Large deviations of divergence measures on partitions

Jan Beirlant * Luc Devroye † László Györfi ‡ Igor Vajda §

July 13, 2000

Abstract

We discuss Chernoff-type large deviation results for the total variation, the I-divergence errors, and the χ^2-divergence errors on partitions. In contrast to the total variation and the I-divergence, the χ^2-divergence has an unconventional large deviation rate. Applications to Bahadur efficiencies of goodness-of-fit tests based on these divergence measures for multivariate observations are given.

AMS Classification: 62G10

Key words and phrases: large deviations, partitions, goodness-of-fit, total variation, I-divergence, χ^2-divergence, Bahadur exact slope

*Department of Mathematics, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3000 Leuven, Belgium. e-mail: jan.beirlant@wis.kuleuven.ac.be

†School of Computer Sciences, McGill University, Montreal, Canada H3A 2K6. e-mail: luc@kriek.cs.mcgill.ca. Supported by a K.U. Leuven Research Fellowship

‡Department of Computer Science and Information Theory, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary. e-mail: gyorfi@inf.bme.hu

§Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, CZ-182 08 Prague, Czech Republic. e-mail: vajda@utia.cas.cz. Supported by the GACR grant 102/99/1137 and the European grant Copernicus 579

1 Introduction

We consider the problem of testing an unknown probability density function. The test statistics are derived from dissimilarity measures of probability measures, like ϕ-divergences introduced by Csiszár (1967). The three most important ϕ-divergences in mathematical statistics and information theory are the total variation distance, the information divergence and the χ^2-divergence. For recent accounts of the theory of ϕ-divergences, see Liese and Vajda (1987) and Vajda (1989).

If μ and ν are probability measures on $I\!R^d$ ($d \geq 1$), then the *total variation distance* between μ and ν is defined by

$$V(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Borel sets A.

The *information divergence* (also called I-divergence, Kullback-Leibler number, relative entropy) of μ and ν is defined by

$$I(\mu, \nu) = \sup_{\{A_j\}} \sum_j \mu(A_j) \log \frac{\mu(A_j)}{\nu(A_j)},$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$. The term "information divergence" is due to Csiszár (1967), who introduced it to identify the Kullback-Leibler mean information for discrimination between two densities f and g (Kullback and Leibler (1951)) within the more general class of ϕ-divergences of f and g (Csiszár (1967), Ali and Silvey (1966)).

The following inequality, termed Pinsker's inequality, gives an upper bound to the total variation in terms of I-divergence (see e.g. Csiszár (1967), Kullback (1967) and Kemperman (1969)):

$$2\{V(\mu, \nu)\}^2 \leq I(\mu, \nu). \tag{1}$$

The χ^2*-divergence measure* between μ and ν is defined by

$$\chi^2(\mu, \nu) = \sup_{\{A_j\}} \sum_j \frac{(\mu(A_j) - \nu(A_j))^2}{\nu(A_j)},$$

where again the supremum is taken over all finite Borel measurable partitions $\{A_j\}$.

By using the inequality $\log t \leq t-1$ one easily obtains that

$$\sum_j \mu(A_j)\log\frac{\mu(A_j)}{\nu(A_j)} \leq \sum_j \frac{\mu(A_j)^2}{\nu(A_j)} - 1 = \sum_j \frac{(\mu(A_j)-\nu(A_j))^2}{\nu(A_j)},$$

from which

$$I(\mu,\nu) \leq \chi^2(\mu,\nu). \tag{2}$$

On the other hand, there are examples where $I(\mu,\nu_n)\to 0$ and $\chi^2(\mu,\nu_n)\to\infty$. Therefore the χ^2-divergence is strictly topologically stronger, i.e. the convergence in χ^2-divergence implies convergence in information divergence, but the converse is not true.

Applications of large deviation results in statistical analysis mainly concern the comparison of test procedures using Bahadur efficiencies. We consider the problem of testing hypotheses

$$H_0 : \nu = \mu \text{ versus } H_1 : \nu \neq \mu$$

by means of test statistics $T_n = T_n(X_1,\ldots,X_n)$ where $X_1, X_2, \ldots$ are independent and identically distributed along ν. Considering two tests rejecting H_0 for large values of the statistics $T_{n,1}$ and $T_{n,2}$, then (see e.g. Bahadur (1971), Groeneboom and Shorack (1981)) the efficiency $e_{1,2}$ of $T_{n,1}$ with respect to $T_{n,2}$ is calculated through the Bahadur exact slopes $2b_1(\nu)$ and $2b_2(\nu)$ for testing against an alternative ν: $e_{1,2} = \frac{b_1(\nu)}{b_2(\nu)}$. The functions b_1 and b_2 are then given by

$$b_j(\nu) = g_j(\psi_j(\nu)), \quad j = 1,2 \tag{3}$$

when ν is absolutely continuous, provided

$$T_{n,j} \to \psi_j(\nu) \text{ a.s. as } n\to\infty \text{ under } H_1, \tag{4}$$

$$\lim_{n\to\infty} \mathbf{E}T_{n,j} = 0 \text{ as } n\to\infty \text{ under } H_0, \tag{5}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}(T_{n,j} > \epsilon) = -g_j(\epsilon) \text{ under } H_0, \quad \epsilon > 0, \tag{6}$$

