Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Alain Berlinet and Igor Vajda:

EXTENSIONS OF A DEVROYE–LUGOSI THEOREM

No. 2240 November 2008

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42) (2) 688 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Extensions of a Devroye-Lugosi theorem

A. Berlinet[1] *and I. Vajda*[2]

Abstract. This paper deals with the problem of selection for each observed data the better of two given estimators of the true probability measure. Such a problem was posed for the first time by Devroye and Lugosi who proposed a feasible suboptimal selection (called Scheffé selection) as an alternative to the optimal but practically nonfeasible selection. The optimality was considered with respect to the estimation error given by the total variation distance. They proved that the Scheffé selection guarantees in typical situations better rate of convergence of the total variation error to zero than any of the two initially given estimates. This result was based on a theorem establishing an inequality between the total variation errors of the Scheffé selection and optimal selection. In this paper we extend this theorem to more general ϕ–divergence distances in two ways. Our first extension estimates the more general ϕ-divergence errors of the Scheffé selection of Devroye and Lugosi. The second one extends the Scheffé selection rule to the more general ϕ-divergence error criteria and estimates the corresponding ϕ-divergence errors. For the space and capacity reasons, we do not deal in this paper with the rates of convergence of the corresponding ϕ–divergence errors.

AMS 1991 subject classification: 62 G 05.

Key Words: Estimation of probability distributions, Selection of the better of two estimates, Divergence error criteria, Optimal and suboptimal selections, Asymptotic optimality of the suboptimal selection.

1 Introduction and basic concepts

Consider observations $X_1, \dots, X_n$ i.i.d. by a probability measure μ on the Borel σ-algebra $\mathcal{B}^d$ of subsets of the Euclidian space $\mathbb{R}^d$ and let μ_n defined by

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{I}\left(X_i \in A\right), \quad A \in \mathcal{B}^d \tag{1}$$

be the standard empirical distribution on $\mathcal{B}^d$ based on these observations and $\mu_n^{(1)}$, $\mu_n^{(2)}$ any two probability measures on $\mathcal{B}^d$ based on these observations and used to estimate the unknown probability measure μ (e.g. a histogram estimate and a kernel estimate). This paper studies the ϕ-divergence error criteria $D_\phi(\mu_n^{(k)}, \mu)$ and the rules

$$\mu_n^{\mathcal{C}} = \begin{cases} \mu_n^{(1)} & \text{if } \mathcal{C}(X_1, \dots, X_n) \text{ is satisfied} \\ \mu_n^{(2)} & \text{otherwise.} \end{cases} \tag{2}$$

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

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

for selection of better of the estimates $\mu_n^{(1)}$, $\mu_n^{(2)}$ based on given criteria $\mathcal{C}(X_1, \dots, X_n)$ depending on observations $X_1, \dots, X_n$. Obviously, the optimal selection is

$$\mu_n^{(0)} = \begin{cases} \mu_n^{(1)} & \text{if } D_\phi\left(\mu_n^{(1)}, \mu\right) < D_\phi\left(\mu_n^{(2)}, \mu\right) \\ \mu_n^{(2)} & \text{otherwise.} \end{cases} \tag{3}$$

Let us now turn attention to a given pair of probability measures μ_0, μ on $\mathcal{B}^d$. It is known (see e.g. Liese and Vajda (1987, 2006) that the restrictions $\mu_{0,\mathcal{S}}, \mu_{\mathcal{S}}$ of these measures on a sub-σ-algebra $\mathcal{S} \subset \mathcal{B}^d$ decreases the ϕ-divergence. Our present paper deals with the special situations where the restrictions $\mu_{n,\mathcal{S}_n}, \mu_{\mathcal{S}_n}$ of the measures μ_n, μ on a given sequence of sub-σ-algebras $\mathcal{S}_n \subset \mathcal{B}^d$ satisfy the asymptotic relation

$$D_\phi\left(\mu_{n,\mathcal{S}_n}, \mu_{\mathcal{S}_n}\right) = o\left(D_\phi\left(\mu_n^{(0)}, \mu_n\right)\right) \quad \text{for } \; n \to \infty \tag{4}$$

i.e. where $D_\phi\left(\mu_{n,\mathcal{S}_n}, \mu_{\mathcal{S}_n}\right)$ tends to zero faster than $D_\phi\left(\mu_n^{(0)}, \mu_n\right)$. Alongside with the practically unfeasible optimal selection (3) we study the feasible suboptimal selection

$$\mu_n^{*} = \begin{cases} \mu_n^{(1)} & \text{if } D_\phi\left(\bar{\mu}_n^{(1)}, \bar{\mu}_n\right) < D_\phi\left(\bar{\mu}_n^{(2)}, \bar{\mu}_n\right) \\ \mu_n^{(2)} & \text{otherwise.} \end{cases} \tag{5}$$

Our main result is the inequality

$$D_\phi(\mu_n^*, \mu) \leq 3D_\phi(\mu_n^{(0)}, \mu) + 2D_\phi\left(\mu_{n,\mathcal{S}_n}, \mu_{\mathcal{S}_n}\right). \tag{6}$$

proved here for all initial estimates $\mu_n^{(1)}$, $\mu_n^{(2)}$ dominated by the Lebesgue measure on $\mathbb{R}^d$ and all metric divergences D_ϕ. This inequality demonstrates that in the situations under consideration the practically feasible suboptimal estimates μ_n^* achieve the same rate of convergence of the error to zero as the practically unfeasible optimal estimates $\mu_n^{(0)}$.

Devroye and Lugosi (2001) proved that there exist situations where (4) holds for any initial estimates $\mu_n^{(1)}$, $\mu_n^{(2)}$ with Lebesque densities on $\mathbb{R}^d$ and they proved the inequality (6) for the total variation error criterion

$$V(\mu_n^{(k)}, \mu) = \int_{\mathbb{R}^d} |f_n^{(k)}(\boldsymbol{x}) - f(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}$$

where $f_n^{(k)}, f$ are Lebesque densities of $\mu_n^{(k)}, \mu$. The total variation $V(\mu_n^{(k)}, \mu)$ is nothing but the special ϕ-divergence criterion $D_\phi(\mu_n^{(k)}, \mu)$ for the function $\phi(t) = |t-1|$. Hence in this paper we extend the inequality of Devroye and Lugosi to arbitrary metric ϕ-divergences. For this purpose also the so far known set of metric divergences was extended in Vajda (2008).

