f-DIVERGENCES: SUFFICIENCY, DEFICIENCY AND TESTING OF HYPOTHESES

Friedrich Liese and Igor Vajda

ABSTRACT. This paper deals with f -divergences of probability measures considered in the same general form as e.g. in [12] or [45], where f is an arbitrary (not necessarily differentiable) convex function. Important particular cases or subclasses are mentioned, including those introduced by Bhattacharyya [3], Kakutani [32], Shannon [61] and Kullback with Leibler [38], Chernoff [7], Kraft [37], Matusita [47], Rényi [57] and De Groot [15]. Some important relations between these subclasses are reproduced or reestablished in a new manner. The main result is a new proof of the representation of general f -divergence $I_{\mathsf{f}}(P_0, P_1)$ by means of the information gains $G_\pi(P_0, P_1), 0 \leq \pi \leq 1$ of De Groot. This proof uses the generalized Taylor formula applied to arbitrary convex functions derived in this paper. The basic known properties of general f -divergences are deduced in a new manner from this representation, among them the convergence for increasing sequences of σ -algebras of observation events. Further, statistical sufficiency and ε -deficiency (approximate sufficiency) are considered in the model of testing the hypothesis $H_0 : P_0$ against the alternative $H_1 : P_1$. New characterizations of these properties by means of f -divergences are given. Finally, the above mentioned results about convergence for increasing sequences of σ -algebras are assessed by evaluating the rate of this convergence in some important cases.

1 INTRODUCTION

Shannon [61] introduced the information between two random variables by comparing the joint distribution $P_{(X,Y)}$ with the product distribution $P_X \otimes P_Y$ of the marginal distributions P_X and P_Y with the help of the divergence

$$\mathsf{K}_1(P_{(X,Y)}, P_X \otimes P_Y) = \int \ln \left(\frac{dP_{(X,Y)}}{d(P_X \otimes P_Y)} \right) dP_{(X,Y)}.$$

The divergence

$$\mathsf{K}_1(P_0, P_1) = \begin{cases} \int \ln \left(\frac{dP_0}{dP_1} \right) dP_0 & \text{if} \quad P_0 \ll P_1 \\ \infty & \text{otherwise} \end{cases},$$

1991 Mathematics Subject Classification. 62B05; 62B10, 62B15, 62G10.

Key words and phrases. divergences, information gain, sufficiency, testing hypotheses, error probabilities, exponential rate.

This paper was prepared with the support of the ASCR grant A1075104 and the MSMT grant 1M0572.

of arbitrary distributions P_0, P_1 was systematically studied by Kullback and Leibler [38], Gelfand et al. [22] and others who recognized its importance in information theory, statistics and probability theory. Rényi [57] introduced a class of measures of divergences of distributions P_0, P_1 with properties similar to $\mathsf{K}_1(P_0, P_1)$ and containing $\mathsf{K}_1(P_0, P_1)$ as a special case. Csiszár [11] (and independently also Ali and Silvey [1]) introduced the f-divergence

$$\mathsf{I}_\mathsf{f}(P_0, P_1) = \int \frac{dP_1}{d\mu} \mathsf{f}\left(\frac{dP_0/d\mu}{dP_1/d\mu}\right) d\mu,$$

for convex $\mathsf{f} : (0, \infty) \to \mathbb{R}$ where μ is a σ-finite measure which dominates P_0 and P_1 and the integrand is appropriately specified at the points where the densities $dP_0/d\mu$ and $dP_1/d\mu$ are zero.

For $\mathsf{f}(t) = t \ln t$ the f-divergence reduces to the classical $\mathsf{K}_1(P_0, P_1)$ which is sometimes denoted by $\mathsf{I}(P_0, P_1)$ and called information divergence or Kullback-Leibler divergence. For the convex or concave function $\mathsf{f}(t) = t^s$, $s > 0$, we obtain the so-called Hellinger integrals $\mathsf{H}_s(P_0, P_1)$ which are related to the divergences $\mathsf{R}_s(P_0, P_1)$ of Rényi [57] by $\mathsf{R}_s(P_0, P_1) = (s-1)^{-1} \ln \mathsf{H}_s(P_0, P_1)$. Note that the divergence measures $\ln \mathsf{H}_s(P_0, P_1)$ were considered for $0 < s < 1$ already by Chernoff [7] and Kraft [37], and the special case for $s = 1/2$ also by Bhattattacharyya [3], Kakutani [32] and Matusita [47]. As pointed out in LeCam [42, p. 29], the expression $\mathsf{H}_s(P_0, P_1)$ does not seem to have been considered by Hellinger. He considered integrals of the type $\int \frac{dP_1 dP_2}{d\mu}$ for P_0 and P_1 dominated by μ. Similar developments were given by Riesz [60] and by Dieudonné [18] for general "homogenous functions of measures".

Among the f-divergences one can find also the basic divergence measures of probability theory and statistics, such as the total variation $\|P_0 - P_1\|$ for $\mathsf{f}(t) = |t-1|$ and the Pearson divergence $\chi^2(P_0, P_1)$ for $\mathsf{f}(t) = (t-1)^2$.

Statistical applications of f-divergences were considered e.g. by Ali and Silvey [1], Csiszár [11, 12], Nemetz [49], Arimoto [2], Vajda [68, 70] and many others. Decision-theoretic applications can be found e.g. in [35], [50], [56], [41], [9], [66], [65] and Fedotov et al. [17]. Information-theoretic applications of f-divergences were studied e.g. in [31], [64], [6], [13], [4], [29], [25], [10] and [14].

Due to the growing importance of divergences in information theory, statistics and probability theory, each possibility to simplify or extend the general theory of f-divergences deserves attention. The first half of this paper is devoted to a considerably simplified derivation of the most important basic properties of f-divergences. The classical approach to these properties is based on the stability of the Jensen inequalities for the expectation and the conditional expectation. These inequalities are quite convoluted if they are rigorously established for all desirable functions f, cf. [12] and [45]. The approach of this paper is based on an extension of the classical Taylor formula to all convex or concave (not necessarily differentiable) functions f. The second derivative of f is replaced by a σ-finite measure one-one related to $\mathsf{f} - \mathsf{f}(1)$. The support of this measure is given by the areas where f is strictly convex.

In the first section of this paper we collect more or or less known properties of convex functions and establish the new generalized Taylor expansion that will be systematically used in the sequel.

f-divergences are introduced in Section 2 and discussed for special convex functions f in order to demonstrate that this type of functional covers many important divergences from different areas of probability theory, mathematical statistics and information sciences. The main result of this section is an integral representation of f-divergences. In this representation the Bayes' errors corresponding to different priors are weighted according to the curvature measure of the convex function f. This representation allows us to prove the information processing theorem for f-divergences and the continuity of f-divergences (approximability by f-divergences on finite sub-σ-algebras) in a much simpler way than was achieved in the previous literature.

The integral representation of f-divergences allows a simple discussion of the stability in the information processing theorem. The curvature measure of the convex function f plays a crucial role in this respect. Our approach allows a unification of different characterizations of the sufficiency of a statistic. These characterizations are the classical factorization theorem of Neyman, the information-theoretic characterization of Csiszár [12], the characterization by the testing problem which is due to Pfanzagl [55] and the characterization with the help of the variational distance which was established by Mussmann [48] and Torgersen [66]. When dealing with the sufficiency of a statistic one compares the original model with the model reduced by the statistic. More generally, in decision theory the statistical models are related by means of a less known tool called $\varepsilon-$deficiency and the related so-called concave function criterion. Using the integral representation we show that this criterion is equivalent to a comparison of certain f-divergences. This equivalence characterizes the meaning of f-divergences in statistical decision theory by connecting the dissimilarity of statistical models specified by f-divergences with basic decision-theoretic concepts.

In the last section we apply Hellinger integrals to establish the exponential rate of convergence to zero for testing a simple null hypothesis versus a simple alternative if the sample size tends to infinity. This leads to the results known in the literature as theorems of Chernoff and Stein. It also allows us to find the exponential rates for a classification problem and obtain in this manner the results of Krafft and Puri [36] in this area.

2 Convex Functions

We introduce and study classes of distances in the space of probability distributions which originated from different roots. Some of them were introduced in information theory to describe the amount of information or the amount of uncertainty in a random sample. Others are information functionals obtained by investigating the rate of convergence of error probabilities when the sample size tends to infinity. Still others resulted from the Cramér-Rao inequality and its generalizations. Hellinger integrals are the Laplace transforms of loglikelihood ratios and thus completely describe the structure of binary statistical models. Information functionals were also used to characterize the sufficiency and the approximate sufficiency of a statistic.

First we summarize well-known properties of convex functions which will be used in the sequel. A function $\mathsf{f} : (0, \infty) \to \mathbb{R}$ is called *convex* if for every $x, z \in (0, \infty)$ and

$0 \le \alpha \le 1$ we have,

$$\mathsf{f}(\alpha x + (1-\alpha)z) \le \alpha \mathsf{f}(x) + (1-\alpha)\mathsf{f}(z). \tag{2.1}$$

If $\mathsf{f} : (0,\infty) \to \mathbb{R}$ is convex and $0 < x < y < z$ then, for $\alpha = (z-y)/(z-x)$,

$$\frac{\mathsf{f}(y)-\mathsf{f}(x)}{y-x} \le \frac{\mathsf{f}(z)-\mathsf{f}(x)}{z-x} \le \frac{\mathsf{f}(z)-\mathsf{f}(y)}{z-y}. \tag{2.2}$$

The properties of convex functions stated in the next lemma are known. For the sake of completeness we give their short direct proofs.

Lemma 2.1. *Every convex function $\mathsf{f} : (0,\infty) \to \mathbb{R}$ is continuous in $(0,\infty)$ and has at each $x \in (0,\infty)$ a derivative from the left $D^- f(x)$ which is left continuous and a derivative from the right $D^+\mathsf{f}(x)$ that is right continuous. These derivatives satisfy for all $0 < a < b < \infty$, the relations,*

$$\mathsf{f}(b) - \mathsf{f}(a) \ge (b-a) D^+\mathsf{f}(a), \tag{2.3}$$

$$D^-\mathsf{f}(a) \le D^+\mathsf{f}(a) \le D^-\mathsf{f}(b) \le D^+\mathsf{f}(b) \tag{2.4}$$

$$\mathsf{f}(b) - \mathsf{f}(a) = \int_a^b D^+\mathsf{f}(s)ds = \int_a^b D^-\mathsf{f}(s)ds. \tag{2.5}$$

Proof. The continuity is implied by the existence of the one sided derivatives which we deduce from (2.2). Namely, the second inequality in (2.2) implies that the difference quotient is nondecreasing in the increment so that the derivative from the right exists and satisfies (2.3). The proof of the existence of the left hand derivative and of (2.4) is similar. The monotonicity follows from (2.2). To prove the right continuity, add $\varepsilon_n \downarrow 0$ to a and b in (2.3) and get, for $a < b$, the relation $(b-a)^{-1}[\mathsf{f}(b+\varepsilon_n) - \mathsf{f}(a+\varepsilon_n)] \ge D^+\mathsf{f}(b+\varepsilon_n)$. If first $n \to \infty$ and then $b \downarrow a$ we get $D^+\mathsf{f}(b) \ge \lim_{n\to\infty} D^+\mathsf{f}(b+\varepsilon_n)$. Since $D^+\mathsf{f}$ is nondecreasing, this gives the right continuity. The left continuity of $D^-\mathsf{f}$ is proved similarly. Finally, the inequality (2.2) yields

$$D^+\mathsf{f}(u) \le \frac{\mathsf{f}(v)-\mathsf{f}(u)}{v-u} \le D^+\mathsf{f}(v), \quad 0 < u < v. \tag{2.6}$$

For $h_n = (b-a)/n$ we get from (2.6)

$$\int_a^b D^+\mathsf{f}(s)ds = \sum_{i=1}^n \int_{a+(i-1)h_n}^{a+ih_n} D^+\mathsf{f}(s)ds \le \sum_{i=1}^n h_n D^+\mathsf{f}(a+ih_n)$$

$$\sum_{i=1}^n [\mathsf{f}(a+(i+1)h_n) - \mathsf{f}(a+ih_n)] \le \mathsf{f}(b+h_n) - \mathsf{f}(a+h_n),$$

and analogously $\int_a^b D^+\mathsf{f}(s)ds \ge \mathsf{f}(b-h_n) - \mathsf{f}(a-h_n)$. The continuity of f completes the proof. ∎

The statement (2.5) implies that the limits $\lim_{x\downarrow 0}\mathsf{f}(x)$ and $\lim_{x\uparrow\infty}\mathsf{f}(x)$ exist. We extend f by setting $\mathsf{f}(0)=\lim_{x\downarrow 0}\mathsf{f}(x)$ and $\mathsf{f}(\infty)=\lim_{x\uparrow\infty}\mathsf{f}(x)$, where $\mathsf{f}(0)$ may attain the value ∞ and $\mathsf{f}(\infty)$ may attain the values $-\infty$ or ∞.

As $D^+\mathsf{f}$ is continuous from the right there is a uniquely determined σ-finite measure γ_{f} on the Borel sets of $(0,\infty)$ that satisfies, for every $0<a<b$,

$$\gamma_{\mathsf{f}}((a,b])=D^+\mathsf{f}(b)-D^+\mathsf{f}(a). \tag{2.7}$$

If f is twice continuously differentiable then $D^+\mathsf{f}=\mathsf{f}'$ and this function is continuously differentiable so that,

$$D^+\mathsf{f}(b)-D^+\mathsf{f}(a)=\int_a^b \mathsf{f}''(t)dt,\quad \text{and}\quad \gamma_{\mathsf{f}}(B)=\int_B \mathsf{f}''(t)dt.$$

Therefore γ_{f} can be viewed as a measure of the curvature of f. We use this curvature measure to establish a generalized second order Taylor expansion.

Lemma 2.2. *If $\mathsf{f}:(0,\infty)\to\mathbb{R}$ is convex then, for $a,b>0$,*

$$\mathsf{f}(b)-\mathsf{f}(a)-D^+\mathsf{f}(a)(b-a)=\begin{cases}\int (b-t)I_{(a,b]}(t)\gamma_{\mathsf{f}}(dt) & \text{if}\quad a<b\\ \int (t-b)I_{(b,a]}(t)\gamma_{\mathsf{f}}(dt) & \text{if}\quad b<a.\end{cases} \tag{2.8}$$

Moreover, the function,

$$\mathsf{f}_0(x)=\mathsf{f}(x)-\mathsf{f}(1)-(x-1)D^+\mathsf{f}(1) \tag{2.9}$$

has the representation,

$$\mathsf{f}_0(x)=\begin{cases}\int (x-t\wedge x)I_{(1,\infty)}(t)\gamma_{\mathsf{f}}(dt) & \text{if}\quad x>1\\ \int (t-t\wedge x)I_{(0,1]}(t)\gamma_{\mathsf{f}}(dt) & \text{if}\quad 0<x\le 1.\end{cases} \tag{2.10}$$

Proof. For $a<b$ we have, from (2.5) and the theorem of Fubini,

$$\begin{aligned}\mathsf{f}(b)-\mathsf{f}(a)-D^+\mathsf{f}(a)(b-a)&=\int_a^b (D^+\mathsf{f}(s)-D^+\mathsf{f}(a))ds\\ &=\int\left(\int I_{(a,b]}(s)I_{(a,s]}(t)\gamma_{\mathsf{f}}(dt)\right)ds\\ &=\int (b-t)I_{(a,b]}(t)\gamma_{\mathsf{f}}(dt).\end{aligned}$$

By interchanging the role of a and b we get, for $a>b$,

$$\begin{aligned}&\mathsf{f}(b)-\mathsf{f}(a)-D^+\mathsf{f}(a)(b-a)\\ &\quad=-(\mathsf{f}(a)-\mathsf{f}(b)-D^+\mathsf{f}(b)(a-b))+(D^+\mathsf{f}(a)-D^+\mathsf{f}(b))(a-b)\\ &\quad=-\int (a-t)I_{(b,a]}(t)\gamma_{\mathsf{f}}(dt)+\int (a-b)I_{(b,a]}(t)\gamma_{\mathsf{f}}(dt)\\ &\quad=\int (t-b)I_{(b,a]}(t)\gamma_{\mathsf{f}}(dt).\end{aligned}$$

The statement (2.10) follows from (2.8) as $\mathsf{f}_0(1) = D^+\mathsf{f}_0(1) = 0$. ∎

A convex function f is said to be *strictly convex at* $x_0 \in (0,\infty)$ if for no $\varepsilon > 0$ the function f is linear in $(x_0-\varepsilon, x_0+\varepsilon)$. It is called *strictly convex on* $(0,\infty)$ if it is strictly convex at every $x_0 \in (0,\infty)$. The representation (2.8) shows that $\mathsf{f} : (0,\infty) \to \mathbb{R}$ is strictly convex at $x_0 \in (0,\infty)$ if and only if $\gamma_f((x_0-\varepsilon, x_0+\varepsilon)) > 0$ for every $\varepsilon > 0$. If f is twice continuously differentiable and $\mathsf{f}''(x) > 0$, $0 < x < \infty$ then the function f is strictly convex on $(0,\infty)$.

We inspect the function f_0 in (2.9) in more detail. The representation (2.8) shows that $\mathsf{f}_0(x) \geq 0$ and

$$\begin{aligned} &\mathsf{f} \text{ is strictly convex at } 1 \iff (A) \vee (B), \text{ where} \\ &(A) \quad \mathsf{f}_0(x) > 0 \text{ for } 0 < x < 1 \\ &(B) \quad \mathsf{f}_0(x) > 0 \text{ for } 1 < x < \infty. \end{aligned} \tag{2.11}$$

For later purposes we define the *-conjugate function by,

$$\mathsf{f}^*(x) = x\mathsf{f}\left(\frac{1}{x}\right), \qquad x > 0. \tag{2.12}$$

If $\mathsf{f} : (0,\infty) \to \mathbb{R}$ is convex, then $\mathsf{f}^* : (0,\infty) \to \mathbb{R}$ is convex as well. Indeed, for $0 < \alpha < 1$ and $0 < x_1 < x_2$, $x_0 = \alpha x_1 + (1-\alpha)x_2$,

$$\begin{aligned} \mathsf{f}^*(x_0) &= x_0\mathsf{f}\left(\frac{1}{x_0}\right) \\ &= x_0\mathsf{f}\left(\frac{\alpha x_1}{x_0}\frac{1}{x_1} + \frac{(1-\alpha)x_2}{x_0}\frac{1}{x_2}\right) \\ &\leq x_0\frac{\alpha x_1}{x_0}\mathsf{f}\left(\frac{1}{x_1}\right) + x_0\frac{(1-\alpha)x_2}{x_0}\mathsf{f}\left(\frac{1}{x_2}\right) \\ &= \alpha\mathsf{f}^*(x_1) + (1-\alpha)\mathsf{f}^*(x_2). \end{aligned} \tag{2.13}$$

Furthermore, it follows from the definition of f^* that,

$$(\mathsf{f}^*)^* = \mathsf{f}, \quad \mathsf{f}^*(0) = \lim_{x\to\infty}\frac{1}{x}\mathsf{f}(x),$$

where $\mathsf{f}^*(0) \in (-\infty, \infty]$. For later purposes we need also the boundary values $\mathsf{f}(0)$ and $\mathsf{f}^*(0)$ expressed in terms of the curvature measure γ_{f} as follows.