for $j = 1, 2$.
Such a limit assumption on the tail of the distribution of $T_{n,j}$ means that

$$\mathbf{P}(T_{n,j} > \epsilon) = e^{-n[g_j(\epsilon)+o(1)]}.$$

Each of the divergence measures defined above, when applied to the distance between the null-hypothesis distribution μ and the empirical distribution μ_n, both restricted to a partition (see Tusnády (1977) and Barron (1989)), does lead to a test procedure. Hence the large deviation results given in the subsequent sections offer the possibility to calculate exact Bahadur slopes for these tests. Quine and Robinson (1985) derived large deviation results for test statistics for uniformity based on I and χ^2, i.e. for the classical likelihood ratio and chi-square goodness-of-fit tests. These authors derived these results establishing probability inequalities. Here, we show how such results can be obtained using a classical result of Sanov (1957). Also we extend these results to a more general setting and derive large deviation results for other test statistics such as the L_1-test statistic. We also refer to Nikitin (1995) for a survey on Bahadur efficiencies of different well-known tests.

2 The L_1 error

We now consider some goodness of fit tests for H_0 given in the Introduction. Suppose that μ is nonatomic. Assume a sample of independent random vectors $X_1, \ldots, X_n$, distributed according to a probability measure μ, and let μ_n denote the empirical measure.

Györfi and van der Meulen (1991) introduced the test statistic

$$J_n = \sum_{j=1}^{m_n} |\mu(A_{n,j}) - \mu_n(A_{n,j})|,$$

based on a finite partition $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$, $(n \geq 2)$, of $I\!R^d$. These authors also showed that under H_0

$$\mathbf{P}(J_n \geq \epsilon) \leq e^{-n(\frac{\epsilon^2}{8}+o(1))}.$$

Moreover, the asymptotic normality of this test statistic under the null hypothesis in case μ has a density was discussed in Beirlant, Györfi and Lugosi (1995). Let $\bar{\mu}_n$ and μ_n^* be the restrictions of μ and μ_n to the partition $\mathcal{P}_n$, then

$$J_n = 2V(\bar{\mu}_n, \mu_n^*).$$

Theorem 1 *Assume that*

$$\lim_{n\to\infty} \max_j \mu(A_{n,j}) = 0 \tag{7}$$

and

$$\lim_{n\to\infty} \frac{m_n \log n}{n} = 0. \tag{8}$$

Then for all $0 < \epsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{J_n > \epsilon\} = -g(\epsilon), \tag{9}$$

where

$$g(\epsilon) = \inf_{0<p<1-\epsilon/2} \left(p \log \frac{p}{p+\epsilon/2} + (1-p) \log \frac{1-p}{1-p-\epsilon/2} \right). \tag{10}$$

In the proofs of our theorems we shall use the function

$$D(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta} \tag{11}$$

and the following lemma.

Lemma 1 (Sanov (1957), see p. 16 in Dembo, Zeitouni (1992), or Problem 1.2.11 in Csiszár and Körner (1981)). *Let* Σ *be a finite set of measurable sets (alphabet),* $\mathcal{L}_n$ *be a set of types (possible empirical distributions) on* Σ*, and let* Γ *be a set of distributions on* Σ*. Then*

$$\left| \frac{1}{n} \log \mathbf{P}\{\mu_n^* \in \Gamma\} + \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{|\Sigma| \log(n+1)}{n} \tag{12}$$

where $|\Sigma|$ *denotes the cardinality of* Σ*.*

PROOF OF THEOREM 1. We apply (12) for

$$\Sigma = \{A_{n,1}, \dots, A_{n,m_n}\}$$

such that

$$\Gamma = \{\tau : 2V(\bar{\mu}_n, \tau) \geq \epsilon\}.$$

Then, according to (12),

$$\left|\frac{1}{n}\log \mathbf{P}\{J_n \geq \epsilon\} + \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n)\right| \leq \frac{m_n \log(n+1)}{n}$$

and therefore, under (8),

$$\lim_{n\to\infty}\frac{1}{n}\log \mathbf{P}\{J_n > \epsilon\} = -\lim_{n\to\infty}\inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n).$$

It now remains to show that

$$\lim_{n\to\infty}\inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = g(\epsilon).$$

The distributions in $\mathcal{L}_n$ are possible empirical distributions, having components of the form $\frac{r}{n}$, where r is integer. Because of (7) we have that

$$m_n \to \infty,$$

therefore because of the continuity of $V(\tau, \bar{\mu}_n)$ and $I(\tau, \bar{\mu}_n)$

$$\lim_{n\to\infty}\inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = \lim_{n\to\infty}\inf_{2V(\tau,\bar{\mu}_n)\geq\epsilon} I(\tau, \bar{\mu}_n).$$

Here

$$I(\tau, \bar{\mu}_n) = \sum_{j=1}^{m_n} \tau(A_{n,j}) \log \frac{\tau(A_{n,j})}{\mu(A_{n,j})}.$$