Devroye and Lugosi (2001) considered the estimates $\mu_n^{(1)}$, $\mu_n^{(2)}$ represented by densities

$$f_n^{(i)} = f_n^{(i)}(\cdot; X_1, \dots, X_n), \quad i \in \{1, 2\} \tag{7}$$

leading to the optimal practically unfeasible selection density

$$f_n^{(0)} = \begin{cases} f_n^{(1)} & \text{if } \displaystyle\int |f_n^{(1)} - f| \ < \displaystyle\int |f_n^{(2)} - f|, \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{8}$$

They proposed a practically feasible approximation to this selection called *Scheffé estimate* obtained by the rule

$$f_n^{*} = \begin{cases} f_n^{(1)} & \text{if } \left| \displaystyle\int_{A_n} f_n^{(1)} - \mu_n(A_n) \right| < \left| \displaystyle\int_{A_n} f_n^{(2)} - \mu_n(A_n) \right|, \\ f_n^{(2)} & \text{otherwise} \end{cases} \tag{9}$$

where

$$A_n = A\left(f_n^{(1)}; f_n^{(2)}\right) = \left\{x : f_n^{(1)}(x) > f_n^{(2)}(x)\right\} \tag{10}$$

is the so-called *Scheffé set* for the ordered pair ($f_n^{(1)}$, $f_n^{(2)}$) and μ_n is the empirical probability measure (1). Chapter 6 of Devroye and Lugosi (2001) contains a number of arguments in favour of the Scheffé selection rule (9). However, the next example demonstrates that the use of the Scheffé rule is problematic in some cases. As above, $\boldsymbol{I}(\cdot)$ denotes the indicator function.

Example 1. Consider the uniform density f on the closed interval $[c, c+1] \subset \mathbb{R}$ with unknown parameter c and independent ordered sample $X_{n:1}, \dots, X_{n:n}$ generated by f. For the estimates

$$f_n^{(1)} = \boldsymbol{I}(X_{n:1} \le x \le X_{n:1} + 1) \text{ and } f_n^{(2)} = \boldsymbol{I}(X_{n:n} - 1 \le x \le X_{n:n})$$

of f it holds

$$A_n = (X_{n:n},\ X_{n:1} + 1], \quad \mu_n(A_n) = 0$$
$$\int_{A_n} f_n^{(1)} = X_{n:1} + 1 - X_{n:n} \text{ and } \int_{A_n} f_n^{(2)} \ = \ 0$$

so that

$$\left| \int_{A_n} f_n^{(1)} - \mu_n(A_n) \right| = |X_{n:1} + 1 - X_{n:n}| > X_{n:1}$$

exceeds with probability 1 the absolute deviation

$$\left| \int_{A_n} f_n^{(2)} - \mu_n(A_n) \right| = 0.$$

Consequently the Scheffé rule selects the estimate $f_n^{(2)}$ achieving the L_1-error $\int |f_n^{(2)} - f| =$

$2\,(1-X_{n:n})$ whereas the estimate $f_n^{(1)}$ achieves the error $\int|f_n^{(1)}-f| = 2X_{n:1}$, so that is strictly better in the L_1-sense with the probability $\Pr(X_{n:1}+X_{n:n}<1)=1/2$ for all $n=1,2,\ldots.$ Hence in this situation the Scheffé rule selects the better of the estimates $f_n^{(1)}, f_n^{(2)}$ with probability 1/2, i.e. it does not bring the selection closer to the optimality than the tossing of a coin.

The book of Devroye and Lugosi (2001) presents a systematic theory dealing with properties and applications of the Scheffé selection f_n^*. This theory is based on Theorem 6.1 which compares the errors

$$\int|f_n^{(0)}-f| \;= \min\left\{\int|f_n^{(1)}-f|, \int|f_n^{(2)}-f|\right\} \quad \text{and} \quad \int|f_n^*-f|.$$

This fundamental theorem can be given the form of the inequality

$$\int|f_n^*-f| \le 3\int|f_n^{(0)}-f| + 4\left|\int_{A_n} f - \mu_n(A_n)\right| \tag{11}$$

where A_n is the Scheffé set for $(f_n^{(1)},\, f_n^{(2)})$. This inequality states that the selection f_n^* can achieve the error level $3\int|f_n^{(0)}-f|$ up to the universal error term appearing on the right. This inequality was applied not only in Chapters 7 − 17 of the Devroye-Lugosi book, but also in subsequent papers, among them in Berlinet, Biau and Rouvière (2005 a, b).

The latter papers observed that the L_1-error criterion $\int|f-g|$ for the estimates g being formally probability densities is a special case of the more general ϕ-divergence criterion $D_\phi(f,g)$ defined for arbitrary probability densities f, g by the formula

$$D_\phi(f,g) = \int g\,\phi\left(\frac{f}{g}\right). \tag{12}$$

Here $\phi(t)$ is nonnegative and convex in the domain $t\in(0,\infty)$, strictly convex and vanishing at the point $t=1$ (for details about formula (12) and the basic properties of ϕ-divergences used below, see Csiszár (1967a) or Liese and Vajda (1987, 2006).

The L_1-error is the ϕ-divergence for $\phi(t)=|t-1|$, called *total variation* and denoted by $V(f,g)$, i. e.

$$V(f,g) = \int|f-g| = 2\sup_{A\in\mathcal{B}^d}\left|\int_A f - \int_A g\right|. \tag{13}$$

Other examples are the *squared Hellinger distance*

$$H^2(f,g) = 2\int\left(\sqrt{f}-\sqrt{g}\right)^2 \quad \text{for } \phi(t) = 2\left(\sqrt{t}-1\right)^2, \tag{14}$$

the *squared Le Cam distance*

$$LC^2(f,g) = \frac{1}{2}\int\frac{(f-g)^2}{f+g} \quad \text{for } \phi(t) = \frac{(t-1)^2}{2(t+1)}, \tag{15}$$

and the *information divergence*

$$I(f,g) = \int f \ln \frac{f}{g} \quad \text{for } \phi(t) = t \ln t. \tag{16}$$

A natural motivation for the alternative ϕ-divergence error criteria is the need to work with estimates convergent in topologies stronger than that induced by the total variation (cf. Csiszár 1967b and Österreicher and Vajda (2003)). This paper introduces a new motivation achieved in Example 3 below by extending the framework of Example 1 through admitting non-uniform densities with unit supports on $\mathbb{R}$. In this extended setting Example 3 demonstrates that for some densities f the alternative ϕ-divergence error criteria exhibit with positive probabilities optimality of the estimate $g = f^{(1)}$ at the same time when the L_1-error exhibits the optimality of $g = f^{(2)}$.

Since the optimality of the Scheffé estimates f_n^* is perceived differently by different ϕ-divergence error criteria, it is important to see whether or how the fundamental Devroye–Lugosi inequality (11) can be extended from the total variation criteria

$$V(f, f_n^*) = \int |f_n^* - f| \quad \text{and} \quad V(f, f_n^{(0)}) = \int |f_n^{(0)} - f| \tag{17}$$

to the more general ϕ-divergence criteria

$$D_\phi(f, f_n^*) = \int f_n^* \, \phi\left(\frac{f}{f_n^*}\right) \quad \text{and} \quad D_\phi(f, f_n^{(0)}) = \int f_n^{(0)} \, \phi\left(\frac{f}{f_n^{(0)}}\right). \tag{18}$$

This problem is solved in Section 3.

