On Divergences and Informations in Statistics and Information Theory

Friedrich Liese and Igor Vajda, *Fellow, IEEE*

***Abstract*—The paper deals with the f-divergences of Csiszár generalizing the discrimination information of Kullback, the total variation distance, the Hellinger divergence, and the Pearson divergence. All basic properties of f-divergences including relations to the decision errors are proved in a new manner replacing the classical Jensen inequality by a new generalized Taylor expansion of convex functions. Some new properties are proved too, e.g., relations to the statistical sufficiency and deficiency. The generalized Taylor expansion also shows very easily that all f-divergences are average statistical informations (differences between prior and posterior Bayes errors) mutually differing only in the weights imposed on various prior distributions. The statistical information introduced by De Groot and the classical information of Shannon are shown to be extremal cases corresponding to $\alpha = 0$ and $\alpha = 1$ in the class of the so-called Arimoto α-informations introduced in this paper for $0 < \alpha < 1$ by means of the Arimoto α-entropies. Some new examples of f-divergences are introduced as well, namely, the Shannon divergences and the Arimoto α-divergences leading for $\alpha \uparrow 1$ to the Shannon divergences. Square roots of all these divergences are shown to be metrics satisfying the triangle inequality. The last section introduces statistical tests and estimators based on the minimal f-divergence with the empirical distribution achieved in the families of hypothetic distributions. For the Kullback divergence this leads to the classical likelihood ratio test and estimator.**

***Index Terms*—Arimoto divergence, Arimoto entropy, Arimoto information, deficiency, discrimination information, f-divergence, minimum f-divergence estimators, minimum f-divergence tests, Shannon divergence, Shannon information, statistical information, sufficiency.**

I. Introduction

SHANNON [46] introduced the information $I(X;Y)$ as the divergence

$$D(P_{XY}, P_X \otimes P_Y) = \int \ln \frac{dP_{XY}}{d(P_X \otimes P_Y)} dP_{XY}$$

of the joint distribution P_{XY} of random variables X, Y and the product $P_X \otimes P_Y$ of the marginal distributions. The divergence

$$D(P,Q) = \begin{cases} \int \ln\left(\frac{dP}{dQ}\right) dP, & \text{if } P \ll Q \\ \infty, & \text{otherwise} \end{cases}$$

of arbitrary distributions P, Q was systematically studied by Kullback and Leibler [30], Gel'fand *et al.* [21] and others who recognized its importance in information theory, statistics, and probability theory. Rényi [44] introduced a class of measures of divergence of distributions P, Q with properties similar to $D(P,Q)$ and containing $D(P,Q)$ as a special case. Csiszár [11] (and independently also Ali and Silvey [1]) introduced the f-divergence

$$D_f(P,Q) = \int \frac{dQ}{d\mu} f\left(\frac{dP/d\mu}{dQ/d\mu}\right) d\mu$$

for convex $f : (0,\infty) \longmapsto \mathbb{R}$, where μ is a σ–finite measure which dominates P and Q and the integrand is appropriately specified at the points where the densities $dP/d\mu$ and/or $dQ/d\mu$ are zero.

For $f(t) = t \ln t$, the f-divergence reduces to the classical "information divergence" $D(P,Q)$ (denoted sometimes also by $I(P,Q)$). For the convex or concave functions $f(t) = t^\alpha, \alpha > 0$ we obtain the so-called Hellinger integrals

$$\mathcal{H}_\alpha(P,Q) = \int (dP/d\mu)^\alpha \, (dQ/d\mu)^{1-\alpha} \, d\mu, \qquad \alpha > 0.$$

For the convex functions

$$f(t) = (\alpha - 1)^{-1} (t^\alpha - 1), \qquad \alpha > 0, \ \alpha \neq 1$$

we obtain the Hellinger divergences

$$H_\alpha(P,Q) = (\alpha - 1)^{-1} (\mathcal{H}_\alpha(P,Q) - 1)$$

which are strictly increasing functions of the Rényi divergences $R_\alpha(P,Q) = (\alpha - 1)^{-1} \ln \mathcal{H}_\alpha(P,Q)$ for $\alpha > 0$, $\alpha \neq 1$. The limits of these divergences for $\alpha \to 1$ may not exist but, as proved in [33], the limit from the left does exist and both the Hellinger and Rényi divergences tend for $\alpha \uparrow 1$ to the information divergence $D(P,Q)$. Note that the divergence measures $B_\alpha(P,Q) = -\ln \mathcal{H}_\alpha(P,Q)$ were considered for $0 < \alpha < 1$ already by Chernoff [8] and the special case for $\alpha = 1/2$ by Bhattacharyya [5] and Kailath [27].

Among the f-divergences one can find also the basic divergence measures of probability theory and mathematical statistics, such as the total variation $V(P,Q)$ (for $f(t) = |t-1|$), the Pearson divergence $\chi^2(P,Q) = H_2(P,Q) - 1$ (for $f(t) = (t-1)^2$ or, equivalently, $f(t) = t^2 - 1$) or, more generally, the likelihood ratio cumulants $\chi^\alpha(P,Q)$ (for $f(t) = |t-1|^\alpha, \alpha \geq 1$) systematically studied in [52].

Statistical applications of f-divergences were considered, e.g., by Ali and Silvey [1], Csiszár [12], Arimoto [2], Barron *et al.* [3], Berlinet *et al.* [4], Gyorfi *et al.* [23], and Vajda [54]. Decision-theoretic applications of f-divergences can be found,

Manuscript received October 26, 2005; revised June 22, 2006. This work was supported by the MSMT under Grant 1M0572 and the GAAV under Grant A100750702.

F. Liese is with the Department of Mathematics, University of Rostock, Rostock 18051, Germany (e-mail: friedrich.liese@uni-rostock.de).

I. Vajda is with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague 18208, Czech Republic (e-mail: vajda@utia.cas.cz)

Communicated by K. Kobayashi, Associate Editor for Shannon Theory..

Digital Object Identifier 10.1109/TIT.2006.881731

e.g., in Kailath [27], Poor [43], LeCam [31], Read and Cressie [45], Clarke and Barron [9], Longo *et al.* [35], Torgersen [50], Österreicher and Vajda [41], Topsøe [49], and Fedotov *et al.* [18]. Applications of f-divergences in the channel and source coding can be found, e.g., in Topsøe [48], Buzo *et al.* [7], Blahut [6], Jones and Byrne [26], Cover and Thomas [10], Csiszár [14], and Harremoës and Topsøe [24].

Due to the growing importance of divergences in information theory, statistics, and probability theory, a possibility to simplify and extend the general theory of f-divergences deserves attention. About half of the present papers are devoted to a considerably simplified derivation of the most important basic properties of f-divergences. The classical derivation of these properties is based on the Jensen inequalities for general expectations and conditional expectations which are quite complicated if they are rigorously formulated for all desirable functions $f(t)$. This concerns especially the stability of these inequalities for convex but not necessarily twice differentiable functions (cf. [33]). The approach of this paper is based on an extension of the classical Taylor formula to all convex or concave (not necessarily differentiable) functions $f(t)$. The remaining papers present new relations between f-divergences and some classical concepts of information theory, probability theory, and statistical decision theory, as well as new applications of f-divergences in the statistics and information theory.

The generalized Taylor formula is introduced in Section II and represents a new tool for the analysis of convex functions (t). In fact, it extends the classical Taylor formula

$$f(t) = f(t_0) + f'(t_0)(t - t_0) + R_f(t, t_0)$$

valid for the twice continuously differentiable functions $f : (u, v) \longmapsto \mathbb{R}$ with the remainder in the in- tegral form

$$R_f(t, t_0) = \int_{t_0}^{t} (t - s)\, f''(s) ds$$

to all convex functions $f : (u, v) \longmapsto \mathbb{R}$ by replacing the derivative $f'(t_0)$ by the right-hand derivative $f'_+(t_0)$ and the remainder in the Riemann integral form by the remainder in the Lebesque–Stieltjes integral form

$$R_f(t, t_0) = \int \mathbf{1}_{(t_0,t]}(s)(t - s) d\, f'_+(s)$$

with the convention $\mathbf{1}_{(t_0,t]}(s) = -\mathbf{1}_{(t,t_0]}(s)$ if $t < t_0$ (see [25]). It is known that the derivative $f'_+(t)$ exists and is right-continuous and nondecreasing on (u, v). The right-continuity and monotonicity means that $f'_+(t)$ defines a Lebesque–Stieltjes measure on the Borel subsets of (u, v). Therefore, the Lebesgue–Stieltjes integrals $\int \phi(s) d\, f'_+(s)$ are well defined for all bounded measurable functions $\phi(s)$ on (u, v), in particular for $\phi(s) = \mathbf{1}_{(t_0,t]}(s)(t - s)$. Proof of the extended Taylor formula

$$f(t) = f(t_0) + f'_+(t_0)(t - t_0) + \int \mathbf{1}_{(t_0,t]}(s)(t - s) d\, f'_+(s)$$

is given in Section II.

In Section III, we introduce the f-divergences, characterize their ranges of values, and present the most important families of examples.

In Section IV, the Shannon information $I(X, Y)$ in a general output X of a channel with binary input Y is shown to be an f-divergence $\mathbb{I}_\pi(P, Q) = D_f(P, Q)$ of the conditional output distributions P, Q where the convex function $f(t) = f_\pi(t)$ depends on the input probabilities $\pi, 1 - \pi$. Further, the Shannon information is shown to be the limit for $\alpha \uparrow 1$ of the informations $I_\alpha(X, Y)$ introduced by Arimoto [2] for $\alpha \neq 1$. Similarly, the Shannon entropy $H(Y)$ and the conditional entropy $H(Y|X)$ are the limits for $\alpha \uparrow 1$ of the Arimoto entropies $H_\alpha(Y)$ and $H_\alpha(Y|X)$, respectively. We consider the Arimoto informations and entropies for $\alpha \in (0, 1)$ and prove that the Arimoto informations are f-divergences $\mathbb{I}_{\pi,\alpha}(P, Q) = D_f(P, Q)$ where the convex function $f(t) = f_{\pi,\alpha}(t)$ depends on π and α. Consequently, the above mentioned *Shannon divergence* $\mathbb{I}_\pi(P, Q)$ is the limit for $\alpha \uparrow 1$ of the *Arimoto divergences* $\mathbb{I}_{\pi,\alpha}(P, Q)$. Since the square roots $\sqrt{\mathbb{I}_{1/2,\alpha}(P, Q)}$ of the Arimoto divergences will be shown to be metrics in the space of probability distributions P, Q, we deduce that the square roots of the Shannon informations in binary channels with equiprobable inputs are metrics in the space of output conditional distributions. Applicability of this is illustrated in Section IV.

In Section V, we show that the limits $H_0(Y)$ and $H_0(Y|X)$ of the Arimoto entropies for $\alpha \downarrow 0$ are the prior Bayes error $B_\pi = \pi \wedge (1 - \pi) := \min\{\pi, 1 - \pi\}$ and posterior Bayes error $B_\pi(P, Q)$ in the decision problem with *a priori* probabilities $\pi, 1 - \pi$ and conditional probabilities P, Q, respectively. The difference $\mathcal{I}_\pi(P, Q) = B_\pi - B_\pi(P, Q)$ is nothing but the statistical information first introduced by De Groot [16]. This information coincides with the limit $I_0(P, Q) = \lim_{\alpha \downarrow 0} I_\alpha(P, Q)$ of the Arimoto informations, i.e., the Shannon and statistical informations are the extreme forms of the Arimoto informations $I_\alpha(P, Q)$ on the interval $\alpha \in (0, 1)$. At the same time, the statistical information $\mathcal{I}_\pi(P, Q)$ coincides with the statistical divergence $\mathbb{I}_{\pi,0}(P, Q)$ which is the limit for $\alpha \downarrow 0$ of the Arimoto divergences $\mathbb{I}_{\pi,\alpha}(P, Q)$. However, the main result of Section V is the representation of an arbitrary f-divergence $D_f(P, Q)$ as an average statistical information $\int_{(0,1)} \mathcal{I}_\pi(P, Q) d\Gamma(\pi)$ where $\Gamma = \Gamma_f$ is a measure on the interval $(0, 1)$ of *a priori* probabilities π depending on the convex function f. This representation follows from the generalized Taylor expansion of convex functions proved in Section II in a manner that is more transparent and simpler than that presented previously in Österreicher and Vajda [41].