Lemma 2.3. *The following result holds:*

$$\lim_{x\downarrow 0}\mathsf{f}^*(x) = \gamma_{\mathsf{f}}((1,\infty)) + D^+\mathsf{f}(1), \tag{2.14}$$

$$\lim_{x\downarrow 0}\mathsf{f}(x) = \int tI_{(0,1]}(t)\gamma_{\mathsf{f}}(dt) + f(1) - D^+\mathsf{f}(1). \tag{2.15}$$

Proof. It is enough to consider the function f_0 defined in (2.9). The representation (2.10), the monotone convergence theorem and $\gamma_{\mathsf{f}} = \gamma_{\mathsf{f}_0}$ give

$$\lim_{x\to\infty} \frac{1}{x}\mathsf{f}_0(x) = \lim_{x\to\infty} \int \frac{1}{x} I_{(1,x]}(t)(x-t)\gamma_{\mathsf{f}_0}(dt) = \gamma_{\mathsf{f}}((1,\infty)),$$
$$\lim_{x\downarrow 0} \mathsf{f}_0(x) = \lim_{x\downarrow 0} \int I_{(x,1]}(t)(t-x)\gamma_{\mathsf{f}_0}(dt) = \int I_{(0,1]}(t)t\gamma_{\mathsf{f}}(dt).$$

∎

3 f-Divergences and Related Distances

Now we introduce a general class of information functionals. Let P_0, P_1 be probability measures defined on $(\mathcal{X}, \mathfrak{A})$ which are dominated by the σ-finite measure μ and denote by p_0 and p_1 the respective μ-densities, i.e. let

$$p_0 = \frac{dP_0}{d\mu} \quad \text{and} \quad p_1 = \frac{dP_1}{d\mu}.$$

Definition 2. *For every convex function* $\mathsf{f} : (0,\infty) \to \mathbb{R}$ *the functional,*

$$\mathsf{I}_{\mathsf{f}}(P_0, P_1) := \int \mathsf{f}(p_0/p_1)p_1 I_{\{p_0>0, p_1>0\}} d\mu + \mathsf{f}(0)P_1(p_0 = 0) + \mathsf{f}^*(0)P_0(p_1 = 0) \tag{3.1}$$

is called the f*-divergence of* P_0 *with respect to* P_1.

The right hand term is well defined because $\mathsf{f}(0)$, $\mathsf{f}^*(0) > -\infty$ and the inequality (2.3) implies,

$$\mathsf{f}(p_0/p_1)I_{\{p_0>0\}}p_1 \geq \mathsf{f}(1)p_1 + (D^+\mathsf{f}(1))(p_0 - p_1).$$

As the right hand function is integrable we see that the integral in (3.1) is well defined but may take on the value $+\infty$. Note that $P_1(p_0 = 0)$ and $P_0(p_1 = 0)$ are the weights of the singular parts of P_1 and P_0 with respect to P_0 and P_1, respectively. They are independent of the special choice of the dominating measure μ. This follows also for the integral in (3.1) by the chain rule of measure theory. Therefore the definition of $\mathsf{I}_{\mathsf{f}}(P_0, P_1)$ is independent of the special choice of μ.

The functional $\mathsf{I}_{\mathsf{f}}(P_0, P_1) - \mathsf{f}(1)$ depends only on the nonlinear part of f in the following sense.

Proposition 3.1. *If* $\mathsf{g}(x) = \mathsf{f}(x) + ax + b$ *then,*

$$\mathsf{I}_{\mathsf{f}}(P_0, P_1) - \mathsf{f}(1) = \mathsf{I}_{\mathsf{g}}(P_0, P_1) - \mathsf{g}(1), \tag{3.2}$$

and, for $\mathsf{g} = f_0$ *from (2.9), we have* $\mathsf{I}_{\mathsf{f}}(P_0, P_1) - \mathsf{f}(1) = \mathsf{I}_{\mathsf{f}_0}(P_0, P_1)$.

Proof. By the definition of $\mathsf{I}_\mathsf{f}(P_0, P_1)$ in (3.1),

$$\begin{aligned}\mathsf{I}_\mathsf{g}(P_0,P_1) &:= \int \mathsf{g}(p_0/p_1)p_1 I_{\{p_0>0,p_1>0\}}d\mu \\ &\qquad + \mathsf{g}(0)P_1\,(p_0=0) + \mathsf{g}^*(0)P_0\,(p_1=0) \\ &= \int \mathsf{f}(p_0/p_1)p_1 I_{\{p_0>0,p_1>0\}}d\mu + \int a(p_0/p_1)p_1 I_{\{p_0>0,p_1>0\}}d\mu \\ &\qquad + \int bp_1 I_{\{p_0>0,p_1>0\}}d\mu \\ &\qquad + (\mathsf{f}(0)+b)P_1\,(p_0=0) + (\mathsf{f}^*(0)+a)P_0\,(p_1=0) \\ &= \mathsf{I}_\mathsf{f}(P_0,P_1) + a + b,\end{aligned}$$

so that $\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) = \mathsf{I}_\mathsf{g}(P_0, P_1) - \mathsf{g}(1)$. ∎

Although the functional $\mathsf{I}_\mathsf{f}(P_0, P_1)$ does not satisfy the axioms of a metric for general f, it has several properties that allow us to interpret this functional as a "distance measure".

Proposition 3.2. *For every convex function* f *we have* $\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) \geq 0$, *with equality for* $P_0 = P_1$. *If* f *is strictly convex at* $x_0 = 1$ *then* $\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) = 0$ *implies* $P_0 = P_1$. *Moreover, the functional* $\mathsf{I}_{\mathsf{f}^*}$ *is conjugate to* I_f *in the sense that,*

$$\mathsf{I}_\mathsf{f}(P_0, P_1) = \mathsf{I}_{\mathsf{f}^*}(P_1, P_0), \tag{3.3}$$

so that $\mathsf{I}_{\mathsf{f}+\mathsf{f}^*}(P_0, P_1) = \mathsf{I}_\mathsf{f}(P_0, P_1) + \mathsf{I}_{\mathsf{f}^*}(P_0, P_1)$ *is symmetric in* P_0, P_1.

Proof. The function f_0 is nonnegative so that the expression, $\mathsf{I}_{\mathsf{f}_0}(P_0, P_1) = \mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1)$, is nonnegative as well. If $P_0 = P_1$ then $p_0 = p_1$ μ-*a.e.* and $P_1(p_0 = 0) = P_0(p_1 = 0) = 0$ so that the integral on the right hand side of (3.1) has the value $\mathsf{f}(1)$. Assume now that f is strictly convex at $x_0 = 1$. In view of Proposition 3.1 it is sufficient to consider f_0. By (2.11) $\mathsf{f}_0(x) > 0$ for every $x > 1$ or $\mathsf{f}_0(x) > 0$ for every $0 < x < 1$. Assume the first condition holds, then $\mathsf{f}_0(x) \geq 0$ and $\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) = 0$ together with (3.1) for $\mathsf{f} = \mathsf{f}_0$ show that $\mu(p_0 > p_1) = 0$. This implies,

$$0 = \int (p_1 - p_0)d\mu = \int_{\{p_1>p_0\}} (p_1 - p_0)d\mu,$$

and, therefore, $\mu(p_1 > p_0) = 0$. Hence $\mu(p_1 \neq p_0) = 0$ and $P_1 = P_0$. The case when $\mathsf{f}_0(x) > 0$ for every $x > 1$ is similar. The statement (3.3) is an immediate consequence of (2.12) and (3.1). ∎

$\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1)$ does not satisfy the triangular inequality and is not symmetric in (P_0, P_1), in general. From Definition 2 it follows that the symmetry in (P_0, P_1) holds if $\mathsf{f}(x) = \mathsf{f}^*(x) := x\mathsf{f}(1/x)$. To keep the notation simple we will use the symbol $\mathsf{I}_\mathsf{f}(P_0, P_1)$ also if f is concave. The next display presents special parametrized classes

of functions which are either convex or concave and provide well-known information functionals.

(3.4)

f	$\mathsf{I}_{\mathsf{f}}(P_0,P_1)$
$\chi_s \quad = \lvert x^s-1\rvert^{\frac{1}{s}}$ if $0<s<1$	$\chi_s(P_0,P)$
$\chi_s \quad = \lvert x-1\rvert^s$ if $1\le s<\infty$	$\chi^s(P_0,P_1)$
$\mathsf{h}_s \quad = \begin{cases} x^s & \text{if } s\neq 0,\neq 1 \\ 1 & \text{if } s=0 \text{ or } s=1 \end{cases}$	$\mathsf{H}_s(P_0,P_1)$
$\mathsf{k}_s \quad = \begin{cases} \frac{x^s-sx-(1-s)}{s(s-1)} & \text{if } s\neq 0,\neq 1 \\ x\ln x-x+1 & \text{if } s=1 \\ -\ln x+x-1 & \text{if } s=0 \end{cases}$	$\mathsf{K}_s(P_0,P_1)$
$\mathsf{g}_\pi \quad = \pi\wedge(1-\pi)-(\pi x)\wedge(1-\pi), 0<\pi<1$	$\mathsf{G}_\pi(P_0,P_1)$

The functionals $\chi^s(P_0,P_1)$ were called χ^s-*divergences* in [68]. The particular

$$\chi^2(P_0,P_1)=\int\frac{(p_1-p_0)^2}{p_1}I_{\{p_1>0\}}d\mu+\infty P_0(p_1=0) \tag{3.5}$$

is the well known χ^2-*divergence*. Furthermore, $\chi^{\frac{1}{2}}(P_0,P_1)=\int(\sqrt{p_0}-\sqrt{p_1})^2d\mu$ is the squared *Hellinger distance*,

$$\mathsf{D}(P_0,P_1)=\left[\int\left(\sqrt{p_0}-\sqrt{p_1}\right)^2d\mu\right]^{\frac{1}{2}}. \tag{3.6}$$

It is clear that $\mathsf{D}(P_0,P_1)$ is a metric on the space of all distributions since it is the $\mathbb{L}_2(\mu)$-distance of the roots of the densities. For $s=1$ we get the *variational distance*,

$$\chi^1(P_0,P_1)=\int|p_1-p_0|d\mu=:\|P_0-P_1\|, \tag{3.7}$$

which is a metric on the space of all distributions as well.

Since one has often to approximate one statistical model by another in the strong sense of variational distance, then inequalities relating the variational distance $\|P_0-P_1\|$ to the better tractable Hellinger or Kullback-Leibler distance are useful. Such inequalities were applied in different areas of probability theory, information theory, and statistics, and thus were independently established by various authors, see [49], [31], [67], [40], [63], [58], [30] and [17].

Proposition 3.3. *The following results hold.*

$$\mathsf{D}^2(P_0,P_1) \le \|P_0-P_1\| \le \left[4-\mathsf{D}^2(P_0,P_1)\right]^{\frac{1}{2}} \mathsf{D}(P_0,P_1) \le 2\mathsf{D}(P_0,P_1),$$

$$\mathsf{D}^2(P_0,P_1) \le 2\left(1-\exp\left\{-\frac{1}{2}\mathsf{I}(P_0,P_1)\right\}\right), \tag{3.8}$$

$$\|P_0-P_1\| \le 2\sqrt{\mathsf{I}(P_0,P_1)}. \tag{3.9}$$

Proof. The first inequality follows from $(\sqrt{a_1}-\sqrt{a_2})^2 \le |a_1-a_2|$. To prove the remaining two, let us start by applying the Schwarz inequality to

$$|p_0-p_1| = |\sqrt{p_0}-\sqrt{p_1}|\,|\sqrt{p_0}+\sqrt{p_1}|$$

and obtain,

$$\int |\sqrt{p_0}+\sqrt{p_1}|^2\,d\mu = 4-\mathsf{D}^2(P_0,P_1) \le 4.$$

To get (3.8) we may assume $P_0 \ll P_1$ as otherwise $\mathsf{K}_1(P_0,P_1)=\infty$ and the inequality becomes trivial. But if $P_0 \ll P_1$ then (3.1) implies $\mathsf{I}(P_0,P_1) = \mathsf{E}_{P_0}\ln(dP_0/dP_1)$. By the convexity of the exponential function and Jensen's inequality,

$$\begin{aligned}
\mathsf{D}^2(P_0,P_1) &= 2\left(1-\int\sqrt{P_0P_1}d\mu\right)\\
&= 2(1-\mathsf{E}_{P_0}(dP_0/dP_1)^{-1/2})\\
&= 2\left(1-\mathsf{E}_{P_0}\exp\left\{-\frac{1}{2}\ln(dP_0/dP_1)\right\}\right)\\
&\le 2\left(1-\exp\left\{-\frac{1}{2}\mathsf{E}_{P_0}\ln(dP_0/dP_1)\right\}\right).
\end{aligned}$$

The last statement follows from $1-\exp\{-x\} \le x$, $x \ge 0$. ∎

The inequality $\|P_0-P_1\| \le c\sqrt{\mathsf{K}_1(P_0,P_1)}$ with some constant c was independently established by many authors and has a long history, see [17] where one can also find improved bounds. An important application of inequality (3.9) can be found in [58], where the following result has been established.

Proposition 3.4. *The Hellinger distance and the variational distance of the binomial distribution and the Poisson distribution satisfy,*

$$\begin{aligned}
&\mathsf{D}(\mathsf{B}(n,\lambda/n),\mathsf{Po}(\lambda)) \le \sqrt{3}\lambda/n \quad \textit{and}\\
&\|\mathsf{B}(n,\lambda/n)-\mathsf{Po}(\lambda)\| \le 2\lambda/n.
\end{aligned}$$

These inequalities imply, as a special case, the well known convergence of $\mathsf{B}(n,\lambda/n)$ to $\mathsf{Po}(\lambda)$ for fixed λ. But they also allow an approximation of the binomial distribution $\mathsf{B}(n,p_n)$ by Poisson distributions if p_n tends to 0 at a lower rate than $1/n$. For

applications and extensions of Proposition 3.4 to binomial processes and curve estimation we refer to [59]. Other applications of the inequalities in Proposition 3.3 can be found in [27] and in [44], where Hellinger integrals of distributions of stochastic processes have been evaluated and used to examine the variational distance between such distributions.

The functionals $\mathsf{M}_s(P_0, P_1)$ are termed *Matusita distances* and have been introduced in [47] and reintroduced by many authors, cf. [26]. A closer look at the families $\chi^s(P_0, P_1)$, $\mathsf{M}_s(P_0, P_1)$, $\mathsf{H}_s(P_0, P_1)$ and $\mathsf{K}_s(P_0, P_1)$ reveals that they intersect at several important special cases and contain many well known functionals. First of all, notice that the functionals $\chi^s(P_0, P_1)$, $\mathsf{M}_s(P_0, P_1)$, $\mathsf{H}_s(P_0, P_1)$ and $\mathsf{K}_s(P_0, P_1)$ are symmetric in (P_0, P_1) for $s = 1/2$ in each family. The functionals,

$$\mathsf{H}_s(P_0, P_1) = \begin{cases} \int p_0^s p_1^{1-s} I_{\{p_0>0\}} d\mu + \infty P_1(p_0 = 0) & \text{if} \quad s < 0, \\ \int p_0^s p_1^{1-s} d\mu & \text{if} \quad 0 < s < 1, \\ \int p_0^s p_1^{1-s} I_{\{p_1>0\}} d\mu + \infty P_0(p_1 = 0) & \text{if} \quad 1 < s, \\ 1 & \text{if} \quad s = 0 \text{ or } 1, \end{cases} \tag{3.10}$$

are called *Hellinger integrals* and are mainly used in the literature for $0 < s < 1$. For some purposes their extension to $s \leq 0$ and $s \geq 1$ is useful. These extensions are f-divergences but for $0 < s < 1$ they are f-divergences only up to the sign because the functions h_s are concave when $0 < s < 1$.

In general, f-divergences or simple transformations of them do not satisfy the axioms of a metric. The reflexivity $\mathsf{I}_\mathsf{f}(P_0, P_1) = 0$ if and only if $P_0 = P_1$ is easily obtained for f strictly convex at $t = 1$, see Proposition 3.2. The symmetry $\mathsf{I}_\mathsf{f}(P_0, P_1) = \mathsf{I}_\mathsf{f}(P_1, P_0)$ holds for $\mathsf{f} = \mathsf{f}^*$. If this equality is not fulfilled we could turn to $\tilde{\mathsf{f}} = \mathsf{f} + \mathsf{f}^*$. The main problem is the triangular inequality which can be verified only for special f. Examples are the Hellinger distance and the variational distance. The next example contains another class of divergences that satisfy the triangular inequality.