Section 4 introduces a replacement of the Scheffé L_1-based selection rule by a more general ϕ-divergence selection rule and solves a problem parallel to that of Section 3, namely whether or how the Devroye–Lugosi inequality (11) can be extended to the alternatively selected estimates and to the more general ϕ-divergence criteria.

It remains to be seen whether the statistical applications of the new selection rules introduced in Sections 3 and 4 are as rich as those given by Devroye and Lugosi in their book : selection from an infinite class, minimum distance estimates using Vapnik–Chervonenkis classes, in particular Yatracos classes with finite Vapnik–Chervonenkis dimension, choice of kernels, partitions or bandwidths, wavelet systems or other orthonormal basis.

For obvious reasons, in this paper the attention is restricted to the estimates (7) which are a.s. probability densities themselves.

2 Metric divergence criteria of errors

Let us start with the following basic properties of ϕ-divergences needed in the sequel:

(i) The *range of values* is

$$0 \leq D_\phi(f,g) \leq \phi(0) + \phi^*(0) \tag{19}$$

where $\phi(0)$, $\phi^*(0)$ are smooth extensions of $\phi(t)$, $\phi^*(t) = t\,\phi(1/t)$ to the point $t = 0$. In (19) $D_\phi(f,g) = 0$ if and only if $f = g$ a. s. and $D_\phi(f,g) = \phi(0) + \phi^*(0)$ if (for finite $\phi(0) + \phi^*(0)$ if and only if) $f \perp g$ (disjoint supports).

(ii) The *symmetry* $D_\phi(f,g) = D_\phi(g,f)$ for all f,g holds if and only if $\phi = \phi^*$ for the adjoint function ϕ^* defined in **(i).**

(iii) The *monotonicity property* deals with relations between the ϕ-divergences

$$D_\phi(\mu,\nu) \equiv D_\phi(f,g)$$

of distributions

$$\mu(A) = \int_A f, \quad \nu(A) = \int_A g, \quad A \in \mathcal{B}^d$$

and the ϕ-divergences of restrictions of these distributions on sub-σ-algebras $\mathcal{S} \subset \mathcal{B}^d$ of the Borel σ-algebra $\mathcal{B}^d$ defined by formula

$$D_\phi(\mu,\nu|\mathcal{S}) = D_\phi(f_\mathcal{S}, g_\mathcal{S}) = \int g_\mathcal{S}\,\phi\left(\frac{f_\mathcal{S}}{g_\mathcal{S}}\right)$$

for $\mathcal{S}$-measurable versions $f_\mathcal{S}$, $g_\mathcal{S}$ of densities f,g. It states that the ordering

$$D_\phi(f,g|\mathcal{S}) \equiv D_\phi(\mu,\nu|\mathcal{S}) \leq D_\phi(\mu,\nu) \equiv D_\phi(f,g) \tag{20}$$

holds. If the equality in (20) takes place then we say that $\mathcal{S}$ *preserves the ϕ-divergence* $D_\phi(f,g)$. It is known (see e.g. Corollary 1.29 in Liese and Vajda (1987)) that if a sub-σ-algebra $\mathcal{S}$ is sufficient for the pair $\{f,g\}$ then the equality takes place in (20), i.e. the sufficient $\mathcal{S}$ always preserves the ϕ-divergence $D_\phi(f,g)$.

(iv) Finally, the *spectral representation* says that if a sub-σ-algebra $\mathcal{S} \subset \mathcal{B}^d$ is generated by a finite or countable $\mathcal{B}^d$-measurable partition $\mathcal{P}$ of $\mathbb{R}^d$ (spectrum of $\mathcal{S}$, in symbols we write $\mathcal{S} = \mathcal{S}(\mathcal{P})$) then

$$D_\phi(f,g|\mathcal{S}) = \sum_{A\in\mathcal{P}} \int_A g\,.\phi\left(\frac{\int_A f}{\int_A g}\right). \tag{21}$$

Example 2. Consider for every $A \in \mathcal{B}^d$ the partition $\mathcal{P} = (A, A^c)$ of $\mathbb{R}^d$ and the $\mathcal{P}$-generated (or, more simply, A-generated) algebra

$$\mathcal{S}_A := \mathcal{S}\,(A, A^c) \subset \mathcal{B}^d \tag{22}$$

consisting of the sets $\mathbb{R}^d, A, A^c, \emptyset$. Then the general spectral representation (21) implies

$$V(f,g|\mathcal{S}_A) = \sum_{B\in\{A,A^c\}} \left|\int_B f - \int_B g\right| = 2\left|\int_A f - \int_A g\right|. \tag{23}$$

From (13) and (23) we see that the fundamental Devroye–Lugosi inequality (11) can be given the form

$$V(f,f_n^*) \le 3V(f,f_n^{(0)}) + 2V\left(\mu,\mu_n|\mathcal{S}_{A_n}\right) \tag{24}$$

for the Scheffé set A_n of the estimates $f_n^{(1)}$ and $f_n^{(2)}$.

If A in (23) is the Scheffé set $A(f;g)$ of f and g then the absolute difference on the right of (23) can be replaced by the ordinary difference. Moreover, it is seen from (13) that then $\mathcal{S}_A$ preserves $V(f,g)$ so that the formula (23) can be extended and specified as follows

$$V\left(f,g|\mathcal{S}_A\right) = V(f,g) = 2\left(\int_A f - \int_A g\right). \tag{25}$$

The following sections extend the Devroye–Lugosi theorem (11), or equivalently (24), to the error criteria $D(f,g)$ for probability densities f, g on $(\mathbb{R}^d,\mathcal{B}^d)$ satisfying similar metric properties as the total variation criterion $V(f,g)$ namely

the *reflexivity*

$$D(f,g) = 0 \quad \text{if and only if } f = g \text{ a.s.}, \tag{26}$$

the *symmetry*

$$D(f,g) = D(g,f) \quad \text{for all } f,\ g \tag{27}$$

and the *triangle inequality*

$$D(f,g) \le D(f,h) + D(h,g) \quad \text{for all } f,\ g,\ h. \tag{28}$$

We restrict ourselves to the *metric divergence criteria* defined as powers

$$D(f,g) = D_\phi(f,g)^\pi, \quad \pi > 0$$

of ϕ-divergences $D_\phi(f,g)$ satisfying (26)-(28). These ϕ-divergences achieve finite upper bounds

$$\phi(0) + \phi^*(0) = 2\phi(0) < \infty \tag{29}$$

(see **(ii)** above for the equality and Csiszár (1967b) for the finiteness).