Put

$$L = \{j : \mu(A_{n,j}) > \tau(A_{n,j})\}$$

and

$$A_n = \cup_{j\in L} A_{n,j}.$$

Then

$$2V(\tau, \bar{\mu}_n) = 2(\mu(A_n) - \tau(A_n))$$

and, by the Information Processing Theorem of Csiszár (1967),

$$I(\tau, \bar{\mu}_n) \geq D(\tau(A_n) \| \mu(A_n)),$$

where the equality holds iff $\frac{\tau(A_{n,j})}{\mu(A_{n,j})}$ is constant both on L and L^c. Thus

$$\begin{aligned}
& \lim_{n\to\infty} \inf_{2V(\tau,\bar{\mu}_n)\geq\epsilon} I(\tau, \bar{\mu}_n) \\
= \; & \inf_{0<p<1-\epsilon/2 : \tau(A_n)=p, \mu(A_n)=p+\epsilon/2} D(\tau(A_n) \| \mu(A_n)), \\
= \; & \inf_{0<p<1-\epsilon/2} \left(p \log \frac{p}{p+\epsilon/2} + (1-p) \log \frac{1-p}{1-p-\epsilon/2} \right) \\
= \; & g(\epsilon),
\end{aligned}$$

and Theorem 1 is proved.

Remark 1. Note that a lower bound for g follows from Pinsker's inequality (1) since

$$\inf_{2V(\tau,\bar{\mu}_n)\geq\epsilon} I(\tau, \bar{\mu}_n) \geq \inf_{I(\tau,\bar{\mu}_n)\geq\epsilon^2/2} I(\tau, \bar{\mu}_n),$$

and therefore

$$g(\epsilon) \geq \epsilon^2/2.$$

The best known lower bound is due to Toussaint (1975):

$$g(\epsilon) \geq \epsilon^2/2 + \epsilon^4/36 + \epsilon^6/280.$$

An upper bound $\hat{g}(\epsilon)$ of $g(\epsilon)$ can be obtained substituting p by $\frac{1-\epsilon/2}{2}$ in definition of $g(\epsilon)$. Then

$$\hat{g}(\epsilon) = \frac{\epsilon}{2} \log \frac{2+\epsilon}{2-\epsilon} \geq g(\epsilon)$$

(Vajda (1970)). Further bounds can be found on p. 294-295 in Vajda (1989). Remark that also in Lemma 5.1 in Bahadur (1971) it was observed that

$$g(\epsilon) = \frac{\epsilon^2}{2}(1 + o(1))$$

as $\epsilon \to 0$.

Local Bahadur efficiencies of the test J_n with respect to other goodness-of-fit tests can now be computed on the basis of the above Theorem.

EXAMPLE 1 (Testing for uniformity). Let μ be the restriction of the Lebesgue measure on (0,1). For $\theta \neq 0$ consider alternatives ν with density $f_\theta(x) = 1 + \theta h(x)$ $(x \in (0,1))$ where $\int_0^1 h = 0$. Then, using (3), (9), the asymptotic formula for $g(\epsilon)$ in Remark 1, and the fact that the ψ-function for this L_1-test equals $\int_0^1 |f_\theta(x) - 1| dx$, we find that the exact Bahadur slope for the J_n-statistic behaves as

$$\theta^2 (\int_0^1 |h(x)| dx)^2 \text{ when } \theta \to 0.$$

Remark that condition (5) can be verified here using Theorem 2.1 in Beirlant and Györfi (1998).

From Groeneboom and Shorack (1981) it follows that the exact Bahadur slope of the Kolmogorov-Smirnov test, respectively the Anderson-Darling test, behaves as $4\theta^2 (\sup_{x \in (0,1)} | \int_0^x h(t) dt|)^2$, respectively $2\theta^2 \int_0^1 (\int_0^x h)^2 [x(1-x)]^{-1} dx$, when $\theta \to 0$. Remark that

$$\begin{aligned} \left| \int_0^x h(t) dt \right| &= \left| \int_0^1 h(t) \left(1_{[0,x]}(t) - \frac{1}{2} \right) dt \right| \\ \leq & \int_0^1 |h(t)| \left| 1_{[0,x]}(t) - \frac{1}{2} \right| dt = \frac{1}{2} \left| \int_0^1 h(t) dt \right| . \end{aligned}$$

Hence the J_n-test is more Bahadur efficient than the Kolmogorov-Smirnov test in the given setting. It is also more efficient than the Anderson-Darling test (AD) for several examples of h. For example when $h(x) = \mathrm{sgn}(\frac{1}{2} - x)$ $(x \in (0,1))$, then $e_{J,AD} = 1/(4\ln(2) - 2) \sim 1.29$.

REMARK 2. The above result obviously can be generalized to a large deviation result for the L_1-inaccuracy rate of the histogram density estimator f_n based on a sample of independent random vectors $X_1, \ldots, X_n$, distributed according to a probability measure μ with density f with respect to the Lebesgue measure λ.

Introducing a partition $\mathcal{P}_n = \{A_{n,j}; \quad j \geq 1\}$, $(n \geq 2)$ of $I\!R^d$ such that $\sup_{j \geq 1} \lambda(A_{n,j}) < \infty$, then the *histogram density estimator* is defined by

$$f_n(x) = \frac{\mu_n(A_n(x))}{\lambda(A_n(x))}. \tag{13}$$

where $A_n(x) = A_{n,i}$ if $x \in A_{n,i}$.
If μ and ν are absolutely continuous with respect to a σ-finite measure λ with densities f and g respectively, then

$$\|f - g\| := \int |f(x) - g(x)| \lambda(dx) = 2\, V(\mu, \nu).$$

Hence for any density estimator the L_1-consistency implies the consistency in total variation.

