Kybernetika

Friedrich Liese

ϕPHI-divergences, sufficiency, Bayes sufficiency, and deficiency

Kybernetika, Vol. 48 (2012), No. 4, 690--713

Persistent URL: http://dml.cz/dmlcz/143056

KYBERNETIKA — VOLUME 48 (2012), NUMBER 4, PAGES 690–713

ϕ-DIVERGENCES, SUFFICIENCY, BAYES SUFFICIENCY, AND DEFICIENCY

Friedrich Liese

In memory of my dear friend and highly regarded colleague Igor Vajda.

The paper studies the relations between ϕ-divergences and fundamental concepts of decision theory such as sufficiency, Bayes sufficiency, and LeCam's deficiency. A new and considerably simplified approach is given to the spectral representation of ϕ-divergences already established in Österreicher and Feldman [28] under restrictive conditions and in Liese and Vajda [22], [23] in the general form. The simplification is achieved by a new integral representation of convex functions in terms of elementary convex functions which are strictly convex at one point only. Bayes sufficiency is characterized with the help of a binary model that consists of the joint distribution and the product of the marginal distributions of the observation and the parameter, respectively. LeCam's deficiency is expressed in terms of ϕ-divergences where ϕ belongs to a class of convex functions whose curvature measures are finite and satisfy a normalization condition.

Keywords: divergences, sufficiency, Bayes sufficiency, deficiency

Classification: 62B05, 62B10, 62B15, 62G10

1. INTRODUCTION

Csiszár [8] (and independently also Ali and Silvey [1]) introduced the ϕ-divergence

$$D_\phi(P,Q) = \int \phi\left(\frac{p}{q}\right) q \,\mathrm{d}\mu$$

for a convex $\phi : (0,\infty) \longmapsto \mathbb{R}$ where μ is a σ-finite measure which dominates the distributions P and Q and the integrand is appropriately specified at the points where the densities $p = \mathrm{d}P/\mathrm{d}\mu$ and/or $q = \mathrm{d}Q/\mathrm{d}\mu$ are zero. For $\phi(t) = t\ln t$ the ϕ-divergence reduces to the classical information divergence

$$I(P,Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P,$$

which was systematically studied by Kullback and Leibler [18] and others who recognized its importance in information theory. For the convex or concave functions $\phi(t) = t^\alpha$,

$\alpha > 0$ we obtain the so-called Hellinger integrals

$$H_\alpha(P,Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)^\alpha \left(\frac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1-\alpha} \mathrm{d}\mu, \quad \alpha > 0,$$

that for $\alpha > 0$, $\alpha \neq 1$ are closely related to the divergences

$$R_\alpha(P,Q) = (\alpha - 1)^{-1} \ln H_\alpha(P,Q)$$

introduced by Rényi [33]. Note that the divergence measures $-\ln H_\alpha(P,Q)$ were considered for $0 < \alpha < 1$ already in Chernoff [6] and the special case for $\alpha = 1/2$ in Bhattacharyya [5] and Kakutani [17].

Among the ϕ-divergences one can find the basic divergence measures of probability theory and mathematical statistics, such as the total variation $\|P - Q\|$ for $\phi(t) = |t-1|$, the Pearson divergence $\chi^2(P,Q)$ for $\phi(t) = (t-1)^2$ or, more generally, the likelihood ratio cumulants $\chi^\alpha(P,Q)$ for $\phi(t) = |t-1|^\alpha, \alpha \geq 1$, systematically studied in Vajda [42].

Statistical applications of ϕ-divergences were considered e. g. by Ali and Silvey [1], Csiszár [9], Arimoto [2], Barron et al. [3], Berlinet et al. [4], and Vajda [45]. Decision-theoretic applications of ϕ-divergences can be found e. g. in Kailath [16], Poor [31], LeCam [20], Read and Cressie [32], Clarke and Barron [7], Guntuboyina [13], Nguyen et al. [27], Torgersen [40], Österreicher and Vajda [29], and Topsoe [39]. Jager and Wellner [15] used ϕ-divergences to construct goodness of fit statistics and studied their asymptotic behavior.

Due to the growing importance of divergences in information theory, statistics and probability theory the investigation of the structure of ϕ-divergences and their relations to fundamental concepts of statistics and decision theory deserves attention. In this sense the present paper is a continuation of Liese and Vajda [22] and [23] where a representation of ϕ-divergences in terms of the minimum Bayes error $b_\pi(P,Q)$ was used to simplify the general theory of ϕ-divergences. A crucial point was a second order generalized Taylor formula for convex functions. In this point we go one step further in this paper. Let

$$\widetilde{\phi}(t) := \phi(t) - \phi(1) - (t-1)D^+\phi(1),$$

where $D^+\phi$ denotes the right derivative of ϕ. The convex function $\widetilde{\phi}$ is centered in the sense that it is zero at $t_0 = 1$ and has a vanishing right hand derivative at $t_0 = 1$. It is easy to see that

$$D_\phi(P,Q) - \phi(1) = D_{\widetilde{\phi}}(P,Q).$$

Hence it suffices to deal with $\widetilde{\phi}$ which is represented as a spectral representation

$$\widetilde{\phi}(t) = \int \psi_\pi(t)\gamma_\phi(\mathrm{d}\pi)$$

in terms of the convex functions ψ_π that are elementary convex functions in the sense that ψ_π is piecewise linear and strictly convex only at the point $t_0 = (1-\pi)/\pi$. The weight measure γ_ϕ is closely related to the curvature measure of the convex function ϕ. The above representation of $\widetilde{\phi}$ is new and reduces the proof of the integral representation of ϕ-divergences (spectral representation) to the application of the Fubini theorem.

Similarly as in Liese and Vajda [22] and [23] the spectral representation of ϕ-divergences as averaged statistical information allows to prove the general form of the information processing theorem for ϕ-divergences (cf. Csiszár [9]) in a much simpler way than this was achieved in the previous literature (see Csiszár [8] and [9], Ali and Silvey [1], and Liese and Vajda [21]). The spectral representation of ϕ-divergences, applied to suitably chosen convex functions ϕ, provides a unified and considerably simplified approach to the different characterizations of sufficiency in the literature. For a Bayes model (X, Θ) consisting of the observation X and the random and unobservable Θ we show that the Bayes sufficiency of a statistic T is equivalent with the sufficiency of (T, Θ) for the binary model $\{\mathcal{L}(X, \Theta), \mathcal{L}(X) \otimes \mathcal{L}(\Theta)\}$ that consists of the joint distribution $\mathcal{L}(X, \Theta)$ and the product of the marginal distributions $\mathcal{L}(X) \otimes \mathcal{L}(\Theta)$. We also establish the relation between Bayes sufficiency and a information processing theorem for the Bayes model.

Since for binary models $\{P, Q\}$ sufficiency is equivalent to the equality in the information processing theorem for at least one strictly convex function the question arise how LeCam's deficiency can be characterized in terms of ϕ-divergences. Using the known relation between deficiency and the minimum Bayes error probabilities $b_\pi(P, Q)$ we are able to show that the deficiency of two binary models is identical with the maximum error for ϕ-divergences if we replace the distributions from the one model by the distributions of the other model. The formulation "maximum error" means that in contrast to the sufficiency we have to consider not only one convex function but some family of convex functions. This family is characterized by a normalization condition to the curvature measures. The established relation between ϕ-divergences and deficiency generalizes the concave function criterion of decision theory.

This paper is devoted to my dear friend and colleague Igor Vajda, with whom I collaborated for more than 30 years. I will remember Igor as an exceptional person, a careful listener and a person with whom you could discuss any topic. He was an outstanding mathematician who was full of innovative ideas. He enjoyed working with others and was a highly appreciated colleague.

2. SPECTRAL REPRESENTATION OF ϕ-DIVERGENCES

In the first part of this section we collect well known facts on convex functions defined on $(0, \infty)$ and establish some new technical results. A function $\phi : (0, \infty) \to \mathbb{R}$ is called convex if for every $s, t \in (0, \infty)$ and $0 \le \alpha \le 1$ it holds

$$\phi(\alpha s + (1 - \alpha)t) \le \alpha\phi(s) + (1 - \alpha)\phi(t).$$

Every convex function $\phi : (0, \infty) \to \mathbb{R}$ is continuous, the derivative from the right $D^+\phi(x)$ exists for every $x \in (0, \infty)$, the function $D^+\phi$ is nondecreasing and continuous from the right and the fundamental theorem of analysis holds

$$\phi(t) - \phi(s) = \int_s^t D^+\phi(\tau)\, \mathrm{d}\tau, \quad 0 < s < t < \infty, \tag{2.1}$$

see Roberts and Varberg [34]. As $D^+\phi$ is continuous from the right and nondecreasing there is a uniquely determined σ-finite measure γ_ϕ on the Borel sets of $(0, \infty)$ that

satisfies

$$\rho_\phi((s,t]) = D^+\phi(t) - D^+\phi(s), \quad 0 < s < t < \infty. \tag{2.2}$$

The measure ρ_ϕ is called the *curvature measure* of ϕ. This notation origins from the fact that

$$\rho_\phi(B) = \int_B \phi''(t)\,\mathrm{d}t \tag{2.3}$$

for every twice continuously differentiable convex functions ϕ.