Example 6. *The functional*

$$\mathsf{A}_\alpha(P_0, P_1) = \frac{1}{2} \int \left[p_0^{1/\alpha} + p_1^{1/\alpha} \right]^\alpha d\mu - 1, \quad 0 < \alpha \leq 1$$

was introduced by Österreicher and Vajda [53] who proved that it is an $\mathsf{f}-$*divergence satisfying the triangular inequality. Since this divergence was motivated by a previous work of Arimoto [2], it was called the Arimoto divergence by these authors.*

From (3.4) and (3.10) we see that,

$$\mathsf{K}_s(P_0, P_1) = \frac{1}{s(1-s)} (1 - \mathsf{H}_s(P_0, P_1)), \quad s \neq 0, s \neq 1, \tag{3.11}$$

$$\mathsf{K}_1(P_0, P_1) = \int \left[\ln \frac{p_0}{p_1} \right] I_{\{p_1>0\}} dP_0 + \infty P_0(p_1 = 0), \tag{3.12}$$

$$2\mathsf{K}_2(P_0, P_1) = \mathsf{H}_2(P_0, P_1) - 1 = \chi^2(P_0, P_1). \tag{3.13}$$

Since $\mathsf{k}_s(x) \geq 0$, by construction we get $0 \leq \mathsf{K}_s(P_0, P_1) \leq \infty$. As mentioned above, $\mathsf{K}_1(P_0, P_1)$ is sometimes called the *Kullback-Leibler divergence.*

We illustrate by examples that Hellinger integrals, and consequently the divergences $\mathsf{K}_s(P_0, P_1)$, can be explicitly evaluated for a large variety of distributions that are important in statistics. We consider the class of exponential families which plays a central role in mathematical statistics. Let $(\mathcal{X}, \mathfrak{A})$ be a given measurable space and $T : \mathcal{X} \to \mathbb{R}^d$ be a statistic. For any σ-finite measure μ we put

$$\Delta = \left\{\theta : \int \exp\{\langle \theta, T\rangle\} d\mu < \infty\right\} \subseteq \mathbb{R}^d,$$
$$K(\theta) = \ln\left(\int \exp\{\langle \theta, T\rangle\} d\mu\right), \quad \theta \in \Delta.$$

If $0 < \alpha < 1$ and $\theta_1, \theta_2 \in \Delta$ then Hölder's inequality yields

$$\begin{aligned}\exp\{K(\alpha\theta_1 + (1-\alpha)\theta_2)\} &= \int \exp\{\langle \alpha\theta_1, T\rangle\} \exp\{\langle (1-\alpha)\theta_2, T\rangle\} d\mu \\ &\leq \left(\int \exp\{\langle \theta_1, T\rangle\} d\mu\right)^{\alpha} \left(\int \exp\{\langle \theta_2, T\rangle\} d\mu\right)^{1-\alpha} \\ &= \exp\{\alpha K(\theta_1) + (1-\alpha) K(\theta_2)\}.\end{aligned}$$

This means that the set Δ is convex and that the function K is convex. Further, for every $\theta \in \Delta$,

$$P_\theta(A) = \int_A \exp\{\langle \theta, T\rangle - K(\theta)\} d\mu, \quad A \in \mathfrak{A} \tag{3.14}$$

is a probability measure on $(\mathcal{X}, \mathfrak{A})$, and the family of distributions $(P_\theta)_{\theta \in \Delta}$ is called an *exponential family with generating statistic T and natural parameter space Δ.*

Example 7. *Assume that $P_\theta, \theta \in \Delta \subseteq \mathbb{R}^d$ is an exponential family with natural parameter $\theta \in \Delta$ and generating statistic $T : \mathcal{X} \to \mathbb{R}^d$. The μ-density of P_θ is then given by $p_\theta = \exp\{\langle T, \theta\rangle - K(\theta)\}$. Consequently, (3.10) implies, for all s and $\theta_1, \theta_2 \in \Delta$ satisfying $s\theta_1 + (1-s)\theta_2 \in \Delta$ and $s \neq 0, s \neq 1$, that,*

$$\begin{aligned}\mathsf{H}_s(P_{\theta_1}, P_{\theta_2}) &= \int p_{\theta_1}^s p_{\theta_2}^{1-s} d\mu \\ &= \exp\{-(sK(\theta_1) + (1-s)K(\theta_2) - K(s\theta_1 + (1-s)\theta_2))\}.\end{aligned} \tag{3.15}$$

Due to the convexity of Δ, the condition $s\theta_1 + (1-s)\theta_2 \in \Delta$ is fulfilled by $0 \leq s \leq 1$. To calculate the Kullback-Leibler divergence $\mathsf{K}_1(P_{\theta_1}, P_{\theta_2})$ we note that, by (3.1) and the fact that P_{θ_1} and P_{θ_2} are measure theoretically equivalent and $\mathsf{E}_{\theta_1} T$ exists for $\theta_1 \in \Delta^0$,

$$\begin{aligned}\mathsf{K}_1(P_{\theta_1}, P_{\theta_2}) &= \mathsf{E}_{\theta_1} \ln \frac{p_{\theta_1}}{p_{\theta_2}} = -K(\theta_1) + K(\theta_2) + \mathsf{E}_{\theta_1} \langle T, \theta_1 - \theta_2\rangle \\ &= \langle \mathsf{E}_{\theta_1} T, \theta_1 - \theta_2\rangle - K(\theta_1) + K(\theta_2), \quad \theta_1 \in \Delta^0, \theta_2 \in \Delta.\end{aligned}$$

It is well known, see e.g. [5], that the moments of T can be calculated by taking the derivative on both sides of the identity,

$$\int \exp\{\langle\theta, T\rangle - K(\theta)\}d\mu = 1, \tag{3.16}$$

where on the left hand side the derivative can be carried out under the integral sign. This yields $\mathsf{E}_\theta T = \nabla K(\theta)$. Together with (3.15) we get,

$$\begin{aligned}\mathsf{K}_1(P_{\theta_1}, P_{\theta_2}) &= \langle\nabla K(\theta_1), \theta_1 - \theta_2\rangle - K(\theta_1) + K(\theta_2)\\ &= \lim_{s\to 1}\frac{sK(\theta_1) + (1-s)K(\theta_2) - K(s\theta_1 + (1-s)\theta_2))}{s(1-s)}\\ &= \lim_{s\to 1}\frac{(1 - \mathsf{H}_s(P_{\theta_1}, P_{\theta_2}))}{s(1-s)}.\end{aligned}$$

A similar statement holds for $s = 0$.

The family $\left\{\mathsf{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\right\}$ of normal distributions is an exponential family on the real line so that we could obtain the corresponding Hellinger integrals by applying (3.15). In this case, however, a direct calculation is simpler. Let φ_{μ,σ^2} be the density of the normal distribution then,

$$\begin{aligned}&\mathsf{H}_s(\mathsf{N}(\mu_1,\sigma_1^2), \mathsf{N}(\mu_2, \sigma_2^2))\\ &\quad= \int \varphi^s_{\mu_1,\sigma_1^2}(x)\varphi^{1-s}_{\mu_2,\sigma_2^2}(x)dx\\ &\quad= \left[\frac{\sigma_1^{2(1-s)}\sigma_2^{2s}}{s\sigma_2^2 + (1-s)\sigma_1^2}\right]^{\frac{1}{2}} \exp\left\{-\frac{1}{2}s(1-s)\frac{(\mu_1-\mu_2)^2}{s\sigma_2^2 + (1-s)\sigma_1^2}\right\}.\end{aligned} \tag{3.17}$$

Further, by (3.12),

$$\begin{aligned}\mathsf{K}_1(\mathsf{N}(\mu_1, \sigma_1^2), \mathsf{N}(\mu_2, \sigma_2^2)) &= \int \varphi_{\mu_1,\sigma_1^2}(x)\ln\left(\varphi_{\mu_1,\sigma_1^2}(x)/\varphi_{\mu_2,\sigma_2^2}(x)\right)dx\\ &= \frac{1}{2}\left(\sigma_1^2/\sigma_2^2 - 1 - \ln(\sigma_1^2/\sigma_2^2)\right) + (\mu_1-\mu_2)^2/\sigma_2^2).\end{aligned} \tag{3.18}$$

Similarly, for $-\infty < s < \infty$,

$$\mathsf{H}_s(\mathsf{Po}(\lambda_1), \mathsf{Po}(\lambda_2)) = \exp\{\lambda_1^s\lambda_2^{1-s} - s\lambda_1 - (1-s)\lambda_2\}, \tag{3.19}$$

for the family of Poisson distributions $\{\mathsf{Po}(\lambda) : \lambda > 0\}$.

To give a statistical interpretation of the $\mathsf{f}-$divergences $\mathsf{G}_\pi(P_0, P_1)$, introduced for $0 \le \pi \le 1$ in (3.4), we consider the problem of testing the simple null hypothesis $\mathsf{H}_0 : P_0$ versus the alternative $\mathsf{H}_1 : P_1$. A statistical test φ is then a measurable mapping $\varphi : \mathcal{X} \to [0, 1]$ where the value $\varphi(x)$ represents the conditional probability of rejecting H_0 when the observation is x. Consequently, $\int \varphi dP_0$ is the probability of rejecting H_0 when H_0 is true, called *the error probability of the first kind.* Similarly $\int(1-\varphi)dP_1$

is the probability of rejecting H_1 when H_1 is true, called *the error probability of the second kind*. The mix,

$$\pi \int \varphi dP_0 + (1-\pi) \int (1-\varphi) dP_1$$

of the error, probabilities taken for the prior probability of the hypothesis $0 \leq \pi \leq 1$, is the *Bayes' error probability* or the *Bayes' risk*. Each test which minimizes the Bayes' error probability is a *Bayes' test*. Next we present a well known result on the Bayes' test and the minimal Bayes' error probability.

Lemma 3.5. *In the binary model* $(\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ *the test* $\varphi_B : \mathcal{X} \to [0,1]$ *defined by*

$$\varphi_B = \begin{cases} 1 & \text{if} \quad cp_0 < p_1 \\ \text{arbitrary} & \text{if} \quad cp_0 = p_1 \\ 0 & \text{if} \quad cp_0 \geq p_1 \end{cases}, \tag{3.20}$$

for $c = \frac{\pi}{1-\pi}$, *is a Bayes' test and the minimal Bayes' error probability,*

$$\mathsf{b}_\pi(P_0, P_1) = \inf_\varphi \left(\pi \int \varphi dP_0 + (1-\pi) \int (1-\varphi) dP_1 \right)$$

is given by,

$$\mathsf{b}_\pi(P_0, P_1) = \int (\pi p_0) \wedge ((1-\pi) p_1) d\mu, \tag{3.21}$$

where μ *is a* σ*-finite dominating measure and* $p_i = dP_i/d\mu$.

Proof. We have,

$$\begin{aligned} \pi \int \varphi dP_0 + (1-\pi) \int (1-\varphi) dP_1 &= \int (\pi \varphi p_0 + (1-\pi)(1-\varphi) p_1) d\mu \\ &= (1-\pi) + \int \varphi (\pi p_0 - (1-\pi) p_1) d\mu. \end{aligned}$$

The right hand side becomes minimal if we set $\varphi_B = 1$ if $\pi p_0 < (1-\pi) p_1$, $\varphi_B = 0$ if $\pi p_0 > (1-\pi) p_1$ and let φ_B be arbitrary if $\pi p_0 = (1-\pi) p_1$. The Bayes' error of this test is given by,

$$\int (\pi \varphi_B p_0 + (1-\pi)(1-\varphi_B) p_1) d\mu = \int (\pi p_0) \wedge ((1-\pi) p_1) d\mu.$$

∎

We see from (3.21) that the minimal Bayes' error probability is related to the divergence $\mathsf{G}_\pi(P_0, P_1)$ defined by means of g_π in (3.4) as follows:

$$\mathsf{G}_\pi(P_0, P_1) = \mathsf{I}_{\mathsf{g}_\pi}(P_0, P_1) = \pi \wedge (1-\pi) - \mathsf{b}_\pi(P_0, P_1). \tag{3.22}$$

The functional $\mathsf{G}_\pi(P_0, P_1)$ admits the following interpretation proposed by De Groot [15, 16]: the first term $\pi \wedge (1-\pi)$ is the minimal Bayes' error probability that can be achieved before the observation in the model $\{P_0, P_1\}$ is made and the second term $\mathsf{b}_\pi(P_0, P_1)$ is the minimal Bayes' error probability achievable after this observation is made. The non-negative difference $\mathsf{G}_\pi(P_0, P_1)$ between these two errors thus represents an information gain achieved by taking the observation.

We will show in the next theorem that $\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1)$ is a superposition of the information gains $\mathsf{G}_\pi(P_0, P_1)$ with respect to a curvature measure ρ_f on $(0,1)$ defined by,

$$\rho_\mathsf{f}(B) = \int (1+t) I_B\left(\frac{1}{1+t}\right) \gamma_\mathsf{f}(dt). \tag{3.23}$$

The measures ρ_f and γ_f satisfy, for every measurable $h : (0,1) \longrightarrow [0,\infty)$, the relation,

$$\int h(\pi)\rho_\mathsf{f}(d\pi) = \int (1+t) h\left(\frac{1}{1+t}\right) \gamma_\mathsf{f}(dt). \tag{3.24}$$

We denote by,

$$S_{\rho_\mathsf{f}} = \{x \in (0,1) : \ \rho_\mathsf{f}(x-\varepsilon, x+\varepsilon) > 0 \quad \text{for all } \varepsilon > 0\} \tag{3.25}$$

the support of the measure ρ_f.

Theorem 3.6. *For every convex function $\mathsf{f} : (0,\infty) \to \mathbb{R}$ and arbitrary distributions P_0, P_1 we have,*

$$\mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) = \int_{(0,1)} \mathsf{G}_\pi(P_0, P_1)\rho_\mathsf{f}(d\pi). \tag{3.26}$$

Corollary 3.7. *The following holds.*

$$\mathsf{K}_s(P_0, P_1) = \int_{(0,1)} \frac{\mathsf{G}_\pi(P_0, P_1)}{(1-\pi)^{1+s}\pi^{2-s}} d\pi, \quad -\infty < s < \infty,$$
$$\mathsf{H}_s(P_0, P_1) = s(1-s) \int_{(0,1)} \frac{\mathsf{b}_\pi(P_0, P_1)}{(1-\pi)^{1+s}\pi^{2-s}} d\pi, \quad 0 < s < 1.$$

Proof. Due to the invariance property (3.2), the left hand term in (3.26) remains unchanged if f is replaced by f_0 in (2.9). As $\gamma_\mathsf{f} = \gamma_{\mathsf{f}_0}$, the right hand term also remains unchanged. Hence we may assume $\mathsf{f}(1) = D^+\mathsf{f}(1) = 0$ without loss of generality. We

see from (2.10) that,

$$\begin{aligned}
&\int I_{(0,\infty)}(p_0 \wedge p_1)\mathsf{f}\left(\frac{p_0}{p_1}\right) dP_1 \\
&= \int \left(\int [I_{(1,\infty)}(t)(p_0 - (tp_1) \wedge p_0) + I_{(0,1]}(t)(tp_1 - (tp_1) \wedge p_0)]\gamma_{\mathsf{f}}(dt) \right) \\
&\qquad \times I_{\{p_0 \wedge p_1 > 0\}} d\mu \\
&= \int (P_0(p_1 > 0) - (1+t)\mathsf{b}_{1/(1+t)}(P_0, P_1)) I_{(1,\infty)}(t)\gamma_{\mathsf{f}}(dt) \\
&\qquad + \int (tP_1\,(p_0 > 0) - (1+t)\mathsf{b}_{1/(1+t)}(P_0, P_1)) I_{(0,1]}(t)\gamma_{\mathsf{f}}(dt).
\end{aligned}$$

Now we use (2.14) and (2.15) to obtain,

$$\begin{aligned}
\mathsf{I}_{\mathsf{f}}(P_0, P_1) &= \int I_{(1,\infty)}(t)(1 - (1+t)\mathsf{b}_{1/(1+t)}(P_0, P_1))\gamma_{\mathsf{f}}(dt) \\
&\qquad + \int I_{(0,1]}(t)(t - (1+t)\mathsf{b}_{1/(1+t)}(P_0, P_1))\gamma_{\mathsf{f}}(dt) \\
&= \int I_{(0,\infty)}(t)(1+t)\left(\frac{1}{1+t} \wedge \frac{t}{1+t} - \mathsf{b}_{1/(1+t)}(P_0, P_1)\right)\gamma_{\mathsf{f}}(dt).
\end{aligned}$$

To complete the proof of the theorem we need only employ (3.22) and (3.24). In order to prove the corollary we use $\mathsf{k}_s(x)$ from (3.4), then $\gamma_{\mathsf{h}_s}(dx) = x^{s-2}dx$ and hence, for every Borel set $B \subseteq (0,1)$,

$$\rho_{\mathsf{k}_s}(B) = \int I_B\left(\frac{1}{1+t}\right)(1+t)t^{s-2}dt = \int I_B(\pi)(1-\pi)^{s-2}\pi^{-1-s}d\pi,$$

which proves the first statement of the corollary. For the second statement, we use,

$$s(1-s)\int_{(0,1)} (\pi \wedge (1-\pi))(1-\pi)^{s-2}\pi^{-1-s}d\pi = 1, \quad 0 < s < 1,$$

and (3.11). ∎

The representation of the f-divergence in Theorem 3.6 was established by Österreicher and Feldman [51] for twice differentiable functions f, and by Torgersen [66] for the special case of Hellinger integrals. Extensions of these representations were studied later by Guttenbrunner [23] and Österreicher and Vajda [53]. We see that this representation connects the concept of the distance of distributions measured by the f-divergence with decision theoretic concepts represented by the minimal Bayes' probability of error.

We now establish the monotonicity property of f-divergences. The basic idea is as follows. Suppose we are faced with two distributions P_0 and P_1 and employ a statistic T for data compression. By doing so we are aware of the fact that the distance between P_0 and P_1 may be reduced, in the sense that it is harder to distinguish

between $P_0 \circ T^{-1}$ and $P_1 \circ T^{-1}$ than between P_0 and P_1. The question then arises as to how much information has been lost, and how to quantify it. An answer is given by the monotonicity theorem (also called the data processing theorem, see [12] or [10]) which is the next object of our interest.

Consider the binary statistical model $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$, suppose that $(\mathcal{Y}, \mathfrak{B})$ is another measurable space and that $\mathsf{K} : \mathfrak{B} \times \mathcal{X} \to [0,1]$ is a stochastic kernel from $(\mathcal{X}, \mathfrak{A})$ to $(\mathcal{Y}, \mathfrak{B})$. We obtain the reduced model $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$ for $Q_i = \mathsf{K}P_i$ defined by,

$$(\mathsf{K}P_i)(B) = \int \mathsf{K}(B|x) P_i(dx), \quad B \in \mathfrak{B}.$$

It is intuitively clear that the model $\mathcal{N}$ is less informative than $\mathcal{M}$ as it is harder to distinguish between $\mathsf{K}P_0$ and $\mathsf{K}P_1$ than between P_0 and P_1 and we can anticipate the inequality $\mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) \leq \mathsf{I}_\mathsf{f}(P_0, P_1)$. This inequality is the content of the following theorem that goes back to Csiszár [11]. Preparatory to this theorem we study the information gain $\mathsf{G}_\pi(P_0, P_1)$ in (3.22). For any test $\varphi : \mathcal{Y} \to [0,1]$ in the reduced model $\mathcal{N}$ we get,

$$\int \varphi d(\mathsf{K}P_0) = \int \varphi(y) \left(\int \mathsf{K}(dy|x) P_0(dx) \right) = \int \left(\int \varphi(y) \mathsf{K}(dy|x) \right) P_0(dx),$$

where $x \longmapsto \int \varphi(y) \mathsf{K}(dy|x)$ is a test in the original model $\mathcal{M}$. As $\mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1)$ is the minimal Bayes' error in the testing problem $\mathsf{H}_0 : \mathsf{K}P_0$ versus $\mathsf{H}_1 : \mathsf{K}P_1$ we obtain,

$$\begin{aligned}
&\mathsf{G}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) \\
&= \sup_\varphi \left(\pi \wedge (1-\pi) - \pi \int \varphi d(\mathsf{K}P_0) - (1-\pi) \int (1-\varphi) d(\mathsf{K}P_1) \right) \\
&\leq \sup_\psi \left(\pi \wedge (1-\pi) - \pi \int \psi dP_0 - (1-\pi) \int (1-\psi) dP_1 \right),
\end{aligned}$$

where the supremum is extended over all tests for $\mathcal{M}$ on the right hand side. Hence,

$$\mathsf{G}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) \leq \mathsf{G}_\pi(P_0, P_1). \tag{3.27}$$

This inequality says that the information gain decreases when observation is taken in the reduced model. Now we are ready to formulate the main result of this section, the so called information processing theorem.

