On asymptotic sufficiency and optimality of quantizations[1]

A. Berlinet[2] *and I. Vajda*[3]

Abstract

It is known that quantizations of primary sources of information reduce the information available for statistical inference. We are interested in the quantizations for which the loss of statistical information can be controlled by the number of cells in the observation space used to quantize observations. If the losses for increasing numbers of cells converge to zero then we speak about asymptotically sufficient quantizations. Optimality is treated on the basis of rate of this convergence. The attention is restricted to the models with continuous real valued observations and to the interval partitions. We give easily verifiable necessary and sufficient conditions for the asymptotic sufficiency and, for a most common measure of statistical information, we study also the rate of convergence to the information in the original non-quantized models. Applications of the results in concrete models are illustrated by examples.

AMS 1991 subject classification: Primary 62 B 10. Secondary 94 A 17.

Key Words: Quantization. Information divergence. Asymptotic sufficiency. Rate of convergence.

1 Introduction and basic concepts

Let us start with the statistical model described by a σ-finite measure space $(\mathcal{X}, \mathcal{S}, \mu)$ and by two probability measures P, Q on $(\mathcal{X}, \mathcal{S})$ dominated by μ with densities

$$f = \frac{dP}{d\mu}, \quad g = \frac{dQ}{d\mu}.$$

Dissimilarity between P and Q is measured by ϕ-divergences

$$D_\phi(P,Q) = \int g\,\phi\left(\frac{f}{g}\right)\,d\mu \tag{1.1}$$

[1] Supported by the MSMT grant 1M 679 855 5601.

[2] I3M, UMR CNRS 5149, University of Montpellier II, Place Bataillon,
34095 Montpellier Cedex, France

[3] Institute of Information Theory and Automation, Czech Academy of Sciences,
182 08 Prague, Czech Republic

where the integral extends over all $\mathcal{X}$ and $\phi(t)$ is strictly convex in the domain $t > 0$ with $\phi(1) = 0$. Then, introducing the $*$-conjugated function $\phi^*(t) = t\,\phi(1/t)$ in the same domain and setting

$$(\phi(0), \phi^*(0)) = \lim_{t \downarrow 0} \, (\phi(t), \phi^*(t)) \, ,$$

we have

$$0 \leq D_\phi(P, Q) \leq \phi(0) + \phi^*(0). \tag{1.2}$$

The left equality holds if and only if $P = Q$. The right equality holds if and only if $P \perp Q$ unless $\phi(0) + \phi^*(0) = \infty$. For $\phi^*(0) = \infty$ or $\phi(0) = \infty$ the right equality holds if $P \not\ll Q$ or $Q \not\ll P$ respectively but these conditions are not necessary. For details about the definition (1.1) and for the basic properties of ϕ-divergences used in this paper, we refer to Liese and Vajda (1987).

By Theorems 1 and 2 in Österreicher and Vajda (1993), there is one-to-one relation between the ϕ-divergences (1.1) and the measures of statistical information introduced by De Groot (1970). Namely, for every ϕ figuring in (1.1) there exists an experiment with the sample family $\{P, Q\}$ such that $D_\phi(P, Q)$ is the statistical information in the experiment and, conversely, for every experiment with the sample family $\{P, Q\}$ there exists ϕ of the type assumed in (1.1) such that the statistical information in the experiment is $D_\phi(P, Q)$. This motivates our use of the ϕ-divergences as measures of statistical information.

Research on quantizations is motivated by the fact that computers process information through discrete methods requiring quantizations of primary sources of the above mentioned statistical information $D_\phi(P, Q)$. Quantizations in the space $(\mathcal{X}, \mathcal{S})$ can be represented as finite or infinite $\mathcal{S}$-measurable partitions

$$\mathcal{P}_k = \{A_{kj} : 1 \leq j \leq k\} \tag{1.3.A}$$

or

$$\mathcal{P}_k = \{A_{kj} : j = 1, 2 \ldots\} \tag{1.3.B}$$

of $\mathcal{X}$. The quantization states are the events $A_{kj} = A_{k,j}$ or their respective indices $(kj) = (k, j)$. These states are supporting the discrete distributions

$$\mathbf{p}_k = (p_{kj}) \quad \text{and} \quad \mathbf{q}_k = (q_{kj})$$

with (1.4)

$$p_{kj} = P(A_{kj}) \quad \text{and} \quad q_{kj} = Q(A_{kj}).$$

The ϕ-divergences $D_\phi(P_k, Q_k)$ for restrictions P_k, Q_k of P, Q to the algebras $\mathcal{S}_k \subset \mathcal{S}$ generated by $\mathcal{P}_k$ are denoted in this paper by $D_\phi(P, Q|\mathcal{P}_k)$, i.e.

$$D_\phi(P, Q|\mathcal{P}_k) = D_\phi(\mathbf{p}_k, \mathbf{q}_k) = \sum_j q_{kj}\phi\left(\frac{p_{kj}}{q_{kj}}\right). \tag{1.5}$$

It is known that

$$D_\phi(P, Q|\mathcal{P}_k) \leq D_\phi(P, Q) \tag{1.6}$$

and that this inequality is strict if $0 < D_\phi(P, Q) < \infty$ (i.e. if $P \neq Q$ and $\phi(0) + \phi^*(0) < \infty$), unless the likelihood ratio $f(x)/g(x)$ is constant on each $A_{kj} \in \mathcal{P}_k$. If this is the case then in our statistical model one needs not distinguish between different observations $x \in A_{kj}$, i.e. the events $A_{kj} \in \mathcal{P}_k$ can be replaced without loss of information by singletons. More rigorously, then the sub-σ-algebra of $\mathcal{S}$ generated by $\mathcal{P}_k$ is sufficient for the pair P, Q. This means that the model is in fact discrete and no quantization is needed. The difference

$$D_\phi(P, Q) - D_\phi(P, Q|\mathcal{P}_k)$$

represents a loss of discernibility of distributions P and Q based on statistical observations. It is a loss of statistical information due to the quantization. We are interested in the quantizations for which this loss can be held under control by the partition parameter k. Therefore one component of our model is a sequence of partitions $\{\mathcal{P}_k\} = \{\mathcal{P}_k : k = 1, 2, \ldots\}$, and the problem is to find conditions on $\{\mathcal{P}_k\}$ under which

$$\lim_{k\to\infty} D_\phi(P, Q|\mathcal{P}_k) = D_\phi(P, Q) \tag{1.7}$$

for all ϕ-divergences under consideration.

Csiszár (1973) proved that if the sequence $\{\mathcal{S}_k\}$ of sub-σ-algebras of $\mathcal{S}$ generated by the sequence $\{\mathcal{P}_k\}$ satisfies for every $A \in \mathcal{S}$ the condition

$$\lim_{k\to\infty} \inf_{B\in\mathcal{S}_k} [P(A\Delta B) + Q(A\Delta B)] = 0 \tag{1.8}$$

where $A\Delta B$ is the symmetric difference $(A \setminus B) \cup (B \setminus A)$, then (1.7) holds. Vajda (2002) studied the model with the Euclidean space $\mathcal{X} = \mathbb{R}^d$, the σ-field $\mathcal{S}$ of Borel sets in $\mathbb{R}^d$, Lebesgue measure μ and the rectangle partitions $\{\mathcal{P}_k\}$. He proved that if for every $x \in \mathbb{R}^d$

$$\lim_{k\to\infty} Q(B_k(x)) = 0 \tag{1.9}$$

where $B_k(x) = A_{kj} \in \mathcal{P}_k$ if $x \in A_{kj}$ for $1 \leq j \leq k$, then (1.7) holds for all P and ϕ such that $D_\phi(P, Q) < \infty$. The condition (1.9) is simpler and more easily verifiable than (1.8).

The convergence in (1.7) is a basic desirable asymptotic property of sequences of quantizations $\{\mathcal{P}_k\}$ in the statistical problems involving measures P and Q. We call this property an *asymptotic sufficiency* of quantizations. In typical applications this optimality is required for all pairs $\{P, Q\}$ from a family of probability measures dominated by μ. The asymptotically sufficient sequences $\{\mathcal{P}_k\}$ maximizing the rate of convergence in (1.7) are *asymptotically optimal* in the obvious sense.