The classical Taylor formula can be obtained by a successive partial integration. We use this idea and apply the integration be parts for measures. The representation (2.1) yields for $a < b$

$$\begin{aligned}\phi(b) - \phi(a) - D^+\phi(a)(b-a) &= \int_a^b (D^+\phi(\tau) - D^+\phi(a))\,\mathrm{d}\tau \\ &= \int_{(a,b]} (b-x)\rho_\phi(\mathrm{d}x).\end{aligned}$$

Similarly, for $b < a$

$$\begin{aligned}\phi(b) - \phi(a) - D^+\phi(a)(b-a) &= \int_b^a (D^+\phi(\tau) - D^+\phi(a))\,\mathrm{d}\tau \\ &= \int_{(b,a]} (x-b)\rho_\phi(\mathrm{d}x).\end{aligned}$$

From here we get the generalized Taylor formula

$$\phi(t) - \phi(1) - (t-1)D^+\phi(1) = \begin{cases} \displaystyle\int_{(1,t]} (t-\tau)\rho_\phi(\mathrm{d}\tau) & \text{if} \quad 1 < t < \infty \\ \displaystyle\int_{(t,1]} (\tau-t)\rho_\phi(\mathrm{d}\tau) & \text{if} \quad 0 < t \le 1, \end{cases} \tag{2.4}$$

that appears already in Liese and Vajda [22] and [23]. As the right hand terms are nonnegative it follows that

$$\widetilde{\phi}(t) := \phi(t) - \phi(1) - (t-1)D^+\phi(1) \ge 0. \tag{2.5}$$

The application of the monotone convergence theorem to the right hand side of (2.4) yields that the limits

$$\phi(0) := \lim_{t\downarrow 0} \phi(t) \quad \text{and} \quad \frac{\phi(\infty)}{\infty} := \lim_{t\uparrow\infty} \frac{\phi(t)}{t}$$

exist but may take the value ∞. Subsequently we use the convention $0\cdot\infty = 0$. A crucial point for all further considerations is a representation of $\widetilde{\phi}$ in terms of elementary convex functions. Let

$$\psi_\pi(t) = \begin{cases} \pi t - (\pi t) \wedge (1-\pi) & \text{if} \quad 0 < \pi \le \frac{1}{2}, t > 0 \\ (1-\pi) - (\pi t) \wedge (1-\pi) & \text{if} \quad \frac{1}{2} < \pi < 1, t > 0. \end{cases} \tag{2.6}$$

The function ψ_π is a nonnegative, it holds $D^+\psi_\pi(t) = 0$ for $t \neq (1-\pi)/\pi$ and $D^+\psi_\pi$ has a jump of high π at $(1-\pi)/\pi$ so that the curvature measure is

$$\rho_{\psi_\pi} = \pi\delta_{(1-\pi)/\pi}. \tag{2.7}$$

We see that ψ_π elementary in the sense that it is piecewise linear and strictly convex only at $t_0 = (1-\pi)/\pi$. Later these functions appear when studying the Bayes error in hypothesis testing problems.

For every convex function we introduce the modified curvature measure on the Borel sets $B \subseteq (0,1)$ by

$$\gamma_\phi(B) = \int I_B\left(\frac{1}{1+\pi}\right)(1+\pi)\rho_\phi(\mathrm{d}\pi), \tag{2.8}$$

where I_B denotes the indicator function of the set B. Later we will use γ_ϕ as weight measure for the Bayes error in binary models with prior $\pi, 1-\pi$. The definition of γ_ϕ implies

$$\begin{aligned}\int_{(0,1)} g(\pi)\gamma_\phi(\mathrm{d}\pi) &= \int_{(0,\infty)} g\left(\frac{1}{1+\tau}\right)(1+\tau)\rho_\phi(\mathrm{d}\tau),\\ \int_{(0,\infty)} h(\tau)\rho_\phi(\mathrm{d}\tau) &= \int_{(0,1)} \pi h\left(\frac{1-\pi}{\pi}\right)\gamma_\phi(\mathrm{d}\pi),\end{aligned} \tag{2.9}$$

for every measurable functions $g : (0,1) \to [0,\infty), h : (0,\infty) \to [0,\infty)$. If ϕ is twice continuously differentiable then $\rho_\phi(\mathrm{d}\tau) = \phi''(\tau)\,\mathrm{d}\tau$ by (2.3) and

$$\begin{aligned}\int_{(0,1)} g(\pi)\gamma_\phi(\mathrm{d}\pi) &= \int_{(0,\infty)} g\left(\frac{1}{1+\tau}\right)(1+\tau)\phi''(\tau)\,\mathrm{d}\tau\\ &= \int_{(0,1)} g(\pi)\frac{1}{\pi^3}\phi''\left(\frac{1-\pi}{\pi}\right)\mathrm{d}\pi,\end{aligned} \tag{2.10}$$

where the last equality follows from the change of variables $\tau = (1-\pi)/\pi$. To illustrate γ_ϕ we consider ψ_π in (2.6). Then by (2.7) and the first equation in (2.9) with $g = 1$

$$\begin{aligned}\gamma_{\psi_\pi}((0,1)) &= \int (1+\tau)(\pi\delta_{(1-\pi)/\pi})(\mathrm{d}\tau)\\ &= \pi\left(1 + \frac{1-\pi}{\pi}\right) = 1,\end{aligned} \tag{2.11}$$

so that the total mass of the modified curvature measure γ_{ψ_π} of the elementary convex functions ψ_π is one.

Now we use the modified curvature measure γ_ϕ to establish a generalized second order Taylor expansion which may be considered as a spectral decomposition that gives a decomposition of a convex function into the piecewise linear functions in (2.6).

Theorem 2.1. If $\phi : (0,\infty) \to \mathbb{R}$ is convex, then $\widetilde{\phi}$ in (2.5) has the spectral representation

$$\widetilde{\phi}(t) = \int_{(0,1)} \psi_\pi(t)\gamma_\phi(\mathrm{d}\pi), \quad 0 < t < \infty, \tag{2.12}$$

$$\widetilde{\phi}(0) = \int_{(0,1)} \psi_\pi(0)\gamma_\phi(\mathrm{d}\pi), \tag{2.13}$$

$$\frac{\widetilde{\phi}(\infty}{\infty} = \int_{(0,1)} \left(\frac{\psi_\pi(\infty)}{\infty}\right) \gamma_\phi(\mathrm{d}\pi), \tag{2.14}$$

where the functions $\psi_\pi,\ 0 < \pi < 1$, are defined in (2.6).

P r o o f. For fixed $t \geq 1$ we set $h(\tau) = (t-\tau)I_{(1,t]}(\tau) = t - t\wedge\tau$. Then by (2.4) and (2.9)

$$\begin{aligned}\widetilde{\phi}(t) &= \int_{(0,\infty)} h(\tau)\rho_\phi(\mathrm{d}\tau) \\ &= \int_{(0,\frac{1}{2}]} \pi\left(t - t\wedge\left(\frac{1-\pi}{\pi}\right)\right)\gamma_\phi(\mathrm{d}\pi) \\ &= \int_{(0,\frac{1}{2}]} (\pi t - (\pi t)\wedge(1-\pi))\gamma_\phi(\mathrm{d}\pi).\end{aligned}$$

For $0 < t < 1$ we set $h(\tau) = (\tau - t)I_{[t,1)}(\tau) = \tau - t\wedge\tau$. It follows

$$\begin{aligned}\widetilde{\phi}(t) &= \int_{(\frac{1}{2},1)} \pi\left(\frac{1-\pi}{\pi} - t\wedge\left(\frac{1-\pi}{\pi}\right)\right)\gamma_\phi(\mathrm{d}\pi) \\ &= \int_{(\frac{1}{2},1)} ((1-\pi) - (\pi t)\wedge(1-\pi))\,\gamma_\phi(\mathrm{d}\pi).\end{aligned}$$

To prove (2.13) we note that the family of functions $\pi \mapsto \psi_\pi(t)$ is nondecreasing in t if $1 \geq t \downarrow 0$ so that (2.13) follows from the monotone convergence theorem. The proof of (2.14) is similar. □

Let P, Q be distributions on $(\mathcal{X}, \mathfrak{A})$. Suppose that μ is any σ-finite dominating measure and denote by p, q their respective μ-densities.

Definition 2.2. The functional

$$D_\phi(P,Q) := \int_{\{p>0,q>0\}} \phi\left(\frac{p}{q}\right) q\,\mathrm{d}\mu + \phi(0)\int_{\{p=0,q>0\}} q\,\mathrm{d}\mu + \frac{\phi(\infty)}{\infty}\int_{\{p>0,q=0\}} p\,\mathrm{d}\mu$$

is called the *ϕ-divergence of P* with respect to Q.

To see that the first right-hand integral is well-defined we refer to the inequality (2.5). We remark that $D_\phi(P,Q)$ may take on the value ∞. Moreover, the definition of $D_\phi(P,Q)$ is independent of the special choice of μ.

The concept of ϕ-divergence was independently introduced by Csiszár [8] and Ali and Silvey [1]. This general class of functionals includes special cases which appeared in Bhattacharyya [5], Kakutani [17], Kullback and Leibler [18], Chernoff [6], Matusita [25], Rényi [33]), and others. ϕ-divergences have been systematically studied in Vajda [43], Liese and Vajda [21], [22], [23] and in Liese and Miescke [24].