Theorem 3.8. *If $(\mathcal{X}, \mathfrak{A})$ and $(\mathcal{Y}, \mathfrak{B})$ are measurable spaces and $\mathsf{K} : \mathfrak{B} \times \mathcal{X} \to [0,1]$ is a stochastic kernel, then for $P_0, P_1 \in \mathcal{P}(\mathfrak{A})$ and every convex function $f\colon (0, \infty) \to \mathbb{R}$ we have,*

$$\mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) \leq \mathsf{I}_\mathsf{f}(P_0, P_1). \tag{3.28}$$

If

$$\mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{b}_\pi(P_0, P_1), \quad 0 < \pi < 1 \tag{3.29}$$

then $\mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{I}_\mathsf{f}(P_0, P_1)$. *Conversely, if* $\mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{I}_\mathsf{f}(P_0, P_1) < \infty$ *then,*

$$\mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{b}_\pi(P_0, P_1),\ \pi \in S_{\rho_\mathsf{f}}, \tag{3.30}$$

where S_{ρ_f} *is the support of* ρ_f *in (3.25).*

Corollary 3.9. *If* $T : \mathcal{X} \to \mathcal{Y}$ *is a statistic then,*

$$\mathsf{I}_\mathsf{f}(P_0 \circ T^{-1}, P_1 \circ T^{-1}) \le \mathsf{I}_\mathsf{f}(P_0, P_1).$$

For strictly convex f *and* $\mathsf{I}_f(P_0, P_1) < \infty$ *the equality holds if and only if*

$$\mathsf{b}_\pi(P_0 \circ T^{-1}, P_1 \circ T^{-1}) = \mathsf{b}_\pi(P_0, P_1), \quad 0 < \pi < 1.$$

Proof. The inequality (3.28) follows directly from (3.27) and Theorem 3.6 where equality holds if (3.29) is satisfied. Suppose now, $\mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{I}_\mathsf{f}(P_0, P_1) < \infty$, then by Theorem 3.6,

$$0 = \mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{I}_\mathsf{f}(\mathsf{K}P_0, \mathsf{K}P_1) = \int [\mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) - \mathsf{b}_\pi(P_0, P_1)]\, \rho_\mathsf{f}(d\pi).$$

The integrand is nonnegative in view of (3.27). Consequently,

$$\rho_\mathsf{f}(\{\pi : \mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) \ne \mathsf{b}_\pi(P_0, P_1)\}) = 0. \tag{3.31}$$

It follows from the Lebesgue theorem and (3.21) that the function $\pi \mapsto \mathsf{b}_\pi(P_0, P_1)$ is continuous. As P_i are arbitrary, we may replace them by $\mathsf{K}P_i$ obtaining that $\pi \mapsto \mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1)$ is continuous too. This implies the continuity of $\pi \mapsto \mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) - \mathsf{b}_\pi(P_0, P_1)$. This continuity, in conjunction with the definition of the support S_{γ_f}, implies $\mathsf{b}_\pi(\mathsf{K}P_0, \mathsf{K}P_1) = \mathsf{b}_\pi(P_0, P_1)$ for all $\pi \in S_{\rho_\mathsf{f}}$, which completes the proof. The corollary follows from the fact that measurable mappings are special kernels. ∎

There are many approaches to reduction of a large sample $X_1, \ldots, X_n$. One of them is to use a partition $\mathfrak{p} = \{A_1, \ldots, A_n\}$ of the sample space $\mathcal{X}$ and to replace the observations by the relative frequencies of these observations in the partition cells. Here, and in the sequel, a partition $\mathfrak{p}$ means a collection $\{A_1, \ldots, A_n\}$, of subsets of $\mathcal{X}$ such that,

$$A_i \in \mathfrak{A}, \quad A_i \cap A_j = \emptyset \quad \text{for} \ \ i \ne j, \quad \text{and} \quad A_1 \cup \cdots \cup\ A_n = \mathcal{X}. \tag{3.32}$$

Instead of the original sample space $(\mathcal{X}, \mathfrak{A})$ we now use the sample space $(\mathcal{X}, \sigma(\mathfrak{p}))$, where $\sigma(\mathfrak{p})$ is the algebra generated by the partition $\mathfrak{p}$. Assume now that we have an increasing sequence of partitions $\mathfrak{p}_n$ so that the sequence of σ-algebras $\mathfrak{A}_n$ generates $\mathfrak{A}$, then we can approximate $\mathfrak{A}$-measurable tests by $\mathfrak{A}_n$-measurable tests. We, therefore, achieve the minimal Bayes' risk approximately, registering the cells visited by observations instead of the observations themselves, provided n is large enough. Denote by $P_i^{\mathfrak{A}_n}$ the restriction of P_i to the sub σ-algebra $\mathfrak{A}_n$.

Lemma 3.10. *If* $\mathfrak{A}_1 \subseteq \mathfrak{A}_2 \subseteq \cdots$ *is a nondecreasing sequence of sub-*σ*-algebras of* $\mathfrak{A}$ *which generates* $\mathfrak{A}$ *then* $\mathsf{G}_\pi(P_0^{\mathfrak{A}_n}, P_1^{\mathfrak{A}_n}) \uparrow \mathsf{G}_\pi(P_0, P_1)$.

Proof. The monotonicity follows from (3.27). Set $\overline{P} = \frac{1}{2}(P_0 + P_1)$ and consider the densities $p_i = dP_i/d\overline{P}$, $i = 0, 1$, as random variables on $(\mathcal{X}, \mathfrak{A}, \overline{P})$. The conditional expectation $p_{i,n} = \mathsf{E}_{\overline{P}}(p_i|\mathfrak{A}_n)$ with respect to $\overline{P}$ satisfies, for every $A \in \mathfrak{A}_n$,

$$\int_A \mathsf{E}_{\overline{P}}(p_i|\mathfrak{A}_n)d\overline{P} = \int_A p_i d\overline{P} = P_i^{\mathfrak{A}_n}(A),$$

which implies $p_{i,n} = dP_i^{\mathfrak{A}_n}/d\overline{P}^{\mathfrak{A}_n}$. Hence, by the Martingale convergence theorem of Levy, see [33], $\mathsf{E}_{\overline{P}}|p_i - p_{i,n}| \to 0$, as $n \to \infty$. Using the elementary inequality $|a \wedge b - c \wedge d| \leq |a - b| + |c - d|$ we arrive at the relation,

$$\int |(\pi p_0) \wedge ((1-\pi)p_1)d\overline{P} - (\pi p_{0,n}) \wedge ((1-\pi)p_{1,n}|d\overline{P} \to 0, \quad \text{as } n \to \infty.$$

Combining this statement with (3.21) we get $\mathsf{b}_\pi(P_0^{\mathfrak{A}_n}, P_1^{\mathfrak{A}_n}) \uparrow \mathsf{b}_\pi(P_0, P_1)$. The desired result is clear from here and (3.22). ∎

Theorem 3.11. *If $\mathfrak{A}_1 \subseteq \mathfrak{A}_2 \subseteq \cdots$ is a nondecreasing sequence of sub-σ-algebras of $\mathfrak{A}$ which generates $\mathfrak{A}$, then,*

$$\lim_{n\to\infty} \mathsf{I}_\mathsf{f}\left(P_0^{\mathfrak{A}_n}, P_1^{\mathfrak{A}_n}\right) = \mathsf{I}_\mathsf{f}(P_0, P_1), \quad \textit{as} \quad n \to \infty. \tag{3.33}$$

Corollary 3.12. *We have,*

$$\mathsf{I}_\mathsf{f}(P_0, P_1) = \sup_{\mathfrak{p}} \sum_{A\in\mathfrak{p}} \mathsf{f}\left(\frac{P_0(A)}{P_1(A)}\right) P_1(A), \tag{3.34}$$

where the supremum is taken over all partitions $\mathfrak{p}$ with $\mathfrak{p} \subseteq \mathfrak{A}$ and the conventions $f(\frac{0}{0})0 = 0$ and $f(\frac{a}{0})0 = af^*(0)$ for $a > 0$ are used.

Proof. The statement (3.33) follows from Lemma 3.10, the representation (3.26) and the monotone convergence theorem. To prove the corollary we first notice that,

$$\mathsf{I}_\mathsf{f}\left(P_0^{\sigma(\mathfrak{p})}, P_1^{\sigma(\mathfrak{p})}\right) = \sum\nolimits_{A\in\mathfrak{p}} \mathsf{f}\left(\frac{P_0(A)}{P_1(A)}\right) P_1(A) \leq \mathsf{I}_\mathsf{f}(P_0, P_1),$$

where the last inequality follows from (3.28). To show that here the equality can be achieved by taking the supremum we set $\mathfrak{B} = \sigma(p_0, p_1)$. If $P_i^{\mathfrak{B}}$ and $\overline{P}^{\mathfrak{B}}$ denote the restrictions of P_i and $\overline{P}$ on $\mathfrak{B}$ then, by the definition of $\mathfrak{B}$, $p_i = dP_i^{\mathfrak{B}}/d\overline{P}^{\mathfrak{B}}$. Hence $\mathsf{B}_\pi(P_0, P_1) = \mathsf{B}_\pi(P_0^{\mathfrak{B}}, P_1^{\mathfrak{B}})$ by (3.21) and (3.22) so that $\mathsf{I}_\mathsf{f}(P_0^{\mathfrak{B}}, P_1^{\mathfrak{B}}) = \mathsf{I}_\mathsf{f}(P_0, P_1)$ by (3.26). As the open intervals (a, b) with rational endpoints generate the σ-algebra of Borel sets of the real line and the complete images of (a, b) under p_0 and p_1 generate $\mathfrak{B}$, we see that $\mathfrak{B}$ is countably generated. This means that we find a nondecreasing sequence of algebras $\mathfrak{A}_n$ that generate $\mathfrak{B}$. If $\mathfrak{p}_n$ is the system of atoms of $\mathfrak{A}_n$ then $\mathfrak{p}_n$ form a nondecreasing sequence of partitions with $\mathfrak{A}_n = \sigma(\mathfrak{p}_n)$. The rest of the proof follows from (3.33). ∎

4 f-Divergences, Sufficiency and $\varepsilon-$Deficiency

If a model $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta})$ is given and $(\mathcal{Y}, \mathfrak{B})$ is a measurable space then a measurable mapping $T : \mathcal{X} \to \mathcal{Y}$ is called a *statistic* and the model $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, (Q_\theta)_{\theta\in\Delta})$ with $Q_\theta = P_\theta \circ T^{-1}$ is said to be reduced by the statistic T. Recall that the statistic $T : \mathcal{X} \to \mathcal{Y}$ is said to be *sufficient* if for every $A \in \mathfrak{A}$ there is a function $k_A : \mathcal{Y} \to \mathbb{R}$ with the property,

$$\mathsf{E}_\theta(I_A|T) = k_A(T), \quad P_\theta\text{-}a.s., \theta \in \Delta. \tag{4.1}$$

The statistic T is called *pairwise sufficient* if *it* is sufficient for each binary model $(\mathcal{X}, \mathfrak{A}, \{P_{\theta_1}, P_{\theta_2}\}), \theta_1, \theta_2 \in \Delta$. It is well known, see e.g. [63], that for dominated models a statistic T is sufficient if and only if T is pairwise sufficient.

The independence of the conditional probability on the parameter θ extends easily to the independence of the conditional expectation of any random variable. Suppose that $S : \mathcal{X} \to \mathbb{R}$ is a random variable with $\mathsf{E}_\theta|S| < \infty, \theta \in \Delta$. If T is sufficient, then there is some measurable function $k_S : \mathcal{T} \to \mathbb{R}$ such that,

$$\mathsf{E}_\theta(S|T) = k_S(T), \quad P_\theta - a.s., \theta \in \Delta. \tag{4.2}$$

The independence of the conditional probabilities of the parameter was historically the starting point of the concept of sufficiency. This concept can be traced back to Fisher [11] who considered a statistic T to be sufficient if the conditional distribution of any other statistic S given T is independent of the parameter so that T contains the complete information. This means that if the value $T = t$ is observed then the knowledge of the value x leading to the observation t contains no additional information about the parameter.

As pairwise sufficiency and sufficiency are equivalent for dominated models, in the sequel we deal only with binary models. Let $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ be a binary model $(\mathcal{Y}, \mathfrak{B})$ a measurable space and $T : \mathcal{X} \to \mathcal{Y}$ a statistic, then $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$, $Q_i = P_i \circ T^{-1}$ is the reduced model. We set,

$$\begin{aligned} &\overline{P} = \frac{1}{2}(P_0 + P_1), \quad \text{and} \quad \overline{Q} = \frac{1}{2}(Q_0 + Q_1), \\ &L_i := \frac{dP_i}{d\overline{P}}, \quad \text{and} \quad M_i := \frac{dQ_i}{d\overline{Q}}, \quad i = 0, 1. \end{aligned} \tag{4.3}$$

The first consequence of the sufficiency of a statistic T is that the hypotheses testing problems

$$\mathsf{H}_0 : P_0 \quad \text{versus} \quad \mathsf{H}_1 : P_1 \qquad \text{and} \qquad \mathsf{H}_0 : Q_0 \quad \text{versus} \quad \mathsf{H}_1 : Q_1 \tag{4.4}$$

are equivalent in the sense that for each test in one of these problems achieving certain error probabilities of the first and second kind there is a test in the other problem achieving the same error probabilities of the first and second kinds. Indeed, if ψ is a test for $\mathsf{H}_0 : Q_0$ versus $\mathsf{H}_A : Q_1$ then $\psi(T)$ is a test for $\mathsf{H}_0 : P_0$ versus $\mathsf{H}_A : P_1$ and we obtain,

$$\int \psi(T)dP_0 = \int \psi dQ_0 \quad \text{and} \quad \int (1 - \psi(T))dP_1 = \int (1 - \psi)dQ_1.$$

Therefore, in the model $\{P_0, P_1\}$, we find at least as good a test as in the model $\{Q_0, Q_1\}$. If T is sufficient we establish the converse statement. If φ is a test for the model $\mathcal{M}$ then we set $\psi(t) = \mathsf{E}_{P_i}(\varphi|T = t)$, which is independent of i according to (4.2) and put $k_\varphi(T) = \psi(T)$. We then have,

$$\mathsf{E}_{Q_i}\psi = \mathsf{E}_{P_i}(\mathsf{E}_{P_i}(\varphi|T)) = \mathsf{E}_{P_i}\varphi, \tag{4.5}$$

which completes the proof of the equivalence of the two testing problems in (4.4).

The densities L_i and M_i are related by the conditional expectation. Indeed, by the definition of the conditional expectation and $\overline{Q} = \overline{P} \circ T^{-1}$, for every $B \in \mathfrak{B}$,

$$\int_B \mathsf{E}_{\overline{P}}(L_i|T = y)d\overline{Q} = \int I_B(T)\mathsf{E}_{\overline{P}}(L_i|T)d\overline{P} = \int I_B(T)dP_i = Q_i(B).$$

This implies,

$$M_i(y) = \mathsf{E}_{\overline{P}}(L_i|T = y), \quad \overline{Q}\text{-}a.s. \tag{4.6}$$

The next lemma studies the stability of the Schwarz inequality for the conditional expectation.

Lemma 4.1. *Let X_0, X_1 be nonnegative random variables on $(\Omega, \mathfrak{F}, P)$ with $\mathsf{E}X_i < \infty$, $i = 0, 1$. If $\mathfrak{F}_0 \subseteq \mathfrak{F}$ is a sub-σ-algebra of $\mathfrak{F}$ then,*

$$\mathsf{E}((X_0X_1)^{1/2}|\mathfrak{F}_0) \le (\mathsf{E}(X_0|\mathfrak{F}_0)\mathsf{E}(X_1|\mathfrak{F}_0))^{1/2}, \quad P\text{-}a.s. \tag{4.7}$$

where the P-a.s. equality holds if and only if

$$X_0\mathsf{E}(X_1|\mathfrak{F}_0) = X_1\mathsf{E}(X_0|\mathfrak{F}_0), \ P\text{-}a.s. \tag{4.8}$$

Proof. Put $Y_i = \mathsf{E}(X_i|\mathfrak{F}_0)$, $A_i = \{Y_i = 0\}$, then $\mathsf{E}I_{A_i}X_i = \mathsf{E}(I_{A_i}X_i|\mathfrak{F}_0) = 0$ and

$$\mathsf{E}I_{A_i}\mathsf{E}((X_0X_1)^{1/2}|\mathfrak{F}_0) = \mathsf{E}I_{A_i}(X_0X_1)^{1/2} = 0,$$

so that both sides of (4.7) are P-a.s. zero on $A_0 \cup A_1$. Hence we may assume $Y_i > 0$ P-a.s. for $i = 0, 1$, giving $\mathsf{E}((X_i/Y_i)|\mathfrak{F}_0) = 1$, P-a.s. so that the inequality (4.7) is equivalent to,

$$\mathsf{E}(((X_0/Y_0)(X_1/Y_1))^{1/2}|\mathfrak{F}_0) \le \frac{1}{2}(\mathsf{E}((X_0/Y_0)|\mathfrak{F}_0) + \mathsf{E}((X_1/Y_1)|\mathfrak{F}_0)),$$

or, equivalently,

$$\mathsf{E}\left(\left((X_0/Y_0)^{1/2} - (X_1/Y_1)^{1/2}\right)^2|\mathfrak{F}_0\right) \ge 0,$$

which completes the proof. ∎

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

Theorem 4.2. *Given $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ a statistic $T : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$ with $Q_i = P_i \circ T^{-1}$, $i = 0, 1$, the following statements are equivalent*

A) T *is sufficient for* $\{P_0, P_1\}$,

B) $L_i = M_i(T) \qquad \overline{P}$*-a.s. for* $i = 0, 1$,

C) $\mathsf{I}_{\mathsf{f}}(P_0 \circ T^{-1}, P_1 \circ T^{-1}) = \mathsf{I}_{\mathsf{f}}(P_0, P_1)$ *for every convex function* f,

D) $\mathsf{I}_{\mathsf{f}}(P_0 \circ T^{-1}, P_1 \circ T^{-1}) = \mathsf{I}_{\mathsf{f}}(P_0, P_1) < \infty$ *for a strictly convex function* f,

E) $\mathsf{b}_\pi(Q_0, Q_1) = \mathsf{b}_\pi(P_0, P_1)$ *for every* $0 < \pi < 1$.