Note that the states of measurable quantizations $\mathcal{P} = \{A_1, ..., A_k\}$ of random observations X from Euclidean probability spaces $(\mathbb{R}^d, \mathcal{B}^d, \mu)$ are often the conditional expectations

$$E_\mu(X|A_1), ..., E_\mu(X|A_k)$$

taking on values in the sets $A_1, ..., A_k$ provided these are convex. There is an extensive literature dealing with quantizations $\mathcal{P}_k = \{A_{k1}, ..., A_{kk}\}$ of random observations $X \sim (\mathbb{R}^d, \mathcal{B}^d, \mu)$ which are *non-asymptotically optimal* in the sense that for a fixed k they minimize over all $\mathcal{P} = \{A_1, ..., A_k\}$ the average quantization errors

$$e_\mu(\mathcal{P}) = \sum_{A \in \mathcal{P}} \int_A \parallel x - E_\mu(X|A) \parallel^2 d\mu(x),$$

(see the recent monograph of Graf and Luschgy (2000) and references therein). We show below that the the asymptotic optimality studied in this paper is fully consistent with this non-asymptotic optimality.

In the present paper we study the simple variant of the model with absolutely continuous observations from $\mathcal{X}$ where $\mathcal{X} = (x_-, x_+)$ is an open interval in $\mathbb{R}$ with $-\infty \leq x_- < x_+ \leq \infty$, $\mathcal{S}$ is the σ-field of Borel subsets of $\mathcal{X}$ and μ is the Lebesgue measure. By $\{\mathcal{P}_k\}$ we denote a sequence of interval partitions of $\mathcal{X}$ and by P, Q two different measures with densities

$$f \geq 0, \qquad g > 0$$

and distribution functions

$$F(x) = \int_{x_-}^{x} f(t)d\mu(t), \qquad G(x) = \int_{x_-}^{x} g(t)d\mu(t), \qquad x \in \mathcal{X} \tag{1.10}$$

respectively. Since $g > 0$ on $\mathcal{X}$, the integrals in (1.1) and the sums in (1.5) are well defined. For this model we prove in Section 2 that (1.9) is equivalent to

$$\lim_{k\to\infty} \sup_j Q(A_{kj}) = 0 \tag{1.11}$$

and that (1.9) or (1.11) is sufficient for (1.7) even if $D_\phi(P,Q) = \infty$. We prove also that any of these conditions is necessary for (1.7) when $D_\phi(P,Q) < \infty$, *e.g.* when $\phi(0) + \phi^*(0) < \infty$.

However, the main results of the paper are in Section 3 where we study the rate of convergence in (1.7). In this respect we restrict ourselves to the χ^2-divergences defined by (1.1) for $\phi(t) = (t-1)^2$ or, equivalently, for $\phi(t) = t^2 - 1$, i.e. to

$$\chi^2(P,Q) = \int \frac{(f-g)^2}{g}\, d\mu = \int \frac{f^2}{g}\, d\mu - 1, \tag{1.12}$$

and to the finite partitions (1.3.A) so that, by (1.5),

$$\chi^2(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \frac{(p_{kj} - q_{kj})^2}{q_{kj}} = \sum_{j=1}^{k} \frac{p_{kj}^2}{q_{kj}} - 1. \tag{1.13}$$

The simplicity of the quadratic function $\phi(t) = t^2 - 1$ makes the analysis of the rate of the convergence

$$\lim_{k\to\infty} \chi^2(P,Q|\mathcal{P}_k) = \chi^2(P,Q) \tag{1.14}$$

easier than the analysis of the rate in (1.7) for a general function ϕ. The χ^2-divergences are thus convenient for a deeper insight into the problem of asymptotic optimality of quantizations $\mathcal{P}_k$.

The asymptotic optimality studied in the present paper is intimately connected with the non-asymptotic optimality mentioned above. For P and Q under consideration denote by $\mathcal{L} = f/g$ the likelihood ratio, by $\mu = Q\mathcal{L}^{-1}$ the corresponding probability measure induced by $\mathcal{L}$ on $(\mathbb{R},\mathcal{B})$, and consider the random variable X $\sim(\mathbb{R},\mathcal{B},\mu)$. As noticed by Bock (1992) (see also Poetzelberger and Strasser (2001)), if $E_\mu(X^2) < \infty$ then the above introduced error $e_\mu(\mathcal{P})$ is minimized by $\mathcal{P}_k = \{A_{k1},...,A_{kk}\}$ if and only if the divergence $\chi^2(P,Q|\mathcal{P})$ is maximized by $\tilde{\mathcal{P}}_k = \{\mathcal{L}^{-1}(A_{k1}),...,\mathcal{L}^{-1}(A_{k1})\}$. Thus the non-asymptotic optimality of quantizations $\mathcal{P}_k$ implies the asymptotic optimality of the corresponding sequence $\{\tilde{\mathcal{P}}_k = \mathcal{L}^{-1}\mathcal{P}_k\}$. On the other hand, $\chi^2(P,Q|\mathcal{P})$ is the larger, the closer it is to $\chi^2(P,Q)$ Therefore if $\{\tilde{\mathcal{P}}_k = \{\tilde{A}_{k1},...,\tilde{A}_{kk}\}\}$ is a sequence of interval partitions of $\mathcal{X}$ =(x_-,x_+) which is asymptotically optimal for $\{P,Q\}$ and the likelihood ratio $\mathcal{L}$ is monotone on $\mathcal{X}$ then one can expect that $\mathcal{P}_k = \{\mathcal{L}(\tilde{A}_{k1}), ..., \mathcal{L}(\tilde{A}_{kk})\}$ will not be too far from the k-state non-asymptotically optimal interval quantization of $X \sim(\mathbb{R}, \mathcal{B}, \mu)$.

The results about the convergence (1.14) and the rate of convergence there have also some other direct statistical applications. By Mayoral et al. (2003), the Fisher informations $I(\theta_0)$ and $I_k(\theta_0)$ in parametrized families $\{P_\theta : \theta \in \Theta\}$ and their restrictions $\{P_{k,\theta} : \theta \in \Theta\}$ due to the quantizations $\mathcal{P}_k$ characterize the powers of $\mathcal{P}_k$-based Pearson-type tests of the hypothesis $\theta = \theta_0$ against local alternatives. Kallenberg et al. (1985) used the fact that $I(0) = \chi^2(P,Q)$ and $I_k(0) = \chi^2(P,Q|\mathcal{P}_k)$ are the Fisher informations at $\theta_0 = 0$ in the families

$$\{P_\theta = (1-\theta)Q + \theta P : 0 \le \theta \le 1\}$$

and (1.15)

$$\{P_{k,\theta} = (1-\theta)Q_k + \theta P_k : 0 \le \theta \le 1\}.$$

We show that the partitions (1.3.A) satisfying natural assumptions fulfil the convergence (1.14) and that, under some restrictions on P, Q, the rate of this convergence is quadratic in $1/k$. For the G-uniform partitions we evaluate the constant at the asymptotic term $(1/k)^2$ and demonstrate by an example that this constant is not maximized by the standard uniform partitions. In other words, the standard uniform quantizations widely used in the modern electronic devices are not always asymptotically optimal in the sense of convergence in (1.14).

Note that the rate of convergence of information functionals and its statistical consequences were studied in a number of previous papers. In addition to those already mentioned above, see e.g. Ghurye and Johnson (1981), Zografos et al. (1986), Menéndez et al (2001) and further references therein. A key fact in the context of the present paper is that Kallenberg et al. (1985) have shown that the rate of convergence in (1.14) is also important when $\chi^2(P,Q) = \infty$. In this case the slower rate than $\sqrt{k}$ in (1.14) means that any $\mathcal{P}_k$-based Pearson test of the hypothesis Q is asymptotically for the sample size $n \to \infty$ more powerful against the local alternatives $(1-1/\sqrt{n})Q + P/\sqrt{n}$ than any $\mathcal{P}_{k_n}$-based test with $k_n \to \infty$ for $n \to \infty$. If the rate is faster than $\sqrt{k}$ then the tests with increasing numbers of partition sets are asymptotically more powerful than the test with any partition $\mathcal{P}_k$ of a fixed size k.