If $\phi : (0, \infty) \to \mathbb{R}$ is a convex function and $\psi(t) = \phi(t) + at + b$ then ψ is again convex and it holds

$$\psi(0) = \phi(0) + b, \quad \psi(1) = \phi(1) + a + b, \quad \frac{\psi(\infty)}{\infty} = \frac{\phi(\infty)}{\infty} + a.$$

These relations in combination with Definition 2.2 yield the invariance

$$D_\phi(P, Q) - \phi(1) = D_\psi(P, Q) - \psi(1), \tag{2.15}$$

and especially for $\psi = \widetilde{\phi}$ in (2.5)

$$D_\phi(P, Q) - \phi(1) = D_{\widetilde{\phi}}(P, Q). \tag{2.16}$$

Recall that a convex function ϕ is called strictly convex at $t_0 = 1$ if ϕ not linear in every interval $(1 - \varepsilon, 1 + \varepsilon), \varepsilon > 0$. It follows from (2.4) that this condition is equivalent with $\rho_\phi((1 - \varepsilon, 1 + \varepsilon)) > 0$ for every $\varepsilon > 0$. As $\widetilde{\phi} \geq 0$ we see $D_\phi(P, Q) - \phi(1) \geq 0$ where equality holds for $P = Q$. Conversely, if ϕ is strictly convex at $t_0 = 1$ the equality $D_\phi(P, Q) - \phi(1) = 0$ implies $P = Q$.

In general, the functional $D_\phi(P, Q)$ is not symmetric in the pair (P, Q). It is easy to see that the adjoint function $\phi^*(t) = t\phi(\frac{1}{t}), t > 0$, is convex and it holds

$$D_\phi(P, Q) = D_{\phi^*}(Q, P). \tag{2.17}$$

This means that the selfadjointness condition

$$\phi(t) = t\phi\left(\frac{1}{t}\right), \quad t > 0, \tag{2.18}$$

implies the symmetry

$$D_\phi(P, Q) = D_\phi(Q, P). \tag{2.19}$$

Even if the condition (2.18) is satisfied, in general, $D_\phi(P, Q)$ does not define a metric in the space of distributions. The problem is that the triangle inequality

$$D_\phi(P, R) \leq D_\phi(P, Q) + D_\phi(Q, P)$$

is not satisfied, in general. For $\phi(t) = |t - 1|$ the special ϕ-divergence

$$\|P - Q\| = D_\phi(P, Q) = \int |p - q| \, \mathrm{d}\mu$$

is the variational distance that satisfies (2.19) and fulfils the triangle inequality. It should be noted that the variational distance is the only ϕ-divergence that is a metric, see Vajda [44]. The *Hellinger distance*

$$D(P, Q) = \left[\int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\mu\right]^{1/2}$$

is also a metric, but it is not a ϕ-divergence; the square of the Hellinger distance is a ϕ-divergence. The *Vincze–LeCam distance* $LC(P,Q)$, with

$$LC^2(P,Q) = \frac{1}{2}\int \frac{(p-q)^2}{p+q}\,\mathrm{d}\mu,$$

was independently introduced by Vincze [46] and LeCam [20]. The functional $LC^2(P,Q)$ is obviously symmetric in P and Q and a ϕ-divergence for the strictly convex function

$$\phi(t) = \frac{(t-1)^2}{2(t+1)}. \tag{2.20}$$

Moreover, $LC(P,Q)$ satisfies the triangle inequality, see Vajda [44] for details. We consider $LC^2(P,Q)$ for the special dominating measure $R = \frac{1}{2}(P+Q)$. Then

$$L(x) := \frac{\mathrm{d}P}{\mathrm{d}R}(x) \quad \text{and} \quad 2 - L(x) = \frac{\mathrm{d}Q}{\mathrm{d}R}(x), \tag{2.21}$$

and with $p = L, q = 2 - L$,

$$LC^2(P,Q) = \int (L-1)^2\,\mathrm{d}R. \tag{2.22}$$

For some purposes it is useful to turn to the symmetrized and normalized version

$$\begin{aligned} &D_\phi\left(P, \frac{1}{2}(P+Q)\right) + D_\phi\left(Q, \frac{1}{2}(P+Q)\right) \\ &= \int \left[\phi\left(\frac{2p}{p+q}\right) + \phi\left(\frac{2q}{p+q}\right)\right]\left[\frac{p+q}{2}\right]\mathrm{d}\mu = D_{\widehat{\phi}}(P,Q), \end{aligned}$$

where

$$\widehat{\phi}(t) = \frac{1+t}{2}\left(\phi\left(\frac{2t}{1+t}\right) + \phi\left(\frac{2}{1+t}\right)\right). \tag{2.23}$$

The convexity of ϕ gives for $0 < \alpha < 1$ and $t_1 \neq t_2$

$$\begin{aligned} &\frac{1+\alpha t_1 + (1-\alpha)t_2}{2}\phi\left(\frac{\alpha(1+t_1)}{1+\alpha t_1+(1-\alpha)t_2}\frac{2t_1}{1+t_1} + \frac{(1-\alpha)(1+t_2)}{1+\alpha t_1+(1-\alpha)t_2}\frac{2t_2}{1+t_2}\right) \\ &\leq \alpha\frac{1+t_1}{2}\phi\left(\frac{2t_1}{1+t_1}\right) + (1-\alpha)\frac{1+t_2}{2}\phi\left(\frac{2t_2}{1+t_2}\right), \end{aligned} \tag{2.24}$$

where the smaller sign stands for a strictly convex ϕ. A similar inequality for $\frac{1+t}{2}\phi\left(\frac{2}{1+t}\right)$ shows that $\widehat{\phi}$ is convex and even strictly convex if ϕ does. If we take $\phi(t) = (t-1)^2$ then

$$D_\phi(P,Q) = \int_{\{q>0\}} \frac{(p-q)^2}{q}\,\mathrm{d}\mu + \infty \cdot P(q=0)$$

is the *Pearson divergence* and

$$\widehat{\phi}(t) = \frac{1+t}{2}\left(\left(\frac{2t}{1+t}-1\right)^2 + \left(\frac{2}{1+t}-1\right)^2\right) = \frac{(t-1)^2}{2(t+1)}$$

is just the convex function in (2.20) that appears in the definition of the Vincze–LeCam distance. This means that the Vincze–LeCam distance is the symmetrized and normalized version of the Pearson divergence.

Now we study a class of divergences which is closely related to the problem of testing simple statistical hypotheses. For distributions P, Q on $(\mathcal{X}, \mathfrak{A})$ we want to find a test $\varphi : \mathcal{X} \to [0,1]$ that minimizes the Bayes error

$$\int (\pi\varphi p + (1-\pi)(1-\varphi)q)\,\mathrm{d}\mu = (1-\pi) + \int \varphi(\pi p - (1-\pi)q)\,\mathrm{d}\mu,$$

where $0 \le \pi \le 1$. Obviously, a test φ_B, called a Bayes test, minimizes the Bayes error if and only if μ-a.s.

$$\begin{aligned} \varphi_B &= 1 \quad \text{if} \quad \pi p < (1-\varphi)q \\ \varphi_B &= 0 \quad \text{if} \quad \pi p > (1-\varphi)q, \end{aligned}$$

and the minimal Bayes error is

$$b_\pi(P,Q) = \int ((\pi p) \wedge (1-\pi)q)\,\mathrm{d}\mu. \tag{2.25}$$

The value $\pi \wedge (1-\pi)$ is the *a priori loss* before making the experiment and $b_\pi(P,Q)$ is the *a posteriori loss* of the experiment so that the difference

$$B_\pi(P,Q) = \pi \wedge (1-\pi) - b_\pi(P,Q) \tag{2.26}$$

is the *statistical information* in the Bayes model $((\mathcal{X}, \mathfrak{A}), \{P,Q\}, \{\pi, 1-\pi\})$, see De Groot [11]. It is easy to see that the statistical information $B_\pi(P,Q)$ has the structure of a ϕ-divergence for the function ψ_π, i. e.

$$B_\pi(P,Q) = D_{\psi_\pi}(P,Q). \tag{2.27}$$

Note that the definition of $B_\pi(P,Q)$ implies $B_\pi(P,Q) = B_{1-\pi}(Q,P)$. The representation (2.12) leads to the conjecture that every ϕ-divergence can be represented as a similar superposition of the statistical informations $B_\pi(P,Q), 0 < \pi < 1$. Such representations connect the concept of the distance of distributions measured by the ϕ-divergence with decision theoretic concepts based on the minimum Bayes risk.

Theorem 2.3. If $\phi : (0,\infty) \to \mathbb{R}$ is convex and P, Q are distributions on $(\mathcal{X}, \mathfrak{A})$ then

$$D_\phi(P,Q) - \phi(1) = \int_{(0,1)} B_\pi(P,Q)\gamma_\phi(\mathrm{d}\pi). \tag{2.28}$$

Corollary 2.4. If ϕ is twice continuously differentiable then

$$D_\phi(P,Q) - \phi(1) = \int_{(0,1)} B_\pi(P,Q)\frac{1}{\pi^3}\phi''\left(\frac{1-\pi}{\pi}\right)\mathrm{d}\pi.$$

P r o o f . Definition 2.2 for ϕ replaced with $\widetilde{\phi}$ and (2.16) yield

$$D_\phi(P,Q) - \phi(1) = \int_{\{p>0, q>0\}} \widetilde{\phi}\left(\frac{p}{q}\right) q\,\mathrm{d}\mu + \widetilde{\phi}(0)Q(p=0) + \frac{\widetilde{\phi}(\infty)}{\infty}P(q=0).$$

It follows from (2.12) and the Theorem of Fubini that

$$\int_{\{p>0,q>0\}} \widetilde{\phi}\left(\frac{p}{q}\right) q \,\mathrm{d}\mu = \int_{(0,1)} \left(\int_{\{p>0,q>0\}} \psi_\pi \left(\frac{p}{q}\right) q \,\mathrm{d}\mu \right) \gamma_\phi(\mathrm{d}\pi).$$

The statements (2.13) and (2.14) yield

$$\widetilde{\phi}(0)Q(p=0) = \int_{(0,1)} \psi_\pi(0)Q(p=0)\gamma_{_\phi}(\mathrm{d}\pi),$$
$$\frac{\widetilde{\phi}(\infty)}{\infty}P(q=0) = \int_{(0,1)} \frac{\psi_\pi(\infty)}{\infty}P(q=0)\gamma_{_\phi}(\mathrm{d}\pi).$$

The sum of the left hand terms is $D_\phi(P,Q) - \phi(1)$ whereas the sum of the terms on the right hand side gives

$$\int D_{\psi_\pi}(P,Q)\gamma_{_\phi}(\mathrm{d}\pi)$$

which is the right hand term in (2.28) in view of (2.27). The proof of the Corollary follows from (2.10). □

The representation of ϕ-divergence in the previous theorem has been established by Österreicher and Feldman [28], Österreicher and Vajda [29], Guttenbrunner [14] for twice differentiable functions ϕ, and by Torgersen [40] for the special case of Hellinger integrals. The general case was treated in Liese and Vajda [22] and [23]. The proof given here is a considerable simplification of the approach given in [22] and [23] and is based on Theorem 2.1.