To provide a sufficiently rich class of such criteria, let us introduce the class of ϕ_α-divergences

$$\mathcal{D}_\alpha(f,g) = D_{\phi_\alpha}(f,g), \quad \alpha \in \mathbb{R}. \tag{30}$$

Here the convex functions $\phi_\alpha(t)$ are given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{\mid \alpha \mid}{\alpha(\alpha-1)} \left(2^{\alpha-1}(t+1) - (t^{1/\alpha}+1)^\alpha\right) \tag{31}$$

if $\alpha(\alpha-1) \neq 0$, and by the corresponding limits

$$\phi_0(t) = \quad \mid t-1 \mid /2, \tag{32}$$

$$\phi_1(t) = t \ln t + (t+1) \ln \frac{2}{t+1} \tag{33}$$

otherwise. The subclass of these divergences for $\alpha \geq 0$ was proposed (with a different parametrization) by Österreicher and Vajda (2003). The extension to $\alpha < 0$ was proposed recently by Vajda (2008). It is easy to verify for all f, g the formulas

$$\mathcal{D}_0(f,g) = \frac{1}{2} V(f,g) \qquad \text{(total variation, (13))}, \tag{34}$$

$$\mathcal{D}_2(f,g) = \frac{1}{2} H^2(f,g) \qquad \text{(Hellinger, (14))}, \tag{35}$$

$$\mathcal{D}_{-1}(f,g) = \frac{1}{4} LC^2(f,g) \quad \text{(Le Cam, (15))} \tag{36}$$

and

$$\mathcal{D}_1(f,g) \;=\; I\left(f,(f+g)/2\right) + I(g,(f+g)/2). \tag{37}$$

In the Appendix we demonstrate that the powers

$$D(f,g) := \mathcal{D}_\alpha(f,g)^{1/\max\{2,\alpha\}}, \quad \alpha \in \mathbb{R} \tag{38}$$

of the divergences (30) satisfy (26) – (28), i. e. that they are metric divergence criteria.

3 Scheffé selection rule

This section extends the fundamental Devroye–Lugosi inequality (11) from the total variation error criteria (17) to the more general ϕ-divergence criteria (18). A strong motivation for this extension is the fact that the optimality of the Scheffé selection $f_n^* \in \left\{f_n^{(1)}, f_n^{(2)}\right\}$ is evaluated differently by different ϕ-divergence error criteria. The next example demonstrates that the anisotony between two such criteria may be total in the sense that f_n^* is worse of $f_n^{(1)}, f_n^{(2)}$ in ϕ_1-divergence error (as e.g. in Example 1) and at the same time better of them in ϕ_2-divergence error.

Example 3. Consider the setting of Example 1 with arbitrary densities f supported by $[0,1]$, and self-adjoint nonnegative convex functions

$$\phi(t) = t\,\phi\left(\frac{1}{t}\right) \equiv \phi^*(t)$$

leading to symmetric ϕ-divergences the powers of which satisfy the triangle inequality. Without loss of generality we can assume $\phi(0) = 1$. Then

$$\begin{aligned}
D_\phi(f_n^{(1)}, f) = D_\phi(f, f_n^{(1)}) &= \int f_n^{(1)}\, \phi\left(\frac{f}{f_n^{(1)}}\right) \\
&= \int_0^{X_{n:1}} 0\, \phi\left(\frac{f}{0}\right) + \int_{X_{n:1}}^1 \phi\,(f) + \int_1^{X_{n:1}+1} 1\, \phi\left(\frac{0}{1}\right) \\
&= \int_0^{X_{n:1}} f\, \phi^*\left(\frac{0}{f}\right) + \int_{X_{n:1}}^1 \phi\,(f) + \int_1^{X_{n:1}+1} \phi\,(0) \\
&= \phi\,(0) \int_0^{X_{n:1}} f + \int_{X_{n:1}}^1 \phi\,(f) + \phi\,(0)\;\, X_{n:1} \\
&= X_{n:1} + F(X_{n:1}) + \int_{X_{n:1}}^1 \phi\,(f)\,.
\end{aligned}$$

Similarly,

$$\begin{aligned}
D_\phi(f_n^{(2)}, f) = D_\phi(f, f_n^{(2)}) &= \int f_n^{(2)}\, \phi\left(\frac{f}{f_n^{(2)}}\right) \\
&= 2 - (X_{n:n} + F(X_{n:n})) + \int_0^{X_{n:n}} \phi\,(f)\,.
\end{aligned}$$

Therefore

$$D_\phi(f_n^{(1)}, f) \lesseqgtr D_\phi(f_n^{(2)}, f)$$

if and only if

$$X_{n:1} + F(X_{n:1}) + \int_{X_{n:1}}^1 \phi\,(f) \lesseqgtr 2 - (X_{n:n} + F(X_{n:n})) + \int_0^{X_{n:n}} \phi\,(f)\,. \tag{39}$$

We shall present a density f satisfying for two concrete functions ϕ the conflicting inequalities (39).

Restrict ourselves for simplicity to the family of self-adjoint convex functions

$$\phi_\alpha(t) = |t^\alpha - 1|^{1/\alpha}\,, \qquad 0 < \alpha \leq 1,$$

leading to the so-called Matusita divergences

$$M_\alpha(f, g) = \int |f^\alpha - g^\alpha|^{1/\alpha}\,, \qquad 0 < \alpha \leq 1,$$

among them

$$M_1(f, g) = \int |f - g| \equiv V(f, g).$$

The powers $M_\alpha(f, g)^\alpha$ are metrics. Further, consider the density function $f : \mathbb{R} \longrightarrow \mathbb{R}$, defined by

$$f(x)=\begin{cases} 1 & \text{if} \ \ 0\le x<1/2 \\ 2(2x-1) & \text{if} \ \ 1/2\le x\le 1 \\ 0 & \text{if} \ \ x\notin[0,1] \end{cases}$$

for which (39) takes on the form

$$X_{n:1}+F(X_{n:1})+\int_{X_{n:n}}^{1}\phi_\alpha\left(f\right)\lesseqgtr 2-\left(X_{n:n}+F(X_{n:n})\right)+\int_0^{X_{n:1}}\phi_\alpha\left(f\right). \tag{40}$$

Finally, restrict ourselves to the class of the Δ-samples $X_1,\dots,X_n$ defined by the property

$$1/2-X_{n:1}=X_{n:n}-1/2\equiv\Delta\in(0,1/4)$$

and denote respectively $\mathcal{L}(\alpha,\Delta)$ and $\mathcal{R}(\alpha,\Delta)$ the above left and right hand side in (40). Then

$$\mathcal{L}(\alpha,\Delta)=1-2\Delta+\int_\Delta^{1/4}\left(1-(4u)^\alpha\right)^{1/\alpha}du+\int_{1/4}^{1/2}\left((4u)^\alpha-1\right)^{1/\alpha}du$$

so that

$$\lim_{\alpha\downarrow 0}\mathcal{L}(\alpha,\Delta)=1-2\Delta \qquad \text{and} \qquad \mathcal{L}(1,\Delta)=2\Delta^2-3\Delta+\frac{5}{4}.$$