Our results about the rate of convergence in (1.14) for $\chi^2(P,Q) < \infty$ are new. For $\chi^2(P,Q) = \infty$ they extend the results of Kallenberg et al.

(1985) by employing similar arguments as theirs. Our convergence conditions are formulated in terms of the moment function of the likelihood ratio f/g (moment generating function of the log-likelihood ratio)

$$M_a(P,Q) = \int \left(\frac{f}{g}\right)^a g \, d\mu = \int \exp\left\{a \ln \frac{f}{g}\right\} d\mu \tag{1.16}$$

and its $\mathcal{P}_k$-reduced version

$$M_a(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \left(\frac{p_{kj}}{q_{kj}}\right)^a q_{kj} = \sum_{j=1}^{k} \exp\left\{a \ln \frac{p_{kj}}{q_{kj}}\right\} q_{kj}. \tag{1.17}$$

The convergence conditions of Kallenberg et al. (1985) were formulated in terms of the χ^a-divergence

$$\chi^a(P,Q) = \int \left|\frac{f-g}{g}\right|^a g \, d\mu$$

and its $\mathcal{P}_k-$reduced version

$$\chi^a(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \left|\frac{p_{kj}-q_{kj}}{q_{kj}}\right|^a q_{kj}.$$

The moment generating function is more familiar to the statisticians than the χ^a-divergence. It is also smoother and therefore simpler for analysis. Moreover, if f, g or $\mathbf{p}_k = (p_{kj}), \mathbf{q}_k = (q_{kj})$ are from an exponential statistical family then $M_a(P,Q)$ or $M_a(P,Q|\mathcal{P}_k)$ can be explicitly evaluated for all real a while for $\chi^a(P,Q)$ or $\chi^a(P,Q|\mathcal{P}_k)$ this is true only when a is a positive even integer.

Notice that if $a \neq 0$ and $a \neq 1$ then

$$I_a(P,Q) = \frac{M_a(P,Q)-1}{a(a-1)} \quad \text{and} \quad I_a(P,Q|\mathcal{P}_k) = \frac{M_a(P,Q|\mathcal{P}_k)-1}{a(a-1)} \tag{1.18}$$

are the power divergences of orders a defined by (1.1) and (1.5) for $\phi(t) = (t^a - 1)/(a(a-1))$. Hence in accordance with (1.6)

$$M_a(P,Q|\mathcal{P}_k) \leq M_a(P,Q) \quad \text{for} \quad a \geq 1 \tag{1.19}$$

and the inequality is strict when $a > 1, M_a(P,Q) < \infty$ and f/g is not piecewise constant in $\mathcal{X}$.

Let us note that the results of the paper obtained for sequences of finite interval partitions $\mathcal{P}_k = \{A_{k1}, \ldots, A_{kk}\}$, $k = 1, 2, \ldots$ remain valid for the subsequences $\mathcal{P}_{k_n}, n = 1, 2, \ldots$ Therefore they can be directly extended to arbitrary sequences of interval partitions $\tilde{\mathcal{P}}_n = \{\tilde{A}_{n1}, \ldots, \tilde{A}_{n,k_n}\}$, $n = 1, 2, \ldots$ for $\lim_{n \longrightarrow \infty} k_n = \infty$ by replacing k in the results with k_n and $k \longrightarrow \infty$ with $n \longrightarrow \infty$, and by taking $\mathcal{P}_{k_n} = \widetilde{\mathcal{P}}_n$ for all n.

2 Convergence

In this section we study the model considered in the second half of Section 1, with $\mathcal{X} = (x_-, x_+) \subset \mathbb{R}$ and the Lebesgue measure μ on the Borel subsets of $\mathcal{X}$. We consider a sequence $\{\mathcal{P}_k\}$ of quantizations (interval partitions) of $\mathcal{X}$ and measures P, Q with Lebesgue densities f, g where $g > 0$ on $\mathcal{X}$, and with distribution functions F, G on $\mathcal{X}$. To avoid trivial situations where the equality in (1.6) can take place, we suppose that the likelihood ratio f/g is not piecewise constant on any open interval in X. This in particular implies $P \neq Q$, *i. e.* the trivial case $P = Q$ is excluded. The partitions $\mathcal{P}_k$ may be either finite of the type (1.3.A) or infinite of the type (1.3.B).

Definition 1 *If*

$$\lim_{k \to \infty} D_\phi(P, Q|\mathcal{P}_k) = D_\phi(P, Q) \tag{2.1}$$

for all ϕ under consideration then the sequence $\{\mathcal{P}_k\}$ of quantizations is said to be asymptotically sufficient for $\{P, Q\}$.

Notice that the convergence (2.1) is required irrespectively of whether $D_\phi(P, Q) < \infty$ or $D_\phi(P, Q) = \infty$. In the following theorem dealing with necessary conditions and sufficient conditions for (2.1) and for the asymptotic sufficiency of sequences $\{\mathcal{P}_k\}$, it is useful to take into account that there exist ϕ-divergences with $D_\phi(P, Q) < \infty$ for all P, Q. By (1.2), for this the condition $\phi(0) + \phi^*(0) < \infty$ suffices. An example is the Hellinger divergence defined by $\phi(t) = (1 - \sqrt{t})^2$ for which

$$0 \le H(P, Q) = \int (\sqrt{f} - \sqrt{g})^2 d\mu \le 2,$$

or the power divergences (1.18) of the orders $0 < a < 1$, for which

$$0 \le I_a(P, Q) \le \frac{1}{a(1-a)}.$$

Theorem 1 *If $\{\mathcal{P}_k\}$ satisfies the condition*

$$\lim_{k\to\infty} \sup_j Q(A_{kj}) = 0 \tag{2.2}$$

then it is asymptotically sufficient for $\{P, Q\}$. If (2.2) does not hold then

$$\liminf_{k\to\infty} D_\phi(P, Q|\mathcal{P}_k) < D_\phi(P, Q) \tag{2.3}$$

for all ϕ under consideration with $D_\phi(P, Q) < \infty$. Therefore (2.2) and (1.9) are equivalent conditions and each of them is necessary and sufficient for the asymptotic sufficiency of $\{\mathcal{P}_k\}$ for $\{P, Q\}$.

Proof. Since (2.2) implies (1.9), it implies (2.1) for all convex $\phi : (0, \infty) \mapsto \mathbb{R}$ with $\phi(0) + \phi^*(0) < \infty$, or with $\phi(0) + \phi^*(0) = \infty$ and $D_\phi(P, Q) < \infty$, according to Theorem 2 in Vajda (2002). If $D_\phi(P, Q) = \infty$ then $\phi(0) + \phi^*(0) = \infty$ and we can use the convex functions $\phi_i : (0, \infty) \mapsto \mathbb{R}$ defined for all integers $i \geq 2$ by

$$\phi_i(t) = \begin{cases} \phi(i) + \phi'_+(i)(t - i) & \text{for} \quad t \geq i \\ \phi(t) & \text{for} \quad (1/i) < t < i \\ \phi(1/i) + \phi'_+(1/i)(t - (1/i)) & \text{for} \quad 0 \leq t \leq (1/i) \end{cases}$$

where ϕ'_+ stands for the right-hand derivative. Obviously,

$$\phi_i(0) + \phi_i^*(0) = \phi(1/i) - \phi'_+(1/i)/i + \phi'_+(i) < \infty$$

and the functions ϕ_i are ordered in the sense $\phi_2 \leq \phi_3 \leq \ldots \leq \phi$, tending pointwise to ϕ for i tending to infinity. Hence for every i and k

$$D_\phi(P, Q|\mathcal{P}_k) \geq D_{\phi_i}(P, Q|\mathcal{P}_k)$$

and (2.2) implies that

$$\lim_{k\to\infty} D_{\phi_i}(P, Q|\mathcal{P}_k) = D_{\phi_i}(P, Q)$$

where

$$\lim_{i\to\infty} D_{\phi_i}(P, Q) = D_\phi(P, Q) = \infty$$

by the monotone convergence theorem. This implies the desired relation

$$\lim_{k\to\infty} D_\phi(P, Q|\mathcal{P}_k) = \infty.$$

The proof of necessity of (2.2) for (2.1) when $D_\phi(P, Q) < \infty$ is based on Lemma A.1. in the Appendix. By this lemma, there exists an interval