Example 2.5. Set

$$\rho_\alpha(t) = \begin{cases} t\ln t - t + 1 & \text{if} \quad \alpha = 1 \\ \frac{1}{\alpha(a-1)}(t^\alpha - \alpha(t-1) - 1) & \text{if} \quad \alpha \neq 0, \alpha \neq 1 \\ -\ln t & \text{if} \quad \alpha = 0 \end{cases} \tag{2.29}$$

and $I_\alpha(P,Q) = D_{\rho_\alpha}(P,Q)$. Then $I(P,Q) = I_1(P,Q)$ is the Kullback–Leibler information divergence. Furthermore

$$I_{\frac{1}{2}}(P,Q) = 4\int \left(\frac{1}{2}p + \frac{1}{2}q - \sqrt{pq}\right) \mathrm{d}\mu = 2D^2(P,Q),$$

where $D(P,Q)$ is the Hellinger distance. Another example from the family $I_\alpha(P,Q)$ is the Pearson divergence.

$$\chi^2(P,Q) = 2I_2(P,Q) = \int_{\{q>0\}} \frac{(p-q)^2}{q} \,\mathrm{d}\mu + \infty \cdot P(q=0).$$

The convex functions ρ_α in (2.29) are twice continuously differentiable and $\rho''_\alpha(t) = t^{\alpha-2}$, $\rho_\alpha(1) = 0$. Corollary 2.4 yields

$$I_\alpha(P,Q) = \int_{(0,1)} \frac{(1-\pi)^{\alpha-2}}{\pi^{\alpha+1}} B_\pi(P,Q)\,\mathrm{d}\pi.$$

For the most often considered $\alpha = 1/2, \alpha = 1$ and $\alpha = 2$ we get

$$D^2(P,Q) = \frac{1}{2}\int_{(0,1)} \frac{B_\pi(P,Q)}{[(1-\pi)\pi]^{3/2}}\,\mathrm{d}\pi, \quad I(P,Q) = \int_{(0,1)} \frac{B_\pi(P,Q)}{(1-\pi)\pi^2}\,\mathrm{d}\pi,$$
$$\chi^2(P,Q) = 2\int_{(0,1)} \frac{B_\pi(P,Q)}{\pi^3}\,\mathrm{d}\pi.$$

Example 2.6. For the convex function $\phi(t) = -t^\alpha, 0 < \alpha < 1$, we have

$$D_\phi(P,Q) = -H_\alpha(P,Q).$$

Applying Corollary 2.4 to $D_\phi(P,Q)$ and using $\phi(1) = -1$ we get

$$H_\alpha(P,Q) + 1 = a(1-\alpha)\int_{(0,1)} \frac{B_\pi(P,Q)}{(1-\pi)^{2-\alpha}\pi^{1+\alpha}}\,\mathrm{d}\pi.$$

Recall that $B_\pi(P,Q) = \pi \wedge (1-\pi) - b_\pi(P,Q)$ with $b_\pi(P,Q)$ from (2.25) and note that

$$a(1-\alpha)\int_{(0,1)} \frac{\pi \wedge (1-\pi)}{(1-\pi)^{2-\alpha}\pi^{1+\alpha}}\,\mathrm{d}\pi = 1.$$

Hence

$$H_\alpha(P,Q) = a(1-\alpha)\int_{(0,1)} \frac{b_\pi(P,Q)}{(1-\pi)^{2-\alpha}\pi^{1+\alpha}}\,\mathrm{d}\pi.$$

This formula was already established in Torgersen [40].

3. ϕ-DIVERGENCES, SUFFICIENCY, AND BAYES SUFFICIENCY

In addition to the binary model $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P,Q\})$ let now $(\mathcal{Y}, \mathfrak{B})$ be another measurable space and $K(B|x), B \in \mathfrak{B}, x \in \mathcal{X}$, be a stochastic kernel. Put

$$(KP)(B) = \int K(B|x)P(\mathrm{d}x), \quad B \in \mathfrak{B},$$

and introduce KQ in a similar way. The model $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{KP, KQ\})$ is called the *randomization* of $\mathcal{M}$. Intuitively it is clear that the model $\mathcal{N}$ is less informative than $\mathcal{M}$ as it is harder to distinguish between KP and KQ than to distinguish between P and Q. Thus we can anticipate the inequality $D_\phi(KP, KQ) \le D_\phi(P,Q)$, which is the content of the information processing theorem firstly established by Csiszár [8] in the general form. In preparation of this theorem we study the special ϕ-divergence $B_\pi(P,Q)$. For every test ψ for $\mathcal{N}$ the function

$$\varphi(x) = \int \psi(y)K(\mathrm{d}y|x)$$

is a test for $\mathcal{M}$ and it holds

$$\int \psi\,\mathrm{d}(KP) = \int \varphi\,\mathrm{d}P \quad \text{and} \quad \int \psi\,\mathrm{d}(KQ) = \int \varphi\,\mathrm{d}Q.$$

As $b_\pi(KP, KQ)$ is the minimal Bayes risk we arrive at

$$b_\pi(KP,KQ) = \inf_\psi \left[\pi \int \psi \,\mathrm{d}(KP) - (1-\pi) \int (1-\psi)\,\mathrm{d}(KQ)\right]$$
$$\geq \inf_\varphi \left[\pi \int \varphi \,\mathrm{d}P - (1-\pi) \int (1-\varphi)\,\mathrm{d}Q\right],$$

where the first and second supremum are taken over all tests for $\mathcal{N}$ and $\mathcal{M}$, respectively. Hence

$$\begin{aligned} B_\pi(KP, KQ) &= \pi \wedge (1-\pi) - b_\pi(KP, KQ) \\ &\leq \pi \wedge (1-\pi) - b_\pi(P, Q) = B_\pi(P, Q). \end{aligned} \tag{3.1}$$

This inequality says that the original Bayes model $\mathcal{M} = ((\mathcal{X}, \mathfrak{A}), \{P, Q\})$ contains more information than the randomized model $\mathcal{N} = ((\mathcal{Y}, \mathfrak{B}), \{KP, KQ\})$.

Theorem 3.1. If $(\mathcal{X}, \mathfrak{A})$ and $(\mathcal{Y}, \mathfrak{B})$ are measurable spaces and $K(B|x), B \in \mathfrak{B}, x \in \mathcal{X}$, is a stochastic kernel then for every distributions P, Q on $(\mathcal{X}, \mathfrak{A})$ and every convex function $\phi : (0, \infty) \to \mathbb{R}$,

$$D_\phi(KP, KQ) \leq D_\phi(P, Q), \tag{3.2}$$

with equality holding for

$$B_\pi(KP, KQ) = B_\pi(P, Q),\ 0 < \pi < 1. \tag{3.3}$$

Conversely, if ϕ is strictly convex in $(0, \infty)$ then $D_\phi(KP, KQ) = D_\phi(P, Q) < \infty$ implies (3.3).

P r o o f . The inequality (3.2) follows directly from (3.1) and Theorem 2.3 where equality holds if (3.3) is satisfied. Conversely, if $D_\phi(KP, KQ) = D_\phi(P, Q) < \infty$ then by Theorem 2.3

$$0 = D_\phi(P, Q) - D_\phi(KP, KQ) = \int [B_\pi(P, Q) - B_\pi(KP, KQ)]\, \gamma_\phi(\mathrm{d}\pi).$$