Using the assumption that for each sphere S centered at the origin

$$\lim_{n\to\infty} \sup_{A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0. \tag{14}$$

and

$$\lim_{n\to\infty} \frac{|\{A_{n,j} \cap S \neq \emptyset\}|}{n} = 0 \tag{15}$$

then

$$\lim_{n\to\infty} \|f_n - f\| = 0 \tag{16}$$

a.s. and

$$\lim_{n\to\infty} \mathbf{E}\|f_n - f\| = 0 \tag{17}$$

(see Devroye and Györfi (1985)). Combining this result with Theorem 1 then yields the following result.

Corollary 1 *Assume (14). If there exists a sequence of spheres S_n centered at the origin such that $S_n \uparrow I\!R^d$ and*

$$\lim_{n\to\infty} card\{A_{n,j} \cap S_n \neq \emptyset\} \frac{\log n}{n} = 0 \tag{18}$$

then for all $0 < \epsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{\|f - f_n\| > \epsilon\} = -g(\epsilon), \tag{19}$$

where $g(\epsilon)$ is defined by (10).

Note that in this statement there is no condition on f, and so (19) holds for all f, and the rate function $g(\epsilon)$ does not depend on f.

PROOF OF COROLLARY 1. Without loss of generality assume that

$$A_{n,j} \cap S_n \neq \emptyset, \text{ for } j = 1, 2, \ldots m_n - 1$$

and

$$A_{n,j} \cap S_n = \emptyset, \text{ for } j = m_n, m_n + 1, \ldots.$$

Put

$$A'_{n,m_n} = \cup_{j=m_n}^{\infty} A_{n,j}$$

and

$$\Delta_n = \sum_{j=1}^{m_n-1} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A'_{n,m_n}) - \mu_n(A'_{n,m_n})|.$$

Then

$$\begin{aligned}
& \|f_n - f\| \\
= \;& \sum_{j=1}^{\infty} \int_{A_{n,j}} |f_n(x) - f(x)| \lambda(dx) \\
\geq \;& \sum_{j=1}^{m_n-1} \int_{A_{n,j}} |f_n(x) - f(x)| \lambda(dx) \\
& + \left| \sum_{j=m_n}^{\infty} \int_{A_{n,j}} f_n(x) \lambda(dx) - \sum_{j=m_n}^{\infty} \int_{A_{n,j}} f(x) \lambda(dx) \right| \\
\geq \;& \sum_{j=1}^{m_n-1} \left| \int_{A_{n,j}} f_n(x) \lambda(dx) - \int_{A_{n,j}} f(x) \lambda(dx) \right| \\
& + \left| \int_{A'_{n,m_n}} f_n(x) \lambda(dx) - \int_{A'_{n,m_n}} f(x) \lambda(dx) \right| \\
= \;& \sum_{j=1}^{m_n-1} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu_n(A'_{n,m_n}) - \mu(A'_{n,m_n})| \\
= \;& \Delta_n.
\end{aligned}$$

On the other hand

$$\begin{aligned}
& \|f_n - f\| \\
\leq \;& \|f_n - \mathbf{E} f_n\| + \|\mathbf{E} f_n - f\| \\
= \;& \sum_{j=1}^{\infty} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + \|\mathbf{E} f_n - f\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{m_n-1} |\mu(A_{n,j}) - \mu_n(A_{n,j})| + |\mu(A'_{n,m_n}) - \mu_n(A'_{n,m_n})| \\
&\quad +2\mu(A'_{n,m_n}) + \|\mathbf{E}f_n - f\| \\
&= \Delta_n + 2\mu(A'_{n,m_n}) + \|\mathbf{E}f_n - f\|.
\end{aligned}
$$

By definition $A'_{n,m_n} \subset S_n^c$, and by assumption $S_n \uparrow I\!R^d$, and therefore $\mu(A'_{n,m_n}) \to 0$. So together with (17)

$$2\mu(A'_{n,m_n}) + \|\mathbf{E}f_n - f\| \to 0,$$

and consequently Corollary 1 is proved if

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{\Delta_n > \epsilon\} = -g(\epsilon), \tag{20}$$

which follows from Theorem 1 with $J_n = \Delta_n$.

Remark 3. Louani (1998) derived large deviation results for a kernel density estimator in terms of the sup-norm distance. Several assumptions are required for this case, such as the boundedness of the density, and the results depend strongly on the bandwidth parameter. Louani (1999) showed a large deviation limit for the L_1-error of the kernel estimate, with the same rate function g as given in Theorem 1.