The representation of f-divergences as average statistical informations allows to prove in Section VI the general form of the information processing theorem for f-divergences and in Section VII the continuity of f-divergences (approximability by f-divergences on finite subalgebras) in a much simpler way than this was achieved in the previous literature (see [11], [1], [12], [33]). The general information processing theorem is compared in Section VI also with the classical statistical sufficiency and with the deficiency of statistical models studied by Torgersen [50], Strasser [47], and LeCam [31].

In Section VIII, we present applications of f-divergences in statistical estimation and testing. We show in particular that the general maximum-likelihood estimation (MLE) and maximum-likelihood testing can be obtained as a special minimum f-divergence estimation and testing. This is established using

an inequality proved already in [33] but, again, this inequality is obtained from the generalized Taylor expansion in a simpler way than in [33].

II. Convex Functions

Let $(u,v) \subseteq \mathbb{R}$ be a finite or infinite interval. The basic tool in the analysis and applications of twice continuously differentiable functions $f : (u,v) \longmapsto \mathbb{R}$ is the Taylor formula

$$f(b) = f(a) + f'(a)(b-a) + R_f(a,b) \tag{1}$$

for $a, b \in (u,v)$, where

$$R_f(a,b) = \int_a^b (b-s)\, f''(s)\, ds \tag{2}$$

is the remainder in the integral form and f', f'' are the derivatives of f. In this paper, we deal with convex functions. If $f : (u,v) \longmapsto \mathbb{R}$ is convex then the right derivative

$$f'_+(s) = \lim_{t \downarrow s} \frac{f(t) - f(s)}{t - s}$$

always exists and is finite on the whole domain (u,v), see [25]. Since this derivative is right continuous and monotone (nondecreasing, typically increasing), there is unique measure λ_f on the Borel subsets of (u,v) such that

$$\lambda_f((a,b]) = f'_+(b) - f'_+(a), \qquad \text{for } (a,b] \subset (u,v). \tag{3}$$

Note that here, and in the sequel, we denote the Lebesgue integrals $\int \phi(s)\, d\lambda_f(s)$ for measurable $\phi : (u,v) \longmapsto \mathbb{R}$ as the Lebesgue–Stieltjes integrals $\int \phi(s) df'_+(s)$ where $f'_+ = f'$ if f is differentiable. Moreover, it is known that

$$\int \phi(s) df'_+(s) = \int \phi(s) f''(s)\, ds \tag{4}$$

when f' is absolutely continuous with the a.e. derivative f''. This paper is based on the following extension of the Taylor formula to convex (or concave) functions.

Theorem 1: If $f : (u,v) \longmapsto \mathbb{R}$ is convex then

$$f(b) = f(a) + f'_+(a)(b-a) + R_f(a,b) \tag{5}$$

for $a, b \in (u,v)$ where $0 = R_f(a,a) \le R_f(a,b)$ and

$$R_f(a,b) = \int \mathbf{1}_{(a,b]}(s)(b-s)\, d\, f'_+(s) \tag{6}$$

or

$$R_f(a,b) = \int \mathbf{1}_{(b,a]}(s)(s-b)\, d\, f'_+(s) \tag{7}$$

depending on whether $b > a$ or $b < a$, respectively.

Proof: As we have already seen, the convexity of f implies that f has a right-hand derivative of locally bounded variation. Hence, by Theorem 18.16 in Hewitt and Stromberg [25]

$$f(b) - f(a) = \int_a^b f'_+(s)\, ds, \qquad a, b \in (u,v). \tag{8}$$

It suffices to prove (5) and (6) for $b > a$ because for $b = a$ the assertion is trivial and for $b < a$ the proof is similar. If $b > a$ then, using (8) and the equality $\mathbf{1}_{(a,b]}(s)\, \mathbf{1}_{(a,s]}(t) = \mathbf{1}_{(a,b]}(t)\, \mathbf{1}_{[t,b]}(s)$, we obtain

$$\begin{aligned} f(b) - f(a) - f'_+(a)(b-a) &= \int_a^b (f'_+(s) - f'_+(a))\, ds \\ &= \int \left(\int \mathbf{1}_{(a,b]}(s) \mathbf{1}_{(a,s]}(t) df'_+(t) \right) ds \\ &= \int \left(\int \mathbf{1}_{(a,b]}(t) \mathbf{1}_{[t,b]}(s) ds \right) df'_+(t) \\ &= \int \mathbf{1}_{(a,b]}(t)(b-t)\, df'_+(t)\,. \end{aligned}$$

Thus, the last integral is the remainder $R_f(a,b)$ in (6). □

In view of (4), Theorem 1 implies that the Taylor expansion in the classical form (1), (2) remains valid for each f with absolutely continuous derivative f'.

In the paper, we deal with the class $\mathcal{F}$ of convex functions $f : (0,\infty) \longmapsto \mathbb{R}$ and the subclass $\mathcal{F}_1 \subset \mathcal{F}$ such that $f \in \mathcal{F}_1$ satisfies $f(1) = 0$. The shift by the constant $-f(1)$ sends every $f \in \mathcal{F} - \mathcal{F}_1$ to $\mathcal{F}_1$.

By (8), each $f \in \mathcal{F}$ is piecewise monotone. Hence, the limit

$$f(0) = \lim_{t \downarrow 0} f(t) \in (0, \infty] \tag{9}$$

exists and (9) extends f into a convex function on $[0,\infty)$ which may eventually be infinite at 0.

For every $f \in \mathcal{F}$, we define the $*$-adjoint function

$$f^*(t) = t\, f(1/t), \qquad t \in (0,\infty)\,. \tag{10}$$

We shall need the following properties of $f \in \mathcal{F}$ and their adjoints f^*.

Theorem 2: If $f \in \mathcal{F}$ then

$$f^* \in \mathcal{F}, \quad (f^*)^* = f, \quad f^*(0) = \lim_{t \to \infty} \frac{f(t)}{t}\,. \tag{11}$$

If $f \in \mathcal{F}_1$ then

$$\tilde{f}(t) := f(t) - f'_+(1)(t-1) \ge 0, \qquad \tilde{f}(1) = \tilde{f}'_+(1) = 0 \tag{12}$$

so that $\tilde{f} \in \mathcal{F}_1$, and also

$$f^* \in \mathcal{F}_1, \quad \widetilde{(f^*)} = (\tilde{f})^*, \quad \tilde{f}(0) + \tilde{f}^*(0) = f(0) + f^*(0)\,. \tag{13}$$

Proof: The first two relations in (11) are clear from the definition of f^* in (10). The third one follows from the definition of f^* and (9). The nonnegativity of $\tilde{f}$ in (12) is clear from the generalized Taylor expansion of $\tilde{f}(t)$ around $t = 1$ and from the nonnegativity of the remainder in Theorem 1. The equalities in (12) are trivial and those in (13) follow directly from the corresponding definitions. □

Special attention is paid to the functions $f \in \mathcal{F}$ which are strictly convex at $t = 1$. As these functions may not be twice differentiable in an open neighborhood of $t = 1$, this concept deserves clarification.

Definition 1: We say that $f : (u,v) \longmapsto \mathbb{R}$ is locally linear at $t \in (u,v)$ if it is linear in an open neighborhood of t. We say that $f \in \mathcal{F}$ is strictly convex at $t \in (0,\infty)$ if it is not locally linear at t. We say that $f \in \mathcal{F}$ is strictly convex if it is strictly convex at all $t \in (0,\infty)$.

Remark 1: It is clear from the definition of λ_f in (3) and the representation of the remainder term in Theorem 1 that $f \in \mathcal{F}$ is strictly convex at $t \in (0,\infty)$ if and only if t belongs to the support of the measure λ_f, i.e., if and only if

$$\lambda_f((t-\varepsilon, t+\varepsilon)) > 0 \quad \text{for every} \quad 0 < \varepsilon < t. \tag{14}$$

Remark 2: One easily obtains from (5) the classical Jensen inequality

$$\pi f(t_2) + (1-\pi) f(t_1) \geq f(\pi t_2 + (1-\pi) t_1)$$

for any $\alpha < t_1 < t_2 < \beta$ and $0 < \pi < 1$. To this end, it suffices to put in (5) $a = \pi t_2 + (1-\pi) t_1$ and first $b = t_1$ and then $b = t_2$. Multiplying (5) in the first case by π and in the second case by $1-\pi$ we get

$$\begin{aligned}\pi f(t_2) + (1-\pi) f(t_1) - f(\pi t_2 + (1-\pi) t_1) \\ = \pi R_f(a, t_2) + (1-\pi) R_f(a, t_1) \geq 0.\end{aligned}$$

The definition of R_f shows that equality in the last inequality holds if and only if $\lambda_f((t_1,t_2)) = 0$. By the previous remark, the last condition means that $f(t)$ is strictly convex at no $t \in (t_1,t_2)$, i.e., that $f(t)$ is locally linear everywhere on (t_1,t_2). The condition $\lambda_f((t_1,t_2)) = 0$ in fact means that the right-hand derivative $f'_+(t)$ is constant on (t_1,t_2), i.e., that $f(t)$ is differentiable on this interval with the usual derivative $f'(t)$ constant. Therefore, $\lambda_f((t_1,t_2)) = 0$ is equivalent to the linearity of $f(t)$ on (t_1,t_2).

Example 1: The function $f(t) = |t-1| \in \mathcal{F}$ has the right-hand derivative $f'_+(t) = \mathbf{1}_{[1,\infty)}(t) - \mathbf{1}_{(-\infty,1)}(t)$ and the measure $\lambda_f = 2\delta_1$, where δ_1 is the Dirac measure concentrated at 1. Therefore, (14) holds for $t = 1$ and $f(t)$ is strictly convex at 1.

By Theorem 2, if $f \in \mathcal{F}$ and $f(1) = f'_+(1) = 0$ then $f \geq 0$ and, by (10), also $f^* \geq 0$. This can be sharpened as follows.

Theorem 3: Let $f \in \mathcal{F}$. i) If $f(1) = f'_+(1) = 0$ then both f and f^* are nonnegative and nonincreasing on $(0,1)$, nondecreasing on $(1,\infty)$. ii) f is strictly convex at 1 if and only if f^* is strictly convex at 1. iii) If f is strictly convex at 1 and $f(1) = f'_+(1) = 0$ then

$$f \text{ or } f^* \text{ is strictly decreasing on } (0,1). \tag{15}$$

Proof: The first statement follows from monotonicity of the remainders in Theorem 1. Furthermore, the function f is linear in (a,b) if and only if f^* is linear in $(1/b, 1/a)$. In view of Theorem 1, this takes place if and only if $\lambda_f((a,b)) = 0$ or, equivalently, $\lambda_{f^*}((1/b, 1/a)) = 0$. Suppose that f is strictly convex at 1. Then Remark 1 implies $\lambda_f((1-\varepsilon, 1+\varepsilon)) > 0$ for $0 < \varepsilon < 1$. If $\lambda_f((1-\varepsilon, 1]) = 0$ then $\lambda_f((1, 1+\varepsilon)) > 0$ and

$$\lambda_{f^*}((1/(1+\varepsilon), 1]) \geq \lambda_{f^*}((1/(1+\varepsilon), 1)) > 0.$$

Hence, the statement ii) follows from Remark 1 and moreover

$$\min\{\lambda_f((1-\varepsilon, 1]), \lambda_{f^*}((1-\varepsilon, 1]\} > 0$$

for $0 < \varepsilon < 1$. Under the assumptions of iii) it holds that

$$f(t) = \int \mathbf{1}_{(t,1]}(s)(s-t)\lambda_f(ds)$$

which is strictly decreasing in $t \in (0,1)$ if $\lambda_f((1-\varepsilon, 1]) > 0$ due to strict local ordering of the integrands in the left neighborhood of 1. The case $\lambda_{f^*}((1-\varepsilon, 1]) > 0$ is similar. □

Next follow examples of functions from $\mathcal{F}_1$ strictly convex at 1. Important examples which are not strictly convex at 1 will be studied in Section IV.