The integrand is nonnegative in view of (3.1). Consequently,

$$\gamma_\phi(\{\pi : B_\pi(P, Q) - B_\pi(KP, KQ) = 0\}) = 0. \tag{3.4}$$

It follows from (2.4) that ϕ is strictly convex in $(0, \infty)$ if and only if $\rho_\phi((a, b)) > 0$ for every $0 < a < b < \infty$ which is equivalent with $\gamma_\phi((c, d)) > 0$ for every $0 < c < d < 1$. The continuity of the function

$$\pi \mapsto B_\pi(P, Q) - B_\pi(KP, KQ)$$

and relation (3.4) provide (3.3). □

Now we specialize the kernel K. For a measurable mapping $T : \mathcal{X} \to \mathcal{Y}$ we consider the special kernel

$$K(B|x) = \delta_{T(x)}(B),$$

where δ_a is the Delta measure on a. Then

$$\begin{aligned}(KP)(B) &= \int \delta_{T(x)}(B)P(\mathrm{d}x)\\ &= P(T^{-1}(B)) = (P\circ T^{-1})(B),\end{aligned}$$

so that $KP = P\circ T^{-1}$, $KQ = Q\circ T^{-1}$ are the induced distributions. It turns out that the equality in (3.2) is closely related to the sufficiency of T. We briefly recall to the classical concept of sufficiency. Let $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta})$ be a statistical model and $(\mathcal{Y}, \mathfrak{B})$ be another measurable space. If $T : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping then the statistic T is called *sufficient* for $\mathcal{M}$ if for every $A \in \mathfrak{A}$ there is a measurable function $k_A : \mathcal{Y} \to [0, 1]$ such that for every $\theta \in \Delta$

$$E_\theta(I_A|T) = k_A(T), \quad P_\theta\text{-a. s.} \tag{3.5}$$

If the family $(P_\theta)_{\theta\in\Delta}$ is dominated by the σ-finite measure μ, $f_\theta(x) := \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x)$, $\theta \in \Delta$, are the corresponding densities and T is sufficient then by the Neyman factorization criterion there are measurable functions $g_\theta(y)$ and $h(x)$ such that

$$f_\theta(x) = g_\theta(T(x))h(x). \tag{3.6}$$

This means for a binary model $((\mathcal{X}, \mathfrak{A}), \{P, Q\})$ and

$$f_1 = \frac{\mathrm{d}P}{\mathrm{d}\mu} = g_1(T(x))h(x), \quad f_2 = \frac{\mathrm{d}Q}{\mathrm{d}\mu} = g_2(T(x))h(x),$$

which implies that the density L in (2.21) satisfies

$$L = \frac{f_1}{\frac{1}{2}(f_1 + f_2)} = \frac{g_1(T)}{\frac{1}{2}(g_1(T) + g_2(T))}.$$

Therefore L is a measurable function of T. Otherwise, if this condition holds then the Neyman criterion, applied to the dominating measure $R = (P + Q)/2$, yields the sufficiency of T. Hence we have the following reformulation of the Neyman criterion.

Clonclusion 3.2. $T : \mathcal{X} \to \mathcal{Y}$ is sufficient for $((\mathcal{X}, \mathfrak{A}), \{P, Q\})$ if and only if $L = \frac{\mathrm{d}P}{\mathrm{d}\left(\frac{1}{2}(P+Q)\right)}$ is a measurable function of T.

Now we are ready to give an information theoretic characterization of sufficiency.

Theorem 3.3. Suppose $T : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping. Then the condition (3.3), with $K = \delta_T$, is equivalent to each of the following conditions

A) T is sufficient for the model $(\mathcal{X}, \mathfrak{A}, \{P, Q\})$,

B) $LC(P\circ T^{-1}, Q\circ T^{-1}) = LC(P, Q)$,

C) $D_\phi(P\circ T^{-1}, Q\circ T^{-1}) = D_\phi(P, Q) < \infty$ for a strictly convex ϕ,

D) $D_\phi(P\circ T^{-1}, Q\circ T^{-1}) = D_\phi(P, Q)$ for every convex ϕ.

P r o o f . The proof is carried out according to the following scheme

$$A) \quad \longleftrightarrow \quad B) \quad \rightarrow \quad C) \quad \rightarrow \quad (3.3) \quad \rightarrow \quad D) \quad \rightarrow \quad B)$$

$A) \longleftrightarrow B)$: Recall that L is defined in (2.21). Then for every $B \in \mathfrak{B}$

$$\begin{aligned}\int_B E_R(L|T=y)(R\circ T^{-1})(\mathrm{d}y) &= \int_{T^{-1}(B)} LR(\mathrm{d}x)\\ &= (P\circ T^{-1})(B),\end{aligned}$$

which gives the known relation

$$\frac{\mathrm{d}(P\circ T^{-1})}{\mathrm{d}(R\circ T^{-1})}(y) = E_R(L|T=y), \quad R\text{-a. s.}$$

As in (2.22) we express the Vincze–LeCam distance of P, Q and $P\circ T^{-1}, Q\circ T^{-1}$ in terms of L and $E_R(L|T)$, respectively, and obtain

$$\begin{aligned}LC^2(P,Q) &= E_R(L-1)^2,\\ LC^2(P\circ T^{-1}, Q\circ T^{-1}) &= \int \left(E_R(L|T=y)-1\right)^2 (P\circ T^{-1})(\mathrm{d}y)\\ &= E_R(E_R(L|T)-1)^2.\end{aligned} \tag{3.7}$$

Using

$$\begin{aligned}E_R L &= E_R(E_R(L|T)) = 1,\\ E_R(LE_R(L|T)) &= E_R(E_R(L|T))^2,\end{aligned}$$

we get

$$E_R(L-E_R(L|T))^2 = LC^2(P,Q) - LC^2(P\circ T^{-1}, Q\circ T^{-1}).$$

From here we see that the condition $B)$ holds if and only if $L = \mathrm{d}P/\mathrm{d}R$ is a measurable function of T. Applying Conclusion 3.2 we see that the conditions $A)$ and $B)$ are equivalent.

$B) \rightarrow C)$ is clear.

$C) \rightarrow (3.3)$: If $C)$ holds for some convex function, say ϕ_0, then by the second part of Theorem 3.1, applied to ϕ_0, we get (3.3).

$(3.3) \rightarrow D)$: For every convex function ϕ the first part of Theorem 3.1 yields $D)$.

$D) \rightarrow B)$ is clear. □

The equivalence of the conditions $A)$ and $C)$ in Theorem 3.3 is an information theoretic characterization of sufficiency that goes back to Csiszár [8].

To relate Theorem 3.3 to another testing theoretic characterizations of sufficiency let $g_\alpha(P,Q)$ denote the second error probability for the best level α test for testing $H_0 : P$ versus $H_A : Q$. Then by Torgersen [40], pp. 590–591:

$$\begin{aligned}b_\pi(P,Q) &= \min_{0<\alpha<1}[\pi\alpha + (1-\pi)g_\alpha(P,Q)], \quad \pi\in(0,1),\\ g_\alpha(P,Q) &= \max_{0<\pi<1} \tfrac{1}{1-\pi}[b_\pi(P,Q) - \pi\alpha], \quad \alpha\in(0,1).\end{aligned} \tag{3.8}$$

Theorem 3.4. Suppose $T : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping. Then the following conditions are equivalent

$$\begin{aligned} &A) \quad T \text{ is sufficient for the model } (\mathcal{X}, \mathfrak{A}, \{P, Q\}), \\ &B) \quad b_\pi(P \circ T^{-1}, Q \circ T^{-1}) = b_\pi(P, Q), \quad \pi \in (0, 1), \\ &C) \quad g_\alpha(P \circ T^{-1}, K \circ T^{-1}) = g_\alpha(P, Q), \quad \alpha \in (0, 1). \end{aligned}$$

P r o o f . Using (2.26) we see from Theorem 3.3 that the condition B) is equivalent with condition A). The equivalence of B) and C) follows from (3.8). □

The equivalence of A) and B) is due to Torgersen [40] whereas the equivalence of A) and C) is due to Pfanzagl [30].

Now we use the integral representation of the ϕ-divergence to give a simplified characterization of sufficiency in terms of the variational distance of the measures P and aQ. The following statement is due to Mussmann [26].

Theorem 3.5. Let Δ be dense in $(0, \infty)$. T is sufficient for the model $(\mathcal{X}, \mathfrak{A}, \{P, Q\})$ if and only if

$$\left\| P \circ T^{-1} - aQ \circ T^{-1} \right\| = \|P - aQ\|, \quad a \in \Delta. \tag{3.9}$$

P r o o f . Put $\phi_a(t) = |t - a|$. Then

$$D_{\phi_a}(P, Q) = \|P - aQ\| \quad \text{and} \quad D_{\phi_a}(KP, KQ) = \|KP - aKQ\|,$$

and the necessity of (3.9) follows from condition D) in Theorem 3.3. To establish the converse statement we fix a countable dense subset $\Delta_0 \subseteq \Delta$ and $\beta(a) > 0, a \in \Delta_0$ with $\sum_{a \in \Delta_0} \beta(a) < \infty$ and put $\phi(t) = \sum_{a \in \Delta_0} \beta(a)|t - a|$. The function ϕ is convex and has the curvature measure

$$\gamma_\phi = \sum\nolimits_{a \in \Delta_0} \beta(a) \gamma_{\phi_a} = \sum\nolimits_{a \in \Delta_0} \beta(a) 2\delta_a.$$

As Δ_0 is dense we have $\gamma_\phi((s, t)) > 0$ for every $0 < s < t < \infty$ so that ϕ is strictly convex. If (3.9) is fulfilled then

$$\begin{aligned} &D_\phi(P, Q) - D_\phi(P \circ T^{-1}, Q \circ T^{-1}) \\ &= \sum\nolimits_{a \in \Delta_0} \beta(a)(D_{\phi_a}(P, Q) - D_{\phi_a}(P \circ T^{-1}, K \circ T^{-1})) \\ &= \sum\nolimits_{a \in \Delta_0} \beta(a)(\|P - aQ\| - \left\| P \circ T^{-1} - aQ \circ T^{-1} \right\|) = 0, \end{aligned}$$

and the statement follows from the strict convexity of ϕ and Theorem 3.3. □

Now we deal with sufficiency in Bayes models. Given the model $(\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta \in \Delta})$ we now suppose that Δ is equipped with a σ-algebra $\mathfrak{B}_\Delta$ that contains all one point sets

