Veselý and Vajda [22] considered a Bayes classification of signals from two classes distributed by μ or ν. They described a neural network realization of this classification under the assumption that there exist "class etalons" μ_* and ν_* such that for all μ and ν under consideration

$$I(\mu, \mu_*) < I(\mu, \nu_*) \quad \text{and} \quad I(\nu, \mu_*) > I(\nu, \nu_*).$$

This condition can easily be verified under the hypotheses $\mu \in \{F_\theta : \theta \in \Theta_1\}$ and $\nu \in \{F_\theta : \theta \in \Theta_2\}$, where Θ_1 and Θ_2 are disjoint intervals of real numbers and F_θ are exponential distributions with natural parameters $\theta \in R$. Indeed, then $\psi(\theta) = I(F_\theta, F_{\theta_0})$ is for every θ_0 nonnegative and convex on R. It follows from here that $\psi(\theta)$ is isotone with $|\theta - \theta_0|$ so that the "class etalons" may be any distributions P_{θ_1} and P_{θ_2} with θ_1 and θ_2 from the interiors of intervals Θ_1 and Θ_2. Thus the statistical verification of the parametric hypotheses $\mu \in \{F_\theta : \theta \in \Theta_1\}$ and $\nu \in \{F_\theta : \theta \in \Theta_2\}$ by using a test of higher statistical power increases the probability that the neural net classification will be optimal in the Bayes sense.

2. BASIC LEMMA

We denote by $|\cdot|$ the absolute value of a number and also the absolute norm of an M-vector. For μ, ν figuring in (3) the norm

$$|\mu - \nu| = \sum_{j=1}^{M} |\mu_j - \nu_j|$$

means the total variation of μ and ν (the ϕ-divergence for $\phi(u) = |u-1|$). Further, for $\varepsilon_n > 0, o(\varepsilon_n)$ or $O(\varepsilon_n)$ denotes a sequence of real numbers such that

$$\lim_{n\to\infty} \frac{o(\varepsilon_n)}{\varepsilon_n} = 0 \quad \text{or} \quad \sup_n \left| \frac{O(\varepsilon_n)}{\varepsilon_n} \right| < \infty$$

respectively.

The results of this paper are based on the following lemma describing a linear approximability of all smooth ϕ-divergences by relatively simpler X^2-divergences.

Lemma 1. Let $P_n = (p_{n,1}, \ldots, p_{n,M})$, $Q_n = (q_{n,1}, \ldots, q_{n,M})$ be sequences of probability vectors for which there exists a probability vector $P = (p_1, \ldots, p_M)$ with

$$\prod_{j=1}^{M} p_j > 0 \tag{9}$$

and a sequence $\varepsilon_n > 0$ such that

$$|P_n - P| = O(\varepsilon_n), \quad |Q_n - P| = O(\varepsilon_n). \tag{10}$$

Then for all ϕ-divergences with $\phi(u)$ twice continuously differentiable in an open neighborhood of $u = 1$

$$D_\phi(P_n, Q_n) = \phi(1) + \frac{\phi''(1)}{2} X^2(P_n, Q_n) + o(\varepsilon_n^2). \tag{11}$$

Proof. By the Taylor expansion of $\phi(u)$ around $u = 1$ we obtain for every $1 \leq j \leq m$ and all sufficiently large n

$$\phi\left(\frac{p_{n,j}}{q_{n,j}}\right) = \phi(1) + \phi'(1)\frac{p_{n,j} - q_{n,j}}{q_{n,j}} + \frac{1}{2}\phi''(r_{n,j})\left(\frac{p_{n,j} - q_{n,j}}{q_{n,j}}\right)^2,$$

where

$$|r_{n,j} - 1| \leq \left|\frac{p_{n,j} - q_{n,j}}{q_{n,j}}\right|.$$

It follows from here and (3)

$$D_\phi(P_n, Q_n) = \phi(1) + \frac{1}{2}\sum_{j=1}^{M} \phi''(r_{n,j})\frac{(p_{n,j} - q_{n,j})^2}{q_{n,j}}.$$