3 The information divergence

In the literature on goodness-of-fit testing two statistics are related to the information divergence, namely the *I-divergence statistic*

$$\tilde{I}_n = I(\bar{\mu}_n, \mu_n^*) = \sum_{j=1}^{m_n} \mu(A_{n,j}) \log \frac{\mu(A_{n,j})}{\mu_n(A_{n,j})},$$

and the *reversed I-divergence statistic*

$$I_n = I(\mu_n^*, \bar{\mu}_n) = \sum_{j=1}^{m_n} \mu_n(A_{n,j}) \log \frac{\mu_n(A_{n,j})}{\mu(A_{n,j})}.$$

If there are empty cells ($\mu_n(A_{n,j}) = 0$) then $\tilde{I}_n = \infty$, and hence I_n is more common. We show that

$$\mathbf{P}\{I_n > \epsilon\} = e^{-n(\epsilon+o(1))}$$

and

$$\mathbf{P}\{\tilde{I}_n > \epsilon\} = e^{-\frac{n}{m_n}(1+o(1))}.$$

Here we can refer to Tusnády (1977) and Barron (1989) who first discussed the exponential character of the tails of I_n.
Again a large deviation result concerning I_n can be derived proceeding similarly as in the proof of Theorem 1.

Theorem 2 (Corollary 2.4 in Kallenberg (1985), Theorem 2 in Quine and Robinson (1985)). *Under (7) and (8), for all* $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{I_n > \epsilon\} = -\epsilon.$$

This leads to the following observation concerning the goodness-of-fit tests based on I_n.

EXAMPLE 2. Again considering alternatives with density $f_\theta(x) = 1 + \theta h(x)$ ($x \in (0,1)$) where $\int_0^1 h = 0$, when testing for uniformity, the above result implies that the test based on I_n has an exact Bahadur slope $2\int_0^1 (1 + \theta h(x)) \log(1 + \theta h(x))dx \sim \theta^2 \int_0^1 h^2$ for $\theta \to 0$. Hence the Cauchy-Schwarz inequality implies that this test is more efficient than the test based on J_n.

From this point on we suppose that for all n

$$\mu(A_{n,j}) = \frac{1}{m_n}, \quad j = 1, \ldots, m_n. \tag{21}$$

This condition can be satisfied for any nonatomic distribution μ on $\mathbb{R}^d$, e.g. by using tree partitions Σ described in Chapter 20 in Devroye, Györfi and Lugosi (1996).

Theorem 3 *Assume that (21) holds. If, in addition,*

$$\lim_{n\to\infty} \frac{m_n^2 \log n}{n} = 0, \tag{22}$$

then we have for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{m_n}{n} \log \mathbf{P}\{\tilde{I}_n > \epsilon\} = -1$$

PROOF. Again apply (12) for $\Sigma = \{A_{n,1}, \ldots, A_{n,m_n}\}$ such that

$$\Gamma = \{\tau : \ I(\bar{\mu}_n, \tau) \geq \epsilon\}.$$

Then, according to (12),

$$\left| \frac{m_n}{n} \log \mathbf{P}\{\tilde{I}_n \geq \epsilon\} + m_n \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{m_n^2 \log(n+1)}{n}$$

Therefore, because of (21) and (22),

$$\begin{aligned} \lim_{n\to\infty} \frac{m_n}{n} \log \mathbf{P}\{\tilde{I}_n > \epsilon\} &= -\lim_{n\to\infty} m_n \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \\ &= -\lim_{n\to\infty} m_n \inf_{I(\bar{\mu}_n, \tau) \geq \epsilon} I(\tau, \bar{\mu}_n). \end{aligned}$$

Using the notation

$$\tau = (\tau_1, \ldots, \tau_{m_n})$$

we obtain

$$I(\tau, \bar{\mu}_n) = \sum_{j=1}^{m_n} \tau_j \log(m_n \tau_j) \tag{23}$$

and

$$I(\bar{\mu}_n, \tau) = -\frac{1}{m_n} \sum_{j=1}^{m_n} \log(m_n \tau_j).$$

Considering the distribution

$$\tau^* = \left(0, \frac{1}{m_n - 1}, \ldots, \frac{1}{m_n - 1}\right) \tag{24}$$

we find that

$$m_n \inf_{I(\bar{\mu}_n, \tau) \geq \epsilon} I(\tau, \bar{\mu}_n) \leq m_n I(\tau^*, \bar{\mu}_n) = m_n \log(m_n/(m_n - 1)) \to 1.$$

Concerning the lower bound we show that for all but finitely many n, the minimizing distribution is of the form

$$\tilde{\tau} = \left(\tau_1, \frac{1 - \tau_1}{m_n - 1}, \ldots, \frac{1 - \tau_1}{m_n - 1}\right), \tag{25}$$

where τ_1 (possibly depending on n) is bounded by

$$0 < \tau_1 \leq \frac{e^{-\epsilon m_n}}{m_n}.$$

To this end fix n and suppose that τ is the minimizing distribution, with coordinates ordered in the sense $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{m_n}$.

(a) It holds $\tau_1 > 0$. Indeed, if $\tau_1 = 0$ then, since the Shannon entropy is maximum for the uniform distribution, it follows that $I(\tau, \overline{\mu}_n) \geq I(\tau^*, \overline{\mu}_n) = \log(1/(1-1/m_n))$. This however contradicts the fact that the functions

$$I(\tilde{\tau}, \overline{\mu}_n) = D(\tau_1 \| 1/m_n), \quad I(\overline{\mu}_n, \tilde{\tau}) = D(1/m_n \| \tau_1), \tag{26}$$