Proof. The proof is carried out according to the following scheme

$$\begin{array}{ccccccccc} C) & \to & D) & \to & E) & \to & C) & & \\ A) & \to & E) & \text{and} & C) & \to & B) & \to & A) \end{array} .$$

In a first step we proof the equivalence of the conditions $C)$, $D)$, and $E)$. $C) \to D)$ is clear. $D) \to E)$: If f is strictly convex then the support $S_{\rho_{\mathsf{f}}}$ in (3.25) is the interval $(0, 1)$. Hence $E)$ follows from (3.30). $E) \to C)$ follows from (3.26).

Now we relate the conditions $C)$, $D)$ and $E)$ to the conditions $A)$ and $B)$. To prove $A) \to E)$ let φ_B be a Bayes test for $\mathsf{H}_0 : P_0$ versus $\mathsf{H}_A : P_1$. Then by (4.2) there is some ψ such that $\psi(T) = \mathsf{E}_{P_i}(\varphi_B|T)$ P_i*-a.s.* Hence $\mathsf{E}_{Q_i}\psi = \mathsf{E}_{P_i}\varphi_B$ by (4.5) and therefore

$$\begin{aligned} \mathsf{b}_\pi(Q_0, Q_1) &= \inf_\varphi (\pi \int \varphi dQ_0 + (1-\pi) \int (1-\varphi) dQ_1) \\ &\le \pi \int \psi dQ_0 + (1-\pi) \int (1-\psi) dQ_1 \\ &= \pi \int \varphi_B dP_0 + (1-\pi) \int (1-\varphi_B) dP_1 = \mathsf{b}_\pi(P_0, P_1). \end{aligned}$$

The converse inequality is trivial as the set of tests $\varphi : \mathcal{X} \to [0, 1]$ which are functions of T is a subset of all tests.

For $C) \to B)$ we use the strictly convex function $\mathsf{f}(t) = -t^{1/2}$. Then the condition $C)$ reads

$$\mathsf{H}_{1/2}(P_0 \circ T^{-1}, P_1 \circ T^{-1}) = \mathsf{H}_{1/2}(P_0, P_1).$$

Using (4.6) and (3.10) for $s = 1/2$ we see the last equality is equivalent to

$$\begin{aligned} \mathsf{H}_{1/2}(P_0, P_1) &= \mathsf{E}_{\overline{P}}(L_0 L_1)^{1/2} = \mathsf{E}_{\overline{P}}(\mathsf{E}_{\overline{P}}((L_0 L_1)^{1/2}|T)) \\ &= \mathsf{E}_{\overline{P}}((L_0^{1/2}|T)\mathsf{E}_{\overline{P}}((L_1^{1/2}|T)), \end{aligned}$$

or

$$\mathsf{E}_{\overline{P}} \left(\mathsf{E}_{\overline{P}}((L_0^{1/2}|T)\mathsf{E}_{\overline{P}}((L_1^{1/2}|T)) - \mathsf{E}_{\overline{P}}((L_0 L_1)^{1/2}|T) \right) = 0$$

for L_0, L_1 defined by (4.3). Using Lemma 4.1 and the fact $L_1 = 2 - L_0$ which follows from the definition of L_0 and L_1, we get

$$L_0 \mathsf{E}_{\overline{P}}((2 - L_0)|T) = (2 - L_0)\mathsf{E}_{\overline{P}}(L_0|T), \quad \overline{P}\text{-}a.s.$$

This together with (4.6) implies $L_0 = \mathsf{E}_{\overline{P}}(L_0|T) = M_0(T)$. To complete the proof of B) it suffices to notice that $2 - M_0(T) = M_1(T)$. $B) \to A)$: We show that $k_A(T) = \mathsf{E}_{\overline{P}}(I_A|T)$ is a version of the conditional expectation $\mathsf{E}_{P_i}(I_A|T)$ for $i = 0, 1$. It holds for every $B \in \mathfrak{B}$

$$\begin{aligned}
&\int I_B(T)\mathsf{E}_{\overline{P}}(I_A|T)dP_i = \int I_B(T)\mathsf{E}_{\overline{P}}(I_A|T)M_i(T)d\overline{P} \\
&= \int I_B(T)\mathsf{E}_{\overline{P}}(I_A M_i(T)|T)d\overline{P} \\
&= \int I_B(T)I_A M_i(T)d\overline{P} = \int I_B(T)I_A dP_i.
\end{aligned}$$

∎

Remark 4.1. *Condition B) is the factorization criterion of Neyman. The equivalence of A) and D) is an information-theoretic characterization of sufficiency which for general divergence goes back to [11] and for the special Kullback-Leibler divergence back to Kullback and Leibler [38]. For the Hellinger distance this relation can also be found in [41]. The equivalence of condition A) and E) in Theorem 4.2 is a testing-theoretic characterization of sufficiency which is due to Pfanzagl [55], who, however, used the $\alpha-$level tests instead of the Bayes tests.*

In the previous theorem we used strictly convex functions to establish the sufficiency of a statistic T. The question remains open whether this function has to be strictly convex at each point. It is clear from the continuity of the function $\pi \to \mathsf{b}_\pi(P_0, P_1)$ and the condition D) that we need only strict convexity at sufficiently many points, say on a dense subset. An equivalent formulation is obtained if we turn from one convex function to a family of convex functions. More precisely, let $\mathsf{g}_t(x)$, $x \in \mathbb{R}_+$, $t \in \mathbb{R}$ be any family of functions convex in the variable x. Further, let γ be any measure on the Borel sets of the real line. We assume that the following conditions are satisfied:

$$\begin{gathered}
(t, x) \mapsto \mathsf{g}_t(x) \text{ is measurable,} \\
\int |\mathsf{g}_t(x)|\gamma(dt) < \infty \text{ for every } x > 0.
\end{gathered}$$

Obviously $\mathsf{f}(x) := \int \mathsf{g}_t(x)\gamma(dt)$ is a convex function and, moreover, it follows from the Fubini theorem that,

$$\mathsf{I}_\mathsf{f}(P_0, P_1) = \int \mathsf{I}_{\mathsf{g}_t}(P_0, P_1)\gamma(dt).$$

The idea of constructing convex functions by a mixture of a given family has been implicitly contained in the representation of f_0 in (2.10). Indeed, if we put

$$\mathsf{g}_t(x) = \begin{cases} x - t \wedge x & \text{if} \quad x > 1 \\ t - t \wedge x & \text{if} \quad 0 < x \leq 1 \end{cases},$$

and $\gamma = \gamma_{\mathsf{f}}$ then $\mathsf{f}_0(x) = \int \mathsf{g}_t(x)\gamma(dt)$ by (2.10). Using the monotone convergence theorem it is not hard to see that,

$$D^+\mathsf{f}(x) = \int D^+\mathsf{g}_t(x)\gamma(dt),$$

which implies $\gamma_{\mathsf{f}}(B) = \int \gamma_{\mathsf{g}_t}(B)\gamma(dt)$. From here we see that f is strictly convex at each $x > 0$ if and only if for every $\varepsilon > 0$ with $x - \varepsilon > 0$,

$$\int (D^+\mathsf{g}_t(x+\varepsilon) - D^+\mathsf{g}_t(x-\varepsilon))\gamma(dt) > 0.$$

The following characterization of sufficiency is due to Mussmann [48] and can also be found in Torgersen [66].

Corollary 4.3. *Given $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ a statistic $T : \mathcal{X} \to \mathcal{Y}$ is sufficient if and only if*

$$\|tP_0 - P_1\| = \|tQ_0 - Q_1\| \tag{4.9}$$

for every $t \in D$, where $D \subseteq \mathbb{R}_+$ is dense in $\mathbb{R}_+$.

Proof. Introduce, for every $t > 0$, the convex function $\mathsf{g}_t(x) = |tx - 1|$, then,

$$I_{\mathsf{g}_t}(P_0, P_1) = \|tP_0 - P_1\|, \quad \text{and} \quad I_{\mathsf{g}_t}(Q_0, Q_1) = \|tQ_0 - Q_1\|,$$

so that the necessity of (4.9) follows from C) in Theorem 4.2. To prove the converse we introduce the measure γ by setting $\gamma = \sum_{k=1}^{\infty} 2^{-k}\delta_{a_k}$, where $D_0 = \{a_1, a_2, \dots\}$ is a subset of D that is dense in $\mathbb{R}_+$. Set $\mathsf{f}(x) = \int \mathsf{g}_t(x)\gamma(dt)$, then the support of γ_{f} is $(0, \infty)$ so that f is strictly convex. As $\|tP_0 - P_1\| \leq 1 + t$ we get,

$$\mathsf{I}_{\mathsf{f}}(P_0, P_1) = \int \mathsf{I}_{\mathsf{g}_t}(P_0, P_1)\gamma(dt) \leq \int (1+t)\gamma(dt) < \infty.$$

If (4.9) holds for every $t \in D$ then $\mathsf{I}_{\mathsf{f}}(P_0, P_1) = \mathsf{I}_{\mathsf{f}}(Q_0, Q_1)$ so that condition D) in Theorem 4.2 is satisfied and the proof is complete. ∎

So far we have compared the reduced model $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$, $Q_i = P_i \circ T^{-1}$ with the original model $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ and we characterized the sufficiency of a statistic by the fact that the two hypothesis testing problems in (4.4) are equivalent. Now we deal with statistics that are only approximately sufficient. This leads to the problem of characterizing the situations where one model is only by ε less informative than the other model. This problem leads to the general theory of ε-deficiency in

statistical decision models. Here we restrict the presentation of this theory to the testing problems in binary statistical models. Model $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$ is said to be ε-*deficient* with respect to $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$, in symbols $\mathcal{M} \succeq^{\varepsilon} \mathcal{N}$, if for every test $\psi : \mathcal{Y} \to [0,1]$ of $\mathsf{H}_0 : Q_0$ versus $\mathsf{H}_1 : Q_1$ there exists a test $\varphi : \mathcal{X} \to [0,1]$ such that,

$$\int \varphi dP_0 \leq \int \psi dQ_0 + \varepsilon \quad \text{and} \quad \int (1-\varphi) dP_1 \leq \int (1-\psi) dQ_1 + \varepsilon.$$

The next theorem clarifies the decision theoretic meaning of the f-divergences. It connects the concept of the distance being defined by the divergence with the purely decision theoretic problem of testing statistical hypotheses.

Choose arbitrary $\alpha \in (0,1)$, set $F_0(t) := P_0(p_1 \leq tp_0)$ and denote by $c_{1-\alpha}$ the $(1-\alpha)-$quantile

$$c_{1-\alpha} = \inf\{t > 0 : F_0(t) \geq 1 - \alpha\}.$$

The test,

$$\varphi_\alpha = \begin{cases} 1 & \text{if} \quad p_1 > c_{1-\alpha} p_0 \\ \gamma_\alpha & \text{if} \quad p_1 = c_{1-\alpha} p_0 \\ 0 & \text{if} \quad p_1 < c_{1-\alpha} p_0 \end{cases} \tag{4.10}$$

with the constant $\gamma_\alpha = [F_0(c_{1-\alpha}) - (1-\alpha)] / P_0(p_1 = c_{1-\alpha} p_0)$ and $0/0 := 0$, is the best $\alpha-$level test in the sense that it minimizes the error probability of the second kind $\int (1-\varphi) dP_1$ in the class Φ_α of all $\alpha-$level tests, i.e. the tests satisfy $\int \varphi dP_0 \leq \alpha$. For the proof of this famous *Neyman-Pearson Lemma* we refer to [43].

Theorem 4.4. *For the two binary models $\mathcal{M} = (\mathcal{X}, \mathfrak{A}, \{P_0, P_1\})$, $\mathcal{N} = (\mathcal{Y}, \mathfrak{B}, \{Q_0, Q_1\})$ and any $\varepsilon \geq 0$ the following conditions are equivalent*

A) $\mathcal{M} \succeq^{\varepsilon} \mathcal{N}$,

B) $\mathsf{b}_\pi(P_0, P_1) \leq \mathsf{b}_\pi(Q_0, Q_1) + \varepsilon$, *for every* $0 < \pi < 1$,

C) $\mathsf{I}_\mathsf{f}(Q_0, Q_1) - \mathsf{f}(1) \leq \mathsf{I}_\mathsf{f}(P_0, P_1) - \mathsf{f}(1) + \varepsilon \rho_\mathsf{f}((0,1))$ *for every convex function $\mathsf{f} : (0, \infty) \to \mathbb{R}$ for the measure ρ_f given by (3.24).*

Proof. The conclusion $B) \to C)$ follows directly from (3.6). Conversely, if we put $\mathsf{f}(t) = \mathsf{g}_\pi(t) := \pi \wedge (1-\pi) - (\pi t) \wedge (1-\pi)$ then g_π is convex and by (3.22)

$$\mathsf{G}_\pi(P_0, P_1) = \pi \wedge (1-\pi) - \mathsf{b}_\pi(P_0, P_1).$$

It follows that $D^+ \mathsf{g}_\pi(t) = -\pi$ if $0 < t < (1-\pi)/\pi$ and $D^+ \mathsf{g}_\pi(t) = 0$ if $t \geq (1-\pi)/\pi$. The measure γ_{g_π} in (2.7) is therefore $\gamma_{\mathsf{g}_\pi} = \pi \delta_{(1-\pi)/\pi}$, where δ_a is the delta measure concentrated at a. Hence ρ_{g_π} in (3.23) with $\mathsf{f} = \mathsf{g}_\pi$ satisfies the equalities,

$$\rho_{\mathsf{g}_\pi}((0,1)) = \int_{(0,1)} (1+t) \pi \delta_{(1-\pi)/\pi}(dt) = 1.$$

This shows that C) implies B). It remains to prove the equivalence of A) and B). This is the statement of Theorem 15.6 in [63] and we follow the proof given there. Denote by $\psi = \varphi_B$ the Bayes' test for the model $\{Q_0, Q_1\}$. If A) is satisfied then,

$$\begin{aligned}\mathsf{b}_\pi(P_0,P_1) &\le \pi\int \varphi dP_0 + (1-\pi)\int (1-\varphi)dP_1\\ &\le \pi\int \varphi_B dQ_0 + (1-\pi)\int (1-\varphi_B)dQ_1 + \varepsilon\\ &= \mathsf{b}_\pi(Q_0,Q_1)+\varepsilon,\end{aligned}$$

so that B) is satisfied. To prove the converse we consider the Lebesgue decomposition $P_0 = \gamma P_0' + (1-\gamma)P_0''$ with distributions P_0' andP_0'' that satisfy $P_0' \ll P_1$ and $P_0'' \perp P_1$ and $0 \le \gamma \le 1$. The case $\gamma = 0$ is trivial. For $\gamma > 0$ we firstly suppose that $\alpha := \int \psi dQ_0 + \varepsilon \le \gamma$. Let φ_α be the test in (4.10) and set $\pi = c_\alpha/(1+c_\alpha)$. We obtain from Lemma 3.5 that,

$$\begin{aligned}\pi\int \varphi_\alpha dP_0 + (1-\pi)\int (1-\varphi_\alpha)dP_1 &= \mathsf{b}_\pi(P_0,P_1)\\ &\le \mathsf{b}_\pi(Q_0,Q_1)+\varepsilon\\ &\le \pi\int \psi dQ_0 + (1-\pi)\int (1-\psi)dQ_1 + \varepsilon,\end{aligned}$$

which yields $\int(1-\varphi_\alpha)dP_1 \le \int(1-\psi)dQ_1$ so that the first case is completed. Suppose now $\int \psi dQ_0 + \varepsilon > \gamma$. As $P_0'' \perp P_1$, we find a test φ with $P_0''(\varphi = 0) = 1$ and $P_1(\varphi = 1) = 1$, then $\int \varphi dP_0 = \gamma < \int \psi dQ_0 + \varepsilon$ and

$$\int (1-\varphi)dP_1 = \int (1-\varphi)dP_0' = 0 \le \int (1-\psi)dQ_1 + \varepsilon$$

which completes the proof. ∎

For special classes of convex functions the total mass $\rho_{\mathsf{f}}((0,1))$ of the curvature measure ρ_{f} appearing in Theorem 3.6 can be directly expressed in terms of the function f and its derivative.