Hence it remains to prove that, for the sequence

$$c_n = \frac{1}{2}\sum_{j=1}^{M} |\phi''(r_{n,j}) - \phi''(1)|\,\frac{(p_{n,j} - q_{n,j})^2}{q_{n,j}}$$

and for ε_n figuring in (10),

$$\lim_{n\to\infty} \frac{c_n}{\varepsilon_n^2} = 0.$$

Since by (10)

$$\lim_{n\to\infty} \frac{1}{2}\sum_{j=1}^{M} \frac{|\phi''(r_{n,j}) - \phi''(1)|}{q_{n,j}} = 0,$$

it suffices to prove that the right-hand side of the obvious inequality

$$\max_{1\leq j\leq M} (p_{n,j} - q_{n,j})^2 \leq |P_n - Q_n|^2$$

is $O(\varepsilon_n^2)$. But this is clear from (10). □

Corollary. For P_n, Q_n satisfying the assumptions of Lemma 1,

$$X^2(Q_n, P_n) - X^2(P_n, Q_n) = o(\varepsilon_n^2). \tag{12}$$

Proof. It follows from (3) and (5) that $X^2(Q_n, P_n) = D_\phi(P_n, Q_n)$ for the convex function $\phi(u) = (1-u)^2/\mathrm{u}$. This function satisfies the assumptions of Lemma 1 with $\phi(1) = 0$ and $\phi''(1) = 2$. Thus (12) follows from (11). □

3. MAIN RESULTS

For $\varepsilon_n > 0$ we denote by $o_p(\varepsilon_n)$ or $O_p(\varepsilon_n)$ sequences of random variables such that

$$\lim_{n\to\infty} P(|o_p(\varepsilon_n)/\varepsilon_n| < \varepsilon) = 1, \quad \text{for all} \quad \varepsilon > 0$$

(the convergence in probability of $o_p(\varepsilon_n)/\varepsilon_n$ to zero), or

$$\lim_{n\to\infty} P(|O_p(\varepsilon_n)/\varepsilon_n| < \infty) = 1$$

(the stochastic boundedness of $O_p(\varepsilon_n)/\varepsilon_n$), respectively. We shall use the easily verifiable fact that if P_n, Q_n in Lemma 1 are sample-depending probability vectors satisfying (10) with $O(\varepsilon_n)$ replaced by $O_p(\varepsilon_n)$, then (11) holds with $o(\varepsilon_n^2)$ replaced by $o_p(\varepsilon_n^2)$. Next we list our assumptions about the source model, estimator and quantization method.

(i) Assumption about $(F_\theta\colon\ \theta \in \Theta)$: All $F_\theta(x)$, $\theta \in \Theta$, are Lipschitz in R^k and all $F_\theta(x), x \in R^k$, are differentiable in θ. Moreover, the gradient

$$\nabla F_\theta(x) = \left(\frac{\partial}{\partial\theta_1} F_\theta(x), \ldots, \frac{\partial}{\partial\theta_m} F_\theta(x)\right)$$

is continuous in θ and x.

(ii) Assumption about $\hat{\theta}_n$: $|\hat{\theta}_n - \theta_0| = O_p(n^{-1/2})$ for each true value $\theta_0 \in \Theta$ (consistency of order $n^{1/2}$).