Example 2: The class of functions $\{f_\alpha : \alpha \in \mathbb{R}\}$ defined on $(0,\infty)$ by

$$f_\alpha(t) = \begin{cases} t \ln t, & \text{if } \alpha = 1 \\ \frac{t^\alpha - 1}{\alpha(\alpha-1)}, & \text{if } \alpha \neq 0, \alpha \neq 1 \\ -\ln t, & \text{if } \alpha = 0 \end{cases} \tag{16}$$

is contained in $\mathcal{F}_1$. The corresponding nonnegative functions obtained by the transformation (12) are

$$\tilde{f}_\alpha(t) = \begin{cases} t \ln t - t + 1, & \text{if } \alpha = 1 \\ \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, & \text{if } \alpha \neq 0, \alpha \neq 1 \\ -\ln t + t - 1, & \text{if } \alpha = 0. \end{cases} \tag{17}$$

It is easy to see that this class is continuous in $\alpha \in \mathbb{R}$ and closed with respect to the $*$-adjoining, namely

$$(\tilde{f}_\alpha)^* = \tilde{f}_{1-\alpha}. \tag{18}$$

Further

$$f_\alpha(0) = \begin{cases} \infty, & \text{if } \alpha \leq 0 \\ \frac{1}{\alpha(1-\alpha)}, & \text{if } \alpha > 0, \alpha \neq 1 \\ 0, & \text{if } \alpha = 1, \end{cases}$$

$$f^*_\alpha(0) = \begin{cases} \infty, & \text{if } \alpha \leq 1 \\ \frac{1}{\alpha}, & \text{if } \alpha > 1 \end{cases}$$

and

$$\tilde{f}_\alpha(0) = \tilde{f}^*_{1-\alpha}(0) = \begin{cases} \infty, & \text{if } \alpha \leq 0 \\ \frac{1}{\alpha}, & \text{if } \alpha > 0. \end{cases}$$

We see that this example achieves the invariance

$$f_\alpha(0) + f^*_\alpha(0) = \tilde{f}_\alpha(0) + \tilde{f}^*_\alpha(0) = \begin{cases} \frac{1}{\alpha(1-\alpha)}, & \text{if } \alpha \in (0,1) \\ \infty, & \text{otherwise} \end{cases} \tag{19}$$

with respect to the transformation (12) predicted in a general form by (13).

III. Divergences

Let P, Q be probability measures on a measurable observation space $(\mathcal{X}, \mathcal{A})$, nontrivial in the sense that $\mathcal{A}$ contains at least one event A different from $\emptyset$ and $\mathcal{X}$. Suppose that P, Q are dominated by a σ-finite measure μ with densities

$$p = \frac{dP}{d\mu} \quad \text{and} \quad q = \frac{dQ}{d\mu}$$

defined on $\mathcal{X}$. Let, as before, $\mathcal{F}$ be the class of convex functions $f : (0, \infty) \to \mathbb{R}$ and $\mathcal{F}_1$ the subclass containing $f \in \mathcal{F}$ normalized by the condition $f(1) = 0$.

Definition 2: For every $f \in \mathcal{F}$ we define f-divergence of probability measures P, Q by

$$D_f(P,Q) = \int_{\{pq>0\}} f\left(\frac{p}{q}\right) dQ + f(0)\, Q(p=0) + f^*(0)\, P(q=0) \quad (20)$$

where $f(0)$ and $f^*(0)$ are given by (9) and (11) and

$$f(0) \cdot 0 = f^*(0) \cdot 0 = 0 \quad (21)$$

even if $f(0) = \infty$ or $f^*(0) = \infty$.

Since $Q(q > 0) = P(p > 0) = 1$, we see that

$$D_f(P,Q) = \int_{\{0<p\le q\}} f\left(\frac{p}{q}\right) dQ + \int_{\{0<q<p\}} f^*\left(\frac{q}{p}\right) dP + f(0)\, Q(p=0) + f^*(0)\, P(q=0). \quad (22)$$

Extension (9) of all $f \in \mathcal{F}$ to the domain $[0, \infty)$ leads to a simpler formula

$$D_f(P,Q) = \int_{\{q>0\}} f\left(\frac{p}{q}\right) dQ + f^*(0)\, P(q=0). \quad (23)$$

This can be further simplified into the form

$$D_f(P,Q) = \int q\, f\left(\frac{p}{q}\right) d\mu \quad (24)$$

where behind the integral are adopted the conventions

$$0\, f\left(\frac{p}{0}\right) = p\, f^*(0) \quad \text{and} \quad 0 \cdot f^*(0) = 0. \quad (25)$$

If $P \ll Q$ (absolute continuity) then $q = 0$ implies $p = 0$ Q-a.s. Since $0 f(0/0) = 0$ by (25), we get from (24) the formula

$$D_f(P,Q) = \int f\left(\frac{dP}{dQ}\right) dQ, \qquad \text{if } P \ll Q\,. \quad (26)$$

If P is not absolutely continuous with respect to Q then $P(q = 0) > 0$ and, under the assumption $f^*(0) = \infty$, we get from (24) and (25)

$$D_f(P,Q) = \infty, \qquad \text{if not } P \ll Q, \text{ i.e., } P \not\ll Q. \quad (27)$$

Equation (26) together with (27) can be taken as an alternative definition of the f-divergence, but only for $f \in \mathcal{F}$ with $f^*(0) = \infty$. Finally, by the definition of $\widetilde{f}$ in (12)

$$D_f(P,Q) - f(1) = D_{\widetilde{f}}(P,Q). \quad (28)$$

The concept of f-divergence was introduced by Csiszár [11] in 1963. In that paper, and also in [12], he used the definition (24) with the convention (25) written in the form

$$0\, f\left(\frac{p}{0}\right) = p\, f^*(0), \quad \text{for } p > 0 \qquad \text{and} \quad 0\, f\left(\frac{0}{0}\right) = 0\,. \quad (29)$$

The first formula of (29) was motivated by continuity of the extension of the continuous function $qf(p/q)$ from the domain $p \ge 0, q > 0$ to the strip $p > 0, q = 0$. The second formula dealing with the remaining point $p = 0, q = 0$ of the closed domain $p \ge 0, q \ge 0$ was imposed in order to achieve the uniqueness of definition and suppression of influence of the events with zero probabilities. Liese and Vajda [33] observed that (29) preserves the very desirable convexity and lower semicontinuity of the extension of $q\, f(p/q)$ to the domain $p \ge 0, q \ge 0$. Vajda [53] noticed that (29) is the unique rule leading to convex and lower semicontinuous extension of $q\, f(p/q)$ to the domain $p \ge 0, q \ge 0$, namely, that the values smaller than those prescribed by (29) break the convexity and the larger ones violate the lower semicontinuity. Note that Csiszár in [11] applied the f-divergences in the problems of probability theory but later in [12] he studied properties of f-divergences important for applications in the statistics and information theory. Ali and Silvey [1] introduced independently the f-divergences and studied their statistical applications.

Example 3: The functions $f_{(1)}(t) := t \ln t$, $f_{(2)}(t) := (t-1)^2$, $f_{(3)}(t) := (\sqrt{t} - 1)^2$, and $f_{(4)}(t) := |t - 1|$ with

$$f_{(1)}(0) = 0, \quad f_{(2)}(0) = f_{(3)}(0) = f_{(4)}(0) = 1$$
$$f^*_{(1)}(0) = f^*_{(2)}(0) = \infty, \quad f^*_{(3)}(0) = f^*_{(4)}(0) = 1$$

belong to $\mathcal{F}_1 \subset \mathcal{F}$. Applying these functions in (23), we get the divergences

$$I(P,Q) = \int_{\{q>0\}} p \ln \frac{p}{q}\, d\mu + \infty\, P(q=0) \quad (30)$$

$$\chi^2(P,Q) = \int_{\{q>0\}} \frac{(p-q)^2}{q}\, d\mu + \infty\, P(q=0) \quad (31)$$

$$H^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2\, d\mu \quad (32)$$

$$V(P,Q) = \int |p - q|\, d\mu. \quad (33)$$

Here, $I(P,Q)$ is the information divergence which is (under this or a different name, and in this or a different notation) one of the basic concepts of information theory and statistics. The Pearson divergence $\chi^2(P,Q)$ plays an important role in statistics. The Hellinger distance $H(P,Q)$ and total variation $V(P,Q)$ are metrics in spaces of probability distributions frequently used in information theory, statistics, and probability theory.

In view of (28), we restrict our attention to the f-divergences with $f \in \mathcal{F}_1$ for which $f(1) = 0$. We pay special attention to the f-divergences with $f \in \mathcal{F}_1$ strictly convex at 1. For these divergences, Csiszár [11], [12] proved the important reflexivity property

$$D_f(P,Q) = 0, \qquad \text{if and only if } P = Q. \quad (34)$$

Sufficient conditions for another important property, the symmetry, are given in the next theorem which follows directly from Definition 2 and from the definition of $*$-adjoint function f^* in (10) and its properties stated in Theorems 2 and 3.

Theorem 4: f^* belongs to $\mathcal{F}$ or $\mathcal{F}_1$ if f does and $D_{f^*}(P,Q) = D_f(Q,P)$. Therefore, the $(f+f^*)$-divergence is symmetric in P,Q.

Liese and Vajda [33] proved that $f,\ f^* \in \mathcal{F}$ satisfy the equality $D_f(P,Q) = D_{f^*}(Q,P)$ for all P,Q under consideration if and only if $f(t) = f^*(t) + const \cdot (t-1)$ where the linear term has no influence on the divergence $D_f(P,Q)$. Thus, the condition $f = f^*$ is in the class $\mathcal{F}_1$ necessary and sufficient for the symmetry of $D_f(P,Q)$ in the variables P,Q. The remaining metric property, the triangle inequality, will be discussed in the next section.

In the rest of this section, we deal with the range of values of f-divergences investigated by Csiszár [11], [12] and Vajda [51]. Here we present a proof based on the generalized Taylor expansion in Theorem 1. It is simpler than the previous proofs.

Theorem 5: If $f \in \mathcal{F}_1$ then

$$0 \le D_f(P,Q) \le f(0) + f^*(0) \tag{35}$$

where the left equality holds for $P = Q$ and the right equality holds for $P \perp Q$ (singularity). If, moreover, f is strictly convex at 1 then the left equality holds only for $P = Q$ and the right equality is attained only for $P \perp Q$ provided $f(0) + f^*(0)$ is finite.