(cfr. (11) and (25)) are continuous, strictly decreasing in the variable $\tau_1 \in (0, 1/m_n]$ with

$$D(0 \| 1/m_n) = \log \frac{1}{1 - 1/m_n}, \quad D(1/m_n \| 0) = \infty \tag{27}$$

and

$$D(1/m_n \| 1/m_n) = 0. \tag{28}$$

(b) There exists $1 \leq r_n < m_n$ such that

$$\tau_1 = \cdots = \tau_{r_n} < \tau_{r_n+1} = \cdots = \tau_{m_n}. \tag{29}$$

If (29) does not hold then either $\tau_1 = \tau_2 = \cdots = \tau_{m_n} = 1/m_n$ or there exist $1 < r_n < s_n \leq m_n$ such that $\tau_1 < \tau_{r_n} < \tau_{s_n}$. The first possibility contradicts the assumption $I(\overline{\mu}_n, \tau) \geq \epsilon > 0$. To see that the second possibility contradicts the definition of τ, suppose for simplicity that $r_n = 2$ and $s_n = 3$, i.e. let (cf. (a))

$$0 < \tau_1 < \tau_2 < \tau_3. \tag{30}$$

Consider a new distribution $\bar{\tau}$ with all coordinates coinciding with those of τ except $\overline{\tau}_1 = \tau_1 - \delta$, $\quad \overline{\tau}_2 = \tau_2 + \delta t$ and $\overline{\tau}_3 = \tau_3 - \delta(t-1)$, where $0 < \delta < \tau_1$ and $0 < t < 1 + \tau_3/\tau_1$ are chosen such that $I(\overline{\mu}_n, \overline{\tau}) = I(\overline{\mu}_n, \tau)$, i.e. such that

$$\sum_{j=1}^{3} \log \frac{\overline{\tau}_j}{\tau_j} = 0.$$

This means that

$$\left(-\frac{1}{\tau_1} + \frac{t}{\tau_2} - \frac{t-1}{\tau_3} \right) \delta + o(\delta) = 0 \quad \text{as } \delta \to 0,$$

i. e.

$$t = \frac{x-1}{y-1} + o(1) \quad \text{for } x = \frac{\tau_3}{\tau_1}, \quad y = \frac{\tau_3}{\tau_2}, \quad \frac{x-1}{y-1} > 1.$$

It is easy to verify that

$$I(\overline{\tau}, \overline{\mu}_n) = I(\tau, \overline{\mu}_n) + \delta\,(\ln x - t \ln y) + o(\delta) \quad \text{as } \delta \to 0,$$

where $x > y > 1$ and $(\ln x)/(x-1) < (\ln y)/(y-1)$. Thus $I(\overline{\mu}_n, \overline{\tau}) < I(\overline{\mu}_n, \tau)$ for sufficiently small $\delta > 0$ while $I(\overline{\mu}_n, \overline{\tau}) \geq \epsilon$ for all $\delta > 0$.

(c) In (29), $r_n = 1$ for all but finitely many n. To see this, we define

$$\alpha_n = \tau_1 r_n$$

so that $\tau_1 = \frac{\alpha_n}{r_n}$, $\tau_{m_n} = \frac{1-\alpha_n}{m_n - r_n}$, and

$$\beta_n = \frac{r_n}{m_n} > \alpha_n.$$

Then, using (11),

$$I(\tau, \overline{\mu}_n) = D(\alpha_n \| \beta_n) \quad \text{and} \quad I(\overline{\mu}_m, \tau) = D(\beta_n \| \alpha_n).$$

Since $\beta_n > \alpha_n$, the assumption $I(\overline{\mu}_n, \tau) \geq \epsilon$ implies

$$\beta_n \log \frac{\beta_n}{\alpha_n} \geq D(\beta_n \| \alpha_n) \geq \epsilon,$$

so that $\alpha_n < \beta_n\, e^{-\epsilon/\beta_n}$. Combining this with the monotonicity of $D(\alpha \| \beta_n)$ in the domain $\alpha \in (0, \beta_n]$, we obtain

$$\begin{aligned} I(\tau, \overline{\mu}_n) &\geq D\left(\beta_n\, e^{-\epsilon/\beta_n} \| \beta_n\right) \\ &= -\epsilon\, e^{-\epsilon/\beta_n} + \left(1 - \beta_n\, e^{-\epsilon/\beta_n}\right) \log \frac{1 - \beta_n\, e^{-\epsilon/\beta_n}}{1 - \beta_n}. \end{aligned} \tag{31}$$

If $r_n = \beta_n m_n \geq 2$ for infinitely many n, then the last inequality contradicts the upper bound

$$I(\tau, \overline{\mu}_n) \leq D(0 \| 1/m_n) = \log \frac{1}{1 - 1/m_n}$$

proved above. This completes the proof of (c).

If (c) holds, then τ under consideration coincides with $\tilde{\tau}$ given by (25) and (3). The desired result $I(\tau, \bar{\mu}_n) \geq 1/m_n + o(1/m_n)$ thus follows from (31) applied to $\alpha_n = \tau_1$ and $\beta_n = 1/m_n$.

REMARK 4. Remark that formula (3) cannot be applied directly for the test based on $\tilde{I}_n$ because (5) does not hold in this case, in contrast to the test based on the I-divergence statistic (see Remark 5 below). The same holds for the tests based on the reversed Pearson statistic discussed in the next section.