$A \subset \mathcal{X}$ such that for some partition intervals $A_{kj_k} \in \mathcal{P}_k$ and a subsequence $\{k_n\}$ of $\{k\}$

$$A \subset A_n : = (A_{kj_k})_{k=k_n}$$

for all $n = 1, 2, \ldots$ Let $\mathcal{S}_A$ be the sub-σ-algebra of the Borel σ-algebra $\mathcal{S}$ consisting of the sets C and $C \cup A$ for all Borel subsets $C \subseteq \mathcal{X} \setminus A$. Further, denote by P_A, Q_A the restrictions of P, Q to $\mathcal{S}_A$. Since A is contained in $A_n \in \mathcal{P}_{k_n}$ and disjoint with the remaining intervals of $\mathcal{P}_{k_n}$ it holds $\mathcal{P}_{k_n} \subset \mathcal{S}_A$ and consequently

$$\mathcal{S}(\mathcal{P}_{k_n}) \subset \mathcal{S}_A \subset \mathcal{S} \quad \text{for all } n = 1, 2, ...$$

By the monotonicity of ϕ-divergences (see Corollary 1.29 in Liese and Vajda (1987)), this implies

$$D_\phi(P, Q|\mathcal{P}_{k_n}) \leq D_\phi(P_A, Q_A) < D_\phi(P, Q) \qquad \text{for all } n = 1, 2, \ldots,$$

where the last inequality is strict because ϕ is strictly convex, $D_\phi(P, Q) < \infty$ and the likelihood ratio f/g is not a. s. constant on A and thus is not $\mathcal{S}_A$-measurable. From here (2.3) follows immediately. □

Example 1. To see that the condition (2.2) is not necessary for (2.1) when $D_\phi(P, Q) = \infty$, consider the case where

$$\int_A g \, \phi\left(\frac{f}{g}\right) \, d\mu = \infty \tag{2.7}$$

and $\mathcal{X} \setminus A$ is an open interval. Take *e. g.* the ϕ-divergence $\chi^2(P, Q)$ defined by $\phi(t) = t^2 - 1$, the doubly exponential P with $f(x) = \exp\{-|x|\}$ on $\mathcal{X} = \mathbb{R}$ and the standard normal Q. Then (2.7) holds for every interval $A = (x, \infty)$, $x \in \mathbb{R}$. Under (2.7) we obtain (2.1) for every sequence of partitions

$$\mathcal{P}_k = \{\mathcal{X} \setminus A, A_{k1}, \ldots, A_{kk}\} \tag{2.8}$$

provided $\mathcal{P}_k^* = \{A_{k1}, \ldots, A_{kk}\}$ is an interval partition of A with the property $Q(A_{kj}) = Q(A)/k$. Since $Q(\mathcal{X} \setminus A) > 0$, the sequence (2.8) does not satisfy (2.2).

Example 2. To see that (2.2) is not necessary for (2.1) with $D_\phi(P, Q) < \infty$ when ϕ is not strictly convex, consider arbitrary ϕ strictly convex in the domain $t \in (0, 2)$ and linear, equal to $at + b$ for $t \geq 2$. Further consider P, Q with the likelihood ratio $f(x)/g(x)$ exceeding 2 on an open interval $\mathcal{X} \setminus A$ (as an example, we can take again the above proposed doubly exponential $f(x)$ and the standard normal $g(x)$). Then

$$D_\phi(P,Q) = \int_A g\,\phi\left(\frac{f}{g}\right)d\mu + a\,P(\mathcal{X}\setminus A) + b\,Q(\mathcal{X}\setminus A)$$

so that (2.1) holds for the partitions (2.8). Similarly as above, $Q(\mathcal{X}\setminus A) > 0$ implies that these partitions do not satisfy (2.2).

3 Rate of convergence

This section is a continuation of Section 2. We study the rate of convergence of the χ^2-divergences $\chi^2(P,Q|\mathcal{P}_k)$ to $\chi^2(P,Q)$ in the cases where $\chi^2(P,Q)$ is finite as well as infinite.
We restrict ourselves to the finite sequences of partitions

$$\mathcal{P}_k = \{A_{k1},\dots,A_{kk}\}$$

such that, for sufficiently large $\Gamma > 0$,

$$k \min_{1\le j\le k} Q(A_{kj}) \ge 1/\Gamma \quad \text{for all } k \tag{3.1}$$

and/or

$$k \max_{1\le j\le k} Q(A_{kj}) \le \Gamma \quad \text{for all } k. \tag{3.2}$$

Special attention is paid to the partitions into Q-equiprobable intervals where

$$Q(A_{kj}) = \frac{1}{k} \quad \text{for all } 1\le j\le k \text{ and all } k, \tag{3.3}$$

so that (3.1) and (3.2) hold for $\Gamma = 1$.

Let $G(x) = Q((-\infty,x)\cap\mathcal{X})$ be the distribution function of Q which is by assumption strictly increasing on $\mathcal{X}$. It transforms the open interval $\mathcal{X}$ onto $\mathcal{Y} = (0,1)$, the distribution Q into the Lebesgue measure on $(0,1)$ and the ϕ-divergences (1.1) into formally simpler integrals on $(0,1)$, namely

$$D_\phi(P,Q) = \int_0^1 \phi(p(y))\,dy \tag{3.4}$$

where

$$p(y) = \frac{f(G^{-1}(y))}{g(G^{-1}(y))}, \qquad y\in(0,1) \tag{3.5}$$

and G^{-1} is the quantile function from $[0,1]$ to the closure $[x_-, x_+]$ of $\mathcal{X}$ (the generalized inverse of the function G). The function G also defines a one to one relation between the partitions $\mathcal{P}_k$ under consideration and interval partitions of $(0,1)$. If $\{x_{kj} : 0 \leq j \leq k\}$ are the cutpoints of $\mathcal{X}$ leading to $\mathcal{P}_k$ (with x_{k0} and x_{kk} being the possibly infinite endpoints of $\mathcal{X}$) and $y_{k0} = 0 < y_{k1} < \ldots < y_{kk} = 1$ are similar cutpoints of $(0,1)$ leading to an interval partition of (0,1) then this relation is represented by the formulas

$$G(x_{kj}) = y_{kj} \quad \text{or} \quad x_{kj} = G^{-1}(y_{kj}), \quad 0 \leq j \leq k. \tag{3.6}$$

Partitions related by (3.6) satisfy the relation

$$Q(A_{kj}) = y_{kj} - y_{k,j-1}, \quad 1 \leq j \leq k \tag{3.7}$$

or, more generally, the probabilities defined in (1.4) satisfy the relations

$$p_{kj} = \int_{y_{k,j-1}}^{y_{kj}} p(y)\, dy \quad \text{and} \quad q_{kj} = \int_{y_{k,j-1}}^{y_{kj}} dy \tag{3.8}$$

where $p(y)$ is given by (3.5). It follows from here for example that the cutpoints x_{kj} of partitions $\mathcal{P}_k$ satisfying (3.3) are uniquely defined by

$$x_{kj} = G^{-1}(y_{kj}) \quad \text{for} \quad y_{kj} = \frac{j}{k}, \quad 0 \leq j \leq k, \tag{3.9}$$

and

$$\frac{1}{\Gamma k} \leq y_{kj} - y_{k,j-1} \leq \frac{\Gamma}{k}, \quad 1 \leq j \leq k, \tag{3.10}$$

for the cutpoints y_{kj} obtained by (3.6) from the partition $\mathcal{P}_k$ satisfying (3.1) and (3.2).