(iii) Assumption about $(\mathcal{D}_M(\theta)\colon\ \theta \in \Theta)$: Consider for $s = 1, \ldots, m$ the x_s-axis partitioned by points

$$-\infty \equiv \xi_{s,0}(\theta) < \ldots < \xi_{s,M_s}(\theta) \equiv \infty,$$

where the finitely-valued functions (i.e. $\xi_{s,j}(\theta)$ for $1 \le j \le M_{s-1}$ and $1 \le s \le m$) are continuously differentiable. Further, consider $M = M_1 \cdots M_m$ and a one-to-one mapping

$$\Psi : \{1, \ldots, M\} \longrightarrow \prod_{s=1}^{m} \{1, \ldots, M_s\}.$$

The jth element $D_j(\theta)$ of $\mathcal{D}_M(\theta)$ is assumed to be the cell in R^k with the partition points defined by $\Psi(j)$, i.e.

$$D_j(\theta) = \prod_{s=1}^{m} (\xi_{s,j_{s-1}}(\theta), \xi_{s,j_s}(\theta)] \quad \text{for } (j_1, \ldots, j_m) = \Psi(j),$$

with] replaced by) if $\xi_{s,j_s}(\theta) \equiv \infty$, i.e. if $j_s = M_s$.

(iv) Assumption about (F_θ: $\theta \in \Theta$) and ($\mathcal{D}_M(\theta)$: $\theta \in \Theta$): All quantizations $\mathcal{D}_M(\theta)$, $\theta \in \Theta$, satisfy the condition (9) for p_j defined by (2).

Lemma 2. Under (i)–(iii), every sequence P_n from the set $\{\hat{P}_n, \tilde{P}_n, \hat{Q}_n, \tilde{Q}_n\}$ satisfies the relation

$$|P_n - P| = O_p(n^{-1/2}). \tag{13}$$

Proof. Assumptions (i)–(iii) imply those of Ruymgaart [19], who proved relation (2.2) of Moore [16] originally established under stronger assumptions. In our notation this relation takes on the form

$$\hat{p}_{n,j} - \tilde{p}_{n,j} + p_j - \tilde{q}_{n,j} = O_p(n^{-1/2}) \quad \text{for all } 1 \le j \le M. \tag{14}$$

Since

$$n^{1/2}(\tilde{p}_{nj} - p_j) = n^{-1/2} \sum_{i=1}^{n} \left(\mathbf{1}_{D_j(\theta_0)}(X_i) - p_j\right),$$

the central limit theorem implies

$$|\tilde{P}_n - P| = O_p(n^{-1/2}). \tag{15}$$

If

$$\tilde{q}_{n,j} - p_j = O_p(n^{-1/2}) \quad \text{for all } 1 \le j \le M, \tag{16}$$

then $|\tilde{Q}_n - P| = O_p(n^{-1/2})$ holds and $|\hat{P}_n - P| = O_p(n^{-1/2})$ follows from (14) and (15). If

$$\hat{q}_{n,j} - \tilde{q}_{n,j} = O_p(n^{-1/2}) \quad \text{for all } 1 \le j \le M, \tag{17}$$

then $|\hat{Q}_n - P| = O_p(n^{-1/2})$ follows from (16). Thus it remains to prove (16) and (17).

By (2) and the definition of $D_j(\theta)$ in (iii), $\tilde{q}_{n,j} - p_j$ is a finite sum of expressions of the form $\pm[F_{\theta_0}(\xi(\hat{\theta}_n)) - F_{\theta_0}(\xi(\theta_0))]$ where ξ is one of the functions supposed to be continuously differentiable in (iii). Denote by $\nabla\xi(\theta)$ the gradient of ξ at θ and by $L(\theta_0)$ the Lipschitz constant for F_{θ_0} (cf. (i)). Then the Lipschitz property and the mean value theorem imply

$$\begin{aligned} |F_{\theta_0}(\xi(\hat{\theta}_n)) - F_{\theta_0}(\xi(\theta_0))| &\le L(\theta_0)|\xi(\hat{\theta}_n) - \xi(\theta_0)| \\ &\le L(\theta_0)|\nabla\xi(\theta_n)(\hat{\theta}_n - \theta_0)^t|, \end{aligned}$$

where θ_n is a function of sample satisfying the condition $||\theta_n - \theta_0|| \le ||\hat{\theta}_n - \theta_0||$. Since $\nabla\xi(\theta)$ is continuous on Θ, it follows from (ii)

$$|\nabla\xi(\theta_n)(\hat{\theta}_n - \theta_0)^t| = O_p(n^{-1/2}).$$

This implies (16).