Proof: Let $f \in \mathcal{F}_1$. Notice that if f is replaced by $\tilde{f}$ transformed by (12) then (20) implies

$$D_f(P,Q) = D_{\tilde{f}}(P,Q). \tag{36}$$

Hence, we may suppose $f(1) = f'_+(1) = 0$ in the sequel. Then the functions f and f^* are nonincreasing on $(0,1)$ in view of Theorem 3. The inequality $0 \le D_f(P,Q) \le f(0) + f^*(0)$ is apparent from (22). Assume now that f is strictly convex at 1. By Theorem 3, at least one of the functions f and f^* is strictly decreasing on $(0,1)$. Therefore, we see from (22) that if $D_f(P,Q) = 0$ then either $Q(p=q) = 1$ or $P(p=q) = 1$ where each of these equalities implies $P = Q$. Similarly, we see from (22) that if $D_f(P,Q) = f(0) + f^*(0) < \infty$ then $Q(p=0) = 1$ or $P(q=0) = 1$ which implies $P \perp Q$. □

Example 4: Applying Definition 2 to the functions f_α of Example 2, we obtain the divergences

$$I_\alpha(P,Q) := D_{f_\alpha}(P,Q) = \begin{cases} I(P,Q), & \text{if } \alpha = 1 \\ \frac{1}{\alpha(\alpha-1)}\,(H_\alpha(P,Q) - 1), & \text{if } \alpha \ne 0, \alpha \ne 1 \\ I(Q,P), & \text{if } \alpha = 0 \end{cases}$$

where $I(P,Q)$ is defined in (30) and the Hellinger integrals $H_\alpha(P,Q)$ of orders $\alpha \in \mathbb{R}$ are defined by

$$H_\alpha(P,Q) = \int p^\alpha q^{1-\alpha}\, d\mu\,. \tag{37}$$

In (37), we suppose $p^\alpha = \infty$ if $p = 0$ and $\alpha < 0$ and $q^{1-\alpha} = \infty$ if $q = 0$ and $\alpha > 1$. From (18) and Theorem 4, we get the skew symmetry

$$I_\alpha(Q,P) = I_{1-\alpha}(P,Q)\,. \tag{38}$$

Since all f_α of Example 2 are strictly convex at 1 with $f_\alpha(1) = 0$, the lower bound 0 for $I_\alpha(P,Q)$ and the reflexivity of $I_\alpha(P,Q)$ at this bound are clear from Theorem 5. The upper bounds $f_\alpha(0) + f^*_\alpha(0)$ for these divergences were evaluated in (19). We see from there and from Theorem 5 that for $\alpha \in (0,1)$ the upper bound $1/[\alpha(1-\alpha)]$ is achieved by $I_\alpha(P,Q)$ if and only if $P \perp Q$. In addition to the information divergences obtained for $\alpha = 1$ and $\alpha = 0$, this class of divergences contains

$$I_2(P,Q) = I_{-1}(Q,P) = \frac{1}{2}\,\chi^2(P,Q)$$

which differs by a factor $1/2$ from the Pearson divergence (31) and

$$I_{1/2}(P,Q) = 2H^2(P,Q)$$

which is the only symmetric divergence in this class, differing by a factor 2 from the squared Hellinger distance.

The convexity in the next theorem implies that $I_\alpha(P,Q)$ as a function of parameter $\alpha \in \mathbb{R}$ is continuous in the effective domain $\{\alpha \in \mathbb{R} : I_\alpha(P,Q) < \infty\}$ which always includes the interval $(0,1)$. Liese and Vajda proved in [33, Proposition 2.14] the continuity from the left and right at the endpoints of this domain. Thus, in particular

$$I(P,Q) = I_1(P,Q) = \lim_{\alpha \uparrow 1} I_\alpha(P,Q)$$

irrespectively of whether $I(P,Q) < \infty$ or $I(P,Q) = \infty$ but, obviously, $\alpha \uparrow 1$ cannot be replaced by $\alpha \to 1$ if $I(P,Q) < \infty$ and $I_\alpha(P,Q) = \infty$ for all $\alpha > 1$. This situation takes place, e.g., if P is doubly exponential on $\mathcal{X} = \mathbb{R}$ with the density $p(x) = e^{-2|x|}$ and Q is standard normal with the density $q(x) = e^{-x^2/2}/\sqrt{2\pi}$.

Theorem 6: The function $\alpha \mapsto I_\alpha(P,Q)$ is convex on $\mathbb{R}$.

For the proof see the argument in Example 7 of Section V.

IV. Divergences and Shannon Information

Consider the communication channel $\langle \mathcal{Y}, \{P,Q\}, (\mathcal{X}, \mathcal{A}) \rangle$ with a binary input $\mathcal{Y}$ consisting of two messages H (hypothesis) and A (alternative), and with conditional distributions P and Q over a measurable output space $(\mathcal{X}, \mathcal{A})$ under H and A, respectively. Channels of this "semicontinuous" type were considered already by Shannon [46] and systematically studied later by Feinstein [19, Ch. 5]. Let $\pi \in (0,1)$ define the input distribution $(\pi, 1-\pi)$ on $\mathcal{Y} = \{H, A\}$, i.e., let $(\mathcal{Y}, (\pi, 1-\pi))$ be the input source generating a random input Y. Then $((\mathcal{X}, \mathcal{A}), W)$ with $W = \pi P + (1-\pi)Q$ is the output source generating a random output X. As in the previous

section, let p and q be densities of P and Q with respect to a σ-finite measure μ on $(\mathcal{X},\mathcal{A})$. Then

$$w=\frac{dW}{d\mu}=\pi p+(1-\pi)q \tag{39}$$

is the output probability density and $\pi(x)$ and $1-\pi(x)$ with

$$\pi(x)=\begin{cases}\frac{\pi p(x)}{w(x)}, & \text{for } w(x)>0\\ \pi, & \text{for } w(x)=0\end{cases} \tag{40}$$

are the conditional probabilities $\Pr(Y=H|X=x)$, $\Pr(Y=A|X=x)$ of the input variable Y under values $x\in\mathcal{X}$ of the output variable X. The Shannon information $I(X;Y)$ in the output X about the input Y is usually defined by the formula

$$I(X;Y)=H(Y)-H(Y|X) \tag{41}$$

where

$$H(Y)=h(\pi):=-\pi\ln\pi-(1-\pi)\ln(1-\pi) \tag{42}$$

is the unconditional input entropy and

$$\begin{aligned}H(Y|X)&=\int H(Y|X=x)\,dW(x)\\&=\int h(\pi(x))\,w(x)\,d\mu(x)\end{aligned} \tag{43}$$

is the conditional input entropy given the output X. Since we use the natural logarithms, the entropies and the information is not given in *bits* but in *nats*.

It deserves to be mentioned here that for any pair of random variables X,Y distributed by $P_{X,Y}$ on a measurable space $(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})$ with marginals P_X,P_Y on $(\mathcal{X},\mathcal{A}),(\mathcal{Y},\mathcal{B})$, the Shannon information $I(X;Y)$ is the information divergence $I(P_{X,Y},P_X\otimes P_Y)$ where $P_X\otimes P_Y$ is the product of P_X and P_Y on $(\mathcal{X}\times\mathcal{Y},\mathcal{A}\otimes\mathcal{B})$. This relation between the Shannon information and the information divergence is well known (see, e.g., Csiszár and Körner [13] or Cover and Thomas [10]). In our situation, the direct verification of the formula

$$I(X;Y)=I(P_{X,Y},P_X\otimes P_Y)$$

is easy. It holds $P_Y=\pi\delta_H+(1-\pi)\delta_A$ and $P_X=W$. Then $P_X\otimes P_Y$ dominates $P_{X,Y}$ with the relative density

$$\frac{dP_{X,Y}}{d(P_X\otimes P_Y)}(x,y)=\mathbf{1}_{\{H\}}(y)\frac{p(x)}{w(x)}+\mathbf{1}_{\{A\}}(y)\frac{q(x)}{w(x)}.$$

Hence, by the definition of $I(P,Q)$ in (30)

$$\begin{aligned}&I(P_{X,Y},P_X\otimes P_Y)\\&=\int\ln\left(\frac{dP_{X,Y}}{d(P_X\otimes P_Y)}\right)dP_{X,Y}\\&=\int\left(\pi p(x)\ln\left(\frac{p(x)}{w(x)}\right)+(1-\pi)q(x)\ln\left(\frac{q(x)}{w(x)}\right)\right)d\mu(x)\end{aligned} \tag{44}$$

$$\begin{aligned}&=h(\pi)-\int h(\pi(x))\,w(x)\,d\mu(x)\\&=H(Y)-H(Y|X).\end{aligned} \tag{45}$$

Consider now for a fixed $\pi\in(0,1)$ the function

$$f_\pi(t)=\pi t\ln t-(\pi t+1-\pi)\ln(\pi t+1-\pi),\ t>0 \tag{46}$$

belonging to the class $\mathcal{F}_1$ of convex functions studied in the previous sections. It is twice differentiable in t with $f_\pi''>0$ which guarantees the strict convexity on $(0,\infty)$. Denote for simplicity by $\mathbb{I}_\pi(P,Q)$ the f_π-divergence of P,Q, i.e., put

$$\mathbb{I}_\pi(P,Q)=D_{f_\pi}(P,Q)\,. \tag{47}$$

By (46), for $w=\pi p+(1-\pi)q$ it holds that

$$f_\pi\left(\frac{p}{q}\right)q=\pi p\ln\left(\frac{p}{w}\right)+(1-\pi)q\ln\left(\frac{q}{w}\right)$$

which implies

$$\mathbb{I}_\pi(P,Q)=\pi\,I(P,W)+(1-\pi)\,I(Q,W). \tag{48}$$

By Theorem 5, $\mathbb{I}_\pi(P,Q)$ takes on values between 0 and $f_\pi(0)+f_\pi^*(0)=h(\pi)$ where $\mathbb{I}_\pi(P,Q)=0$ if and only if $P=Q$ and $\mathbb{I}_\pi(P,Q)=h(\pi)$ if and only if $P\perp Q$.

Theorem 7: For every channel $\langle\{H,A\},\{P,Q\},(\mathcal{X},\mathcal{A})\rangle$ and every input distribution $(\pi,1-\pi)$ with $\pi\in(0,1)$, the Shannon information $I(X;Y)$ coincides with the f-divergence of the output conditional distributions P,Q for $f=f_\pi$ given by (46), i.e.,

$$I(X;Y)=\mathbb{I}_\pi(P,Q)=\pi I(P,W)+(1-\pi)I(Q,W)\,. \tag{49}$$

Proof: By (44) we have

$$I(X;Y)=\pi I(P,W)+(1-\pi)I(Q,W).$$

The rest is clear from (48). □

The equality $I(X;Y)=\pi I(P,W)+(1-\pi)I(Q,W)$ in (49) is well known in information theory, see, e.g., Topsøe [49] who called the divergence $I(P,W)+I(Q,W)=2\mathbb{I}_\pi(P,Q)$ *capacitory discrimination*. The fact that the Shannon information $I(X;Y)$ is the f-divergence for $f=f_\pi$ defined by (46) seems to be overlooked in the earlier literature. The f-divergence interpretation of $I(X;Y)$ has some nontrivial consequences, e.g., the triangle inequality for $\sqrt{I(X;Y)}$ which will be proved later. The equality $I(X;Y)=\mathbb{I}_\pi(P,Q)$ motivates us to call the divergences $\mathbb{I}_\pi(P,Q)$ for $\pi\in(0,1)$ *the Shannon divergences*. The maximal Shannon divergence

$$C(P,Q)=\sup_{\pi\in(0,1)}\mathbb{I}_\pi(P,Q)\le\ln 2$$

is the capacity of the channel under consideration.

In what follows, we consider for $\alpha\in(0,1)$ the entropies $H_\alpha(Y)$ and the conditional entropies $H_\alpha(Y|X)$ of Arimoto [2]. He proved that the Shannon entropies $H(Y)$ and $H(Y|X)$ are limits for $\alpha\uparrow 1$ of $H_\alpha(Y)$ and $H_\alpha(Y|X)$, respectively (cf. in this respect (59) and (60) below). In this paper, we introduce the *Arimoto informations*

$$I_\alpha(X;Y)=H_\alpha(Y)-H_\alpha(Y|X) \tag{50}$$

as natural extensions of the Shannon information

$$I(X;Y) = I_1(X;Y) := \lim_{\alpha \uparrow 1} I_\alpha(X;Y) \tag{51}$$

to the domain $\alpha \in (0,1)$ and find an f-divergence representations of these informations which smoothly extend the representation $\mathbb{I}_\pi(P,Q)$ of $I(X;Y)$ obtained above.