By (3.4) and (3.8), the χ^2-divergences under consideration can be expressed as follows

$$\chi^2(P,Q) = \int_0^1 p^2(y)\, dy - 1,$$

$$\tag{3.11}$$

$$\chi^2(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \frac{1}{y_{kj} - y_{k,j-1}} \left(\int_{y_{k,j-1}}^{y_{kj}} p(y)\, dy \right)^2 - 1.$$

Example 3. Let us consider the situation where $\mathcal{X} = \mathcal{Y} = (0,1)$ and the probability measures P and Q are defined by the distribution functions

$$F(x) = (G(x))^2 \text{ and } G(x) = \begin{cases} 4x/3 & \text{for} \quad 0 < x \leq 1/2 \\ 2x/3 + 1/3 & \text{for} \quad 1/2 < x < 1. \end{cases}$$

By (3.5) and (3.11),

$$p(y) = 2y \qquad \text{and} \qquad \chi^2(P,Q) = 4\int_0^1 y^2\, dy - 1 = \frac{1}{3}.$$

Let $\mathcal{P}_k$ be the uniform partition of $\mathcal{X} = (0,1)$ defined by the cutpoints $x_{kj} = j/k$ for $0 \le j \le k$ and $\mathcal{P}_k^*$ be the Q-uniform partition by the cutpoints $x_{kj}^* = G^{-1}(x_{kj})$ for $0 \le j \le k$. We shall compare $\chi^a(P,Q|\mathcal{P}_k)$ and $\chi^a(P,Q|\mathcal{P}_k^*)$. Assuming for simplicity that k is even we see that

$$y_{kj} = \begin{cases} 4j/(3k) & \text{for} \quad 0 \le j \le k/2 \\ \\ 2j/(3k) + 1/3 & \text{for} \quad k/2 < x \le k \end{cases}$$

and $y_{kj}^* = G(x_{kj}^*) = x_{kj}$ are the cutpoints defined by (3.6). Hence by (3.8) and (3.11),

$$\chi^2(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \frac{\left[\widetilde{F}(y_{kj}) - \widetilde{F}(y_{k,j-1})\right]^2}{y_{kj} - y_{k,j-1}} - 1$$

and

$$\chi^2(P,Q|\mathcal{P}_k^*) = \sum_{j=1}^{k} \frac{\left[\widetilde{F}(x_{kj}) - \widetilde{F}(x_{k,j-1})\right]^2}{x_{kj} - x_{k,j-1}} - 1$$

where $\widetilde{F}(y) = y^2$ is primitive to $p(y)$. Substituting the values of y_{kj} and x_{kj} specified above we get

$$\chi^2(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k/2} \frac{[(4j/(3k))^2 - (4(j-1)/(3k))^2]^2}{4/(3k)} +$$

$$\sum_{j=k/2+1}^{k} \frac{[(2j/(3k) + 1/3)^2 - (2(j-1)/(3k) + 1/3)^2]^2}{2/(3k)} - 1$$

and

$$\chi^2(P,Q|\mathcal{P}_k^*) = \sum_{j=1}^{k} \frac{[(j/k)^2 - ((j-1)/k)^2]^2}{1/k} - 1.$$

Applying the substitution $j = k/2 + i$ to $k/2 < j \le k$ in the formula for $\chi^2(P,Q|\mathcal{P}_k)$ and using repeatedly the formula

$$\sum_{j=1}^{k/2} j^2 = \frac{k(k+1)(k+2)}{24},$$

we obtain for every k under consideration

$$\chi^2(P,Q|\mathcal{P}_k) = \frac{1}{3}\left(1 - \frac{4}{k^2}\right)$$

and

$$\chi^2(P,Q|\mathcal{P}_k^*) = \frac{1}{3}\left(1 - \frac{1}{k^2}\right)$$

Thus $\chi^2(P,Q|\mathcal{P}_k)$ is a four times less accurate approximation of $\chi^2(P,Q) = 1/3$ than $\chi^2(P,Q|\mathcal{P}_k^*)$, *i. e.* the Q-uniform quantization $\mathcal{P}_k^*$ is for all k significantly better than the standard uniform quantization $\mathcal{P}_k$.

In the last example the reduced χ^2-divergences were of the form $\chi^2(P,Q) - \rho/k^2$ where $\rho = 4/3$ or $\rho = 1/3$ depending on whether the reduction was due to the quantization $\mathcal{P}_k$ or $\mathcal{P}_k^*$ respectively. The next theorem shows that if $\chi^2(P,Q) < \infty$ then for the Q-uniform partitions $\mathcal{P}_k$ the difference $\chi^2(P,Q) - \chi^2(P,Q|\mathcal{P}_k)$ tends to zero typically with the rate at least $1/k^2$ for k tending to infinity. For regular P,Q it shows that this rate is exactly $1/k^2$ and we explicitly evaluate the coefficient $\rho = \rho(P,Q)$ at $1/k^2$ in the asymptotic expansion of the difference. Note that in this theorem and in the sequel, the asymptotic formulas are considered for $k \to \infty$ unless otherwise explicitly stated.

Theorem 2 *Let $p(y)$ defined by (3.5) be twice continuously differentiable on (0,1) with first and second derivatives $\dot{p}(y)$ and $\ddot{p}(y)$, and let $\ddot{p}(y)$ be bounded on (0,1). Then $\chi^2(P,Q)$ is finite and*

$$\chi^2(P,Q|\mathcal{P}_k) = \chi^2(P,Q) - O\left(\frac{1}{k^2}\right) \tag{3.12}$$

for all sequences $\{\mathcal{P}_k\}$ satisfying (3.1) and (3.2). If $\{\mathcal{P}_k\}$ satisfies (3.3) then

$$\chi^2(P,Q|\mathcal{P}_k) = \chi^2(P,Q) - \frac{\rho(P,Q)}{k^2} + o\left(\frac{1}{k^2}\right) \tag{3.13}$$

where

$$\rho(P,Q) = \frac{1}{12}\int_0^1 [\dot{p}^2(y) + p(y)\ddot{p}(y)]\, dy > 0. \tag{3.14}$$

Proof. Let us start with a detailed proof of the second assertion which is more complicated. Suppose that $\{\mathcal{P}_k\}$ satisfies (3.3) so that

$$y_{kj} - y_{k,j-1} = \frac{1}{k}. \tag{3.15}$$

If $\overline{y}_{kj} = (y_{k,j-1} + y_{kj})/2$ then for $y \in (y_{k,j-1}, y_{kj})$

$$p(y) = p(\overline{y}_{kj}) + \dot{p}(\overline{y}_{kj})(y - \overline{y}_{kj}) + \frac{\ddot{p}(\overline{y}_{kj})}{2}(y - \overline{y}_{kj})^2 + R_{kj}(y),$$

where $R_{kj}(y)$ is the remainder in the Taylor series expansion. Since y varies in an interval of length $1/k$, the assumptions imply $R_{kj}(y) = o\,(1/k^2)$ uniformly for all $y \in (y_{k,j-1}, y_{kj})$ and $1 \le j \le k$. Therefore

$$p^2(y) = p^2(\overline{y}_{kj}) + 2\dot{p}(\overline{y}_{kj})p(\overline{y}_{kj})(y - \overline{y}_{kj}) + \rho(\overline{y}_{kj})(y - \overline{y}_{kj})^2 + o\left(1/k^2\right)$$

where $\rho(y)$ denotes the integrand of (3.14). Further,

$$\int_{y_{k,j-1}}^{y_{kj}} p(y)\,dy = \frac{p(\overline{y}_{kj})}{k} + o\left(1/k^2\right)$$

and

$$\int_{y_{k,j-1}}^{y_{kj}} p^2(y)\,dy = \frac{p^2(\overline{y}_{kj})}{k} + \frac{\rho(\overline{y}_{kj})}{12k^3} + o\left(\frac{1}{k^3}\right).$$

Consequently

$$\int_0^1 p^2(y)\,dy = \frac{1}{k}\sum_{j=1}^{k} p^2(\overline{y}_{kj}) + \frac{1}{12k^3}\sum_{j=1}^{k} \rho(\overline{y}_{kj}) + o\left(\frac{1}{k^3}\right)$$

and

$$k\sum_{j=1}^{k}\left(\int_{y_{k,j-1}}^{y_{kj}} p(y)\,dy\right)^2 = \frac{1}{k}\sum_{j=1}^{k} p^2(\overline{y}_{kj}) + o\left(\frac{1}{k^2}\right).$$

Since $\rho(y)$ is Riemann integrable it holds

$$\frac{1}{12k}\sum_{j=1}^{k} \rho(\overline{y}_{kj}) = \frac{1}{12}\int_0^1 \rho(y)\,dy + o\,(1) = \rho(P,Q) + o\,(1)$$

and the previous two formulas imply that

$$k\sum_{j=1}^{k}\left(\int_{y_{k,j-1}}^{y_{kj}} p(y)\,dy\right)^2 = \int_0^1 p^2(y)\,dy - \frac{\rho(P,Q)}{k^2} + o\left(\frac{1}{k^2}\right).$$

Now (3.13) follows from (3.11) and (3.15). The first assertion can be proved by repeating similar steps with the formula (3.15) for $y_{kj} - y_{k,j-1} = Q(A_{kj})$ replaced with (3.9). □

Thus in the regular models of Theorem 2, the divergence $\chi^2(P, Q)$ is finite and the quantizations into Q-equiprobable or nearly Q-equiprobable states lead to a quadratic rate of convergence in (2.1).