Lemma 4.5. *If* $\mathsf{f} : (0,\infty) \to \mathbb{R}$ *is a convex function such that* $\lim_{t\downarrow 0}\mathsf{f}(t) = 0$, $\lim_{t\to\infty}\mathsf{f}(t) > -\infty$, $\lim_{t\downarrow 0} D^+\mathsf{f}(t) > -\infty$ *and* f *is nonincreasing then* $\mathsf{f}^*(t) = t\mathsf{f}(1/t)$ *has these properties too and*

$$\begin{aligned}\lim_{t\to\infty}\mathsf{f}(t) = \lim_{t\downarrow 0} D^+\mathsf{f}^*(t), \quad \lim_{t\downarrow 0} D^+\mathsf{f}(t) = \lim_{t\to\infty}\mathsf{f}^*(t),\\ \rho_{\mathsf{f}}((0,1)) = -\lim_{t\to\infty}\mathsf{f}(t) - \lim_{t\downarrow 0} D^+\mathsf{f}(t) = \rho_{\mathsf{f}^*}((0,1)).\end{aligned}$$

Proof. We see from (2.13) that f^* is convex. Further,

$$\begin{aligned}D^+\mathsf{f}^*(s) &= \mathsf{f}\left(\frac{1}{s}\right) - \frac{1}{s}D^+\mathsf{f}\left(\frac{1}{s}\right)\\ &= -\int_0^{1/s}\left(D^+\mathsf{f}\left(\frac{1}{s}\right) - D^+\mathsf{f}(t)\right)dt \le 0\end{aligned} \tag{4.11}$$

because $D^+\mathsf{f}(t)$ is nondecreasing. Application of (2.5) to f^* yields that f^* is nonincreasing. Convexity of f^* shows that $D^+\mathsf{f}^*(s)$ is nondecreasing so that $\lim_{s\downarrow 0} D^+\mathsf{f}^*(s)$ exists. By assumption, $\lim_{t\to\infty}\mathsf{f}(t)$ exists and is finite. Hence,

$$A := \lim_{s\to 0}\frac{1}{s}D^+\mathsf{f}\left(\frac{1}{s}\right) = \lim_{t\to\infty} tD^+\mathsf{f}(t)$$

exists and $-\infty \le A \le 0$. If $A < 0$ then there exists $a < 0$ and t_0 such that $D^+\mathsf{f}(t) \le a/t$ for $t \ge t_0$. In this case,

$$\lim_{t\to\infty}\mathsf{f}(t) = \lim_{t\to\infty}\int_{t_0}^{t} D^+\mathsf{f}(s)ds + \mathsf{f}(t_0) \le \lim_{t\to\infty} a(\ln t - \ln t_0) + \mathsf{f}(t_0) = -\infty$$

which contradicts the assumption. Hence $\lim_{t\to\infty} tD^+\mathsf{f}(t) = 0$ which implies $\lim_{t\to\infty} D^+\mathsf{f}(t) = 0$ and $\lim_{t\downarrow 0}\mathsf{f}^*(t) = \lim_{t\downarrow 0} t\mathsf{f}(1/t) = \lim_{t\downarrow 0} t\int_0^{1/t} D^+\mathsf{f}(s)ds = 0$. The relation (4.11) yields $\lim_{s\downarrow 0} D^+\mathsf{f}^*(s) = \lim_{s\downarrow 0}\int_0^{1/s} D^+\mathsf{f}(t)dt = \lim_{t\to\infty}\mathsf{f}(t)$ and in view of $(\mathsf{f}^*)^* = \mathsf{f}$ this implies $\lim_{t\downarrow 0} D^+\mathsf{f}(t) = \lim_{t\to\infty}\mathsf{f}^*(t)$. Furthermore,

$$\begin{aligned}
&\int_{(a,b]}(1+t)\gamma_{\mathsf{f}}(dt) = (1+a)(D^+\mathsf{f}(b) - D^+\mathsf{f}(a)) + \int_{(a,b]}(\int_{(a,t)} ds)\gamma_{\mathsf{f}}(dt)\\
&= (1+a)(D^+\mathsf{f}(b) - D^+\mathsf{f}(a)) + (b-a)D^+\mathsf{f}(b) - \mathsf{f}(b) + \mathsf{f}(a)\\
&= D^+\mathsf{f}(b) - D^+\mathsf{f}(a) + bD^+\mathsf{f}(b) - aD^+\mathsf{f}(a) - \mathsf{f}(b) + \mathsf{f}(a)\\
&\rho_{\mathsf{f}}((0,1)) = \int_{(0,\infty)}(1+t)\gamma_{\mathsf{f}}(dt) = -\lim_{b\to\infty}\mathsf{f}(b) - \lim_{a\downarrow 0} D^+\mathsf{f}(a)
\end{aligned}$$

because $\lim_{a\downarrow 0}\mathsf{f}(a) = 0$ by assumption and $\lim_{b\to\infty} bD^+\mathsf{f}(b) = \lim_{b\to\infty} D^+\mathsf{f}(b) = 0$ as established above. ∎

If the convex function f satisfies the assumptions of the last theorem then we may transform the condition C) in Theorem 4.4 to the form known in decision theory as the *concave function criterion*, see Theorem 17.1 in [63]. Using L_0 and L_1 from (4.3) we introduce the likelihood ratio of P_1 with respect to P_0 by $L_{0,1} = \frac{L_1}{L_0} I_{(0,\infty)}(L_0) + \infty I_{\{0\}}(L_0)$. $M_{0,1}$ is similarly defined by means of Q_0, Q_1.

Corollary 4.6. *The conditions A), B) and C) in Theorem 4.4 are equivalent to*

$$\mathsf{E}_{P_0}\mathsf{h}(L_{0,1}) \le \mathsf{E}_{Q_0}\mathsf{h}(M_{0,1}) + \varepsilon(D^+\mathsf{h}(0) + \lim_{t\to\infty}\mathsf{h}(t)), \tag{4.12}$$

for every nondecreasing concave function $\mathsf{h} : [0,\infty) \longrightarrow \mathbb{R}$ *with* $\mathsf{h}(0) = 0$ *and the sum* $D^+\mathsf{h}(0) + \lim_{t\to\infty}\mathsf{h}(t)$ *is finite.*

Proof. The necessity follows from C) in Theorem 4.4 and (3.3) if we set $\mathsf{f}^* = -\mathsf{h}$ and use Lemma 4.5 to see that $\mathsf{f}(0) = \mathsf{f}^*(0) = 0$. To prove the sufficiency we use the concave function $\mathsf{h}_\pi(t) := ((1-\pi)t) \wedge \pi$ which is nondecreasing, concave and satisfies

$$\begin{aligned}
&D^+\mathsf{h}_\pi(0) + \lim_{t\to\infty}\mathsf{h}_\pi(t) = 1\\
&\mathsf{E}_{P_0}\mathsf{h}_\pi(L_{0,1}) = \mathsf{b}_\pi(P_0,P_1), \quad \text{and} \quad \mathsf{E}_{Q_0}\mathsf{h}_\pi(M_{0,1}) = \mathsf{b}_\pi(Q_0,Q_1),
\end{aligned}$$

Hence (4.12) implies B) in Theorem 4.4 and the proof is completed. ∎

5 Rate of Error Probabilities for Increasing Sample Sizes

In this section we apply information functionals to characterize the quality of statistical tests for increasing sample size. It is intuitively clear that decisions are the easier the larger is the distinction between the distributions in the statistical model. In this section we study the rate of convergence at which the error probabilities in testing a simple hypothesis versus a simple alternative tends to zero when the sample size tends to infinity. Our aim is to quantify this rate. This will lead to the famous statistical results known as the *Theorems of Chernoff and Stein*.

To describe the convergence of the error probabilities to zero we will use the concept of exponential rate. This means that for any sequence of nonnegative numbers a_n tending to zero we characterize the rate of convergence to zero by the value $\mathsf{R} := -\lim_{n\to\infty} \frac{1}{n} \ln a_n$ provided the limit exists. We call the nonnegative value R the *exponential rate* of the sequence $\{a_n\}$. If $0 < \mathsf{R} < \infty$ then we obtain $a_n = \exp\{-n(\mathsf{R} - \varepsilon_n)\}$ for $\varepsilon_n := \mathsf{R} + \frac{1}{n} \ln a_n$ tending to zero. Although the value R is useful to reflect the convergence rate, it is only a rough measure describing nothing more than the exponential rate. Constants are suppressed, e.g. for any $c > 0$, $\lim_{n\to\infty} \frac{1}{n} \ln(ca_n) = \lim_{n\to\infty} \frac{1}{n} \ln a_n$ holds. Moreover, the exponential rate characterizes only the worst case in the following sense. Let a_n and b_n be two sequences of nonnegative numbers. As $\max(a_n, b_n) \le a_n + b_n \le 2\max(a_n, b_n)$, we see that the exponential rate for $\max(a_n, b_n)$ exists if and only if the exponential rate of $a_n + b_n$ exists and in this case the two rates are identical.

For the problem of testing the hypothesis $\mathsf{H}_0 : P_0$ versus $\mathsf{H}_1 : P_1$ we denote by Φ_α the set of all $\alpha-$level tests, i.e. the set of all tests φ such that $\int \varphi dP_0 \le \alpha$. By

$$\mathsf{d}_\alpha(P_0, P_1) = \inf_{\varphi\in\Phi_\alpha} \left\{ \int (1-\varphi) dP_1, \varphi \in \Phi_\alpha \right\}$$

we denote the second kind error probability of the best $\alpha-$level test. We know from the Neyman-Pearson lemma that the test in (4.10) attains the infimum

$$\mathsf{d}_\alpha(P_0, P_1) = \int (1 - \varphi_\alpha) dP_1.$$

Now we relate the minimal error probability of the second kind to the minimal Bayes' error. This relation is well known and can be found e.g. in [66, p. 590-591].

Lemma 5.1. *We have,*

$$\mathsf{b}_\pi(P_0, P_1) = \min_{0<\alpha<1} [\pi\alpha + (1-\pi)\mathsf{d}_\alpha(P_0, P_1)], \quad \pi \in (0,1), \tag{5.1}$$

$$\mathsf{d}_\alpha(P_0, P_1) = \max_{0<\pi<1} \frac{1}{1-\pi} (\mathsf{b}_\pi(P_0, P_1) - \pi\alpha), \quad \alpha \in (0,1). \tag{5.2}$$

Proof. For fixed $\pi \in (0,1)$ the inequality,

$$\mathsf{b}_\pi(P_0, P_1) \le \pi\alpha + (1-\pi)\mathsf{d}_\alpha(P_0, P_1), \quad \alpha \in (0,1),$$

follows from the definition of $\mathsf{d}_\alpha(P_0, P_1)$. It implies,

$$\mathsf{b}_\pi(P_0, P_1) \leq \inf_{0<\alpha<1}[\pi\alpha + (1-\pi)\mathsf{d}_\alpha(P_0, P_1)], \quad \pi \in (0,1), \tag{5.3}$$

$$\mathsf{d}_\alpha(P_0, P_1) \geq \sup_{0<\pi<1} \frac{1}{1-\pi}(\mathsf{b}_\pi(P_0, P_1) - \pi\alpha), \quad \alpha \in (0,1). \tag{5.4}$$

If $\alpha_0 \in (0,1)$ is fixed, then φ_{α_0} in (4.10) is, according to Lemma 3.5, a Bayes' test for $\pi = c_{1-\alpha_0}/(1 + c_{1-\alpha_0})$. This yields,

$$\pi\alpha_0 + (1-\pi)\mathsf{E}_1(1 - \varphi_{\alpha_0}) = \mathsf{b}_\pi(P_0, P_1),$$

and the proof is complete. ∎

To establish bounds for $\mathsf{d}_\alpha(P_0, P_1)$ we use the Hellinger integrals in (3.10). The following result has been independently established by Kraft and Plachky [35] and by Österreicher [50]. For the proof we need the elementary inequalities,

$$z \vee 1 \leq az^s + 1, \quad s > 1, z \geq 0, a = (s-1)^{s-1}s^{-s}, \tag{5.5}$$

$$z \wedge 1 \leq z^s, \quad 0 < s < 1, z \geq 0. \tag{5.6}$$

Lemma 5.2. *For every $0 < \alpha < 1$ the second kind error probability of the best $\alpha-$level test for testing P_0 versus P_1 satisfies the inequalities,*

$$\mathsf{d}_\alpha(P_0, P_1) \leq (1-s)\left(\frac{s}{\alpha}\right)^{s/(1-s)} (\mathsf{H}_s(P_0, P_1))^{1/(1-s)}, \quad 0 < s < 1, \tag{5.7}$$

$$\mathsf{d}_\alpha(P_0, P_1) \geq (1-\alpha)^{t/(t-1)}(\mathsf{H}_t(P_0, P_1))^{-1/(t-1)}, \quad 1 < t < \infty. \tag{5.8}$$

Proof. We have,

$$\int ((p_1 - cp_0) \vee 0)d\mu = 1 - \int ((cp_0) \wedge p_1)d\mu = \int ((cp_0) \vee p_1)d\mu - c. \tag{5.9}$$

Taking $c = \pi/(1-\pi)$ we get, from (5.2),

$$\begin{aligned}\mathsf{d}_\alpha(P_0, P_1) &= \sup_{c>0}\left[\int (cp_0) \wedge p_1 d\mu - c\alpha\right] \\ &= \sup_{c>0}\left[1 - \int (cp_0) \vee p_1 d\mu + c(1-\alpha)\right].\end{aligned}$$

The application of (5.5) yields,

$$\begin{aligned}\int ((cp_0) \vee p_1)d\mu &\leq c^s(s-1)^{s-1}s^{-s}\int p_0^s p_1^{1-s} I_{(0,\infty)}(p_1)d\mu + 1 \\ &\leq c^s(s-1)^{s-1}s^{-s}\mathsf{H}_s(P_0, P_1) + 1,\end{aligned}$$

and

$$\begin{aligned}\mathsf{d}_\alpha(P_0, P_1) &\geq \sup_{c>0}\left[c(1-\alpha) - c^s(s-1)^{s-1}s^{-s}\mathsf{H}_s(P_0, P_1)\right] \\ &= (1-\alpha)\sup_{c>0}[c - c^s d] = (1-\alpha)(s-1)s^{-s/(s-1)}d^{-1/(s-1)}.\end{aligned}$$

Inserting the value $d = (1-\alpha)^{-1}(s-1)^{s-1}s^{-s}\mathsf{H}_s(P_0,P_1)$ on the right hand side we get (5.8). The proof of (5.7) is similar if we use (5.6) to estimate $\int (cp_0)\wedge p_1)d\mu$. ∎

Now we characterize the exponential rate of the probability of an error of the second kind if the error probabilities of the first kind are supposed to be bounded by a fixed α.

Consider the case of an increasing sample size. Let $X_1,\dots,X_n$ be a sample of size n consisting of independent and identically distributed random variables where the joint distribution is $P_0^{\otimes n}$ or $P_1^{\otimes n}$. Here and in the sequel $\otimes$ is the symbol for product measure. We want to study the hypotheses testing problem $\mathsf{H}_0 : P_0^{\otimes n}$ versus $\mathsf{H}_A : P_1^{\otimes n}$. Our aim is to investigate the rate of convergence to zero of the error probabilities of the second kind of level α tests.

Theorem 5.3. *If the Kullback-Leibler divergence $\mathsf{K}_1(P_0,P_1)$ is finite then for any $0<\alpha<1$ the error probability of the second kind $\mathsf{d}_\alpha(P_0^{\otimes n},P_1^{\otimes n})$ of the best level α-test φ_n for testing $\{P_0^{\otimes n}\}$ versus $\{P_1^{\otimes n}\}$ satisfies*

$$-\lim_{n\to\infty}\frac{1}{n}\ln \mathsf{d}_\alpha(P_0^{\otimes n},P_1^{\otimes n}) = \mathsf{K}_1(P_0,P_1). \tag{5.10}$$

Proof. We give a proof only under the restricted condition that for some $s_0>1$, $\mathsf{H}_{s_0}(P_0,P_1)<\infty$ holds. For the general case we refer to [35]. A simple consequence of the fact that the density of a product measure with respect to another product measure is just the product of the densities, is that the Hellinger integral H_s in (3.10) satisfies

$$\mathsf{H}_s(P_0^{\otimes n},P_1^{\otimes n}) = (\mathsf{H}_s(P_0,P_1))^n. \tag{5.11}$$

Hence, by the application of (5.7) and (5.8) for $0<s<1<t<s_0$,

$$\begin{aligned} -\frac{1}{t-1}\ln \mathsf{H}_t(P_0,P_1) &\le \liminf_{n\to\infty}\frac{1}{n}\ln[\mathsf{d}_\alpha(P_0^{\otimes n},P_1^{\otimes n})] \\ &\le \limsup_{n\to\infty}\frac{1}{n}\ln[\mathsf{d}_\alpha(P_0^{\otimes n},P_1^{\otimes n})] \\ &\le \frac{1}{1-s}\ln \mathsf{H}_s(P_0,P_1). \end{aligned} \tag{5.12}$$

The function k_s in (3.4) satisfies $\mathsf{k}_s(1)=\mathsf{k}_s'(1)=0$ and $\mathsf{k}_s''(x)=x^{s-2}$. Hence,

$$\begin{aligned} &0\le \mathsf{k}_s''(x)\le \mathsf{k}_{1/2}''(x)+\mathsf{k}_{s_0}''(x), \quad \frac{1}{2}\le s\le s_0, x\ge 0, \quad \text{therefore} \\ &0\le \mathsf{k}_s(x)\le \mathsf{k}_{1/2}(x)+\mathsf{k}_{s_0}(x), \quad \frac{1}{2}\le s\le s_0, x\ge 0. \end{aligned}$$

Thus, from the Lebesgue theorem,

$$\lim_{s\uparrow 1}\mathsf{K}_s(P_0,P_1)=\lim_{s\downarrow 1}\mathsf{K}_s(P_0,P_1)=\mathsf{K}_1(P_0,P_1).$$

Using (3.11) we arrive at

$$\lim_{s\uparrow 1}\frac{1}{1-s}\ln \mathsf{H}_s(P_0,P_1)=\lim_{s\downarrow 1}\frac{1}{1-s}\ln \mathsf{H}_s(P_0,P_1)=-\mathsf{K}_1(P_0,P_1).$$

Combining this statement with inequality (5.12) we get the desired result. ∎

Chernoff [8] and Kullback [39] refer for the statement (5.10) to an unpublished paper by C. Stein. This statement is, therefore, known as Stein's theorem.

In the previous theorem we assumed the Kullback-Leibler distance $\mathsf{K}_1(P_0,P_1)$ to be finite. The infinite case $\mathsf{K}_1(P_0,P_1)=\infty$ was studied by Janssen [28].