On the other hand

$$\mathcal{R}(\alpha,\Delta)=-2\Delta^2-\Delta+1$$

so that

$$\mathcal{L}(1,\Delta)-\mathcal{R}(1,\Delta)=\left(2\Delta-\frac{1}{2}\right)^2>0$$

while for $\alpha>0$ small enough we get the opposite inequality

$$\mathcal{L}(\alpha,\Delta)-\mathcal{R}(\alpha,\Delta)<0.$$

The last two inequalities hold also for all the "approximately Δ-samples" $X_1,\dots,X_n$ defined by the condition that $1/2-X_{n:1}$ and $X_{n:n}-1/2$ are sufficiently close elements of the open interval $(0,1/4)$. Under the given f, such samples appear with positive probability. Thus the present example fulfills the promised requirements.

Let us now turn to the main result of this section which is the following theorem. This theorem and its proof refer to the lower and upper error bounds

$$L_\phi(V)\le D_\phi(f,g)\le U_\phi(V) \tag{41}$$

achieved for a given convex ϕ by the ϕ-divergences $D_\phi(f,g)$ on the class of densities f,g

satisfying the total variation condition

$$V(f,g) = V, \quad 0 \leq V \leq 2.$$

By Proposition 8.27 in Liese and Vajda (1987), the upper bound is for general ϕ given by the formula

$$U_\phi(V) = V \cdot c_\phi \quad \text{where} \quad c_\phi = \frac{\phi(0) + \phi^*(0)}{2} \quad \text{(cf. (19))} \tag{42}$$

and the lower bound $L_\phi(V)$ is convex and strictly increasing in the variable V from the minimum $L_\phi(0) = 0$ to the maximum $L_\phi(2) = \phi(0) + \phi^*(0) = 2c_\phi$. Hence the strictly increasing and concave inverse function

$$L_\phi^{-1}(D) : [0, 2c_\phi] \longrightarrow [0, 2] \tag{43}$$

always exists. For ϕ such that the powers $D_\phi(f,g)^\pi$ are metrics on the space of densities f, g (29) implies

$$c_\phi = \phi(0) < \infty. \tag{44}$$

Theorem 1. Let f be an estimated distribution on $\mathbb{R}^d$, $f_n^{(0)}$, $f_n^{(1)}$and $f_n^{(2)}$ the estimates considered in (7), (8) with the corresponding Scheffé set A_n and f_n^* the Scheffé estimate resulting from the selection rule (9). Then for every metric divergence criterion $D(f,g) = D_\phi(f,g)^\pi$

$$D\left(f_n^*, f\right) \leq D\left(f_n^{(0)}, f\right) + 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi} \right) + 2 \left| \int_{A_n} f - \mu_n(A_n) \right| \right]^\pi \tag{45}$$

where L_ϕ^{-1} and c_ϕ are given by (43) and (44)

Proof. Consider the random variables

$$\mathcal{E}_{ij} = \boldsymbol{I}\big(f_n^* = f_n^{(i)},\, f_n^{(0)} = f_n^{(j)}\big) \;\text{ where } \sum_{i,j=1}^{2} \mathcal{E}_{ij} = 1. \tag{46}$$

By the triangle inequality and symmetry of $D(f,g)$, and by the definition of $\mathcal{E}_{ii}$,

$$\begin{aligned} D\left(f_n^*, f\right) &\leq D\left(f_n^{(0)}, f\right) + \sum_{i,j=1}^{2} D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{ij} \\ &= D\left(f_n^{(0)}, f\right) + D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{21} + D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{12}. \end{aligned} \tag{47}$$

It suffices to prove that for $i \neq j$

$$D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{ij} \leq 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi} \right) + 2 \left| \int_{A_n} f - \mu_n(A_n) \right| \right]^\pi \mathcal{E}_{ij}. \tag{48}$$

We restrict ourselves to $\mathcal{E}_{21}$. For $\mathcal{E}_{12}$ the proof is similar. By the definition of $\mathcal{E}_{21}$ and (42), (43),

$$\begin{aligned} D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{21} = D\left(f_n^{(1)}, f_n^{(2)}\right)\mathcal{E}_{21} &\le \left[c_\phi\, V\left(f_n^{(1)}, f_n^{(2)}\right)\right]^\pi \mathcal{E}_{21} \\ &= c_\phi^\pi\, V\left(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_{A_n}\right)^\pi \mathcal{E}_{21} \\ &\le c_\phi^\pi \left[V\left(f_n^{(1)}, f|\mathcal{S}_{A_n}\right) + V\left(f_n^{(2)}, f|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le c_\phi^\pi \left[V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(2)}, \mu_n|\mathcal{S}_{A_n}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le 2^\pi c_\phi^\pi \left[V\left(f_n^{(0)}, f\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21}. \end{aligned}$$

where we bounded the sum of the total variations in the third line above by

$$\begin{aligned} &V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(1)}, \mu_n|\mathcal{S}_{A_n}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right) \\ &\le V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(1)}, \mu|\mathcal{S}_{A_n}\right) + 2V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right) \\ &\le 2V\left(f_n^{(1)}, f\right) + 2V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right). \end{aligned}$$

This completes the proof.

The next corollary reformulates the result of Theorem 1 in a simpler but slightly weaker form.

Corollary 1. For $0 < \pi \le 1$, under the assumptions and notations of Theorem 1,

$$D\left(f_n^*, f\right) \le 2^{1-\pi} c_\phi^\pi \left[3L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right) + 4\left|\int_{A_n} f - \mu_n(A_n)\right|\right]^\pi \tag{49}$$

Proof. Clear from (45) by taking into account the inequalities

$$D\left(f_n^{(0)}, f\right) \le U_\phi\left(V\left(f_n^{(0)}, f\right)\right)^\pi = \left[c_\phi V\left(f_n^{(0)}, f\right)\right]^\pi \le \left[c_\phi L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right)\right]^\pi$$

obtained from (41), (43) and also the inequality

$$\psi_\pi(a) + \psi_\pi(b) \le 2^{1-\pi}\psi_\pi(a+b)$$

obtained from Jensen's inequality for the concave function $\psi_\pi(x) = x^\pi$.

The next example demonstrates that Theorem 1 generalizes the Devroye and Lugosi inequality (11).