Example 4. It is easy to see that the model of Example 3 with the Q-uniform quantization $\mathcal{P}_k^*$ satisfies all assumptions of Theorem 2. In this model (3.5) yields $p(y) = 2y$ so that (3.14) implies $\rho(P, Q) = 1/3$. Thus Example 3 verified by a direct calculation that in this concrete model (3.13) holds with $o(1/k^2) = 0$ for all even $k > 1$. The calculation indicates that (3.13) holds with $o(1/k^2) \neq 0$ also for odd $k > 1$. Let us now illustrate the applicability of Theorem 2 in one of the most familiar statistical models. Namely, let P and Q be from the logistic family on $\mathcal{X} = \mathbb{R}$ with distribution functions

$$F_{\theta_1}(x) = \frac{e^{x-\theta_1}}{1+e^{x-\theta_1}} \quad \text{and} \quad F_{\theta_2}(x) = \frac{e^{x-\theta_2}}{1+e^{x-\theta_2}}, \theta_1 \neq \theta_2$$

and densities

$$f(x) = \frac{e^{x-\theta_1}}{[1+e^{x-\theta_1}]^2} \quad \text{and} \quad g(x) = \frac{e^{x-\theta_2}}{[1+e^{x-\theta_2}]^2}$$

respectively. Here

$$G^{-1}(y) = \theta_2 + \ln\frac{y}{1-y} \quad \text{for } y \in (0,1)$$

so that

$$g(G^{-1}(y)) = y(1-y)$$

and

$$f(G^{-1}(y)) = \frac{\tau y(1-y)}{[1+y(\tau-1)]^2} \quad \text{for } \tau = e^{\theta_2-\theta_1} > 0, \tau \neq 1. \tag{3.16}$$

Therefore the function (3.5) takes on the form

$$p(y) = \frac{\tau}{[1+y(\tau-1)]^2}$$

for τ given in (3.16), and its derivatives are

$$\dot{p}(y) = \frac{-2\tau(\tau-1)}{[1+y(\tau-1)]^3} \quad \text{and} \quad \ddot{p}(y) = \frac{6\tau(\tau-1)^2}{[1+y(\tau-1)]^4}.$$

We see that the assumptions of Theorem 2 are satisfied and that

$$p^2(y) = \frac{\tau^2}{[1+y(\tau-1)]^4}$$

and

$$\dot{p}^2(y) + p(y)\ddot{p}(y) = \frac{10\tau^2(\tau-1)^2}{[1+y(\tau-1)]^6}.$$

Hence the χ^2-divergence of (3.11) is

$$\chi^2(P,Q) = \frac{(\tau-1)(\tau^2+2\tau+3)}{3}$$

for τ given in (3.16). By Theorem 2, the reduced value $\chi^2(P,Q|\mathcal{P}_k)$ of this divergence after the quantization of $\mathcal{X}$ by the cutpoints

$$x_{kj} = G^{-1}(j/k) = \theta_2 + \ln\frac{j}{k-j}, \quad 1 \le j \le k-1$$

satisfies the asymptotic relation (3.13) with

$$\rho(P,Q) = \frac{2(\tau-1)^2(\tau^4+\tau^3+\tau^2+\tau+1)}{\tau^3}$$

for the same τ as above.

The next theorem deals with the rate of convergence in (2.1) in the case where $\chi^2(P,Q)$ is infinite. We use the terminology introduced by the three following definitions. For illustration of the concepts defined there we refer to Example 5 below.

Definition 2 *A nonnegative sequence s_k is said to be of the order of at most k^c (in symbols, $s_k \lesssim k^c$) or at least k^c (in symbols, $s_k \gtrsim k^c$) if $s_k/k^b \to 0$ for all $b > c$, or $s_k/k^b \to \infty$ for all $b < c$, respectively. If $s_k \lesssim k^c$ and also $s_k \gtrsim k^c$ then we say that s_k is of the order of k^c (in symbols $s_k \approx k^c$).*

The following definition deals with the nonnegative functions $p(y)$ of (3.5) leading to infinite divergence $\chi^2(P,Q)$. We see from (3.11) that such functions must be unbounded on the definition domain (0,1).

Definition 3 *We say that the function p is regularly unbounded if its extension in $[0,1]$ is bounded except in neighborhoods of finitely many points. If it is not bounded in a right (left) neighborhood of $y \in [0,1]$ then it is assumed that $h(t) = p(y+1/t)$ (or $h(t) = p(y-1/t)$) varies regularly at infinity, i.e. that for sufficiently large $t > 0$ and some $\rho \in \mathbb{R}$*

$$h(t) = t^\rho \lambda(t) \tag{3.17}$$

where $\lambda(t)$ varies slowly at infinity in the sense that

$$\lim_{t\to\infty} \frac{\lambda(t\alpha)}{\lambda(t)} = 1 \quad \text{for any } \alpha > 0.$$

In the next definition it is useful to take into account that the moment function $M_a(P,Q)$ defined by (1.16) is skew symmetric about $a = 1/2$ in the sense that $M_{1-a}(P,Q) = M_a(Q,P)$ and $0 \leq M_a(P,Q) \leq 1$ for $a \in [0,1]$.

Definition 4 *The values*

$$a_+ = a_+(P,Q) = \sup\{a \geq 1 : M_a(P,Q) < \infty\}$$

and

$$a_- = a_-(P,Q) = \inf\{a \leq 0 : M_a(P,Q) < \infty\}$$

are maximal and minimal effective arguments of the moment function. The value

$$c = c(P,Q) = \frac{2 - a_+}{a_+}$$

assumed to be -1 when $a_+ = \infty$ and taking on values from the interval $(-1,1]$ when $a_+ < \infty$, is said to be a characteristic exponent of P and Q.

Note that Definition 3 summarizes properties of $p(y)$ previously considered by Kallenberg et al. (1985). The following theorem extends Propositions 4.2 and 4.4 of these authors.

Theorem 3 *If the characteristic exponent $c = c(P,Q)$ is negative, i.e. if the maximal effective argument $a_+(P,Q) > 2$, then $\chi^2(P,Q) < \infty$. If c is positive, i.e. if $a_+(P,Q) < 2$ then $\chi^2(P,Q) = \infty$. In the latter case*

$$\chi^2(P,Q|\mathcal{P}_k) \lesssim k^c \tag{3.18}$$

provided $\{\mathcal{P}_k\}$ satisfies (3.1) and

$$\chi^2(P,Q|\mathcal{P}_k) \gtrsim k^c \tag{3.19}$$

provided $\{\mathcal{P}_k\}$ satisfies (3.2) and $p(y)$ is regularly unbounded in the sense of Definition 3. Therefore in the models with positive characteristic exponent c

$$\chi^2(P,Q|\mathcal{P}_k) \approx k^c \tag{3.20}$$

provided all the mentioned conditions hold.