$\{\theta\}$, $\theta \in \Delta$. Furthermore, we assume that $\theta \mapsto P_\theta(A)$ is measurable for every $A \in \mathfrak{A}$. For a probability measure Π on $(\Delta, \mathfrak{B}_\Delta)$, called prior, we set

$$(P \otimes \Pi)(C) = \int \left(\int I_C(x,\theta) P_\theta(\mathrm{d}x) \right) \Pi(\mathrm{d}\theta), \quad C \in \mathfrak{A} \otimes \mathfrak{B}_\Delta,$$
$$(P\Pi)(A) = \int P_\theta(A) \Pi(\mathrm{d}\theta), \quad A \in \mathfrak{A}.$$

To have a canonical probability space, on which the random vector (X, Θ) with distribution $P \otimes \Pi$ is defined, we use the probability space

$$(\mathcal{X} \times \Delta, \mathfrak{A} \otimes \mathfrak{B}_\Delta, \mathbb{P}), \quad \mathbb{P} = P \otimes \Pi, \tag{3.10}$$

and denote by X and Θ the projections of $\mathcal{X} \times \Delta$ on $\mathcal{X}$ and Δ, respectively. Then

$$\mathcal{L}(X, \Theta) = P \otimes \Pi, \quad \mathcal{L}(X) = P\Pi, \quad \mathcal{L}(\Theta) = \Pi.$$

In Bayesian statistics X is observable and we want to make inference on Θ. To study the dependence between the random variables X and Θ we compare the joint distribution $P \otimes \Pi$ with the product $(P\Pi) \otimes \Pi$ of the marginal distributions. It is clear that the smaller this distance is, the weaker is the dependence between X and Θ. To specify the distance between the distributions $P \otimes \Pi$ and $(P\Pi) \otimes \Pi$ we use the divergences introduced in Definition 2.2. Set

$$I_\phi(X, \Theta) := D_\phi(P \otimes \Pi, (P\Pi) \otimes \Pi), \tag{3.11}$$

and call $I_\phi(X, \Theta)$ the *mutual ϕ-information of X and Θ*. If $\phi(x) = x \ln x$ then $I_\phi(X, \Theta)$ becomes the classical mutual information $I(X, \Theta)$ of information theory, see Shannon [36] and Cover and Thomas [10].

If the family $(P_\theta)_{\theta \in \Delta}$ is dominated by the σ-finite measure μ and the density $f_\theta(x) := \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x)$ is measurable in (x, θ), which will be assumed in the sequel, then

$$\frac{\mathrm{d}(P \otimes \Pi)}{\mathrm{d}(\mu \otimes \Pi)}(x, \theta) = f_\theta(x), \quad \mu \otimes \Pi\text{-a.e.},$$
$$m(x) := \frac{\mathrm{d}(P\Pi)}{\mathrm{d}\mu}(x) = \int f_\theta(x) \Pi(\mathrm{d}\theta), \quad \mu\text{-a.e.} \tag{3.12}$$

Furthermore,

$$\pi(\theta|x) = \begin{cases} \frac{f_\theta(x)}{m(x)} & \text{if} \quad m(x) > 0 \\ 1 & \text{if} \quad m(x) = 0 \end{cases}, \tag{3.13}$$
$$\Pi(B|x) = \int_B \pi(\theta|x) \Pi(\mathrm{d}\theta), \quad B \in \mathfrak{B}_\Delta$$

are the posterior density and the posterior distribution, respectively. For the special case of a binary prior $\Pi = \frac{1}{2}(\delta_{\theta_1} + \delta_{\theta_2})$ the posterior distribution is concentrated on $\{\theta_1, \theta_2\}$ and it holds

$$\Pi(\{\theta_1\}|x) = \frac{f_{\theta_1}(x)}{f_{\theta_1}(x) + f_{\theta_2}(x)}, \quad \Pi(\{\theta_2\}|x) = \frac{f_{\theta_2}(x)}{f_{\theta_1}(x) + f_{\theta_2}(x)}. \tag{3.14}$$

Using the notations in (3.12) and (3.13) we may write $I_\phi(X,\Theta)$ in the following form

$$I_\phi(X,\Theta) = \int \left(\int \phi(\pi(\theta|x)) m(x) \mu(\mathrm{d}x) \right) \Pi(\mathrm{d}\theta)$$
$$= \int D_\phi(P_\theta, P\Pi)\Pi(\mathrm{d}\theta). \tag{3.15}$$

If we again specialize the prior to be $\Pi = \frac{1}{2}(\delta_{\theta_1} + \delta_{\theta_2})$ we get

$$I_\phi(X,\Theta) = \frac{1}{2} D_\phi \left(P_{\theta_1}, \frac{1}{2}(P_{\theta_1} + P_{\theta_2}) \right) + \frac{1}{2} D_\phi \left(P_{\theta_2}, \frac{1}{2}(P_{\theta_1} + P_{\theta_2}) \right)$$
$$= D_{\widehat{\phi}}(P_{\theta_1}, P_{\theta_2}), \tag{3.16}$$

where $\widehat{\phi}$ in (2.23) is strictly convex for a strictly convex ϕ.

Definition 3.6. Given the model $(\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta})$ and a family $\mathcal{P}$ of priors on $(\Delta, \mathfrak{B}_\Delta)$ the statistic $T : \mathcal{X} \to \mathcal{Y}$ is called Bayes sufficient for $\mathcal{P}$ if for every $\Pi \in \mathcal{P}$ and every $B \in \mathfrak{B}_\Delta$ there exists a measurable function $k_B : \mathcal{Y} \to [0,1]$ such that

$$\mathbb{P}(\Theta \in B|X) = k_B(T(X)), \quad \mathbb{P}\text{-a. s.} \tag{3.17}$$

where X and Θ are the projections defined on the probability space (3.10).

By $\mathcal{P}_0$ we denote the family of binary priors

$$\mathcal{P}_0 = \left\{ \Pi : \Pi = \frac{1}{2}(\delta_{\theta_1} + \delta_{\theta_2}), \theta_1, \theta_2 \in \Delta \right\}.$$

Theorem 3.7. If the family $(P_\theta)_{\theta\in\Delta}$ is dominated and $\mathcal{P}_0 \subseteq \mathcal{P}$ then the following conditions are equivalent:

- *A)* T is sufficient for the model $(\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta})$,
- *B)* T is Bayes sufficient for $\mathcal{P}$,
- *C)* (T,Θ) is sufficient for the model $(\mathcal{X}\times\Delta, \mathfrak{A}\otimes\mathfrak{B}_\Delta, \{P\otimes\Pi, (P\Pi)\otimes\Pi)\})$ for every $\Pi \in \mathcal{P}$,
- *D)* $I_\phi(T(X),\Theta) = I_\phi(X,\Theta) < \infty$ for a strictly convex function ϕ and every $\Pi \in \mathcal{P}$,
- *E)* $I_\phi(T(X),\Theta) = I_\phi(X,\Theta)$ for every convex function ϕ and every $\Pi \in \mathcal{P}$.

Proof.

The proof is carried out according to the following scheme

$$\begin{array}{ccccccc} A) & \longleftrightarrow & B) & & & & \\ \downarrow & & & & & & \\ C) & \to & E) & \to & D) & \to & A) \end{array}$$

$A) \to B)$: If T is sufficient then by (3.6) and (3.13)

$$\pi(\theta|x) = \left[\int g_\theta(T(x))\Pi(\mathrm{d}\theta)\right]^{-1} g_\theta(T(x))$$

is a functions of T. Then the condition (3.17) is satisfied with

$$k_B(T(X)) = \left[\int g_\theta(T(X))\Pi(\mathrm{d}\theta)\right]^{-1} \int_B g_\theta(T(X))\Pi(\mathrm{d}\theta),$$

and T is Bayes sufficient.
$B) \to A)$: Put $\Pi = \frac{1}{2}(\delta_{\theta_1} + \delta_{\theta_2})$. Then the posterior probabilities $\Pi(\{\theta_i\}|x), i = 1, 2$ in (3.14) are functions of T. Hence in view of Conclusion 3.2 the statistic T is sufficient for the binary model $\{P_{\theta_1}, P_{\theta_2}\}$. As for dominated models the pairwise sufficiency implies the sufficiency we get $A)$.
$A) \to C)$: If $A)$ is satisfied then by (3.6)

$$\frac{\mathrm{d}(P \otimes \Pi)}{\mathrm{d}((P\Pi) \otimes \Pi)}(x, \theta) = \frac{g_\theta(T(x))}{\int g_\theta(T(x))\Pi(\mathrm{d}\theta)}, \quad (P\Pi) \otimes \Pi\text{-a.s.},$$

so that the left hand term and is a measurable function of (T, Θ) and $C)$ holds in view of Neyman's factorization criterion.
$C) \to E)$: If $C)$ is fulfilled then by condition $D)$ in Theorem 3.3 we get

$$\begin{aligned} &D_\phi(P \otimes \Pi, (P\Pi) \otimes \Pi) \\ &= D_\phi((P \otimes \Pi) \circ (T, \Theta)^{-1}, ((P\Pi) \otimes \Pi) \circ (T, \Theta)^{-1}). \end{aligned}$$

If $C \in \mathfrak{B} \otimes \mathfrak{B}_\Delta$ then with $Q_\theta = P_\theta \circ T^{-1}$

$$\begin{aligned} ((P \otimes \Pi) \circ (T, \Theta)^{-1})(C) &= \int \left(\int I_C(T(x), \theta\right) P_\theta(\mathrm{d}x))\Pi(d\theta) \\ &= \int \left(\int I_C(y, \theta) Q_\theta(\mathrm{d}x)\right) \Pi(\mathrm{d}\theta) = (Q \otimes \Pi)(C), \end{aligned}$$

and similarly

$$(Q\Pi) \otimes \Pi) \circ (T, \Theta)^{-1} = (Q\Pi) \otimes \Pi).$$

The statement E) follows from (3.11).
$E) \to D)$ is obvious.
$D) \to A)$: If $\Pi = \frac{1}{2}(\delta_{\theta_1} + \delta_{\theta_2})$ then by (3.16)

$$I_\phi(X, \Theta) = D_{\widehat{\phi}}(P_{\theta_1}, P_{\theta_2}),$$

and similarly

$$I_\phi(T(X), \Theta) = D_{\widehat{\phi}}(P_{\theta_1} \circ T^{-1}, P_{\theta_2} \circ T^{-1}).$$