Let us start with the definitions of the Arimoto entropies for the binary channel input Y distributed unconditionally by $(\pi, 1-\pi)$ and conditionally (for a given $X = x$) by $(\pi(x), 1-\pi(x))$. Consider for $\alpha \in (0,1)$ and $t \in (0,\infty)$ the functions

$$h_\alpha(\pi|t) = \frac{1}{1-\alpha}\left(1 - \left[(\pi t)^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha\right)$$

$h_\alpha(\pi) = h_\alpha(\pi|1)$ of variable $\pi \in [0,1]$. According to [2],

$$H_\alpha(Y) = h_\alpha(\pi) = \frac{1}{1-\alpha}\left(1 - \left[\pi^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha\right) \tag{52}$$

is the Arimoto entropy of the channel input Y and

$$\begin{aligned} H_\alpha(Y|X) &= \int H_\alpha(Y|X=x)\, dW(x) \\ &= \int h_\alpha(\pi(x))\, w(x)\, d\mu(x) \ \text{(cf. (40))} \end{aligned} \tag{53}$$

is the conditional Arimoto entropy of the channel input given the output X. The entropy (52) is a similar measure of uniformity of the input distribution $(\pi, 1-\pi)$ as the Shannon entropy (42). It is concave in π which guarantees, similarly as in the case of Shannon entropy, that the informations (50) are nonnegative (for more about concave entropies see Morales *et al.* [36] and Remark 3 below).

Let us now consider for fixed $\alpha \in (0,1)$ and $\pi \in (0,1)$ the function

$$\begin{aligned} f_{\pi,\alpha}(t) &= h_\alpha(\pi) - h_\alpha(\pi|t) \\ &= \frac{1}{1-\alpha}\left(\left[(\pi t)^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha \right. \\ &\quad \left. - \left[\pi^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha\right) \end{aligned} \tag{54}$$

of variable $t \in (0,\infty)$. Taking the second derivative, it is easy to see that this function is strictly convex and belongs to the class $\mathcal{F}_1$ considered in previous sections. The constants considered in Theorem 5 are

$$\begin{aligned} f_{\pi,\alpha}(0) &= \frac{1 - \pi - \left[\pi^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha}{1-\alpha} \\ f^*_{\pi,\alpha}(0) &= \frac{\pi}{1-\alpha} \end{aligned} \tag{55}$$

so that

$$f_{\pi,\alpha}(0) + f^*_{\pi,\alpha}(0) = h_\alpha(\pi) < \infty. \tag{56}$$

In accordance with Definition 2, the function $f_{\pi,\alpha}$ in (54) defines the $f_{\pi,\alpha}$-divergence

$$\mathbb{I}_{\pi,\alpha}(P,Q) = D_{f_{\pi,\alpha}}(P,Q) \tag{57}$$

$$\begin{aligned} &= \frac{1}{1-\alpha}\left(\int \left[(\pi p)^{1/\alpha} + ((1-\pi)q)^{1/\alpha}\right]^\alpha d\mu \right. \\ &\quad \left. - \left[\pi^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha\right) \end{aligned} \tag{58}$$

which is zero if and only if $P = Q$ and attains the maximal value (56) if and only if $P \perp Q$. The divergences (57) with arbitrary $\alpha, \pi \in (0,1)$ will be called *Arimoto divergences*.

We next formulate a theorem extending Theorem 7. To this end, we need to take into account that the functions $h_\alpha(\pi) = H_\alpha(Y)$ of variable π given in (52) are nonnegative and uniformly bounded above by $h_\alpha(1/2) = (1 - 2^{\alpha-1})/(1-\alpha) \le \ln 2$ for all $\alpha \in (0,1)$ (this bound follows from the inequality $e^{-t} \ge 1 - t$). By the L'Hospital rule, the functions $h_\alpha(\pi)$ converge to $h(\pi) = H(Y)$ for $\alpha \uparrow 1$. In other words

$$H_\alpha(Y) \longrightarrow H_1(Y) := H(Y) \tag{59}$$

for $\alpha \uparrow 1$. The convergence of functions h_α to h together with the dominated convergence theorem implies also the convergence of the integrals (53), i.e.,

$$H_\alpha(Y|X) \longrightarrow H_1(Y|X) := H(Y|X). \tag{60}$$

Consequently, the corresponding differences (50) converge as well

$$I_\alpha(X;Y) \longrightarrow I_1(X;Y) := I(X;Y). \tag{61}$$

By (50) and definitions (52), (53) of the Arimoto entropies, we find that the Arimoto informations $I_\alpha(X;Y)$ coincide with the expression (57) for the $f_{\pi,\alpha}$–divergence $\mathbb{I}_{\pi,\alpha}(P,Q)$. Hence, the convergence (61) implies

$$\mathbb{I}_{\pi,\alpha}(P,Q) \longrightarrow \mathbb{I}_{\pi,1}(P,Q) := \mathbb{I}_\pi(P,Q) \tag{62}$$

for $\alpha \uparrow 1$. Thus, the Shannon entropies, informations, and divergences are special cases of the Arimoto entropies, informations, and divergences smoothly extended to the domain $\alpha \in (0,1]$. In particular, the following assertion holds.

Theorem 8: For every channel $\langle \{H,A\}, \{P,Q\}, (\mathcal{X}, \mathcal{A})\rangle$ with input distribution $(\pi, 1-\pi)$ where $\pi \in (0,1)$, and for every $\alpha \in (0,1]$, the Arimoto information $I_\alpha(X;Y)$ coincides with the f-divergence of the output conditional distributions P, Q for $f = f_{\pi,\alpha}$ given by (54), i.e.,

$$I_\alpha(X;Y) = \mathbb{I}_{\pi,\alpha}(P,Q) \tag{63}$$

where $\mathbb{I}_{\pi,\alpha}(P,Q) = \mathbb{I}_\pi(P,Q)$ is given by (48) if $\alpha = 1$ and by (57) if $\alpha \in (0,1)$.

A pleasing property of the Shannon and Arimoto divergences is that

$$\varrho_\alpha(P,Q) = \sqrt{\mathbb{I}_{1/2,\alpha}(P,Q)}, \qquad \alpha \in (0,1]$$

are metrics in the space of probability measures on any measurable space $(\mathcal{X}, \mathcal{A})$. Since the symmetry of $\varrho_\alpha(P,Q)$ in P, Q is clear from the definition of $\mathbb{I}_{1/2,\alpha}(P,Q)$ and the reflexivity

of $\varrho_\alpha(P,Q)$ follows from Theorem 5, the only open metric problem is the triangle inequality

$$\varrho_\alpha(P,Q) \le \varrho_\alpha(P,\tilde{P}) + \varrho_\alpha(\tilde{P},Q) \tag{64}$$

for probability distributions $\tilde{P}$ on $(\mathcal{X},\mathcal{A})$. For $\alpha \in (0,1)$, this inequality was proved by Österreicher [40]. For $\alpha = 1$, the inequality (64) extends automatically by the continuity (62). The Arimoto divergences $\mathbb{I}_{1/2,\alpha}(P,Q)$ were introduced for all $\alpha \in \mathbb{R}$ by Österreicher and Vajda [42] who proved the triangle inequality (64) for all $\alpha \in (0,1]$ and found also the metric properties of these divergences for $\alpha > 1$. The inequality (64) for the special case $\alpha = 1$ and finite observation space $\mathcal{X}$ was established also in [17].

By combining these results with Theorem 8, we obtain the following surprising property of the Shannon and Arimoto informations.

Theorem 9: The square roots $\sqrt{I_\alpha(X;Y)}$ of the Arimoto and Shannon informations transmitted via channels $\langle \mathcal{Y}, \{P,Q\}, (\mathcal{X},\mathcal{A})\rangle$ with uniformly distributed binary inputs are metrics in the space of conditional output distributions P,Q.

This theorem provides upper bounds for the information about uniform sources transmitted via channels with binary inputs. Let us illustrate this by a simple example.

Example 5: Let us consider a binary-symmetric channel (BSC)$\langle \{0,1\}, \{P,\tilde{P}\}, \{0,1\}\rangle$ with the conditional distributions $P = (1-\varepsilon,\varepsilon), \tilde{P} = (\varepsilon, 1-\varepsilon)$. If P or $\tilde{P}$ is replaced by the uniform distribution $Q = (1/2, 1/2)$ then we obtain a "half-blind" channel with a totally noisy response to the input with the distribution Q. The problem is how many parallel "half-blind" channels are needed to replace one BSC if the inputs are chosen with equal probabilities $(\pi, 1-\pi) = (1/2, 1/2)$. In other words, we ask what is the minimal k such that the Shannon information $I(\tilde{X};Y)$ transmitted by the BSC is exceeded by k times the information $I(X;Y)$ transmitted by the "half-blind" channels. Since the informations transmitted by the half-blind channels

$$\langle \{0,1\}, \{P,Q\}, \{0,1\}\rangle \quad \text{and} \quad \langle \{0,1\}, \{Q,P\}, \{0,1\}\rangle$$

with equiprobable inputs coincide, their common value $I(X;Y)$ must satisfy the triangle inequality $\sqrt{I(\tilde{X};Y)} \le 2\sqrt{I(X;Y)}$, i.e., $I(\tilde{X};Y) \le 4I(X;Y)$. From here we see that $k \le 4$. By a direct calculation one can verify that $k = 3$ is not enough, at least for the error probabilities ε close to 0.

Theorem 9 also provides a lower bound for the information transmitted by two channels $\langle \mathcal{Y}, \{P,\tilde{P}\}, (\mathcal{X},\mathcal{Y})\rangle$ and $\langle \mathcal{Y}, \{\tilde{P},Q\}, (\mathcal{X},\mathcal{Y})\rangle$ with one common conditional distribution $\tilde{P}$ when the input distribution is uniform and P,Q are orthogonal.

Example 6: Let us consider two channels of the mentioned type for $P \perp Q$ so that $\mathbb{I}_{1/2,\alpha}(P,Q)$ is the entropy $h_\alpha(1/2)$. If $I_\alpha^{(1)}(X;Y)$ and $I_\alpha^{(2)}(X;Y)$ are the informations transmitted by the first and second channel then

$$\sqrt{I_\alpha^{(1)}(X;Y)} + \sqrt{I_\alpha^{(2)}(X;Y)} \ge \sqrt{h_\alpha(1/2)} = \sqrt{\frac{1-2^{1-\alpha}}{1-\alpha}}$$

if $\alpha \in (0,1)$ and in the case of Shannon informations $I^{(1)}(X;Y)$ and $I^{(2)}(X;Y)$

$$\sqrt{I^{(1)}(X;Y)} + \sqrt{I^{(2)}(X;Y)} \ge \sqrt{\ln 2}\,.$$

V. Divergences and Statistical Information

The information-theoretic model $\langle \mathcal{Y}, \{P,Q\}, (\mathcal{X},\mathcal{A})\rangle$ of a channel with a binary random input $Y \sim \langle \mathcal{Y}, (\pi, 1-\pi)\rangle$ where $\mathcal{Y} = \{H,A\}$, and with a general output $X \sim \langle (\mathcal{X},\mathcal{A}), W\rangle$ where $W = \pi P + (1-\pi)Q$, can be statistically interpreted in two equivalent ways: I. As the classical statistical model of testing a π-probable hypothesis $H : P$ against a $(1-\pi)$-probable alternative $A : Q$ on the basis of observation X. II. As the model of Bayesian statistics where decisions from the space $\{H,A\}$ are based on observations X conditionally distributed by P if $Y = H$ and by Q if $Y = A$. *A priori* probabilities of $Y = H$ and $Y = A$ are π and $1-\pi$ and the loss is 0 or 1 depending on whether the decision coincides with Y or not.