4 The χ^2-divergence

Of course the best known goodness-of-fit test is based on the χ^2 statistic. We refer to Neyman (1949), Watson (1958) for some important historic references. See also Kallenberg *et al.* (1985) for a more recent discussion. The *Pearson statistic* is given by

$$\chi_n^2 = \chi^2(\mu_n^*, \bar{\mu}_n) = \sum_{j=1}^{m_n} \frac{(\mu(A_{n,j}) - \mu_n(A_{n,j}))^2}{\mu(A_{n,j})},$$

while the reversed Pearson statistic, also known as *Neyman* or *Neyman-modified Pearson statistic* is defined as

$$\tilde{\chi}_n^2 = \chi^2(\bar{\mu}_n, \mu_n^*) = \sum_{j=1}^{m_n} \frac{(\mu(A_{n,j}) - \mu_n(A_{n,j}))^2}{\mu_n(A_{n,j})}.$$

We show that

$$\mathbf{P}\{\chi_n^2 > \epsilon\} = e^{-\frac{n \log m_n}{\sqrt{m_n}}(\sqrt{\epsilon}/2 + o(1))}$$

and

$$\mathbf{P}\{\tilde{\chi}_n^2 > \epsilon\} = e^{-\frac{n}{m_n}(1+o(1))}.$$

Theorem 4 *(Theorem 1 in Quine and Robinson (1985)) Suppose that (21) holds. If, in addition,*

$$\frac{m_n^{3/2} \log n}{n \log m_n} \to 0, \tag{32}$$

then for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{\sqrt{m_n}}{n \log m_n} \log \mathbf{P}\{\chi_n^2 > \epsilon\} = -\sqrt{\epsilon}/2.$$

PROOF. Again apply (12) for $\Sigma = \{A_{n,1}, \ldots, A_{n,m_n}\}$ such that

$$\Gamma = \{\tau : \ \chi^2(\tau, \bar{\mu}_n) \geq \epsilon\}.$$

Then, according to (12),

$$\left| \frac{\sqrt{m_n}}{n \log m_n} \log \mathbf{P}\{\chi_n^2 \geq \epsilon\} + \frac{\sqrt{m_n}}{\log m_n} \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{m_n^{3/2} \log(n+1)}{n \log m_n}$$

Therefore, because of (32) and (21),

$$\begin{aligned} \lim_{n\to\infty} \frac{\sqrt{m_n}}{n \log m_n} \log \mathbf{P}\{\chi_n^2 > \epsilon\} &= -\lim_{n\to\infty} \frac{\sqrt{m_n}}{\log m_n} \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \\ &= -\lim_{n\to\infty} \frac{\sqrt{m_n}}{\log m_n} \inf_{\chi^2(\tau, \bar{\mu}_n) \geq \epsilon} I(\tau, \bar{\mu}_n). \end{aligned}$$

Using the notation

$$\tau = (\tau_1, \ldots, \tau_{m_n})$$

we have (23) and

$$\chi^2(\tau, \bar{\mu}_n) = m_n \sum_{j=1}^{m_n} \tau_j^2 - 1.$$

Concerning the upper bound consider the distribution $\tilde{\tau}$ defined by (25) where τ_1 solves the equation

$$\chi^2(\tilde{\tau}, \bar{\mu}_n) = \epsilon,$$

or,

$$\tau_1 = \frac{1 + \sqrt{\epsilon(m_n - 1)}}{m_n}.$$

Then

$$\frac{\sqrt{m_n}}{\log m_n} \inf_{\chi^2(\tau, \bar{\mu}_n) \geq \epsilon} I(\tau, \bar{\mu}_n) \leq \frac{\sqrt{m_n}}{\log m_n} I(\tilde{\tau}, \bar{\mu}_n) \to \sqrt{\epsilon}/2. \tag{33}$$

Concerning the lower bound one shows along similar lines as in the proof of Theorem 3 that for all but finitely many n, $\tilde{\tau}$ is the minimizing distribution.

REMARK 5. Since $\mathbf{E}(I_n) \leq \mathbf{E}(\chi_n^2) = \frac{m_n-1}{n}$, (5) holds for the statistics I_n and χ_n^2 when $m_n/n \to 0$.

REMARK 6. Theorem 4 means that when $m_n \to \infty$, the Bahadur exact slope of the χ^2-test is identically zero. Another interpretation is that the tail of the χ^2-test statistic is of sub-exponential nature, that is, is heavier than an exponential tail. This is due to cells $A_{n,j}$ with small probabilities, which put too much weight on the squared difference $(\mu(A_{n,j}) - \mu_n(A_{n,j}))^2$. Although there is a widespread believe in literature that the I-divergence test I_n and the χ^2-test have a similar behaviour, their Bahadur slopes are quite different.