Now we prove (17). By the similar argument as above, it suffices to prove

$$F_{\hat{\theta}_n}(\xi(\hat{\theta}_n)) - F_{\theta_0}(\xi(\hat{\theta}_n)) = O_p(n^{-1/2})$$

for the same ξ as above. By the mean value theorem,

$$F_{\hat{\theta}_n}(\xi(\hat{\theta}_n)) - F_{\theta_0}(\xi(\hat{\theta}_n)) = \nabla F_{\theta_n}(\xi(\hat{\theta}_n))\,(\hat{\theta}_n - \theta_0)^t,$$

where $\nabla F_{\theta_n}(x)$ is the gradient of $F_\theta(x)$ at $\theta = \theta_n$ and θ_n is similar as above. By the continuity of gradient function assumed in (i), one obtains from (ii)

$$|\nabla F_{\theta_n}(\xi(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)^t| = O_p(n^{-1/2}). \qquad \square$$

Now we can formulate the main result of the paper.

Theorem. Suppose (i) – (iv) are satisfied and that μ_n, ν_n are two sequences from the set $\{\hat{P}_n, \tilde{P}_n, \hat{Q}_n, \tilde{Q}_n\}$. Then for every ϕ considered in Lemma 1

$$D_\phi(\mu_n, \nu_n) = \phi(1) + \frac{\phi''(1)}{2} X^2(\mu_n, \nu_n) + o_p(n^{-1}). \tag{18}$$

Proof. Clear from Lemmas 1 and 2 and from the observation at the beginning of this section. □

Corollary. If the assumptions of Theorem hold then for all ϕ and μ_n, ν_n considered there

$$D_\phi(\mu_n, \nu_n) - \phi(1) = O_p(n^{-1}). \tag{19}$$

Proof. By assumptions, it holds for all $1 \le j \le M$

$$\mu_{n,j} - p_j = O_p(n^{-1/2}), \quad \nu_{n,j} - p_j = O_p(n^{-1/2}),$$

i.e.

$$\mu_{n,j} - \nu_{n,j} = O_p(n^{-1/2}).$$

Therefore

$$X^2(\mu_n, \nu_n) = \sum_{j=1}^{M} \frac{(\mu_{n,j} - \nu_{n,j})^2}{\nu_{n,j}} = O_p(n^{-1}) \tag{20}$$

and (19) follows from (18). □

4. APPLICATIONS TO SOURCE CODING

In this section we present an optimality of quantization methods $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ from the source coding point of view, in order to demonstrate that the conceptual framework of Example 1 is justified. Then we apply the results of Section 3 to the situation described by this example. For simplicity we restrict ourselves to a parametrized source model $(F_\theta : \theta \in \Theta)$ on the real line R, and to the distortion function (distortion measure)

$$\rho(x_1, x_2) = |x_1 - x_2|.$$

Let $\mathcal{D}_M(\theta) = (D_1(\theta), \ldots, D_M(\theta))$ be a quantization defined by partitioning of R by functions $-\infty \equiv \xi_0(\theta) < \xi_1(\theta) < \ldots < \xi_M(\theta) \equiv \infty$, i.e. let

$$D_j(\theta) = (\xi_{j-1}(\theta), \xi_j(\theta)],$$

with] replaced by) for j=M.