Example 4. Put $D(f,g) = \mathcal{D}_0(f,g) = V(f,g)/2$ (cf. (34)). Then $c_0 = \phi_0(0) = 1/2$,

$$L_0(V) = U_0(V) = \frac{V}{2}, \quad 0 \le V \le 2$$

and $L_0^{-1}(D) = 2D$. Hence Theorem 1 implies

$$\mathcal{D}_0\left(f_n^*, f\right) \leq \mathcal{D}_0\left(f_n^{(0)}, f\right) + 2\mathcal{D}_0\left(f_n^{(0)}, f\right) + 2\left|\int_{A_n} f - \mu_n(A_n)\right|$$

or, equivalently,

$$V\left(f_n^*; f\right) \leq 3V\left(f_n^{(0)}, f\right) + 4\left|\int_{A_n} f - \mu_n(A_n)\right|$$

which coincides with (11) and (24).

The next example illustrates contributions of Theorem 1 and its Corollary 1 beyond the framework of Devroye and Lugosi.

Example 5. Put $D(f,g) = \mathcal{D}_{-1}(f,g)^{1/2}$, i.e. take the Le Cam error criterion $LC(f,g)/2$ (cf. (36)). Then parts (ii) and (iii) of Theorem A1 in the Appendix imply that $c_{-1} = 1/8$, $U_{-1}(V) = V/16$ and

$$\begin{aligned} L_{-1}(V) &= \frac{1}{2}\left(\frac{1}{2} - \left[\frac{1}{1+V/2} + \frac{1}{1-V/2}\right]^{-1}\right) \\ &= \frac{1}{2}\left[\frac{1}{2} - \frac{1-(V/2)^2}{2}\right] = \left(\frac{V}{4}\right)^2. \end{aligned}$$

Therefore $L_{-1}^{-1}(D) = 4\sqrt{D}$ and for the Scheffé selection f_n^* of Devroye and Lugosi we get from Theorem 1 the relation

$$D\left(f_n^*, f\right) \leq D\left(f_n^{(0)}, f\right) + \left[\frac{2}{8}\left(4D\left(f_n^{(0)}, f\right) + 2\left|\int_{A_n} f - \mu_n(A_n)\right|\right)\right]^{1/2}$$

i. e.

$$LC\left(f_n^*, f\right) \leq LC\left(f_n^{(0)}, f\right) + \sqrt{\frac{1}{2}LC\left(f_n^{(0)}, f\right) + \frac{1}{8}\left|f_n - \mu_n(A_n)\right|}$$

where A_n is the Scheffé set of the initial estimates $f_n^{(1)}$ and $f_n^{(2)}$. Corollary 1 implies for the same f_n^* and A_n as before the alternative inequality

$$D\left(f_n^*, f\right) \leq \left(2\,\frac{3}{8}\,4D\left(f_n^{(0)}, f\right) + 2\,\frac{4}{8}\left|\int_{A_n} f - \mu_n(A_n)\right|\right)^{1/2}$$

i. e.

$$LC\left(f_n^*, f\right) \leq \sqrt{\frac{3}{2}\,LC\left(f_n^{(0)}, f\right) + \frac{1}{4}\left|\int_{A_n} f - \mu_n(A_n)\right|}.$$

We see that the rate of convergence of the Le Cam error $LC\left(f_n^*, f\right)$ to zero guaranteed by our theory for the Scheffé estimate is strictly below the rate of the Le Cam error $LC\left(f_n^{(0)}, f\right)$ achieved by the ideal estimate $f_n^{(0)}$. One can deduce from the known properties of the lower bound $L_\phi(V)$ and its inverse $L_\phi^{-1}(D)$ that similar result can be expected also for other divergence errors $D_\phi\left(f_n^*, f\right)$ with ϕ strictly convex everywhere.

4 Divergence selection rule

This section is a continuation of Section 3. Here the estimation errors are still evaluated by the criteria of the type $D(f,g) = D_\phi(f,g)^\pi, \pi > 0$ but the Scheffé selection (9) of Devroye and Lugosi (2001) is replaced by a more general selection rule. One arrives quite naturally at such a generalization if one applies the same metric divergence criteria also to the definitions of the optimal estimate $f_n^{(0)}$ and its practical approximation f_n^*. In other words, the generalization consists in the replacement of the L_1-based definition (8) by the divergence based definition

$$f_n^{(0)} = \begin{cases} f_n^{(1)} & \text{if } D\left(f_n^{(1)}, f\right) < D\left(f_n^{(2)}, f\right) \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{50}$$

and the L_1-based Scheffé selection rule (9) by the *divergence selection rule*

$$f_n^* = \begin{cases} f_n^{(1)} & \text{if } D\left(\mu_n^{(1)}, \mu_n | \mathcal{S}_n\right) < D\left(\mu_n^{(2)}, \mu_n | \mathcal{S}_n\right) \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{51}$$

The latter rule uses the empirical distribution μ_n defined by (**??**), the estimates

$$\mu_n^{(i)}(B) = \int_E f_n^{(i)}, \quad B \in \mathcal{B}^d, \ i \in \{1,2\}$$

of the probability distribution $\mu \sim f$, and the sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ preserving the divergence $D\left(\mu_n^{(1)}, \mu_n^{(2)}\right)$, i. e. satisfying the equality

$$D\left(\mu_n^{(1)}, \mu_n^{(2)}\right) = D\left(\mu_n^{(1)}, \mu_n^{(2)} | \mathcal{S}_n\right) \quad \text{(cf. (20))}, \tag{52}$$

e.g. the intersection of all sub-σ-algebras $\mathcal{S} \subset \mathcal{B}^d$ preserving this divergence.

A strong motivation for this extension is the fact that the Scheffé selection $f_n^* \in \left\{f_n^{(1)}, f_n^{(2)}\right\}$ cannot be universally better of $f_n^{(1)}, f_n^{(2)}$ with respect to all ϕ-divergence error criteria. This was demonstrated by Example 3 in the previous section presenting density f, estimates $f_n^{(i)} = f_n^{(i)}(\cdot; X_1, \dots, X_n)$, and ϕ-divergences $D_{\phi_i}(f,g)$, $i \in \{1,2\}$ admitting with positive probability samples $X_1, \dots, X_n$ for which simultaneously

$$D_{\phi_1}(f_n^{(1)}, f) < D_{\phi_1}(f_n^{(2)}, f) \quad \text{and} \quad D_{\phi_2}(f_n^{(1)}, f) > D_{\phi_2}(f_n^{(2)}, f).$$

Next follows the main result of this section dealing with the concepts just introduced above.