Proof. The first assertion follows directly from Lemma A.2 in the Appendix. To prove (3.18), put $b_0 = 1$ if $c = 1$ and $b_0 \in (c,1)$ otherwise. Then $M_{a_0}(P,Q)$ is finite for $a_0 = 2/(1+b_0)$ and, by Lemma A.3, the sequence $k^{-b_0}\chi^2(P,Q|\mathcal{P}_k)$ is bounded. This means that for all $b > b_0$ (and, consequently, for all $b > c$)

$$k^{-b}\chi^2(P,Q|\mathcal{P}_k) \to 0$$

i.e. (3.18) holds. Relation (3.19) follows directly from Lemma A.4 and (3.20) is clear. □

Example 5. A simple application of Theorem 3 is obtained when P and Q are probability measures on the observation space $\mathcal{X} = (0,1)$ with distribution functions $F(x) = F_\theta(x) = x^{1-\theta}$ for some $0 < \theta < 1$ and $G(x) = x$. Then $f(x) = (1-\theta)x^{-\theta}$ and (3.5) implies that $p(y) = (1-\theta)y^{-\theta}$ for $\mathcal{Y} = (0,1)$ which is a regularly unbounded function on $\mathcal{Y}$ in the sense of Definition 3. By (3.11) and (1.16),

$$\chi^2(P,Q) = \begin{cases} \theta^2/(1-2\theta) & \text{if} \;\; 0 < \theta < 1/2 \\ \infty & \text{if} \;\; 1/2 \le \theta < 1. \end{cases}$$

and

$$M_a(P,Q) = \begin{cases} (1-\theta)^a/(1-a\theta) & \text{if} \;\; a < 1/\theta \\ \infty & \text{if} \;\; a \ge 1/\theta. \end{cases}$$

Hence the maximal effective argument is $a_+ = 1/\theta$ and, by Definition 4, the characteristic exponent is $c = c(P,Q) = 2\theta - 1$. It is in the interval $(-1,0)$ for $0 < \theta < 1/2$ and in the interval $(0,1)$ for $1/2 < \theta < 1$. Let $\{\mathcal{P}_k\}$ be the sequence of uniform partitions of $\mathcal{X}$ into k cells. Since these partitions are also Q-uniform, they satisfy (3.3). We see that all assumptions of Theorem 3 are satisfied. Therefore this theorem says that if $0 < \theta < 1/2$ then $\chi^2(P,Q) < \infty$ and if $1/2 < \theta < 1$ then $\chi^2(P,Q) = \infty$ which agrees with the direct above computations. The case $\theta = 1/2$ is ignored by the theorem, but it also says that if $1/2 < \theta < 1$ then $\chi^2(P,Q|\mathcal{P}_k) \approx k^{2\theta-1}$. By Definition 2, this means that $k^{2\theta-1}$ characterizes the rate of convergence of $\chi^2(P,Q|\mathcal{P}_k)$ to $\chi^2(P,Q) = \infty$ in the sense that, asymptotically for k tending to infinity, $\chi^2(P,Q|\mathcal{P}_k) = k^{2\theta-1+o(1)}$. This is a new fact about the special model under consideration. Its direct verification requires to evaluate

$$\chi^2(P,Q|\mathcal{P}_k) = \sum_{j=1}^{k} \frac{\left[(j/k)^{1-\theta} - ((j-1)/k)^{1-\theta}\right]^2}{1/k}$$

(cf. (3.8)-(3.11)) for $1/2 < \theta < 1$, or at least to prove the asymptotic relation

$$\ln \sum_{j=1}^{k} \left[j^\alpha - (j-1)^\alpha\right]^2 = o(\ln k)$$

for $0 < \alpha < 1/2$. These tasks are not so easy.

The results of this section are relevant to the theory of optimal quantizations $\mathcal{P}_k^*$ that maximize the divergence $\chi^2(P,Q|\mathcal{P}_k)$ over all k-elements interval partitions $\mathcal{P}_k$ of the observation space. Since the Q-uniform interval partitions $\mathcal{P}_k$ satisfy (3.3), and consequently also (3.1) and (3.2), the asymptotic representation (3.20) obtainable for these partitions can be used to estimate from below the maximal divergence $\chi^2(P,Q|\mathcal{P}_k^*)$. Similarly we can use the estimate

$$\chi^2(P,Q) - \chi^2(P,Q|\mathcal{P}_k^*) \leq \frac{\rho(P,Q)}{k^2} + o\left(\frac{1}{k^2}\right)$$

for $\rho(P,Q)$ given by (3.14) when P,Q are regular in the sense of Theorem 2. An analogous idea was recently applied by Mayoral *et al* (2003) to the partitions $\mathcal{P}_k^*$ that maximize the Fisher information in parametric models.

Appendix

Lemma A.1 *Let Q be an absolutely continuous probability measure on an interval $\mathcal{X} \subseteq \mathbb{R}$. If a sequence of intervals $A_k \subset \mathcal{X}$ satisfies the condition*

$$\limsup_{k\to\infty} Q(A_k) > 0$$

then there exists an open interval B and a subsequence $\{A_{k_n}\}$ of $\{A_k\}$ such that B is contained in A_{k_n} for all sufficiently large n and $Q(B) > 0$.

Proof. By assumption, there exists a subsequence $\{A_{k_r} : r \in 1, 2, \ldots\}$ such that

$$\inf_r Q(A_{k_r}) \geq 2\delta \quad \text{for some } \delta > 0. \tag{$A.1$}$$

Let $(a_1, a_2) \subset \mathcal{X}$ be such that $Q((a_1, a_2)) > 1 - \delta$. Define intervals

$$B_r = A_{k_r} \cap (a_1, a_2).$$

By (A.1), these intervals are nonvoid with

$$\inf_r Q(B_r) \geq \delta.$$

Since the endpoints b_{1r} and b_{2r} of B_r (where $b_{1r} \leq b_{2r}$) are in the compact set $[a_1, a_2]$, there exists a subsequence $\{r_n\}$ of $\{r\}$ for which both the limits

$$b_1 = \lim_{n\to\infty} b_{1r_n} \quad \text{and} \quad b_2 = \lim_{n\to\infty} b_{2r_n}$$

exist in $[a_1, a_2]$. By continuity of measure Q with respect to the set-theoretic convergence of events,

$$Q((b_1, b_2)) = \lim_{n\to\infty} Q(B_{r_n}) \geq \delta.$$

Therefore $b_2 > b_1$ and any nonvoid open subinterval $B \subset (b_1, b_2)$ together with the subsequence A_{k_n} where k_n is the index k_r for which $r = r_n$ satisfy the statement of the lemma. □

The remaining lemmas are applied in Section 3. Therefore in these lemmas we consider the same P, Q and $\mathcal{P}_k$ as in Section 3.

Lemma A.2 *The moment generating function (1.16) satisfies for all $0 < a_1 < a_2$ and all P, Q the inequality*

$$(M_{a_1}(P, Q))^{1/a_1} \leq (M_{a_2}(P, Q))^{1/a_2}.$$

In particular, the divergence $\chi^2(P,Q) = M_2(P,Q) - 1$ is bounded for all $0 < a_1 < 2$, $a_2 > 2$ and P,Q as follows

$$(M_{a_1}(P,Q))^{2/a_1} - 1 \leq \chi^2(P,Q) \leq (M_{a_2}(P,Q))^{2/a_2} - 1.$$

Proof. The first assertion follows from the convexity of $\psi(t) = t^{a_2/a_1}$ and the second assertion is a trivial consequence. □

Lemma A.3 *If $\mathcal{P}_k$ satisfies (3.1) then for every $1 \leq a \leq 2$ and for Γ which appears in (3.1)*

$$k^{(a-2)/a} M_2(P,Q|\mathcal{P}_k) \leq \Gamma^{(2-a)/a}(M_a(P,Q))^{2/a}.$$

Consequently for a sequence $\{\mathcal{P}_k\}$ satisfying (3.1) and all $a \in [1,2]$

$$\sup_{k\geq 1} k^{(a-2)/a}\chi^2(P,Q|\mathcal{P}_k) \leq \Gamma^{(2-a)/a}(M_a(P,Q))^{2/a}.$$