A source code is a mapping $\kappa : R \rightarrow R$. The number $\log_2(\mathrm{card}\kappa(R))$ is the code information rate. We are interested in codes with finite information rates $\log_2 M$, i.e. in codes with finite codebooks $\mathcal{C}(R) = (x_1, \ldots, x_M)$. The average distortion

$$\rho_{\theta_0}(\kappa) = \sum_{j=1}^{M} \int_{\kappa^{-1}(x_j)} |x - x_j| \, \mathrm{d}F_{\theta_0}(x) \tag{21}$$

is the measure of quality of the code κ when θ_0 is the true source parameter.

Every code κ with a codebook $\mathcal{C}_M = (x_1, \ldots, x_M)$ defines a source quantization $\mathcal{D}_M = (\kappa^{-1}(x_1), \ldots, \kappa^{-1}(x_M))$. In the sequel we use the fact that, for a given source quantization $\mathcal{D}_M = (D_1, \ldots, D_M)$, a source code of rate $\log_2 M$ is uniquely specified by a codebook $\mathcal{C}_M = (x_1, \ldots, x_M)$ (both sets are considered to be ordered). Analogously, for a given codebook $\mathcal{C}_M = (x_1, \ldots, x_M)$ a source code of rate $\log_2 M$ is uniquely specified by a source quantization $\mathcal{D}_M = (D_1, \ldots, D_M)$. In this sense the optimality considered below is the optimality of source codes.

It is obvious that a source quantization $\mathcal{D}_M^* = (D_1^*, \ldots, D_M^*)$ is optimal for a codebook $\mathcal{C}_M = (x_1, \ldots, x_M)$ in the sense that it minimizes the average distortion

$$\rho_{\theta_0}(\mathcal{D}_M) = \sum_{j=1}^{M} \int_{D_j} |x - x_j| \, \mathrm{d}F_{\theta_0}(x) \qquad \text{(cf. (20))}$$

over all source quantizations $\mathcal{D}_M = (D_1, \ldots, D_M)$ if and only if for all $1 \leq j \leq M$

$$D_j^* = (\xi_{j-1}^*, \xi_j^*],$$

where

$$\xi_j^* = \frac{x_j + x_{j+1}}{2}, \quad j = 1, \ldots, M-1, \tag{22}$$

and $\xi_0^* = -\infty$, $\xi_M^* = \infty$. We see that this optimality is independent of the true source parameter θ_0. It is also known that a codebook

$$\mathcal{C}_M^*(\theta_0) = (x_1^*(\theta_0), \ldots, x_M^*(\theta_0))$$

is optimal for a source quantization $\mathcal{D}_M = (D_1, \ldots, D_M)$ with the property $F_{\theta_0}\{D_j\} > 0$ for all $1 \leq j \leq M$ in the sense that it minimizes the average distortion

$$\rho_{\theta_0}(\mathcal{C}_M) = \sum_{j=1}^{M} \int_{D_j} |x - x_j| \, \mathrm{d}F_{\theta_0}(x) \qquad \text{(cf. (20))}$$

over all codebooks $\mathcal{C}_M = (x_1, \ldots, x_M)$ if and only if

$$x_j^*(\theta_0) = \text{median of } F_{\theta_0}^{(j)},$$

where $F_{\theta_0}^{(j)}$ is the distribution function F_{θ_0} conditioned on D_j. In particular, the codebook $\mathcal{C}_M^*(\theta,\theta_0) = (x_1^*(\theta,\theta_0),\dots,x_M^*(\theta,\theta_0))$ optimal for the above considered $\mathcal{D}_M(\theta)$ is given by

$$x_j^*(\theta,\theta_0) = \frac{1}{2}\left(F_{\theta_0}^{-1}(\xi_j(\theta)) - F_{\theta_0}^{-1}(\xi_{j-1}(\theta))\right) \quad \text{for } 1 \le j \le M, \tag{23}$$

where $F_{\theta_0}^{-1} : [0,1] \to \overline{R}$ is the source quantile function.