Example 8. *We illustrate the above theorem for distributions from the same exponential family, say $P_0=P_{\theta_0}$, $P_1=P_{\theta_1}$. Example 7 gives,*

$$\mathsf{H}_s(P_{\theta_0},P_{\theta_1})=\exp\{-(sK(\theta_0)+(1-s)K(\theta_1)-K(s\theta_0+(1-s)\theta_1))\},$$

for $s\theta_0+(1-s)\theta_1\in\Delta$. Due to the convexity of Δ the condition $s\theta_0+(1-s)\theta_1\in\Delta$ is certainly fulfilled for $0<s<1$. Hence, by Theorem 5.2,

$$\begin{aligned}-\lim_{n\to\infty}\frac{1}{n}\ln \mathsf{d}_\alpha(P_{\theta_0}^{\otimes n},P_{\theta_1}^{\otimes n})&=\mathsf{K}_1(P_{\theta_0},P_{\theta_1})\\&=\langle\nabla K(\theta_0),\theta_0-\theta_1\rangle+K(\theta_1)-K(\theta_0).\end{aligned}$$

Note that in the example under consideration, even the restrictive condition $\mathsf{H}_{s_0}(P_{\theta_0},P_{\theta_1})<\infty$ for some $s_0>1$ is fulfilled if $\theta_0\in\Delta^0$. Then there is some $s_0>1$ such that $s_0\theta_0+(1-s_0)\theta_1\in\Delta$ which yields $\mathsf{H}_{s_0}(P_{\theta_0},P_{\theta_1})<\infty$. If $P_\theta=\mathsf{N}(\theta,\sigma_0^2)$ then,

$$\mathsf{K}_1\left(\mathsf{N}_{\theta_0,\sigma_0^2},\mathsf{N}_{\theta_1,\sigma_0^2}\right)=\frac{(\theta_0-\theta_1)^2}{2\sigma_0^2}$$

by (3.18). This means that

$$-\lim_{n\to\infty}\frac{1}{n}\ln \mathsf{d}_\alpha\left(\mathsf{N}_{\theta_0,\sigma_0^2}^{\otimes n},\mathsf{N}_{\theta_1,\sigma_0^2}^{\otimes n}\right)=\frac{(\theta_0-\theta_1)^2}{2\sigma_0^2}. \tag{5.13}$$

We compare this rate with the exact value of $\mathsf{d}_\alpha\left(\mathsf{N}_{\theta_0,\sigma_0^2}^{\otimes n}\mathsf{N}_{\theta_1,\sigma_0^2}^{\otimes n}\right)$ which is nothing but the probability of an error of the second kind of the Gauss test being given by the formula

$$\mathsf{d}_\alpha\left(\mathsf{N}_{\theta_0,\sigma_0^2}^{\otimes n},\mathsf{N}_{\theta_1,\sigma_0^2}^{\otimes n}\right)=\Phi_{0,1}(u_{1-\alpha}-\sigma_0 n^{-1/2}(\theta_1-\theta_0)),$$

where $\Phi_{0,1}$ is the distribution function of the standard normal distribution. In order to evaluate the right hand term for large n we use Mill's ratio, i.e. the inequality,

$$\frac{|x|}{1+x^2}\le\frac{\Phi_{0,1}(x)}{\varphi_{0,1}(x)}\le\frac{1}{|x|}\quad\text{for } x<0.$$

We get for $a_n = u_{1-\alpha} - \frac{\sqrt{n}}{\sigma_0}(\theta_1 - \theta_0)$ *and all sufficiently large* n,

$$\frac{|a_n|}{(1+a_n^2)\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{a_n^2}{2\sigma_0^2}\right\} \le \mathsf{d}_\alpha\left(\mathsf{N}^{\otimes n}_{\theta_0,\sigma_0^2}, \mathsf{N}^{\otimes n}_{\theta_1,\sigma_0^2}\right) \le \frac{1}{|a_n|\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{a_n^2}{2\sigma_0^2}\right\}.$$

As $\lim_{n\to\infty}\frac{1}{n}\ln|a_n| = 0$ *we again obtain, via another method, the result (5.13). From this example we also see that the exponential rate provides an asymptotic expression ignoring the factor* $(|a_n|\sqrt{2\pi}\sigma_0)^{-1} \sim n^{-1/2}$. *This means that in our example the real error probability of the second kind tends by the factor* $n^{-1/2}$ *faster to zero than indicated by the exponential rate.*

Now we turn to the exponential rate for the Bayesian as well as for the minimax risk. Our aim is to calculate the exponential rates of the sequence of Bayes' risks $\mathsf{b}_\pi\left(P_0^{\otimes n}, P_1^{\otimes n}\right)$, see (3.21), and of the minimax risks given by

$$\mathsf{m}_\rho\left(P_0^{\otimes n}, P_1^{\otimes n}\right) = \inf_\varphi\left(\max\left(\rho\int\varphi dP_0^{\otimes n}, (1-\rho)\int(1-\varphi)dP_0^{\otimes n}\right)\right).$$

Using the inequalities,

$$\mathsf{m}_\rho\left(P_0^{\otimes n}, P_1^{\otimes n}\right) \le \mathsf{b}_\rho\left(P_0^{\otimes n}, P_1^{\otimes n}\right) \le 2\mathsf{m}_\rho\left(P_0^{\otimes n}, P_1^{\otimes n}\right) \tag{5.14}$$

we immediately see that the exponential rates of the two sequences of risks are identical provided they exist. As a preparation of the next result we bound the error probabilities by terms of Hellinger integrals. If φ is any likelihood ratio test for P_0 versus P_1 rejecting the null hypothesis at the critical value $1/c$ then,

$$\int\varphi dP_0 \le c^{1-s}\int p_0^s p_1^{1-s}d\mu = c^{1-s}\mathsf{H}_s(P_0,P_1). \tag{5.15}$$

Similarly, $\mathsf{b}_\pi(P_0,P_1)$ in (3.21) satisfies

$$\mathsf{b}_\pi(P_0,P_1) \le \pi^s(1-\pi)^{1-s}\mathsf{H}_s(P_0,P_1). \tag{5.16}$$

A lower bound for $\mathsf{b}_\pi(P_0,P_1)$ is given by the inequality,

$$\pi^s(1-\pi)^{1-s}\mathsf{H}_s(P_0,P_1) \le (\mathsf{b}_\pi(P_0,P_1))^{\pi\wedge(1-\pi)}(1-\mathsf{b}_\pi(P_0,P_1))^{\pi\vee(1-\pi)} \tag{5.17}$$

established by Vajda [67].

Lemma 5.4. *If* P_0 *and* P_1 *are neither identical nor mutually singular then the functions* $s \mapsto \mathsf{H}_s(P_0,P_1)$ *and* $s \mapsto \ln\mathsf{H}_s(P_0,P_1)$ *are strictly convex and infinitely often differentiable in* $(0,1)$.

Proof. If $T : \mathcal{X} \to \mathbb{R}$ is any statistic then by the same arguments as in Example 7 one can see that $\{s : \int \exp\{sT\}d\mu < \infty\}$ is an interval and one can show (see e.g. [5]) that the function $s \mapsto \int \exp\{sT\}d\mu$ is analytic in the interior of this interval and the derivatives can be carried out under the integral. Applying this fact to $d\nu = I_{\{p_0>0,p_1>0\}}dP_1$ and $T = \ln(p_0/p_1)$ we find that the function $s \mapsto \mathsf{H}_s(P_0, P_1)$ is infinitely often differentiable in $(0, 1)$ and if P_0, P_1 are neither identical nor mutually singular then we have,

$$\frac{d^2}{ds^2}\mathsf{H}_s(P_0, P_1) = \int I_{\{p_0\wedge p_1>0\}} \exp\{s\ln(p_0/p_1)\}(\ln(p_0/p_1))^2 dP_1 > 0.$$

The convexity of $\ln \mathsf{H}_s(P_0, P_1)$ follows from the Hölder inequality applied to $\mathsf{H}_s(P_0, P_1) = \int (p_0/p_1)^s dP_1$. ∎

We introduce a family of distributions $(P_\theta)_{\theta\in(0,1)}$ by

$$dP_\theta = (\mathsf{H}_\theta(P_0, P_1))^{-1} p_0^\theta p_1^{1-\theta} d\mu, \quad 0 < \theta < 1, \tag{5.18}$$

for which

$$\mathsf{H}_s(P_\theta, P_1) = \frac{\mathsf{H}_{s\theta}(P_0, P_1)}{\mathsf{H}^s_\theta(P_0, P_1)}. \tag{5.19}$$

If we have a sample of size n then by (5.11) and (5.16),

$$\limsup_{n\to\infty} \frac{1}{n} \ln(\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n})) \le \inf_{0<s<1}(\ln \mathsf{H}_s(P_0, P_1)). \tag{5.20}$$

Our aim is to show that in (5.20) holds in fact the equality which means that $-\ln \mathsf{H}_{s^*}(P_0, P_1)$ is the exponential rate of $\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n})$. Next follows the corresponding result which is due to Chernoff [7].

Theorem 5.5. *The Bayes' risk* $\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n})$ *for testing* $\{P_0^{\otimes n}\}$ *versus* $\{P_1^{\otimes n}\}$ *with prior* $(\pi,1-\pi)$ *satisfies, for every* $0 < \pi < 1$*, the relation,*

$$\lim_{n\to\infty} \frac{1}{n} \ln(\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n})) = \inf_{0<s<1}(\ln \mathsf{H}_s(P_0, P_1)).$$

Corollary 5.6. *The minimax risk* $\mathsf{m}_\rho(P_0^{\otimes n}, P_1^{\otimes n})$ *satisfies, for every* $0 < \rho < 1$*, the relation,*

$$\lim_{n\to\infty} \frac{1}{n} \ln(\mathsf{m}_\rho(P_0^{\otimes n}, P_1^{\otimes n})) = \inf_{0<s<1}(\ln \mathsf{H}_s(P_0, P_1)).$$

Proof. The case in which P_0 and P_1 are mutually singular is trivial as in this case $\ln \mathsf{H}_s(P_0, P_1) = -\infty$ and $\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n}) = 0$. Similarly, if $P_0 = P_1$ then $\ln \mathsf{H}_s(P_0, P_1) = 0$. Therefore we can assume that P_0 and P_1 are neither mutually singular nor identical. First of all notice that in view of (3.21) we have,

$$((\pi \wedge (1-\pi))\mathsf{b}_{1/2}(P_0^{\otimes n}, P_1^{\otimes n}) \le \mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n}) \le 2\mathsf{b}_{1/2}(P_0^{\otimes n}, P_1^{\otimes n}),$$

so that we may assume without loss of generality $\pi = 1/2$. As $\lim_{s\downarrow 0} \mathsf{H}_s(P_0, P_1) = P_1(p_0 > 0)$ and $\lim_{s\uparrow 1} \mathsf{H}_s(P_0, P_1) = P_0(p_1 > 0)$ are finite the function $s \mapsto \mathsf{H}_s(P_0, P_1)$ can be extended to a continuous function on $[0, 1]$ that attains the infimum in, say $s^* \in [0, 1]$. The strict convexity of $\mathsf{H}_s(P_0, P_1)$ in $(0, 1)$ yields

$$\mathsf{H}_s(P_0, P_1) > \mathsf{H}_{s^*}(P_0, P_1) \quad for \;\; s \neq s^*, 0 < s < 1.$$

If $s^* \in \{0, 1\}$ then we may assume $s^* = 0$ as for $s^* = 1$ we may interchange the role of P_0 and P_1. Denote by φ_n a likelihood ratio test for $P_0^{\otimes n}$ versus $P_1^{\otimes n}$ at $c = 1$, see (3.20), then by the construction of P_θ in (5.18) the test φ_n is also a likelihood ratio test for $P_\theta^{\otimes n}$ versus $P_1^{\otimes n}$ at $c_n = [\mathsf{H}_\theta(P_0^{\otimes n}, P_1^{\otimes n})]^{-1}$. Fix θ with $s^* < \theta < 1$ and $0 < s < 1$ with $s^* < s\theta < \theta$. Then $\mathsf{H}_{s\theta}(P_0, P_1) < \mathsf{H}_\theta(P_0, P_1)$. Hence, by (5.15) and (5.19),

$$(5.21)\;\; \alpha_n := \int \varphi_n dP_\theta^{\otimes n} \leq c_n^{1-s} \mathsf{H}_s(P_\theta^{\otimes n}, P_1^{\otimes n}) = \left[\frac{\mathsf{H}_{s\theta}(P_0, P_1)}{\mathsf{H}_\theta(P_0, P_1)}\right]^n \to 0, \quad \text{as } n \to \infty.$$

The inequality (5.8) yields, for $1 < t < \frac{1}{\theta}$,

$$\frac{1}{n} \ln\left(1 - \int \varphi_n dP_1^{\otimes n}\right) \geq \frac{t}{t-1} \frac{1}{n} \ln(1 - \alpha_n) - \frac{1}{t-1} \ln \mathsf{H}_t(P_\theta, P_1).$$

By (5.21)

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \ln\left(1 - \int \varphi_n dP_1^{\otimes n}\right) & \\ \geq \frac{1}{1-t} \ln \mathsf{H}_t(P_\theta, P_1) & \\ = \frac{1}{1-t}[\ln \mathsf{H}_{t\theta}(P_0, P_1) - t \ln \mathsf{H}_\theta(P_0, P_1)], \quad 1 < t < \frac{1}{\theta}. & \end{aligned}$$

If $s^* = 0$ then we take $\theta \downarrow 0$ and obtain

$$\liminf_{n\to\infty} \frac{1}{n} \ln\left(1 - \int \varphi_n dP_1^{\otimes n}\right) \geq \ln \mathsf{H}_{s^*}(P_\theta, P_1) = \inf_{0<s<1} \ln \mathsf{H}_s(P_0, P_1).$$

Suppose now $0 < s^* < 1$. The function $s \mapsto \mathsf{H}_s(P_0, P_1)$ is continuous in $(0, 1)$. Taking $\theta \downarrow s^*$ we get, for $1 < t < \frac{1}{s^*}$,

$$\liminf_{n\to\infty} \frac{1}{n} \ln\left(1 - \int \varphi_n dP_1^{\otimes n}\right) \geq \frac{1}{1-t}[\ln \mathsf{H}_{ts^*}(P_0, P_1) - t \ln \mathsf{H}_{s^*}(P_0, P_1)].$$

Since s^* is a local minimum point of the differentiable function $s \mapsto \mathsf{H}_s(P_0, P_1)$, the derivative $\frac{d}{ds}\mathsf{H}_s(P_0, P_1)$ vanishes at s^*, therefore,

$$\lim_{t\downarrow 1} \frac{1}{1-t}[\ln \mathsf{H}_{ts^*}(P_0, P_1) - t \ln \mathsf{H}_{s^*}(P_0, P_1)] \;\; = \ln \mathsf{H}_{s^*}(P_0, P_1)$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \ln \left(1 - \int \varphi_n dP_1^{\otimes n}\right) \geq \ln \mathsf{H}_{s^*}(P_0, P_1).$$

The likelihood ratio test φ_n for $P_0^{\otimes n}$ versus $P_1^{\otimes n}$ at 1 is a likelihood ratio test for $P_1^{\otimes n}$ versus $P_0^{\otimes n}$ at 1. Hence, by the last inequality and the trivial fact,

$$\ln \mathsf{H}_{s^*}(P_0, P_1) = \inf_{0<s<1} \ln \mathsf{H}_s(P_0, P_1) = \inf_{0<s<1} \ln \mathsf{H}_{1-s}(P_0, P_1)$$

we get

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \ln \left(\int \varphi_n dP_0^{\otimes n}\right) &= \liminf_{n\to\infty} \frac{1}{n} \ln \left(1 - \int (1-\varphi_n) dP_0^{\otimes n}\right) \\ &\geq \ln \mathsf{H}_{s^*}(P_0, P_1).\end{aligned}$$

As the likelihood ratio test φ_n for $P_0^{\otimes n}$ versus $P_1^{\otimes n}$ at $c = 1$ is a Bayes' test with prior $\pi = \frac{1}{2}$ we obtain, from the concavity of $\ln x$,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \ln(\mathsf{b}_{1/2}(P_0^{\otimes n}, P_1^{\otimes n})) &= \liminf_{n\to\infty} \frac{1}{n} \ln \left(\frac{1}{2}\int \varphi_n dP_0^{\otimes n} + \frac{1}{2}\int (1-\varphi_n) dP_1^{\otimes n}\right) \\ &\geq \inf_{0<s<1} \ln \mathsf{H}_s(P_0, P_1).\end{aligned}$$

The opposite inequality has been already established in (5.20). The proof is complete. The proof of the corollary follows from inequality (5.14). ∎

Remark 5.1. *There is a large number of papers dealing with the exponential rate of convergence for error probabilities for increasing sample sizes. Without giving a complete list we just remark that some of these papers study stochastic processes instead of i.i.d. samples, see [34], [69] or [46]. Linkov used the concept of Hellinger processes to study the error probabilities. In general large deviation theory, arbitrary sequences of distributions are studied and the exponential rate of error probabilities is expressed with the help of the asymptotic behavior of the moment generating functions of the log likelihood, which is nothing but the Hellinger integral, see [62], [19], and other books on large deviations.*

The quantity

$$\mathsf{C}(P_0, P_1) = -\inf_{0<s<1} \ln \mathsf{H}_s(P_0, P_1) \tag{5.22}$$

is called the *Chernoff index* of P_0 and P_1. It gives the exponential rate at which the Bayesian risk $\mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n})$ tends to zero. Using the Chernoff index the statement of Theorem 5.5 and Corollary 5.6 may be written as,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \ln \mathsf{b}_\pi(P_0^{\otimes n}, P_1^{\otimes n}) &= \lim_{n\to\infty} \frac{1}{n} \ln \mathsf{m}_\rho(P_0^{\otimes n}, P_1^{\otimes n}) \\ &= -\mathsf{C}(P_0, P_1).\end{aligned} \tag{5.23}$$

Example 9. *To illustrate the above statement we suppose that $P_\theta, \theta \in \Delta \subseteq \mathbb{R}^d$ is an exponential family with natural parameter θ and generating statistic $T : \mathcal{X} \to \mathbb{R}^d$. It follows then, by (3.15), that,*

$$\mathsf{C}(P_{\theta_1}, P_{\theta_2}) = \inf_{0<s<1}\{sK(\theta_1) + (1-s)K(\theta_2) - K(s\theta_1 + (1-s)\theta_2))\}.$$

If $P_\theta = \mathsf{N}(\theta, \sigma^2)$ then by (3.17),

$$\ln \mathsf{H}_s(\mathsf{N}(\theta_0, \sigma^2), \mathsf{N}(\theta_1, \sigma^2)) = -\frac{1}{2}s(1-s)(\theta_0 - \theta_1)^2)/\sigma^2, \quad \textit{and}$$

$$\mathsf{C}(\mathsf{N}(\theta_0, \sigma^2), \mathsf{N}(\theta_1, \sigma^2)) = (\theta_0 - \theta_1)^2/(8\sigma^2). \tag{5.24}$$

The next example is a simple decision model with a special symmetry.