Theorem 5 *Under the conditions of Theorem 3*

$$\lim_{n\to\infty} \frac{m_n}{n} \log \mathbf{P}\{\tilde{\chi}_n^2 > \epsilon\} = -1.$$

PROOF. Again apply (12) for $\Sigma = \{A_{n,1}, \ldots, A_{n,m_n}\}$ such that

$$\Gamma = \{\tau :\ \chi^2(\bar{\mu}_n, \tau) \geq \epsilon\}.$$

Then, according to (12),

$$\left| \frac{m_n}{n} \log \mathbf{P}\{\tilde{\chi}_n^2 \geq \epsilon\} + m_n \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{m_n^2 \log(n+1)}{n}.$$

Therefore, because of (21) and (22),

$$\begin{aligned} \lim_{n\to\infty} \frac{m_n}{n} \log \mathbf{P}\{\tilde{\chi}_n^2 > \epsilon\} &= -\lim_{n\to\infty} m_n \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) \\ &= -\lim_{n\to\infty} m_n \inf_{\chi^2(\bar{\mu}_n,\tau)\geq\epsilon} I(\tau, \bar{\mu}_n). \end{aligned}$$

Concerning the upper bound consider the distribution (24). Then

$$m_n \inf_{\chi^2(\bar{\mu}_n,\tau)\geq\epsilon} I(\tau, \bar{\mu}_n) \leq m_n I(\tau_0, \bar{\mu}_n) \to 1.$$

Concerning the lower bound apply Theorem 3 and (2) to obtain

$$m_n \inf_{\chi^2(\bar{\mu}_n,\tau)\geq\epsilon} I(\tau, \bar{\mu}_n) \geq m_n \inf_{I(\bar{\mu}_n,\tau)\geq\epsilon} I(\tau, \bar{\mu}_n) \to 1.$$

References

[1] Ali, M. S. and Silvey, S.D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser B*, 28, 131-140.

[2] Bahadur, R.R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

[3] Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.*, 17, 107-124.

[4] Beirlant, J., Györfi, L. and Lugosi, G. (1995). On the asymptotic normality of the L_1- and the L_2-errors in histogram density estimation. *Canadian Journal of Statist.*, 22, 309-318.

[5] Beirlant, J. and Györfi, L. (1998) On the L_1-error in histogram density estimation: the multidimensional case. *Nonparametric Statistics*, 9, 197-216.

[6] Berlinet, A., Devroye, L. and Györfi, L. (1995). Asymptotic normality of L_1 error in density estimation. *Statistics.* 26, 329-343.

[7] Csiszár, I. (1967). Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2, 299-318.

[8] Csiszár, I. and Körner, J. (1981). *Information Theory: Coding Theorems for Memoryless Systems.* Academic Press, New York.

[9] Dembo, A. and Zeitouni, O. (1992). *Large Deviations Techniques and Applications.* Jones and Bartlett Publishers.

[10] Devroye, L. (1983). On arbitrary slow rates of global convergence in density estimation. *Zeitschrift für Wahscheinlichkeitstheorie und verwandte Gebiete*, 62, 475-483.

[11] Devroye, L. (1991). Exponential inequalities in nonparametric estimation. In: *Nonparametric Functional Estimation*, ed. G. Roussas, 31–44, NATO ASI Series, Dordrecht: Kluwer Academic Publishers.

[12] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: the L_1 View.* Wiley, New York.

[13] Devroye, L. and Györfi, L. (1990). No empirical measure can converge in total variation sense for all distributions. *Ann. of Statist.*, 18, 1496-1499.

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

[15] Groeneboom, P. and Shorack, G.R. (1981). Large deviations of goodness of fit statistics and linear combinations of order statistics. *Ann. Probability*, 9, 971-987.

[16] Györfi, L. and van der Meulen, E. C. (1991). A consistent goodness–of–fit test based on the total variation distance. In: *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.), Kluwer, Boston, 631-646.

[17] Kallenberg, W. C. M. (1985). On moderate and large deviations in multinomial distributions. *Ann. Statist.*, 13, 1554-1580.

[18] Kallenberg, W. C. M., Oosterhof, J. and Shriever, B. F. (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the Americal Statistical Association*, 80, 959-968.

[19] Kemperman, J. H. B. (1969). An optimum rate of transmitting information. *Ann. Math. Statist.*, 40, 2156-2177.

[20] Kullback, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Information Theory*, 13, 126-127.

[21] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22, 79-86.

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

[23] Louani, D. (1998). Large deviation limit theorems for the kernel density estimator. *Scand. J. Statistics*, 25, 243-253.

[24] Louani, D. (1999). Large deviations for the L_1-distance in kernel density estimation. *Submitted for publication.*

[25] Neyman, J. (1949). Contribution to the theory of the χ^2 test. Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, 239-273. Berkeley, CA, Berkeley University Press.

[26] Nikitin, Ya. (1995). *Asymptotic efficiency of nonparametric tests.* Cambridge University Press, Cambridge.

[27] Quine, M.P. and Robinson, J. (1985). Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *Ann. Statist.*, 13, 727-742.

[28] Sanov, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sb.*, 42, pp. 11-44 (English translation in *Sel. Transl. Math. Statist. Prob.*, 1, (1961), 213-244).

[29] Toussaint, G. T. (1975). Sharper lower bounds for information in term of variation. *IEEE Trans. Information Theory*, IT-21, 99-103.

[30] Tusnády, G. (1977). On asymptotically optimal tests. *Annals of Statistics*, 5, 385-393.

[31] Vajda, I. (1989). *Theory of Statistical Inference and Information.* Kluwer Academic Publishers.

[32] Vajda, I. (1970). Note on discrimination information and variation. *IEEE Trans. Information Theory*, IT-16, 771-773.

[33] Watson, G. S. (1958). On chi–squared goodness–of–fit tests for continuous distributions. *Journal of the Royal Statistical Society, Series B*, 20, 44-61.