The quantization method $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ is not optimal from the point of view of source coding if there exists $\theta_0 \in \Theta$ and $1 \le j \le M-1$ such that

$$\xi_j(\theta_0) \neq \frac{x_j^*(\theta_0,\theta) + x_{j+1}^*(\theta_0,\theta)}{2}.$$

In this case there exists a better code of the source (R, F_{θ_0}) than any of the codes which are constant on the cells $D_j(\theta_0) = (\xi_{j-1}(\theta_0), \xi_j(\theta_0)]$ of the quantization $\mathcal{D}_M(\theta_0)$. Indeed, by (22), it is the code given by the partitioning $\mathcal{D}_M^*$ of R defined by points

$$\xi_j^* = \frac{x_j^*(\theta_0,\theta_0) + x_{j+1}^*(\theta_0,\theta_0)}{2} \quad \text{for } 1 \le j \le M-1,$$

and $\xi_0^* = -\infty, \xi_M^* = \infty$, and by the corresponding codebook $\mathcal{C}_M^* = (x_1^*,\dots,x_M^*)$, defined by

$$x_j^* = \frac{1}{2}\left(F_{\theta_0}^{-1}(\xi_j^*) - F_{\theta_0}^{-1}(\xi_{j-1}^*)\right) \quad \text{for } 1 \le j \le M \quad \text{(cf. (23))}.$$

Therefore a condition for considering a quantization method $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ defined by functions $-\infty < \xi_1(\theta) < \dots < \xi_{M-1}(\theta) < \infty$ to be optimal from the source coding point of view is

$$\xi_j(\theta) = \frac{x_j^*(\theta,\theta) + x_{j+1}^*(\theta_0,\theta)}{2} \quad \text{for all } \theta \in \Theta \text{ and } 1 \le j \le M-1,$$

where $x_j(\theta,\theta)$ are given by (23). In other words, the condition is

$$\xi_j(\theta) = \frac{F_\theta^{-1}(\xi_{j+1}(\theta)) - F_\theta^{-1}(\xi_{j-1}(\theta))}{4} \quad \text{for all } \theta \in \Theta \text{ and } 1 \le j \le M-1. \tag{24}$$

It is easy to see that for the source models with known quantile functions F_θ^{-1}, $\theta \in \Theta$, like the exponential or logarithmic ones with $\Theta = (0,\infty)$ and with

$$F_\theta(x) = 1 - \exp(-\theta x),\ x \ge 0, \quad \text{or} \quad F_\theta(x) = 1 - (\log_e x)^{-\theta},\ x \ge e,$$

one can construct explicitly for any $M > 0$ the partitioning functions $\xi_1^{(M)}(\theta),\dots$ $\dots,\xi_{M-1}^{(M)}(\theta)$ such that the corresponding quantization method $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ satisfies (24).

Now consider a parametrized source $(F_\theta : \theta \in \Theta)$ quantized by the method $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ under consideration, optimum in the sense of (24). Let the source and the partitioning functions $\xi_j(\theta)$ satisfy assumptions (i) and (iii), (iv) of Section 3, and let $\hat{\theta}_n$ be an estimator of the unknown true source parameter θ_0 satisfying assumption (ii) of Section 3. In accordance with Example 1, suppose that the codebooks $x_j^*(\hat{\theta}_n, \hat{\theta}_n)$ defined by (23) are transmitted by means of a Shannon binary code. In the first case, let the quantized source probabilities be estimated by $\hat{P}_n$ and in the second case by $\hat{Q}_n$, both defined by (2). Using the Corollary of Section 2, we obtain the following formulas for the inaccuracies introduced in Example 1:

$$I(\hat{P}_n, \tilde{Q}_n) = O_p(n^{-1}) \quad \text{and} \quad I(\hat{Q}_n, \tilde{Q}_n) = O_p(n^{-1}) \quad \text{as } n \to \infty.$$

We see that our simple theory is not able to distinguish which of the estimates $\hat{P}_n$ and $\hat{Q}_n$ is better, but it is able to confirm that both these estimates are good in the sense that the corresponding inaccuracies tend to zero fast enough.