In Case II, the (randomized) decision functions are measurable mappings

$$\varphi : \mathcal{X} \longmapsto [0,1] \tag{65}$$

where $\varphi(x)$ and $1-\varphi(x)$ are probabilities of the decisions A, H when the observation is $X = x$. The problem is to characterize the optimal decision functions φ_{opt} called Bayes decision functions achieving the minimal loss

$$B_\pi(P,Q) = \inf_\varphi \left\{ \pi \int \varphi\, dP + (1-\pi) \int (1-\varphi)\, dQ \right\} \tag{66}$$

called Bayes loss. In Case I, the test is a mapping (65) where $\varphi(x)$ is the probability that H is rejected. Therefore, $\int \varphi\, dP$ and $\int (1-\varphi)\, dQ$ are the error probabilities of two kinds of the test. This means that φ_{opt} are Bayes tests achieving the minimal average error (66). To characterize the Bayes tests notice that

$$\begin{aligned} &\pi \int \varphi\, dP + (1-\pi) \int (1-\varphi)\, dQ \\ &\quad = \int (1-\pi) q\, d\mu + \int \varphi(\pi p - (1-\pi) q)\, d\mu \\ &\quad = \int (1-\pi) q\, d\mu + \int \varphi(\pi(x) - (1-\pi(x)))\, dW(x) \end{aligned} \tag{67}$$

where $\pi(x)$ is from (40). Hence, the Bayes tests satisfy W-a.s. the well-known necessary and sufficient condition

$$\varphi_{\text{opt}}(x) = \begin{cases} 1, & \text{if } \pi(x) < 1-\pi(x) \\ 0, & \text{if } \pi(x) > 1-\pi(x) \end{cases} \tag{68}$$

see, e.g., De Groot [16]. We see from (68) that $\varphi(x) = \pi(x) \wedge (1-\pi(x))$ is a Bayes test so that

$$B_\pi(P,Q) = \int \pi(x) \wedge (1-\pi(x))\, dW(x) = \int B_{\pi(x)}\, dW(x) \tag{69}$$

where

$$B_\pi = \pi \wedge (1-\pi), \qquad \pi \in (0,1). \tag{70}$$

If $Q = P$ then $\pi(x)$ is the constant π for all $x \in \mathcal{X}$ (cf. (40)) so that the optimal decision (68) depends on the *a priori* probability π only and not on the observations $x \in \mathcal{X}$. Therefore,

$$B_\pi(P,P) = B_\pi \tag{71}$$

is the *a priori* Bayes loss.

Remark 3: The inequality

$$(\pi x_1 + (1-\pi)y_1) \wedge (\pi x_2 + (1-\pi)y_2) \geq \pi(x_1 \wedge x_2) + (1-\pi)(y_1 \wedge y_2)$$

shows that $(x,y) \mapsto x \wedge y$ is a concave function of two variables. Hence, by Jensen's inequality

$$\begin{aligned} B_\pi(P,Q) &= \int \pi(x) \wedge (1-\pi(x))\, dW(x) \\ &\geq \left(\int \pi(x)\, dW(x)\right) \wedge \left(\int (1-\pi(x))\, dW(x)\right) \\ &= \pi \wedge (1-\pi)\,. \end{aligned}$$

Therefore, $B_\pi(P,Q) \leq B_\pi$ holds for all P, Q, i.e., the *a posteriori* Bayes loss cannot exceed the *a priori* Bayes loss.

The following definition is due to De Groot [15], [16].

Definition 3: The difference

$$\mathcal{I}_\pi(P,Q) = B_\pi - B_\pi(P,Q) \tag{72}$$

between the *a priori* and *a posteriori* Bayes loss is the *statistical information* in the model $\langle (\mathcal{X}, \mathcal{A}), \{P,Q\}, \pi \rangle$.

The next theorem shows that the statistical information is the f-divergence of P, Q for $f = f_{\pi,0}$ defined by

$$f_{\pi,0}(t) = \pi \wedge (1-\pi) - \pi \wedge (1-\pi t), \quad t \in (0,\infty) \tag{73}$$

which is nothing but the limit of $f_{\pi,\alpha}(t)$ defined in (54) for $\alpha \downarrow 0$ as

$$\lim_{\alpha \downarrow 0} (a^{1/\alpha} + b^{1/\alpha})^\alpha = a \vee b := \max\{a,b\}. \tag{74}$$

It also shows that there is an intimate connection between $B_\pi, B_\pi(P,Q), \mathcal{I}_\pi(P,Q)$ and the Arimoto's $H_\alpha(Y)$, $H_\alpha(Y|X), I_\alpha(X;Y)$ for α close to 0.

Theorem 10: The Arimoto entropies $H_\alpha(Y)$, $H_\alpha(Y|X)$ and informations $I_\alpha(X;Y)$ continuously extend to $\alpha = 0$ and the extensions satisfy the relations

$$\begin{aligned} H_0(Y) &= B_\pi, \; H_0(Y|X) = B_\pi(P,Q) \\ I_0(X;Y) &= \mathcal{I}_\pi(P,Q) \end{aligned} \tag{75}$$

i.e., the Bayes losses are extended Arimoto entropies and the statistical information is an extended Arimoto information. Moreover, the divergences $\mathbb{I}_{\pi,\alpha}(P,Q)$ continuously extend to $\alpha = 0$ and $\mathbb{I}_{\pi,0}(P,Q)$ is the f-divergence for $f = f_{\pi,0} \in \mathcal{F}_1$ defined by (73) which coincides with the statistical information, i.e.,

$$\mathbb{I}_{\pi,0}(P,Q) = \mathcal{I}_\pi(P,Q)\,. \tag{76}$$

Proof: From (74), we deduce that the functions

$$\begin{aligned} h_\alpha(\pi) &= \frac{1}{1-\alpha}\left(1 - \left[(\pi)^{1/\alpha} + (1-\pi)^{1/\alpha}\right]^\alpha\right) \\ &= H_\alpha(Y) \end{aligned}$$

tend to the function $\pi \wedge (1-\pi) = B_\pi$ for $\alpha \downarrow 0$. From here, one easily obtains for $\alpha \downarrow 0$ the mirror images of (59)–(61) with $H_1(Y), H_1(Y|X), I_1(X;Y)$ replaced by $H_0(Y)$, $H_0(Y|X), I_0(X;Y)$ and with the Shannon quantities $H(Y)$, $H(Y|X), I(X;Y)$ replaced by the statistical quantities B_π, $B_\pi(P,Q), \mathcal{I}_\pi(P,Q)$. This means that the continuous extensions (75) hold. Further, from (74) we get $\lim_{\alpha \downarrow 0} f_{\pi,\alpha} = f_{\pi,0}$ for $f_{\pi,0}$ defined by (73) as mentioned above. Since these functions are for every π bounded uniformly for all $\alpha \in (0,1)$ on the domain $t \in (0,\infty)$, this implies also the convergence of the corresponding integrals $\lim_{\alpha \downarrow 0} \mathbb{I}_{\pi,\alpha}(P,Q) = \mathbb{I}_{\pi,0}(P,Q)$. The equality (76) follows from the equalities

$$I_\alpha(X;Y) = \mathbb{I}_{\pi,\alpha}(P,Q), \qquad \alpha \in (0,1)$$

and from the already established convergences $I_\alpha(X;Y) \to \mathcal{I}_\pi(P,Q)$ and $\mathbb{I}_{\pi,\alpha}(P,Q) \to \mathbb{I}_{\pi,0}(P,Q)$ for $\alpha \downarrow 0$. □

By (73), (10), and (70), $f_{\pi,0}(0) + f^*_{\pi,0}(0) = B_\pi$. Hence, it follows from Theorem 5 that the statistical information $\mathcal{I}_\pi(P,Q)$ takes on values between 0 and B_π and $\mathcal{I}_\pi(P,Q) = 0$ if and only if $P = Q$, $\mathcal{I}_\pi(P,Q) = B_\pi$ if and only if $P \perp Q$.

The continuity stated in Theorem 10 implies that the triangle inequality (64) extends to $\alpha = 0$, i.e., that if $\pi = 1/2$ then the square root $\sqrt{\mathcal{I}_{1/2}(P,Q)}$ of the statistical information satisfies the triangle inequality on the space of probability measures on $(\mathcal{X}, \mathcal{A})$. In fact, the statistical information $\mathcal{I}_{1/2}(P,Q)$ itself satisfies the triangle inequality and is thus a metric. This result is not surprising because, by the definitions of $B_{1/2}$ and $B_{1/2}(P,Q)$

$$\mathcal{I}_{1/2}(P,Q) = \frac{1}{2} V(P,Q) \tag{77}$$

where $V(P,Q)$ is the total variation (33) which is a well-known metric.

Remark 4: The functions $f_{\pi,0}$ given in (73), defining the statistical informations $\mathcal{I}_\pi(P,Q)$ as $f_{\pi,0}$-divergences, are not strictly convex at 1 unless $\pi = 1/2$. Looking closer at them we find that if the likelihood ratio p/q is on $\mathcal{X}$ bounded, and bounded away from 0, then $\mathcal{I}_\pi(P,Q) = 0$ for all π sufficiently close to 0 and 1. More generally, if $\pi \neq 1/2$ then the statistical information $\mathcal{I}_\pi(P,Q)$ is insensitive to small local deviations of Q from P. This insensitivity increases as π deviates from $1/2$ and grows into a total ignorance of P, Q in the proximity of $\pi = 0$ and $\pi = 1$. Similarly, $\mathcal{I}_\pi(P,Q)$ achieves for $\pi \neq 1/2$ the maximal value B_π at nonsingular P, Q and this effect increases too as π deviates from $1/2$. Thus, $\mathcal{I}_\pi(P,Q)$ exhibits a kind of hysteresis which disappears when $\pi = 1/2$. Nevertheless, the collection $\{\mathcal{I}_\pi(P,Q) : \pi \in (0,1)\}$ of statistical informations in all Bayesian models $\langle (\pi, 1-\pi), \{P,Q\}, (\mathcal{X}, \mathcal{A}) \rangle$ completely characterizes the informativity of the non-Bayesian model $\langle \{P,Q\}, (\mathcal{X}, \mathcal{A}) \rangle$, as it will be seen in the next section.

The final aim of this section is to show that every f-divergence is an average statistical information $\mathcal{I}_\pi(P,Q)$ where the average is taken over *a priori* probabilities $\pi \in (0,1)$ according to the σ-finite measure Γ_f defined on Borel subsets of $(0,1)$ by

$$\Gamma_f((\pi_1,\pi_2]) = \int_{\left(\frac{1-\pi_2}{\pi_2},\frac{1-\pi_1}{\pi_1}\right]} (1+t)\,df'_+(t) = \int_{(\pi_1,\pi_2]} \frac{1}{\pi}\,dg_f(\pi) \tag{78}$$

for the nondecreasing function $g_f(\pi) = -f'_+((1-\pi)/\pi)$, $\pi \in (0,1)$.

The following theorem is not new—representations of f-divergences as average statistical informations of De Groot were already established in Österreicher and Vajda [41]. Representation of f-divergences by means of the Bayes losses was obtained under some restrictions on f or P,Q already by Feldman and Österreicher [20] and Guttenbrunner [22]. The formulation presented here is general and the proof is simpler, obtained on few lines by application of the generalized Taylor expansion (Theorem 1) in Definition 2.

Theorem 11: Let $f \in \mathcal{F}_1$ and let Γ_f be the above defined measure on $(0,1)$. Then for arbitrary probability measures P,Q

$$D_f(P,Q) = \int_{(0,1)} \mathcal{I}_\pi(P,Q)\,d\Gamma_f(\pi). \tag{79}$$

Proof: We obtain from (6) and (7) in Theorem 1 that the integral in Definition 2 is the sum

$$\begin{aligned}&\int_{\{pq>0\}}\int_{(1,\infty)} [p-(tq)\wedge p]\,df'_+(t)\,d\mu \\ &+\int_{\{pq>0\}}\int_{(0,1]} [tq-(tq)\wedge p]\,df'_+(t)\,d\mu \\ &= \int_{(1,\infty)} \left[P(q>0)-(1+t)B_{1/(1+t)}(P,Q)\right]\,df'_+(t) \\ &+\int_{(0,1]} \left[tQ(p>0)-(1+t)B_{1/(1+t)}(P,Q)\right]\,df'_+(t)\end{aligned}$$

where we used the relation

$$B_\pi(P,Q) = \int_{\{pq>0\}} [\pi p \wedge (1-\pi)q]\,d\mu$$

which follows from he definition of $B_\pi(P,Q)$ in (69). As $f \in \mathcal{F}_1$, Theorem 1 implies

$$f(t) = \int_{(1,t]} (t-s)df'_+(s), \qquad \text{for } t>1$$

so that, by the monotone convergence theorem

$$f^*(0) = \lim_{t\to\infty}\frac{f(t)}{t} = \int_{(1,\infty)} df'_+(s).$$

Similarly, by Theorem 1, $f(0) = \int_{(0,1]} s\,df'_+(s)$. Substituting these expressions in (20) we obtain

$$\begin{aligned}D_f(P,Q) &= \int_{(1,\infty)} \left[1-(1+t)B_{1/(1+t)}(P,Q)\right]\,df'_+(t) \\ &\quad+\int_{(0,1]} \left[t-(1+t)B_{1/(1+t)}(P,Q)\right]\,df'_+(t) \\ &= \int_{(0,\infty)} (1+t)\mathcal{I}_{1/(1+t)}(P,Q)\,df'_+(t) \\ &= \int_{(0,1)} \mathcal{I}_\pi(P,Q)\,d\Gamma_f(\pi)\end{aligned}$$

where the last equality follows by substitution $1/(1+t) \mapsto \pi$ from the relation

$$\Gamma_f(B) = \int \mathbf{1}_B\left(\frac{1}{1+t}\right)(1+t)\,df'_+(t)$$

satisfied by the measure Γ_f for every Borel set $B \subseteq (0,1)$. □

Notice that the measure Γ_f on $(0,1)$ may be defined also by the formula

$$\Gamma_f(B) = \int \mathbf{1}_B\left(\frac{1}{1+t}\right)(1+t)\,d\lambda_f(t)$$

where λ_f is the measure on $(0,\infty)$ defined by (3).