Proof. Let $k \geq 1$. If a is equal to 1 then $M_a(P,Q)$ is also equal to 1. As we have

$$\begin{aligned} M_2(P,Q|\mathcal{P}_k) &= \sum_{j=1}^{k} q_{kj}\left(\frac{p_{kj}}{q_{kj}}\right)^2 \\ &\leq \max_{1\leq j\leq k}\left(\frac{p_{kj}}{q_{kj}}\right) \\ &\leq k\,\Gamma \quad \text{(by (3.1))}, \end{aligned}$$

both inequalities easily follow. Now suppose that $1 < a \leq 2$ and put $b = (2-a)/a$. If $b = 0$ then the assertion is trivial. If $b > 0$ then

$$\begin{aligned} M_2(P,Q|\mathcal{P}_k) &= \sum_{j=1}^{k} q_{kj}\left(\frac{p_{kj}}{q_{kj}}\right)^2 \\ &\leq \max_{1\leq j\leq k}\left(\frac{p_{kj}}{q_{kj}}\right)^{ab}\sum_{j=1}^{k} q_{kj}\left(\frac{p_{kj}}{q_{kj}}\right)^a \qquad (ab+a=2) \\ &\leq \left(\max_{1\leq j\leq k}\frac{p_{kj}}{q_{kj}}\right)^{ab} M_a(P,Q). \end{aligned}$$

where the last inequality follows from the formula for $M_a(P,Q|\mathcal{P}_k)$ in (1.17) and from the inequality (1.19). By (3.8) and the Hölder inequality

$$p_{kj} = \int_{y_{k,j-1}}^{y_{kj}} p(y)\,dy \leq q_{kj}^{1-1/a}\left(\int_{y_{k,j-1}}^{y_{kj}} p(y)^a\,dy\right)^{1/a}$$

so that

$$\begin{aligned}\max_{1\le j\le k}\frac{p_{kj}}{q_{kj}} &\le \max_{1\le j\le k} q_{kj}^{-1/a}\left(\int_{y_{k,j-1}}^{y_{kj}} p(y)^a\,dy\right)^{1/a}\\ &\le \Gamma^{1/a}k^{1/a}\left(\max_{1\le j\le k}\int_{y_{k,j-1}}^{y_{kj}} p(y)^a\,dy\right)^{1/a} \quad \text{(see (3.1))}\\ &\le \Gamma^{1/a}k^{1/a}(M_a(P,Q))^{1/a} \quad \text{(see (1.16))}.\end{aligned}$$

Combining this with the previous inequality, we obtain

$$k^{-b}M_2(P,Q|\mathcal{P}_k) \le \Gamma^b(M_a(P,Q))^{2/a}$$

which completes the proof. □

The proof of the previous lemma uses the arguments of the proof of Proposition 4.2 in Kallenberg et al. (1985). The following lemma generalizes Proposition 4.4 of the same paper. Notice that its statement is trivial for $b < 0$ because the assumption (3.2) implies (2.2) so that $\chi^2(P,Q|\mathcal{P}_k) \to \chi^2(P,Q)$ for all P, Q by Theorem 1. Therefore if $b < 0$ then our assumption $P \ne Q$ implies $k^{-b}\chi^2(P,Q|\mathcal{P}_k) \to \infty$ automatically for all P, Q under consideration.

Lemma A.4. *Let P, Q be probability measures with the function $p(y)$ bounded or regularly unbounded in the sense of Definition 3 and with a characteristic exponent specified in Definition 4. If $\{\mathcal{P}_k\}$ satisfies (3.2) then for every $b < c$*

$$\lim_{k\to\infty} k^{-b}\chi^2(P,Q|\mathcal{P}_k) = \infty.$$

Proof. If $p(y)$ is bounded on $(0,1)$ then $c = -1$ so that the statement is trivial. It remains to be trivial unless $c > 0$, i.e. unless the maximal effective argument a_+ of the moment function $M_a(P,Q)$ is below 2. Therefore let $a_+ < 2$ and

$$\int_U p(y)^a\,dy = \infty \quad \text{if } a > a_+$$

for a neighborhood $U \subset [0,1]$ of at least one point $y \in [0,1]$. Let for simplicity the point be $y = 0$ and put for brevity $y_k = y_{k1}$ for y_{k1} defined in (3.6). Since (3.2) holds, we see that $k^{-b} \ge (y_k/\Gamma)^b$. Hence, by (3.11), it suffices to prove that the expression

$$y_k^b\frac{\left(\int_0^{y_k} p(y)\,dy\right)^2}{y_k} = \left(y_k^{(b-1)/2}\int_0^{y_k} p(y)\,dy\right)^2$$

tends to infinity if $0 \leq b < c$. For every $0 \leq b < c$ take a such that

$$a_+ < a < \frac{2}{b+1}.$$

Note that such a exists since $2/(b+1) > a_+$ for all $0 \leq b < c$. By the assumptions,

$$\int_0^{y_k} p(y)^a \, dy = \infty \quad \text{and} \quad \int_0^{y_k} p(y) \, dy < \infty.$$

If we put $t_k = 1/y_k$ then, by (3.8), $p(1/t) = t^\rho \lambda(t)$ for $\lambda(t)$ slowly varying at infinity, so that

$$\int_{t_k}^\infty t^{a\rho-2} \lambda(t)^a dt = \infty \quad \text{and} \quad \int_{t_k}^\infty t^{\rho-2} \lambda(t) dt < \infty.$$

Since $\lambda^a(t)$ is slowly varying at infinity too, the first assertion of the lemma on page 280 of Feller (1966) can be applied to both these relations. The first one implies in this manner that $\rho \geq 1/a$ and the second one implies $\rho \leq 1$. Now there are two possibilities : either $\rho = 1$ in which case

$$\int_{t_k}^\infty t^{-1} \lambda(t) dt < \infty$$

or $\rho < 1$. In both these cases the second assertion of the Feller lemma implies that

$$\int_t^\infty s^{\rho-2} \lambda(s) \, ds = t^{\rho-1} \Lambda(t)$$

where

$$\Lambda(t) = t^{1-\rho} \int_t^\infty s^{\rho-2} \lambda(s) \, ds$$

is slowly varying at infinity. By Lemma 2 on p. 277 of Feller, $\Lambda(t) > t^{-\varepsilon}$ for any fixed $\varepsilon > 0$ and all t sufficiently large. Therefore

$$t_k^\beta \int_{t_k}^\infty s^{\rho-2} \lambda(s) ds = y_k^{-\beta} \int_0^{y_k} p(y) \, dy \to \infty$$

whenever $\beta > 1 - \rho$. By definition of a and the inequality $\rho \geq 1/a$,

$$\frac{1-b}{2} > 1 - \frac{1}{a} \geq 1 - \rho$$

so that the desired relation

$$y_k^{(b-1)/2} \int_0^{y_k} p(y) \, dy \to \infty$$

is proved. □

Acknowledgement. We thank two referees and the Coordinating Editor for stimulating remarks and proposals and for the detection of two errors in the original version of the Appendix.

References

[1] Bock, H. H. (1992) A clustering technique for maximizing ϕ-divergence, noncentrality and discrimination power. *Analyzing and Modelling Data and Knowledge* (Ed. M. Schader), 19-36. Springer Verlag, Berlin.

[2] Cziszár, I. (1973) Generalized entropy and quantization problems. *Trans. 6th Prague Conference Inform. Theory, Statist. Decision Functions, Random Processes*, 159–174. Academia, Prague.

[3] De Groot, M. H. (1970) *Optimal Statistical Decisions.* McGraw Hill, New York.

[4] Feller, W. (1966) *An Introduction to Probability Theory and its Applications*, Vol. 2, Second edition. Wiley, New York.

[5] Ghurye, S. G. and Johnson, B. R. (1981) Discrete approximations to the information integral. *Canadian Journal of Statistics,* 9, 27-37.

[6] Graf, S. and Luschgy, H. (2000) *Foundations of Quantization for Probability Distributions.* Springer Verlag, Berlin.

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

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

[9] Mayoral, A. M., Morales, D., Morales, J. and Vajda, I. (2003) On efficiency of estimation and testing with data quantized to fixed number of cells. *Metrika,* 57, 1–27.

[10] Menéndez, M. L., Morales, D., Pardo, L. and Vajda, I. (2001) Minimum disparity estimators for discrete and continuous models. *Applications of Mathematics,* 46, 439-466.

[11] Österreicher, F. and Vajda, I. (1993) Statistical information and discrimination. *IEEE Transactions on Inform. Theory,* 39, 1036–1039.

[12] Poetzelberger, K. and Strasser, H. (2001) Clustering and quantization by MSP-partitions. *Statistics and Decisions,* 19, 331-371.

[13] Vajda, I. (2002) On convergence of information contained in quantized observations. *IEEE Transactions on Inform. Theory,* 48, 2163–2172.

[14] Zografos, K. , Ferentinos, K. and Papaioannou, T. (1986) Discrete approximations of Cziszár, Rényi and Fisher measures of information. *Canadian Journal of Statistics,* 14, 355-366.