5. APPLICATIONS TO GOODNESS-OF-FIT

In this section we consider applications of our Theorem to the problem of testing statistical hypotheses introduced in Example 2. Under assumptions stronger than ours (i)–(iv), Moore [16] proved for $X^2(\hat{P}_n, \hat{Q}_n)$ more explicite asymptotic result than (20), namely he found the asymptotic distribution of $nX^2(\hat{P}_n, \hat{Q}_n)$. By means of our Theorem, this distribution can easily be adapted to all statistics $D_\phi(\hat{P}_n, \hat{Q}_n)$ under consideration. Thus these statistics can be used as alternatives to the X^2-statistic in the goodness of fit testing.

By Theorem 1 of Moore [16], under assumptions concerning $(F_\theta : \theta \in \Theta), \hat{\theta}_n$, and $(\mathcal{D}_M(\theta) : \theta \in \Theta)$ stated there, $nX^2(\hat{P}_n, \hat{Q}_n)$ tends in distribution to (8). The parameters $\lambda_1, \dots, \lambda_{M-1}$ have formally the same properties as those in (8), but they are given by a more complicated formula than in the case where $\hat{\theta}_n$ is MLE and the quantization $\mathcal{D}_M(\theta)$ is constant for all $\theta \in \Theta$.

By our Theorem, under the mentioned assumptions of [16], it holds for all ϕ considered in Lemma 1 with $\phi''(1) \neq 0$ that the statistic

$$\frac{2n(D_\phi(\hat{P}_n, \hat{Q}_n) - \phi(1)}{\phi''(1)} \tag{25}$$

tends in distribution to (8) as well, for the same $\lambda_1, \dots, \lambda_{M-1}$ as in [16].

Note that in the particular case of quantizations $\mathcal{D}_M(\theta)$ not depending on $\theta \in \Theta$ and convex functions $\phi_\alpha(u) = \text{sign}(\alpha - 1)(u^\alpha - 1), \alpha > 0, \alpha \neq 1$, the asymptotic distribution of statistics (25) has been investigated by Cressie and Read [6], who also proved that, from the point of view of second order properties, some of the statistics (25) are better than the classical $nX^2(\hat{P}_n, \hat{Q}_n)$. Extensions of the results of [6] to all ϕ differentiable on $(0, \infty)$ have been considered in [14] and [17].

Note also that the assumptions of [15] are stronger than those of our Theorem only as regards the family $(F_\theta : \theta \in \Theta)$ and estimator $\hat{\theta}_n$. As regards the quantization, they are equivalent to ours (iii) and (iv).

In information theory the problem of testing considered in this section takes place if a source is described by a family $(F_\theta : \theta \in \Theta_*)$ and the hypothesis $\mathcal{H} : H(F_\theta) \in \mathcal{I}$ about the source entropy $H(F_\theta)$ is to be tested. Here $\mathcal{I}$ is an interval from the entropy domain and the entropy may be considered e. g. in the Shannon sense if the source is discrete, or in the differential sense if the source is continuous. For some models the hypothesis corresponds to a subset of parameters $\Theta \subset \Theta_*$ satisfying the assumptions of our theory, and the alternative corresponds to the complement $\Theta_* - \Theta$.

Applications of the results of this section in statistical decisions based on neural networks have already been indicated in Example 3.

6. FURTHER EXTENSIONS

Sometimes it is convenient to consider functions $h(D_\phi(\mu,\nu))$ of ϕ-divergences instead of ϕ-divergences themselves. For example, the divergences

$$\sum_{i=1}^{M} |\mu_i^a - \nu_i^a|^{1/a}, \quad 0 < a < 1,$$

defined by $\phi_a(u) = |u^a - 1|^{1/a}$ are not metrics but their powers $h_a(D_{\phi_a}(\mu,\nu)) = D_{\phi_a}(\mu,\nu)^a$ are. The results of our Lemmas and Theorem, and their Corollaries, can obviously be extended by replacing $D_\phi(\mu,\nu)$ by $h(D_\phi(\mu,\nu)) - h(\phi(1))$ for h continuously differentiable.