If D) is satisfied then we get

$$D_{\widehat{\phi}}(P_{\theta_1}, P_{\theta_2}) = D_{\widehat{\phi}}(P_{\theta_1} \circ T^{-1}, P_{\theta_2} \circ T^{-1}).$$

If ϕ is strictly convex then $\widehat{\phi}$ is strictly convex as well, see (2.24). Using condition C) in Theorem 3.3 we may conclude that T is sufficient for the model $(\mathcal{X}, \mathfrak{A}, \{P_{\theta_1}, P_{\theta_2}\})$. Since θ_1, θ_2 are arbitrary the statistic T is pairwise sufficient. As the family $(P_\theta)_{\theta\in\Delta}$ is dominated the statement A) follows. □

Remark 3.8. The equivalence of the conditions A) and B) in Theorem 3.7 is known, see e. g. Schervisch [35] p. 86. It should be noted that for A) → B) the assumption of dominance can be removed. The conditions C), D), and E) seem to be new. The condition D) corresponds to the information theoretic characterization of sufficiency in Theorem 3.3.

4. ϕ-DIVERGENCES AND DEFICIENCY OF MODELS

Let $(\mathcal{X}, \mathfrak{A}, \{P_1, P_2\})$ be a binary model and $(\mathcal{Y}, \mathfrak{B})$ be another measurable space. If the statistic $T : \mathcal{X} \to \mathcal{Y}$ is sufficient and $(\mathcal{X}, \mathfrak{A})$ is a standard Borel space then k_A in (3.5) can be chosen as a regular conditional distribution, i. e. there is a stochastic kernel $M(A|y), A \in \mathfrak{A}, y \in \mathcal{Y}$ that satisfies for $Q_i = P_i \circ T^{-1}$

$$P_i(A \cap \{x : T(x) \in B\}) = \int_B M(A|y) Q_i(\mathrm{d}y), \quad i = 1, 2.$$

For $B = \mathcal{Y}$ this implies $P_i = MQ_i, i = 1, 2$, so that the two models

$$\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_1, P_2\}) \quad \text{and} \quad \mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_1, Q_2\})$$

are mutual randomizations and therefore equivalent from the decision theoretic point of view. This concept of sufficiency is referred to as *Blackwell sufficiency*. In order to express that two models are at least approximately equivalent LeCam introduced the concept of deficiency. Set

$$\delta(\mathcal{M}, \mathcal{N}) = \inf_K \max_{i=1,2} \|Q_i - KP_i\|,$$

where the infimum is taken over all kernels $K(B|x), B \in \mathfrak{B}$, $x \in \mathcal{X}$, and put

$$\Delta(\mathcal{M}, \mathcal{N}) = \max(\delta(\mathcal{M}, \mathcal{N}), \delta(\mathcal{N}, \mathcal{M})).$$

Then $\Delta(\mathcal{M}, \mathcal{N})$ becomes a semimetric in the space of all binary models. It is clear that $\Delta(\mathcal{M}, \mathcal{N}) = 0$ if $\mathcal{M}$ and $\mathcal{N}$ are mutual randomizations. Otherwise, two models $\mathcal{M}$ and $\mathcal{N}$ with Δ-distance zero can be shown to be mutual randomizations by the so called randomization theorem, see Strasser [37].

By the definition of $\Delta(\mathcal{M}, \mathcal{N})$ we find, for every $\varepsilon > 0$, a stochastic kernel K such that

$$\|Q_i - KP_i\| \le \Delta(\mathcal{M}, \mathcal{N}) + \varepsilon.$$

If ψ is a test for $\mathcal{N}$ then for $i = 1, 2$,

$$\left| \int \psi(\mathrm{d}Q_i - \mathrm{d}(KP_i)) \right| \le \Delta(\mathcal{M}, \mathcal{N}) + \varepsilon.$$

As $(K^*\psi)(x) := \int \psi(y)K(\mathrm{d}y|x)$ is a test for $\mathcal{M}$ and

$$\int \psi \,\mathrm{d}(KP_i) = \int (K^*\psi)\,\mathrm{d}P_i,$$

we get

$$\begin{aligned} b_\pi(Q_1, Q_2) &= \inf_\psi \left(\pi \int \psi \,\mathrm{d}Q_1 + (1-\pi)\int (1-\psi)\,\mathrm{d}Q_2\right) \\ &\geq \inf_\varphi (\pi \int \varphi \,\mathrm{d}P_1 + (1-\pi)\int (1-\varphi)\,\mathrm{d}P_2) - \Delta(\mathcal{M},\mathcal{N}) - \varepsilon \\ &= b_\pi(P_1, P_2) - \Delta(\mathcal{M},\mathcal{N}) - \varepsilon. \end{aligned}$$

Take $\varepsilon \to 0$, use (2.26) and interchange the role of $\mathcal{M}$ and $\mathcal{N}$ to get

$$\sup_\pi |B_\pi(P_1, P_2) - B_\pi(Q_1, Q_2)| \leq \Delta(\mathcal{M},\mathcal{N}).$$

But one can even show that

$$\sup_\pi |B_\pi(P_1, P_2) - B_\pi(Q_1, Q_2)| = \Delta(\mathcal{M},\mathcal{N}). \tag{4.1}$$

For a proof of this statement we refer to Torgersen [40]. The functionals B_π are the ϕ-divergences that belong to the class of convex functions

$$\mathcal{F}_0 = \{\psi_\pi : 0 < \pi < 1,\ \psi_\pi \text{ defined in (2.6)}\}, \tag{4.2}$$

see (2.27). To study the relation between $\Delta(\mathcal{M},\mathcal{N})$ and other ϕ-divergences different from B_π we recall to the modified curvature measure γ_ϕ introduced in (2.8). We have already seen in (2.11) that $\gamma_{\psi_\pi}((0,1)) = 1$, $0 < \pi < 1$. Next we calculate the total mass $\gamma_\phi((0,1))$ of the modified curvature measure for another class of convex functions ϕ.

Lemma 4.1. For every convex and non-increasing function $\phi : (0,\infty) \to \mathbb{R}$ with

$$\begin{aligned} \phi(0) = 0, \quad \phi(\infty) = \lim_{t\to\infty} \phi(t) > -\infty, \\ D^+\phi(0) = \lim_{t\downarrow 0} D^+\phi(t) > -\infty, \end{aligned} \tag{4.3}$$

it holds

$$\gamma_\phi((0,1)) = -D^+\phi(0) - \phi(\infty).$$

P r o o f . Take $s \downarrow 0$ in (2.1) and use $\phi(0) = 0$ to get

$$\begin{aligned} \int_{(0,t]} (1+s)\rho_\phi(\mathrm{d}s) &= D^+\phi(t) - D^+\phi(0) + \int_{(0,t]} \left(\int I_{(0,s)}(\tau)\,\mathrm{d}\tau \rho_\phi(\mathrm{d}s)\right) \\ &= D^+\phi(t) - D^+\phi(0) + \int_{(0,t]} (D^+\phi(t) - D^+\phi(\tau))\,\mathrm{d}\tau \\ &= D^+\phi(t) - D^+\phi(0) + tD^+\phi(t) - \phi(t). \end{aligned} \tag{4.4}$$

For $t \to \infty$ the limit of the left hand side in (4.4) exists. Furthermore, $D^+\phi(t) \le 0$ and $D^+\phi$ is nondecreasing. Hence $\lim_{t\to\infty} D^+\phi(t)$ exists and is finite. Finally, $\lim_{t\to\infty} \phi(t)$ is finite by assumption and we may conclude that

$$A := \lim_{t\to\infty} tD^+\phi(t) \le 0$$

exists. If $A < 0$ then

$$\phi(t) - \phi(1) = \int_1^t \frac{1}{\tau}(\tau D^+\phi(\tau))\,\mathrm{d}\tau$$

implies $\lim_{t\to\infty} \phi(t) = -\infty$ which contradicts the assumption. Hence $A = 0$ and consequently $\lim_{\tau\to\infty} D^+\phi(\tau) = 0$. Taking $t \to \infty$ in (4.4) we get the statement. □

The next statement shows that the deficiency of two models is small if and only if all distances of the two distributions, measured by special ϕ-divergences, are uniformly close.

Theorem 4.2. For any two binary models

$$\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_1, P_2\}) \quad \text{and} \quad \mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_1, Q_2\})$$

and any convex function ϕ with $\gamma_\phi((0,1)) < \infty$ it holds

$$|D_\phi(P_1, P_2) - D_\phi(Q_1, Q_2)| \le \Delta(\mathcal{M}, \mathcal{N})\gamma_\phi((0,1)). \tag{4.5}$$

If $\mathcal{F}$ is any class of convex functions ϕ with $\gamma_\phi((0,1)) = 1$ that contains $\mathcal{F}_0$ in (4.2) then

$$\sup_{\phi\in\mathcal{F}} |D_\phi(P_1, P_2) - D_\phi(Q_1, Q_2)| = \Delta(\mathcal{M}, \mathcal{N}). \tag{4.6}$$

Corollary 4.3. Let $\mathcal{F}_1$ be the class of all nondecreasing convex functions ϕ on $(0, \infty)$ that satisfy

$$\begin{aligned} &\phi(0) = \lim_{t\downarrow 0} \phi(t) = 0, \quad \lim_{t\to\infty} \phi(t) > -\infty, \\ &-D^+\phi(0) - \phi(\infty) = 1. \end{aligned} \tag{4.7}$$

Then the statement (4.6) holds with $\mathcal{F}$ replaced with $\mathcal{F}_1$.