Theorem 2. The estimate f_n^* resulting from the metric divergence selection rule (51) satisfies the inequality

$$D\left(f_n^*, f\right) \le 3D\left(f_n^{(0)}, f\right) + 2D\left(\mu, \mu_n | \mathcal{S}_n\right). \tag{53}$$

Proof. We can start with the equality (47) valid in the present situation as well. It suffices to prove that for $i \neq j$

$$D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{ij} \leq 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{ij}$$

where $\mathcal{E}_{ij}$ is defined by (46) for $f_n^*, f_n^{(0)}$ given by (50), (51). Using repeatedly the triangle inequality and relations (52) and (20) we obtain

$$\begin{aligned}
D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{21} &= D\left(f_n^{(0)}, f_n^{(2)}\right)\mathcal{E}_{21} = D\left(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_n\right)\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f|\mathcal{S}_n\right) + D\left(f_n^{(2)}, f|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(2)}, \mu_n|\mathcal{S}_n\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(2)}, \mu|\mathcal{S}_n\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(1)}, \mu|\mathcal{S}_n\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[2D\left(f_n^{(1)}, f\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&= 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21}.
\end{aligned}$$

In the same manner we obtain

$$D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{12} \leq 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{12}$$

which completes the proof of (53).

The next corollary presents a different expression of the error term in (53).

Corollary 2. The estimate f_n^* resulting from the selection rule (51) for a metric divergence $D(f,g) = D_\phi(f,g)^\pi$ satisfies the inequality

$$D\left(f_n^*, f\right) \leq 3D\left(f_n^{(0)}, f\right) + 2^{\pi+1}c_\phi^\pi \sup_{B\in\mathcal{S}_n}\left|\int_A f - \mu_n(B)\right|^\pi \tag{54}$$

where $f_n^{(0)}, f$ and $\mathcal{S}_n$ are the same as in Theorem 2 and $c_\phi = \phi(0) < \infty$.

Proof. By Proposition 8.27 in Liese and Vajda (1987) and (13),

$$D_\phi\left(\mu_n, \mu|\mathcal{S}_n\right) \leq c_\phi V\left(\mu_n, \mu|\mathcal{S}_n\right) \quad \text{and} \quad V\left(\mu_n, \mu|\mathcal{S}_n\right) = 2\sup_{A\in\mathcal{S}_n}\left|\mu(A) - \mu_n(A)\right|.$$

The rest is clear from Theorem 2 and (44).

As in the previous section, our first step is to verify that Theorem 1 generalizes the Devroye–Lugosi result (11).

Example 6. Putting $D(f,g) = V(f,g)$ in Theorem 2 and using the fact that by (25) the sub-σ-algebra $\mathcal{S}_{A_n}$ preserves the total variation $V(f_n^{(1)}, f_n^{(2)})$ of the estimates $f_n^{(1)}, f_n^{(2)}$, we get

$$V\left(f_n^*, f\right) \leq 3V\left(f_n^{(0)}, f\right) + 2V\left(\mu, \mu_n | \mathcal{S}_{A_n}\right).$$

This coincides with the equivalent form (24) of the Devroye-Lugosi inequality (11).

Most important from the point of view of applications is the complexity of the sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ which appears in the right-hand error terms of (53) and (54). It depends on the complexities of the used error criterion $D(f,g)$ and estimates $f_n^{(1)}$, $f_n^{(2)}$. In the previous example we have seen that if $D(f,g)$ is as simple as the total variation $V(f,g)$, then $\mathcal{S}_n$ is the simple σ-algebra $\mathcal{S}_{A_n}$ generated by just one set – the Scheffé set A_n of the estimates $f_n^{(1)}$, $f_n^{(2)}$ – irrespective of how complex these estimates are. In the following example we shall see the opposite extreme, namely simple estimates $f_n^{(1)}$, $f_n^{(2)}$ leading to a simple σ-algebra $\mathcal{S}_n = \mathcal{S}_{B_n}$ generated by just one set B_n specified by these estimates, irrespective of how complex the divergence criterion $D(f,g)$ is. More precisely, B_n does not depend on this criterion at all.

Example 7. Let the sample $X_1, \ldots, X_n$ be governed by a bell-shaped density f on $\mathbb{R}$ and consider the sample mean and variance

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)^2,$$

and also the following *central cover set*

$$B_n = \{x : |x - \mu_n| < 3\sigma_n\}. \tag{55}$$

Let f be estimated by Cauchy type densities

$$f_n^{(1)}(x) = \frac{\sigma_n}{\pi\left[\sigma_n^2 + (x-\mu_n)^2\right]}$$

and

$$f_n^{(2)}(x) = \boldsymbol{I}(x \in B_n)\,\frac{b\sigma_n}{\pi[\sigma_n^2 + (x-\mu_n)^2]} \tag{56}$$

where

$$b = \left[1 - 2\left(\frac{1}{2} - \frac{1}{\pi}\operatorname{arctg} 3\right)\right]^{-1} = \frac{\pi}{2\operatorname{arctg} 3}.$$

In (56) we used the fact that the condition $\boldsymbol{I}(x \in B_n)$ cuts away from $f_n^{(1)}(x)$ two tail probabilities of the size

$$\begin{aligned}\int_{-\infty}^{\mu_n - 3\sigma_n} f_n^{(1)} &= \int_{-\infty}^{-3} \frac{\mathrm{d}x}{\pi[1+x^2]} \\ &= \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}(-3) = \frac{1}{2} - \frac{1}{\pi}\operatorname{arctg} 3\end{aligned}$$

so that the $f_n^{(1)}$-probability of the sample central cover set is 1/b. The likelihood ratio $f_n^{(2)}/f_n^{(1)}$ is piecewise constant,

$$\frac{f_n^{(2)}(x)}{f_n^{(1)}(x)} = \begin{cases} b & \text{if } x \in B_n \\ 0 & \text{otherwise,} \end{cases}$$

where b is the normalizing factor used in (56). Therefore the sub-σ-algebra $\mathcal{S}_{B_n} = \{\mathbb{R}, B_n, B_n^c, \emptyset\} \subset \mathcal{B}$ generated by the central cover set B_n of (55) is sufficient for the family $\{f_n^{(1)}, f_n^{(2)}\}$. By what was said in Section 2, this means that $\mathcal{S}_{B_n}$ preserves for every convex ϕ the ϕ-divergence $D_\phi(f_n^{(1)}, f_n^{(2)})$. In other words, the sub-σ-algebra $\mathcal{S}_n$ considered in Theorem 2 and Corollary 2 is $\mathcal{S}_{B_n}$. Hence, by Theorem 1 and formula (21), for every metric divergence criterion $D(f,g) = D_\phi(f,g)^\pi$ with $\pi > 0$

$$D\left(f_n^*, f\right) \leq 3D\left(f_n^{(0)}, f\right) + 2\left[\sum_{B \in \{B_n, B_n^c\}} \int_B f\,\phi\left(\frac{\mu_n(B)}{\int_B f}\right)\right]^\pi. \tag{57}$$

By Corollary 2, simpler but in general weaker variant of the result (57) is the inequality

$$D\left(f_n^*, f\right) \leq 3D\left(f_n^{(0)}, f\right) + 2^{\pi+1} c_\phi^\pi(0) \left|\int_{B_n} f - \mu_n(B_n)\right|^\pi. \tag{58}$$

Next follows a theorem which generalizes and specifies the phenomena observed in the last example.