Example 10. *Assume we are given two distributions Q_0, Q_1 on the sample space $(\mathcal{X}, \mathfrak{A})$ and two samples $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ where the (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. with common distribution which is either $Q_0 \otimes Q_1$ or $Q_1 \otimes Q_0$ and we have to decide between the two cases. Set $P_0 = Q_0 \otimes Q_1, P_1 = Q_1 \otimes Q_0$, then by Corollary 5.6,*

$$\lim_{n\to\infty} \frac{1}{n} \ln(\mathsf{m}_\rho(P_0^{\otimes n}, P_1^{\otimes n})) = -\mathsf{C}(Q_0 \otimes Q_1, Q_1 \otimes Q_0).$$

To evaluate the Chernoff index we use (5.11) and $\mathsf{H}_s(Q_1, Q_0) = \mathsf{H}_{1-s}(Q_0, Q_1)$ giving,

$$\begin{aligned}
\mathsf{C}(Q_0 \otimes Q_1, Q_1 \otimes Q_0) &= -\inf_{0<s<1}[\ln \mathsf{H}_s(Q_0 \otimes Q_1, Q_1 \otimes Q_0)] \\
&= -\inf_{0<s<1}[\ln \mathsf{H}_s(Q_0, Q_1) + \ln \mathsf{H}_{1-s}(Q_0, Q_1)] \\
&= -2 \ln \mathsf{H}_{\frac{1}{2}}(Q_0, Q_1)
\end{aligned}$$

as $\ln \mathsf{H}_s(Q_0, Q_1) + \ln \mathsf{H}_{1-s}(Q_0, Q_1)$ is convex in view of Lemma 5.4 and symmetric around $s = \frac{1}{2}$. Consequently we have proved that,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathsf{m}_\rho((Q_0 \otimes Q_1)^{\otimes n}, (Q_1 \otimes Q_0)^{\otimes n}) = 2 \ln \mathsf{H}_{\frac{1}{2}}(Q_0, Q_1). \tag{5.25}$$

We now apply Theorem 5.5 to the classification problem. Assume that $(\mathcal{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Delta})$ is a statistical model with finite parameter set $\Delta = \{1, \ldots, m\}$ and we want to estimate the parameter θ, i.e. we want to find the true distribution. We call this decision problem a *classification problem*. A *classification rule* is a vector $q(x) = (q_1(x), \ldots, q_m(x))$ consisting of measurable functions $q_i : \mathcal{X} \to [0, 1]$ with $\sum_{i=1}^m q_i(x) = 1$ for every $x \in \mathcal{X}$. Let $q_i(x)$ be the probability of selecting the distribution P_i if x was observed. The probability of a false classification is called the risk and is given by

$$\mathsf{R}(\theta, q) = \int (1 - q_\theta(x)) P_\theta(dx).$$

If $\theta_1 \neq \theta_2$ then $\sum_{i=1}^m q_i(x) = 1$ implies $1 - q_{\theta_2}(x) \geq q_{\theta_1}(x)$ and

$$\begin{aligned}(5.26)\qquad \max(\mathsf{R}(\theta_1, q), \mathsf{R}(\theta_2, q)) &= \max\left(\int (1 - q_{\theta_1}) dP_{\theta_1}, \int (1 - q_{\theta_2}) dP_{\theta_2}\right) \\ &\geq \max\left(\int (1 - q_{\theta_1}) dP_{\theta_1}, \int q_{\theta_1} dP_{\theta_2}\right) \\ &\geq 2\mathsf{m}_{1/2}(P_{\theta_1}, P_{\theta_2}).\end{aligned}$$

Now we suppose that a sample of size n is available so that we deal with the sequence of statistical models $(\mathcal{X}^n, \mathfrak{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta \in \Delta})$. If $q^{(n)}$ is any sequence of classification rules then,

$$\max(\mathsf{R}(\theta_1, q^{(n)}), \mathsf{R}(\theta_2, q^{(n)})) \geq 2\mathsf{m}_{1/2}(P_{\theta_1}^{\otimes n}, P_{\theta_2}^{\otimes n}).$$

If we apply Theorem 5.5 to the right hand term we obtain the following statement:

$$\begin{aligned}(5.27)\quad \liminf_{n\to\infty} \frac{1}{n} \ln\left(\max_{1\leq\theta\leq m} \mathsf{R}(\theta, q^{(n)})\right) &\geq \liminf_{n\to\infty} \frac{1}{n} \ln\left(\max_{1\leq\theta_1\neq\theta_2\leq m} \mathsf{m}_{1/2}(P_{\theta_1}^{\otimes n}, P_{\theta_2}^{\otimes n})\right) \\ &\geq \max_{1\leq\theta_1\neq\theta_2\leq m} (-\mathsf{C}(P_{\theta_1}, P_{\theta_2})) \\ &= -\min_{1\leq\theta_1\neq\theta_2\leq m} \mathsf{C}(P_{\theta_1}, P_{\theta_2}),\end{aligned}$$

where $\mathsf{C}(P_{\theta_1}, P_{\theta_2})$ denotes the Chernoff index of P_{θ_1} and P_{θ_2} from (5.22). The natural desire is to search for a classification rule that attains asymptotically the lower bound for the risk in (5.27). Roughly speaking, we show that the maximum likelihood classification rule has this property. More precisely, fix $\mu \in \mathcal{M}^\sigma(\mathfrak{A})$ which dominates all P_θ and set $p_{n,\theta} = dP_\theta^{\otimes n}/d\mu^{\otimes n}$. Put for any $x \in \mathcal{X}^n$,

$$B_n(x) = \left\{\eta \in \{1, \ldots, m\} : p_{n,\eta}(x) = \max_{1\leq\theta\leq m} p_{n,\theta}(x)\right\}.$$

If $B_n(x) \subseteq \{1, \ldots, m\}$ is a singleton then we choose this value, otherwise we select randomly a point from $B_n(x)$ according to the uniform distribution on $B_n(x)$. This means that the *maximum likelihood classification rule* is given by,

$$(5.28)\qquad q_{\mathrm{ML}}^{(n)}(x) = \frac{1}{|B_n(x)|}(I_{B_n(x)}(1), \ldots, I_{B_n(x)}(m)),$$

where $|B_n(x)|$ is the number of elements of $B_n(x)$. To derive an upper bound for the risk of the maximum likelihood classification rule we note that $I_{B_n(x)}(i) > 0$ implies $p_{n,i}(x) = \max_{\eta\in\Delta} p_{n,\eta}(x) \geq p_{n,\theta}(x)$ for every $\theta \in \{1, \ldots, m\}$. Hence,

$$\begin{aligned}(5.29)\qquad \mathsf{R}(\theta, q_{\mathrm{ML}}^{(n)}) &= \sum\nolimits_{i\neq\theta} \int \frac{1}{|B_n(x)|} I_{B_n(x)}(i) P_\theta^{\otimes n}(dx) \\ &\leq \sum\nolimits_{i\neq\theta} P_\theta^{\otimes n}(p_{n,\theta} \leq p_{n,i}) \\ &\leq \sum\nolimits_{i\neq\theta} \int (p_{n,\theta} \wedge p_{n,i}) d\mu^{\otimes n} \\ &\leq 2 \sum_{i,j:i\neq j} \mathsf{b}_{1/2}(P_j^{\otimes n}, P_i^{\otimes n}),\end{aligned}$$

where the last inequality follows from (3.22). The following statement was first established by Krafft and Puri [36].

Theorem 5.7. *If $q^{(n)}$ is a sequence of classification rules for the respective models $(\mathcal{X}^n, \mathfrak{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\{1,\dots,m\}})$ then,*

$$\liminf_{n\to\infty} \frac{1}{n} \ln \left(\max_{1\le\theta\le m} \mathsf{R}(\theta, q^{(n)}) \right) \ge - \min_{1\le\theta_1\ne\theta_2\le m} \mathsf{C}(P_{\theta_1}, P_{\theta_2}). \tag{5.30}$$

The maximum likelihood classification rule $\mathsf{q}_{ml}^{(n)}$ in (5.28) attains the maximum exponential rate of the probability of incorrect classification, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \ln \left(\max_{1\le\theta\le m} \mathsf{R}(\theta, q_{ML}^{(n)}) \right) = - \min_{1\le\theta_1\ne\theta_2\le m} \mathsf{C}(P_{\theta_1}, P_{\theta_2}).$$

Proof. The first statement follows from (5.27). To prove the optimality of the maximum likelihood classification rule we use the fact that for any sequences $a_n, b_n \ge 0$ we have,

$$\limsup_{n\to\infty} \frac{1}{n} \ln(a_n + b_n) = \max \left(\limsup_{n\to\infty} \frac{1}{n} \ln a_n, \limsup_{n\to\infty} \frac{1}{n} \ln b_n \right). \tag{5.31}$$

Applying this statement to (5.29) we arrive at,

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \ln(\mathsf{R}(\theta, q_{\mathrm{ML}}^{(n)})) &\le \limsup_{n\to\infty} \frac{1}{n} \ln \left(2 \sum\nolimits_{i,j:i\ne j} \mathsf{b}_{1/2}(P_j^{\otimes n}, P_i^{\otimes n}) \right) \\ &\le \max_{i,j:i\ne j} \limsup_{n\to\infty} \frac{1}{n} \ln(\mathsf{b}_{1/2}(P_j^{\otimes n}, P_i^{\otimes n})). \end{aligned}$$

To complete the proof we have only to apply (5.23) to the right hand side. ∎

Example 11. *Suppose $(P_\theta)_{\theta\in\Delta}$ is an exponential family with natural parameter θ and $\Delta^0 = \{\theta_1, \dots, \theta_m\}$ is a finite subset of Δ. If we replace P_j by P_{θ_j} and use Example 9 we obtain the exponential rate of the maximum error probabilities of the asymptotically optimal classification rule.*

$$\begin{aligned} &\inf_{1\le\theta_i\ne\theta_j\le m} \mathsf{C}(P_{\theta_i}, P_{\theta_j}) \\ &\qquad = \inf_{1\le\theta_i\ne\theta_j\le m, 0<s<1} \{sK(\theta_i) + (1-s)K(\theta_j) - K(s\theta_i + (1-s)\theta_j))\}. \end{aligned}$$

If $P_{\theta_j} = \mathsf{N}(\theta_j, \sigma^2)$ then by (5.24)

$$\inf_{1\le\theta_i\ne\theta_j\le m} \mathsf{C}(\mathsf{N}(\theta_i, \sigma^2), \mathsf{N}(\theta_j, \sigma^2)) = \frac{1}{8\sigma^2} \min_{i\ne j} (\theta_i - \theta_j)^2.$$

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. BHATTACHARYYA, On some analogues to the amount of information and their uses in statistical estimation. *Sankhya*, **8** (1946), 1–14.

[4] R.E. BLAHUT, *Principles and Practice of Information Theory.* Reading, MA; Adison-Wesley, (1987).

[5] L.D. BROWN, *Fundamentals of Statistical Exponential Families.* IMS Lecture Notes-Monograph Series, **9** (1986).

[6] A. BUZO, A. H. GRAY JR, R. M. GRAY and J. D. MARKEL, Speech coding based upon vector quantization, *IEEE Trans. Inform. Theory*, **28** (1980), 562–574.

[7] 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.

[8] H. CHERNOFF, Large sample theory: Parametric case. *Ann. Math. Statist.*, **27** (1956), 1–22.

[9] B.S. CLARKE and A.R. BARRON, Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, **36** (1990), 453–471.

[10] T. COVER and J. THOMAS, *Elements of Information Theory*, New York: Wiley (1991).

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

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

[13] I. CSISZÁR and J. KÖRNER, *Information Theory. Coding Theorems for Discrete Memoryless Systems.* Budapest: Akademiai Kaidó (1981).

[14] I. CSISZÁR, Generalized cutoff rates and Rényi information measures. *IEEE Trans. Inform. Theory*, **41** (1995), 26–34.

[15] M.H. DE GROOT, Uncertainty, information and sequential experiments. *Ann. Math. Statist.,* **33** (1962), 404–419.

[16] M.H. DE GROOT, *Optimal Statistical Decisions,* New York: McGraw Hill (1970).

[17] A. A. FEDOTOV, P. HARREMOËS and F. TOPSØE, Refinements of Pinsker's inequality. *IEEE Trans. Inform. Theory*, **49** (2003), 1491–1498.

[18] J. DIEUDONNE, Sur le théorème de Lebesgue-Nikodym II, *Bull. Soc. Math. France*, **72** (1944), 193–239

[19] R.S. ELLIS, *Entropy, Large Deviations, and Statistical Mechanics.* Berlin: Springer (1985).

[20] D. FELDMAN and F. ÖSTERREICHER, A note on f-divergences. *Studia Sci. Math. Hungar.,* **24** (1989), 191–200.

[21] R.A. FISHER, A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error. *Monthly Not. Roy Astr. Soc.,* **80** (1920), 758–770.

[22] I.M. GELFAND, A.N. KOLMOGOROV and A.M. YAGLOM, On the general definition of the amount of information. *Dokl. Akad. Nauk. SSSR,* **11** (1956), 745–748.

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

[24] L. GYORFI, G. MORVAI and I. VAJDA, Information-theoretic methods in testing the goodness of fit. *Proc. IEEE Int. Symp. on Inform. Theory*, **28** (2000), Sorrento, Italy.

[25] P. HARREMOËS and F. TOPSØE, Inequalities between entropy and the index of coincidence derived from information diagrams. *IEEE Trans. Inform. Theory*, **46** (1990), 1602–1609.

[26] I.A. IBRAGIMOV and R.Z. HAS'MINSKII, *Asymptotic Theory of Estimation.* Berlin: Springer (1981),

[27] J. JACOD and A.N. SHIRYAEV, *Limit Theorems for Stochastic Processes.* Berlin: Springer (1987).

[28] A. JANSSEN, Asymptotic properties of Neyman-Pearson tests for infinite Kullback-Leibler information. *Ann. Math. Stat.* **14** (1986), 1068–1079.

[29] L.K. JONES and C.L. BYRNE, Generalized entropy criteria for inverse problems with applications to data compression, pattern classification and cluster analysis. *IEEE Trans. Inform. Theory*, **36** (1990), pp. 23–30.

[30] G. JONGBLOED, Minimax lower bounds and moduli of continuity. *Stat. and Prob. Letters,* **50** (2000), 279–284.

[31] T. KAILATH, The divergence and Bhattacharyya distance in signal selection. *IEEE Trans. on Communications* **5** (1967), 52–60.

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

[33] O. KALLENBERG, *Foundations of Modern Probability.* 2nd edition, New York: Springer (2002).

[34] E.I. KOLOMIETS, On asymptotical behavior of probabilities of the second type error for Neyman-Pearson test. *Theory Prob. Appl.,* **32** (1987), 503–522.

[35] O. KRAFFT and D. PLACHKY, Bounds for the power of likelihood ratio test and their asymptotic properties. *Ann. Math. Statist.,* **41** (1970), 1646–1654.

[36] O. KRAFFT and M.L. PURI, The asymptotic behavior of the minimax risk for multiple decision problems. *Sankhyā,* **36** (1974), 1–12.

[37] C.H. KRAFT, Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publ. Statist.* **1** (1955), 125–142.

[38] S. KULLBACK and R. LEIBLER, On information and sufficiency. *Ann. Math. Statist.*, **22** (1951), 79–86.

[39] S. KULLBACK, *Information Theory and Statistics.* New York: Wiley (1959).

[40] L. LeCAM, On the information contained in additional observations. *Ann. Statist.,* **2** (1974), 630–649.

[41] L. LeCAM, *Asymptotic Methods in Statistical Decision Theory*, Berlin: Springer (1986).

[42] L. LeCAM and G.L. YANG, *Asymptotics in Statistics*, Berlin: Springer (1989).

[43] E.L. LEHMANN, *Testing Statistical Hypotheses.* 2nd edition, New York: Springer (1986).

[44] F. LIESE, Hellinger integrals of diffusion processes. *Statistics*, **17** (1986), 63–78.

[45] F. LIESE and I. VAJDA, *Convex Statistical Distances*, Leipzig: Teubner (1987).

[46] Yu. N. LINKOV, *Asymptotic Statistical Methods for Stochastic Processes*, AMS **196**, Providence, Rhode Island (2001).

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

[48] D. MUSSMANN, Sufficiency and f-divergences. *Studia Sci. Math. Hungar.,* **14** (1979), 37–41.

[49] T. NEMETZ, Information theory and the testing of a hypothesis. Proc. Coll. Inform. Theory, **2** (1967), Debrecen.

[50] F. ÖSTERREICHER, On the dimensioning of tests for composite hypothesis and not necessarily independent observations. *Probl. of Control and Inf. Th.,* **7** (1978), 333–343.

[51] F. ÖSTERREICHER and D. FELDMAN, Divergenzen von Wahrscheinlichkeits - verteilungen - integralgeometrisch betrachtet. *Acta Math. Sci. Hungar.,* **37** (1981), 329–337.

[52] F. ÖSTERREICHER, On a class of perimeter-type distances of probability distributions. *Kybernetika,* **32** (1996), 389–393.

[53] F. ÖSTERREICHER and I. VAJDA, Statistical information and discrimination," *IEEE Trans. Inform. Theory*, **39** (1993), 1036–1039.

[54] F. ÖSTERREICHER and I. VAJDA, A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Statist. Math.*, **55** (2003), 639–653.

[55] J. PFANZAGL, A characterization of su ciency by power functions, *Metrika*, **21** (1974), 197–199.

[56] H. V. POOR, Robust decision design using a distance criterion. *IEEE Trans. Inform. Theory*, **26** (1980), 578–587.

[57] A. RÉNYI, On measures of entropy and information. *Proc. 4th Berkeley Symp. on Probab. Theory and Math. Statist.,* 547–561, Berkeley CA: Berkeley Univ. Press (1961).

[58] R.D. REISS, *Approximate Distributions of Order Statistics*, Springer, Berlin, (1989).

[59] R.D. REISS, *A course on Point Processes,* Springer, Berlin, (1993).

[60] F. RIESZ, Sur quelques notions fondamentales dans la théorie générale des opérateurs linéaires, *Ann. of Math.,* **41** (1940), 174–206.

[61] C.E. SHANNON, A mathematical theory of communication. *Bell. Syst. Tech. J.*, **27** (1948), pp. 379–423, 623-656.

[62] J. STEINEBACH, *Large Deviations and Some Related Topics*. Carleton Math. Lect. Notes 28, Carleton: Carleton Univ.Press (1980).

[63] H. STRASSER, *Mathematical Theory of Statistics*. Berlin: De Gruyter (1985).

[64] F. TOPSØE, Information-theoretical optimization techniques. *Kybernetika*, **15** (1979), 7–17.

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

[66] E. TORGERSEN, *Comparison of Statistical Experiments*. Cambridge: Cambridge Univ. Press (1991).

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

[68] I. VAJDA, $\chi^{\alpha}-$divergence and generalized Fisher's information, *Trans. 6th Prague Conf. Inform. Theory, Statist. Dec. Funct., Random Processes,* 873–886, Prague: Academia (1973).

[69] I. VAJDA, Distances and discrimination rates for stochastic processes. *Stochastic Proc. and Appl.*, **35** (1990), pp.47–57.

[70] I. VAJDA, On convergence of information contained in quantized observations, *IEEE Trans. Inform. Theory*, **48** (2002), 2163–2172.

Friedrich Liese
Department of Mathematics
University of Rostock
Germany
E-mail: friedrich.liese@uni-rostock.de

Igor Vajda
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic.
E-mail: vajda@utia.cas.cz