P r o o f . The spectral representation (2.28) implies

$$|D_\phi(P_1, P_2) - D_\phi(Q_1, Q_2)| \le \int \sup_{0<\pi<1} |B_\pi(P_1, P_2) - B_\pi(Q_1, Q_2)|\gamma_\phi(\mathrm{d}\pi),$$

so that (4.5) follows from (4.1). If $\mathcal{F}_0 \subseteq \mathcal{F}$ then by (4.1) and

$$D_{\psi_\pi}(P_1, P_2) = B_\pi(P_1, P_2) \quad \text{and} \quad D_{\psi_\pi}(Q_1, Q_2) = B_\pi(Q_1, Q_2), \tag{4.8}$$

see (2.27), we get

$$\begin{aligned}\Delta(\mathcal{M},\mathcal{N}) &= \sup_{\phi\in\mathcal{F}_0} |D_\phi(P_1,P_2) - D_\phi(Q_1,Q_2)| \\ &\le \sup_{\phi\in\mathcal{F}} |D_\phi(P_1,P_2) - D_\phi(Q_1,Q_2)| \le \Delta(\mathcal{M},\mathcal{N}),\end{aligned}$$

where the last inequality follows from (4.5) and $\gamma_\phi((0,1)) = 1$ for $\phi \in \mathcal{F}$. To prove the Corollary we set $\psi_\pi^0(t) = -((\pi t) \wedge (1-\pi))$. Then $\psi_\pi^0(0) = 0$ and

$$-D^+\psi_\pi^0(0) - \psi_\pi^0(\infty) = -(-\pi) - (-(1-\pi)) = 1,$$

so that $\psi_\pi^0 \in \mathcal{F}_1$. As ψ_π^0 and ψ_π differ only by a linear function we get from (4.8) and (2.15) that

$$B_\pi(P_1,P_2) - B_\pi(Q_1,Q_2) = D_{\psi_\pi^0}(P_1,P_2) - D_{\psi_\pi^0}(Q_1,Q_2).$$

Hence by (4.1)

$$\begin{aligned}\Delta(\mathcal{M},\mathcal{N}) &= \sup_{0<\pi<1} |D_{\psi_\pi^0}(P_1,P_2) - D_{\psi_\pi^0}(Q_1,Q_2)| \\ &\le \sup_{\phi\in\mathcal{F}_1} |D_\phi(P_1,P_2) - D_\phi(Q_1,Q_2)| \le \Delta(\mathcal{M},\mathcal{N}),\end{aligned}$$

where the last inequality follows from (4.5) and the fact that $\gamma_\phi((0,1)) = -D^+\phi(0) - \phi(\infty) = 1$ by Lemma 4.1. □

Remark 4.4. In a somewhat different formulation that uses concave instead of convex functions the statement (4.6) is the concave function criterion in decision theory, see e. g. Strasser [37]. The above theorem extends this result and clarifies the role of the condition (4.7), which appeared there as a purely technical condition. Now we see that this condition is a normalizing condition for the modified curvature measure of the convex function ϕ.

5. ACKNOWLEDGEMENT

The author sincerely thanks the Associated Editor and the referees for their careful reading of the paper and the constructive suggestions that led to a substantial improvement of the manuscript.

(Received January 21, 2011)

REFERENCES

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

[2] S. Arimoto: Information-theoretical considerations on estimation problems. Inform. Control. *19* (1971), 181–194.

[3] A. R. Barron, L. Györfi, and E. C. van der Meulen: Distribution estimates consistent in total variation and two types of information divergence. IEEE Trans. Inform. Theory *38* (1990), 1437–1454.

[4] A. Berlinet, I. Vajda, and E. C. van der Meulen: About the asymptotic accuracy of Barron density estimates. IEEE Trans. Inform. Theory *44* (1990), 999–1009.

[5] A. Bhattacharyya: On some analogues to the amount of information and their uses in statistical estimation. Sankhya *8* (1946), 1–14.

[6] H. Chernoff: A measure of asymptotic efficiency for test of a hypothesis based on the sum of observations. Ann. Math. Statist. *23* (1952), 493–507.

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

[8] I. Csiszár: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffscher Ketten. Publ. Math. Inst. Hungar. Acad. Sci.*8* (1963), 84–108.

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

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

[11] M. H. De Groot: Optimal Statistical Decisions. McGraw Hill, New York 1970.

[12] D. Feldman and F. Österreicher: A note on f-divergences. Studia Sci. Math. Hungar. *24* (1989), 191–200.

[13] A. Guntuboyina: Lower bounds for the minimax risk using f-divergences, and applications. IEEE Trans. Inform. Theory *57* (2011), 2386–2399.

[14] C. Guttenbrunner: On applications of the representation of f-divergences as averaged minimal Bayesian risk. In: Trans. 11th Prague Conf. Inform. Theory, Statist. Dec. Funct., Random Processes *A*, 1992, pp. 449–456.

[15] L. Jager and J. A. Wellner: Goodness-of-fit tests via phi-divergences. Ann. Statist. *35* (2007), 2018–2053.

[16] T. Kailath: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. Commun. Technol. *15* (1990), 52-60.

[17] S. Kakutani: On equivalence of infinite product measures. Ann. Math. *49* (1948), 214–224.

[18] S. Kullback and R. Leibler: On information and sufficiency. Ann. Math. Statist. *22* (1951), 79–86.

[19] L. LeCam: Locally asymptotically normal families of distributions. Univ. Calif. Publ. *3*, (1960), 37–98.

[20] L. LeCam: Asymptotic Methods in Statistical Decision Theory. Springer, Berlin 1986.

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

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

[23] F. Liese and I. Vajda: f-divergences: Sufficiency, deficiency and testing of gypotheses. In: Advances in Inequalities from Probability Theory and Statistics. (N. S. Barnett and S. S. Dragomir, eds.), Nova Science Publisher, Inc., New York 2008, pp. 113–149.

[24] F. Liese and K. J. Miescke: Statistical Decision Theory, Estimation, Testing and Selection. Springer, New York 2008.

[25] K. Matusita: Decision rules based on the distance, for problems of fit, two samples and estimation. Ann. Math. Statist. *26* (1955), 613–640.

[26] D. Mussmann: Decision rules based on the distance, for problems of fit, two samples and estimation. Studia Sci. Math. Hungar. *14* (1979), 37–41.

[27] X. Nguyen, M. J. Wainwright, and M. I. Jordan: On surrogate loss functions and f-divergences. Ann. Statist. *37* (2009), 2018–2053.

[28] F. Österreicher and D. Feldman: Divergenzen von Wahrscheinlichkeitsverteilungen – integralgeometrisch betrachtet. Acta Math. Sci. Hungar. *37* (1981), 329–337.

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

[30] J. Pfanzagl: A characterization of sufficiency by power functions. Metrika *21* (1974), 197–199.

[31] H. V. Poor: Robust decision design using a distance criterion. IEEE Trans. Inform. Theory *26* (1980), 578-587.

[32] M. R. C. Read and N. A. C. Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, Berlin 1988.

[33] A. Rényi: On measures of entropy and information. In: Proc. 4th Berkeley Symp. on Probab. Theory and Math. Statist. Berkeley Univ. Press, Berkeley 1961, pp. 547–561.

[34] A. W. Roberts and D. E. Varberg: Convex Functions. Academic Press, New York 1973.

[35] M. J. Schervish: Theory of Statistics. Springer, New York 1995.

[36] C. E. Shannon: A mathematical theory of communication. Bell. Syst. Tech. J. *27* (1948), 379–423, 623–656.

[37] H. Strasser: Mathematical Theory of Statistics. De Gruyter, Berlin 1985.

[38] F. Topsøe: Information-theoretical optimization techniques. Kybernetika *15* (1979), 7–17.

[39] F. Topsøe: Some inequalities for information divergence and related measures of discrimination. IEEE Trans. Inform. Theory *46* (2000), 1602–1609.

[40] E. Torgersen: Comparison of Statistical Experiments. Cambridge Univ. Press, Cambridge 1991.

[41] I. Vajda: On the f-divergence and singularity of probability measures. Periodica Math. Hungar. *2* (1972), 223–234.

[42] I. Vajda: $\chi^\alpha-$divergence and generalized Fisher's information. In: Trans. 6th Prague Conf. Inform. Theory, Statist. Dec. Funct., Random Processes, Academia, Prague 1973, pp. 873–886.

[43] I. Vajda: Theory of Statistical Inference and Information. Kluwer Academic Publishers, Dordrecht – Boston – London 1989.

[44] I. Vajda: On metric divergences of probability measures. Kybernetika *45* (2009), 885–900.

[45] I. Vajda: On convergence of information contained in quantized observations. IEEE Trans. Inform. Theory. *48* (1980) 2163–2172.

[46] I. Vincze: On the concept and measure of information contained in an observation. In: Contribution to Probability. (J. Gani and V. F. Rohatgi, eds.) Academic Press, New York 1981, pp. 207–214.

Friedrich Liese, Department of Mathematics, University of Rostock, Ulmenstrasse 69 HS 3, D-18057 Rostock. Germany.

e-mail: friedrich.liese@uni-rostock.de