Theorem 3. If the metric divergence criterion $D(f,g)$ is a ϕ-divergence power with $\phi(t)$ strictly convex in the whole domain $t > 0$ then a sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ preserves $D(f_n^{(1)}, f_n^{(2)})$ in the sense

$$D\left(f_n^{(1)}, f_n^{(2)} | \mathcal{S}_n\right) = D\left(f_n^{(1)}, f_n^{(2)}\right)$$

if and only if $\mathcal{S}_n$ is sufficient for $\{f_n^{(1)}, f_n^{(2)}\}$.

Proof. Let $D(f_n^{(1)}, f_n^{(2)}) = D_\phi(f_n^{(1)}, f_n^{(2)})^\pi$ for some $\pi > 0$. By the Corollary 2 above, the metricity of $D_\phi(f,g)^\pi$ implies $D_\phi(f_n^{(1)}, f_n^{(2)}) \leq 2\phi(0) < \infty$. Hence, by Corollary 1.29 in Liese and Vajda (1987), the equality $D_\phi(f_n^{(1)}, f_n^{(2)}) = D_\phi(f_n^{(1)}, f_n^{(2)} | \mathcal{S}_n)$ takes place if and only if $\mathcal{S}_n$ is sufficient.

From this theorem we see that functions ϕ strictly convex everywhere define the most complex divergence criteria for which the σ-algebra $\mathcal{S}_n$ is simple only if the estimates $f_n^{(1)}$, $f_n^{(2)}$ are simple enough. Example 4 illustrated such situation.

5 Appendix

For practical applications of the results of Sections 3 and 4 one needs concrete metric divergence criteria $D(f,g) = D_\phi(f,g)^\pi$ with known and simple upper and lower bound $U_\phi(V)$ and $L_\phi(V)$ introduced in (41). For this purpose one can use the criteria from the class

$$D(f,g) = \mathcal{D}_\alpha(f,g)^{\pi(\alpha)} \quad \text{for} \quad \pi(\alpha) = \frac{1}{\max\{2,\alpha\}} = \begin{cases} \frac{1}{2} & \text{when } -\infty < \alpha \le 2 \\ \\ \frac{1}{\alpha} & \text{when } \alpha > 2. \end{cases} \tag{59}$$

introduced in (30)−(33). The following theorem summarizes basic relevant properties of the divergences $\mathcal{D}_\alpha(f,g)$. For the proof we refer to Vajda (2008).

Theorem A1.

(i) $\mathcal{D}_\alpha(f,g)$ are ϕ_α-divergences with functions $\phi_\alpha(t)$ strictly convex in the domain $t > 0$ when $\alpha \ne 0$ and self-adjoint in the sense $\phi_\alpha(t) = t\phi_\alpha(1/t)$ on this domain.

(ii) The lower bounds of the divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$ under the constraint $V(f,g) = V$ are given for all $0 \le V \le 2$ by the formulas

$$L_\alpha(V) = \frac{|\alpha|}{\alpha(\alpha-1)} \left(2^\alpha - \left[\left(1 + \frac{V}{2}\right)^{1/\alpha} + \left(1 - \frac{V}{2}\right)^{1/\alpha} \right]^\alpha \right) \tag{60}$$

if $\alpha(\alpha-1) \ne 0$ and otherwise by the corresponding limits

$$L_0(V) = V/2, \quad L_1(V) = \left(1 + \frac{V}{2}\right) \ln\left(1 + \frac{V}{2}\right) + \left(1 - \frac{V}{2}\right) \ln\left(1 - \frac{V}{2}\right). \tag{61}$$

(iii) The upper bounds of the divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$ under the constraint $V(f,g) = V$ are $U_\alpha(V) = c_\alpha V$ where $c_\alpha > 0$ is continuous in the variable $\alpha \in \mathbb{R}$, given by the formula

$$c_\alpha = \phi_\alpha(0) = \begin{cases} \dfrac{2^{\alpha-1}}{|\alpha|+1} & \text{when } \alpha < 0 \\ \ln 2 & \text{when } \alpha = 1 \\ \dfrac{2^{\alpha-1}-1}{\alpha-1} & \text{when } \alpha \ge 0,\ \alpha \ne 1. \end{cases} \tag{62}$$

(iv) The powers $\mathcal{D}_\alpha(f,g)^{\pi(\alpha)}$ given in (59) are metrics in the space of probability densities f, g.

Remark. Putting $\alpha = 0$ in (iv) of Theorem A1 one obtains among other results also the inequality

$$\sqrt{\mathcal{D}_0(f,g)} \le \sqrt{\mathcal{D}_0(f,h)} + \sqrt{\mathcal{D}_0(h,g)}$$

for the particular divergence $\mathcal{D}_0(f,g) = V(f,g)/2$. This inequality is weaker than the classical triangle inequality

$$\mathcal{D}_0(f,g) \le \mathcal{D}_0(f,h) + \mathcal{D}_0(h,g) \tag{63}$$

obtained by applying the L_1-norm argument to the total variation $V(f,g)$. Using the continuity of the divergences $\mathcal{D}_\alpha(f,g)$ in the variable $\alpha \in \mathbb{R}$ we can deduce from (63) that more sophisticated arguments than those used to prove Theorem A1 lead to stronger triangle inequalities also for the remaining divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$, in particular for those with α close to 0.

Acknowledgement. This research was supported by the GA ČR grant 102/07/1131.

References

A. Berlinet, G. Biau and L. Rouvière (2005a): Parameter selection in modified histogram estimates. *Statistics* **39,** 91-105.

A. Berlinet, G. Biau and L. Rouvière (2005b): Optimal L_1 bandwidth selection for variable kernel estimates. *Statistics and Probability Letters* **74,**.116-127.

I. Csiszár (1967a): Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungarica* **2**, 299–318.

I. Csiszár (1967b): On topological properties of f-divergences. *Studia Sci. Math. Hungarica* **2**, 329–339.

L. Devroye and G. Lugosi (2001): *Combinatorial Methods in Density Estimation.* Springer, Berlin.

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

F. Liese and I. Vajda (2006): On divergences and informations in statistics and information theory. *IEEE Trans. Inform. Theory* **52**, 10, 4394–4412.

F. Österreicher and I. Vajda (2003): A new class of metric divergences on probability spaces and its statistical applications. *Ann. Inst. Statist. Math.* **55**, 639–653.

I. Vajda (2008): On metric f-divergences of probability measures. *Kybernetika* (submitted).