(Received August 8, 1995.)

REFERENCES

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

[2] T. Berger: Rate Distortion Theory: A Mathematical Basis for Data Compression. Prentice-Hall, Englewood Cliffs, NJ 1971.

[3] H. Chernoff and E. L. Lehmanm: The use of maximum likelihood estimates in χ^2 tests of goodness of fit. Ann. Math. Statist. *25* (1954), 579–586.

[4] B. S. Clarke and A. R. Barron: Information–theoretic asymptotics and Bayes methods. IEEE Trans. Inform. Theory *36* (1990), 453–471.

[5] T.M. Cover and J. A. Thomas: Elements of Information Theory. New York, Wiley 1991.

[6] N. Cressie and T. R. C. Read: Multinomial goodness of fit tests. J. Roy. Statist. Soc. Ser. A *46* (1984), 440–464.

[7] I. Csiszár: Information–type measures of difference of probability distributions and their indirect observation. Studia Sci. Math. Hungar. *2* (1967), 299–318.

[8] I. Csiszár: Generalized cutoff rates and Rényi's information measures. IEEE Trans. Inform. Theory *41* (1995), 26–34.

[9] R. C. Dahiya and J. Gurland: Pearson chi–squared test of fit with random intervals. Biometrika *59* (1972), 147–153.

[10] A. Gersho and R. M. Gray: Vector Quantization and Signal Compression. Kluwer, Boston 1991.

[11] L. Györfi, I. Vajda and E. van der Meulen: Minimum Hellinger distance point estimates consistent under weak family regularity. Mathem. Methods of Statistics *3* (1994), 25–45.

[12] L. Györfi, I. Vajda and E. van der Meulen: Parameter estimation by projecting on structural families. In: Proc. 5th Prague Symp. on Asympt. Statistics (P. Mandl and H. Hušková, eds.), Physica Verlag, Wien 1994, pp. 261–272.

[13] W. C. M. Kallenberg, J. Oosterhoff and B. F. Schriever: The number of classes in chi–squared goodness of fit tests: J. Amer. Statist. Assoc. *80* (1985), 959–968.

[14] M. Menéndez, D. Morales, L. Pardo and I. Vajda: Divergence-based estimation and testing of statistical models of classification. J. Multivariate Anal. *54* (1995), 329–354.

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

[16] D. S. Moore: A chi–squared statistics with random cell boundaries. Ann. Math. Statist. *42* (1971), 147–156.

[17] D. Morales, L. Pardo and I. Vajda: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *49*, 1995.

[18] F. Österreicher and I. Vajda: Statistical information and discrimination. IEEE Trans. Inform. Theory *39* (1993), 1036–1039.

[19] F. H. Ruymgaart: A note on chi–square statistics with random cell boundaries. Ann. Statist. *3* (1975), 965–968.

[20] M. Teboulle and I. Vajda: Convergence of best ϕ–entropy estimates. IEEE Trans. Inform. Theory *39* (1993), 297–301.

[21] I. Vajda: From perceptron to Boltzman machine: Information processing by cognitive networks. In: Proc. of the Third European School of System Sciences (I. Figuearas, A. Moncho and R. Torres, eds.), Univ. of Valencia, Valencia 1994, pp. 65–68.

[22] A. Veselý and I. Vajda: Classification of random signals by neural networks. In: Proc. of 14th Internat. Congress of Cybernetics, University of Namur, Namur 1996, in print.

Professor Domingo Morales and Professor Leandro Pardo, Facultad de Matemáticas, Universidad Complutense de Madrid, Madrid. Spain.

Ing. Igor Vajda, DrSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.