The representation (79) suggests the interpretation of the f-divergences $D_f(P,Q)$ as wideband statistical informations in which the "narrowband informations" $\mathcal{I}_\pi(P,Q)$ participate with the infinitesimal weights

$$\frac{1}{\pi}\,dg_f(\pi) = \frac{1}{\pi^3}\,f''\left(\frac{1-\pi}{\pi}\right)\,d\pi \tag{80}$$

(if f' is differentiable) depending on f, and the "band" means the interval $(0,1)$ of *a priori* probabilities π. Thus, various f-divergences mutually differ only in the weight attributed to possible *a priori* distributions in statistical models.

Example 7: The divergences $I_\alpha(P,Q)$ of Example 3 can be interpreted for all $\alpha \in \mathbb{R}$ as the wideband statistical informations

$$I_\alpha(P,Q) = \int_0^1 \mathcal{I}_\pi(P,Q) w_\alpha(\pi)\,d\pi \tag{81}$$

where the narrowband components $\mathcal{I}_\pi(P,Q)$ contribute by the weights

$$w_\alpha(\pi) = \left(\frac{1-\pi}{\pi}\right)^\alpha \frac{1}{(1-\pi)^2\pi} \quad \text{(cf. (80))}. \tag{82}$$

For the most often considered $\alpha = 1$ and 0, $\alpha = 2$ and -1, and $\alpha = 1/2$

$$I(P,Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{(1-\pi)\pi^2}\,d\pi,\ I(Q,P) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{(1-\pi)^2\pi}\,d\pi, \tag{83}$$

$$\chi^2(P,Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{\pi^3}\,d\pi,\ \chi^2(Q,P) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{(1-\pi)^3}\,d\pi \tag{84}$$

and

$$H^2(P,Q) = \int_0^1 \frac{\mathcal{I}_\pi(P,Q)}{[(1-\pi)\pi]^{3/2}}\, d\pi. \tag{85}$$

Since $\mathcal{I}_\pi(P,Q)$ is bounded above by $B_\pi = \pi \wedge (1-\pi)$, it tends to zero at the endpoints of $(0,1)$. Thus, the powers of π and $1-\pi$ in the numerator strongly influence the properties of these divergences. Further, we see from (81), (82) that $I_\alpha(P,Q)$ is convex in the variable $\alpha \in \mathbb{R}$ because $((1-\pi)/\pi)^\alpha$ in (82) is convex in α for every fixed $\pi \in (0,1)$.

Example 8: By (77), the total variation $V(P,Q)$ is an example of f-divergence which is narrowband in the sense considered above. The Hellinger integrals of Example 4 are average Bayes losses

$$H_\alpha(P,Q) = \alpha(1-\alpha)\int_0^1 \frac{B_\pi(P,Q)}{(1-\pi)^{2-\alpha}\pi^{1+\alpha}}\, d\pi,\ \alpha \in (0,1). \tag{86}$$

This particular formula was already established by Torgersen [50, p. 45].

In the following corollary of Theorem 11, we use a measure alternative to Γ_f but still related to the measure λ_f of (3).

Corollary 1: Let γ_f be the σ-finite measure defined for every $f \in \mathcal{F}$ on the Borel subsets of $(0,\infty)$ by the condition

$$\gamma_f((a,b]) = \int_{(a,b]} (1+t)\, d\lambda_f(t) = \int_{(a,b]} (1+t)\, df'_+(t) \tag{87}$$

for all $0 < a < b < \infty$. Then

$$D_f(P,Q) = f(1) + \int_{(0,\infty)} \mathcal{I}_{1/(1+t)}(P,Q)\, d\gamma_f(t) \tag{88}$$

is the integral representation of an arbitrary divergence with $f \in \mathcal{F}$.

Proof: Notice that the measure Γ_f remains unchanged if we turn from f to $f - f(1)$. To complete the proof, it suffices to change the integration variable in (79) by setting $\pi = 1/(1+t)$. □

VI. Divergences, Deficiency, and Sufficiency

In this section, we introduce the classical statistical concepts of informativity, namely, the deficiency and sufficiency of observation channels. We study their relations to the divergences and informations investigated in the previous sections. To this end, we need a special subclass of the class of convex functions $\mathcal{F}$ considered above.

Let $\mathcal{F}_0$ be the class of all nonincreasing convex functions $f \in \mathcal{F}$ such that

$$f(0) := \lim_{t \downarrow 0} f(t) = 0 \tag{89}$$

and

$$\lim_{t\to\infty} f(t) > -\infty. \tag{90}$$

In the next lemma, we summarize properties of the class $\mathcal{F}_0$.

Lemma 1: If $f \in \mathcal{F}_0$ then

$$\gamma_f((0,\infty)) = -\lim_{t\to\infty} f(t) - \lim_{t\downarrow 0} f'_+(t). \tag{91}$$

Proof: Consider the $*$-conjugated function f^* defined in (10). By (10)

$$(f^*)'_+(s) = f\left(\frac{1}{s}\right) - \frac{1}{s} f'_+\left(\frac{1}{s}\right). \tag{92}$$

As f is convex, the derivative $(f^*)'_+$ is nondecreasing which implies that the limit of the left-hand side for $s \downarrow 0$ exists. The condition $f \in \mathcal{F}_0$ implies that $\lim_{t\to\infty} f(t)$ exists and is finite. Hence,

$$A = \lim_{s\to 0} f'_+(1/s)/s = \lim_{t\to\infty} t f'_+(t)$$

exists and satisfies the inequalities $-\infty \le A \le 0$. If $A < 0$, then there exist constants $a < 0$ and t_0 such that $f'_+(t) \le at$ for $t \ge t_0$. Then

$$\begin{aligned}\lim_{t\to\infty} f(t) &= \lim_{t\to\infty} \int_{t_0}^t f'_+(s)ds + f(t_0)\\ &\le \lim_{t\to\infty} a(\ln(t) - \ln(t_0)) + f(t_0) = -\infty\end{aligned}$$

which contradicts (90). Hence,

$$\lim_{t\to\infty} t f'_+(t) = 0. \tag{93}$$

Further, for $0 < a < b < \infty$ we get from (3)

$$\begin{aligned}\gamma_f((a,b]) &= \int_{(a,b]} (1+t)\lambda_f(dt)\\ &= f'_+(b) - f'_+(a) + \int_{(a,b]} \left(\int_{(a,t)} ds\right) \lambda_f(dt)\\ &= f'_+(b) - f'_+(a) + \int_{(a,b]} (f'_+(b) - f'_+(t))dt\\ &= f'_+(b) - f'_+(a) + (b-a)f'_+(b) - f(b) + f(a).\end{aligned}$$

But $\lim_{a\downarrow 0} f(a) = 0$ by (89) and $\lim_{b\to\infty} b f'_+(b) = 0$ by (93). Hence, we get (91) by taking $a \downarrow 0$ and $b \to \infty$ in the representation of $\gamma_f((a,b])$. □

Let

$$\mathcal{M}_0 = \langle (\mathcal{X}_0, \mathcal{A}_0), \{P_0, Q_0\}\rangle \text{ and } \mathcal{M}_1 = \langle (\mathcal{X}_1, \mathcal{A}_1), \{P_1, Q_1\}\rangle$$

be two binary statistical models in which we want to test the hypotheses $\mathsf{H} : P_0$ versus the alternative $\mathsf{A} : Q_0$ and $\mathsf{H} : P_1$ versus $\mathsf{A} : Q_1$, respectively. In order to compare the difficulty of the testing problem in these two models, we apply the concept of ε-deficiency introduced by LeCam [31] for general statistical models and $\varepsilon \ge 0$.

Definition 4: $\mathcal{M}_0$ is said to be ε-deficient with respect to $\mathcal{M}_1$ if for every test φ in the model $\mathcal{M}_1$ we find a test ψ in the model $\mathcal{M}_0$ such that

$$\int \psi dP_0 \le \int \varphi dP_1 + \varepsilon$$

and

$$\int (1-\psi) dQ_0 \le \int (1-\varphi) dQ_1 + \varepsilon. \tag{94}$$

Then we write $\mathcal{M}_1 \overset{\varepsilon}{\preceq} \mathcal{M}_0$. In the particular case $\mathcal{M}_1 \overset{0}{\preceq} \mathcal{M}_0$ we say that $\mathcal{M}_0$ is more informative than $\mathcal{M}_1$ or, equivalently, that $\mathcal{M}_1$ is less informative than $\mathcal{M}_0$.

Example 9: To give an example of a model less informative than some $\mathcal{M}_0 = \langle(\mathcal{X}_0, \mathcal{A}_0), \{P_0, Q_0\}\rangle$, consider a stochastic kernel $K : \mathcal{X}_0 \times \mathcal{A}_1 \mapsto [0,1]$, i.e., a family of probability measures $\{K(\cdot|x) : x \in \mathcal{X}_0\}$ on $(\mathcal{X}_1, \mathcal{A}_1)$ such that $K(B|x)$ is $\mathcal{A}_0$-measurable for every $B \in \mathcal{B}$. Then every distribution P_0 on $(\mathcal{X}_0, \mathcal{A}_0)$ defines a new probability measure KP_0 on $(\mathcal{X}_1, \mathcal{A}_1)$ by

$$(KP_0)(B) = \int K(B|x)\, dP_0(x) \tag{95}$$

and Q_0 defines similarly KQ_0. Thus, the kernel defines a new model

$$\mathcal{M}_1 = \langle(\mathcal{X}_1, \mathcal{A}_1), \{KP_0, KQ_0\}\rangle. \tag{96}$$

Special cases of kernels are measurable statistics $T : \mathcal{X} \mapsto \mathcal{Y}$ where $K(B|x) = \mathbf{1}_B(T(x))$, i.e., the kernel measures $K(\cdot|x)$ are the Dirac's $\delta_{T(x)}$. If $\psi : \mathcal{X}_2 \mapsto [0,1]$ is a test in the new model then $\varphi(x_1) = \int \psi(x_2) K(dx_2|x_1)$ defines a test in the original model $\mathcal{M}_0$ such that

$$\int \psi d(KP_0) = \int \left(\int \psi(x_2) K(dx_2|x_1) \right) dP(x_1) = \int \varphi dP_0.$$

Similarly

$$\int (1-\psi) d(KQ_0) = \int (1-\varphi) dQ_0.$$

This means that the new model $\mathcal{M}_1$ obtained by observing the outputs of the original model $\mathcal{M}_0$ through the observation channel $\langle(\mathcal{X}_0, \mathcal{A}_0), \{K(\cdot|x) : x \in \mathcal{X}_0\}, (\mathcal{X}_1, \mathcal{A}_1)\rangle$ is less informative than $\mathcal{M}_0$.

Next we characterize the ε-deficiency by means of the f-divergences. As before, by γ_f we denote the measure defined in (87).

Theorem 12: For any binary statistical models $\mathcal{M}_0 = \langle(\mathcal{X}_0, \mathcal{A}_0), \{P_0, Q_0\}\rangle$, $\mathcal{M}_1 = \langle(\mathcal{X}_1, \mathcal{A}_1), \{P_1, Q_1\}\rangle$, and $\varepsilon \geq 0$ the following statements are equivalent:

i) $\mathcal{M}_0 \overset{\varepsilon}{\preceq} \mathcal{M}_1$,
ii) $B_\pi(P_1, Q_1) \geq B_\pi(P_0, Q_0) + \varepsilon$ for every $0 < \pi < 1$,
iii) $D_f(P_0, Q_0) \leq D_f(P_1, Q_1) + \varepsilon\gamma_f((0,\infty))$, for every $f \in \mathcal{F}_0$.

Proof: Since $B_\pi(P_i, Q_i) = \mathcal{I}_\pi(P_i, Q_i) + B_\pi$, the implication ii) $\Longrightarrow$ iii) follows directly from the integral representation of f-divergences in (88). Further, for the convex function $f(t) = \pi \wedge (1-\pi) - (\pi t) \wedge (1-\pi)$, we obtain $D_f(P_i, Q_i) = B_\pi - B_\pi(P_i, Q_i)$ and

$$\begin{aligned}\gamma_f((0,\infty)) &= \int_{(0,\infty)} (1+t) df'_+(t) \\ &= \int_{(0,\infty)} (1+t)\pi d\delta_{(1-\pi)/\pi}(t) = 1.\end{aligned}$$

Therefore iii) implies ii). It remains to prove the equivalence i) $\Longleftrightarrow$ ii). If i) holds then for every test φ in $\mathcal{M}_0$ there exists a test ψ in $\mathcal{M}_1$ such that

$$\begin{aligned}\pi \int \psi dP_0 + (1-\pi) \int (1-\psi) dQ_0 \\ \leq \pi \int \varphi dP_1 + (1-\pi) \int (1-\varphi) dQ_1 + \varepsilon\end{aligned}$$

which implies ii). Let now ii) hold. Using the inequality $|a_0 \wedge b_0 - a_1 \wedge b_0| \leq |a_0 - a_1|$ we find that

$$|B_\pi(\widetilde{P}_0, Q_0) - B_\pi(P_0, Q_0)| \leq V(\widetilde{P}_0, P_0)$$

where V denotes the total variation (33). Setting $\widetilde{P}_{0,\varepsilon} = (1-\varepsilon)P_0 + \varepsilon Q_0$, we find that $V(\widetilde{P}_{0,\varepsilon}, P_0) \to 0$ as $\varepsilon \to 0$ and $Q_0 \ll \widetilde{P}_{0,\varepsilon}$. Thus, it suffices to prove i) under the absolute continuity $Q_0 \ll P_0$. Let $\alpha = \int \psi dP_1 + \varepsilon < 1$. Since $Q_0 \ll P_0$, by the Neyman–Pearson lemma (see, e.g., De Groot [16]) we find c and γ such that the test

$$\varphi = \begin{cases} 1, & \text{if } dQ_0/dP_0 > c \\ \gamma, & \text{if } dQ_0/dP_0 = c \\ 0, & \text{if } dQ_0/dP_0 < c \end{cases}$$

satisfies $\int \varphi dP_0 = \alpha$. As φ is a likelihood ratio test with a critical value c, it is known to be a Bayes test with the prior probabilities $\pi = c/(1+c)$ and $1-\pi = 1/(1+c)$. Hence,

$$\begin{aligned}&\pi \int \varphi dP_0 + (1-\pi) \int (1-\varphi) dQ_0 \\ &= B_\pi(P_0, Q_0) \\ &\leq B_\pi(P_1, Q_1) + \varepsilon \\ &\leq \pi(\int \psi dP_1 + \varepsilon) + (1-\pi) \int (1-\psi) dQ_1 + (1-\pi)\varepsilon \\ &= \pi \int \varphi dP_0 + (1-\pi) \int (1-\psi) dQ_1 + (1-\pi)\varepsilon.\end{aligned}$$

Since $1 - \pi > 0$, this implies the inequality

$$\int (1-\varphi) dQ_0 \leq \int (1-\psi) dQ_1 + \varepsilon$$

which proves i) because $\int \varphi dP_0 = \alpha = \int \psi dP_1 + \varepsilon$. □

Remark 5: Condition iii) is just another form of the so-called concave function criterion for ε-deficiency in the statistical decision theory, see Strasser [47] or LeCam [31]. Let p_i, q_i be the densities of P_i, Q_i with respect to the σ-finite measure μ_i and note that, in view of $P_i(p_i > 0) = 1$, the likelihood ratio $\ell_{Q_i,P_i} = q_i/p_i$ is P_i-a.s. defined. Denote by R_i the distribution of ℓ_{Q_i,P_i} under P_i. Then the concave function criterion for the models $\mathcal{M}_0$ and $\mathcal{M}_1$ is the condition

$$\int g(t) dR_0 \geq \int g(t) dR_1 + \varepsilon(\lim_{s \downarrow 0} g'_+(s) + \lim_{s \to \infty} g(s)) \tag{97}$$

required for any nondecreasing concave function g on $[0,\infty)$ with $g(0) = 0$ and $g(\infty) = \lim_{t\to\infty} g(t) < \infty$. To relate this condition to iii), we introduce the convex function $f = -g$ which belongs to the class $\mathcal{F}_0$ defined at the beginning of this section. Then we get from the definition of D_f in (20) and $f(0) = f^*(0) = 0$ implied by (89) and (90) that

$$\begin{aligned}D_f(Q_0, P_0) &= -\int g(q_0/p_0) dP_0 = -\int g(t) dR_0, \\ D_f(Q_1, P_1) &= -\int g(q_1/p_1) dP_1 = -\int g(t) dR_1.\end{aligned}$$

Now we use (91) to argue that

$$\gamma_f((0,\infty)) = \lim_{s\downarrow 0} g'_+(s) + \lim_{s\to\infty} g(s).$$

Hence, (97) is equivalent to

$$D_f(Q_0,P_0) \le D_f(Q_1,P_1) + \varepsilon\gamma_f((0,\infty)).$$

If we replace f by f^* then this reduces to condition iii) of the last theorem.

Example 10: By the previous example, the randomization using a stochastic kernel K transforms each model $\mathcal{M} = \langle(\mathcal{X},\mathcal{A}),\{P,Q\}\rangle$ into a less informative model $\mathcal{N} = \langle(\mathcal{Y},\mathcal{B}),\{KP,KQ\}\rangle$. For example, finite or countable quantizations of observations can be represented by the statistics $T:\mathcal{X}\mapsto\mathcal{Y}$ where $\mathcal{Y}=\{1,2,\ldots\}$ with all subsets in $\mathcal{B}$. Then the Dirac's kernels

$$K(\cdot|x) = (\delta_{T(x)}(1),\delta_{T(x)}(2),\ldots)$$

on $\mathcal{Y}$ define KP, KQ as discrete distributions $\boldsymbol{p} = (p_1,p_2,\ldots), \boldsymbol{q} = (q_1,q_2,\ldots)$ on $\mathcal{Y}$ where

$$p_i = P(A_i),\ q_i = Q(A_i), \qquad \text{for } A_i = T^{-1}(i) \in \mathcal{A}. \quad (98)$$

More rigorously, KP and KQ are probability measures on the subsets of $\mathcal{Y}=\{1,2,\ldots\}$ dominated by the counting measure μ and (98) are the densities $d(KP)/d\mu, d(KQ)/d\mu$ at the points $i\in\mathcal{Y}$. Therefore, we get from Definition 2 that $D_f(KP,KQ) = D_f(\boldsymbol{p},\boldsymbol{q})$ where

$$D_f(\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{\infty} \mathbf{1}_{(0,\infty)}(p_i\wedge q_i)\, f\left(\frac{p_i}{q_i}\right) q_i + f(0)q_0 + f^*(0)p_0 \quad (99)$$

for

$$p_0 = \sum_{i=1}^{\infty} \mathbf{1}_{\{0\}}(q_i)p_i, \quad q_0 = \sum_{i=1}^{\infty} \mathbf{1}_{\{0\}}(p_i)q_i. \quad (100)$$

Let us return back to the general situation of Example 9 where a less informative model $\mathcal{N} = \langle(\mathcal{Y},\mathcal{B}),\{KP,KQ\}\rangle \overset{0}{\preceq} \mathcal{M}$ was obtained from a model $\mathcal{M} = \langle(\mathcal{X},\mathcal{A}),\{P,Q\}\rangle$ by means of a stochastic kernel K. The relation $\mathcal{N}\overset{0}{\preceq}\mathcal{M}$ was introduced in a nonstrict sense which means that at the same time the validity of the reversed relation $\mathcal{M}\overset{0}{\preceq}\mathcal{N}$ is possible. If both these relations simultaneously hold, then we say that $\mathcal{M}$ and $\mathcal{N}$ are equally informative. In this case, the kernel K is considered to be sufficient for the model $\mathcal{M}$. Next follows a more formal definition.

Definition 5: We say that a stochastic kernel $K:\mathcal{X}\times\mathcal{B}\to[0,1]$ is sufficient for $\{P,Q\}$ if it preserves the statistical information in the model $\langle(\mathcal{X},\mathcal{A}),\{P,Q\}\rangle$, i.e., if

$$\langle(\mathcal{X},\mathcal{A}),\{P,Q\}\rangle \overset{0}{\preceq} \langle(\mathcal{Y},\mathcal{B}),\{KP,KQ\}\rangle. \quad (101)$$

A statistics $T:\mathcal{X}\to\mathcal{Y}$ is called sufficient if the kernel $K=\delta_T$ is sufficient.

The next proposition is a direct consequence of Theorem 12 for the case $\varepsilon=0$ and of the fact that the procedure of randomization leads to less informative models.

Proposition 1: If $K:\mathcal{X}\times\mathcal{B}\to[0,1]$ is a stochastic kernel then the statistical informations $\mathcal{I}_\pi$ satisfy the inequality

$$\mathcal{I}_\pi(KP,KQ) \le \mathcal{I}_\pi(P,Q), \qquad \text{for all } \pi\in(0,1). \quad (102)$$

The kernel K is sufficient for $\{P,Q\}$ if it preserves the statistical informations, i.e., if

$$\mathcal{I}_\pi(KP,KQ) = \mathcal{I}_\pi(P,Q), \qquad \text{for all } \pi\in(0,1). \quad (103)$$

At the first look it seems that the characterization of sufficiency above is different from those commonly used. In common understanding, a statistics $T:\mathcal{X}\to\mathcal{Y}$ is sufficient for $\{P,Q\}$ if the density $dP/d(P+Q)$ is $T^{-1}(\mathcal{B})$-measurable. The next theorem shows that both these characterizations in fact coincide.

Theorem 13: A statistics $T:\mathcal{X}\to\mathcal{Y}$ is sufficient in the sense of Definition 5 if and only if there is a measurable function $\rho:\mathcal{Y}\to\mathbb{R}$ such that

$$(P+Q)(\rho(T)\ne dP/d(P+Q)) = 0. \quad (104)$$

Proof: Let $\widetilde{\mathcal{A}}$ be the sub-σ-algebra of $\mathcal{A}$ generated by the σ-algebra $T^{-1}(\mathcal{B})\subset\mathcal{A}$ and by the μ-null sets in $\mathcal{A}$. Approximating any $\widetilde{\mathcal{A}}$-measurable real function $\widetilde{g}$ by step functions one can easily see that $\widetilde{g}$ is $\widetilde{\mathcal{A}}$-measurable if and only if there is a $T^{-1}(\mathcal{B})$-measurable function g such that $g=\widetilde{g}$ μ-a.s. This shows that $\widetilde{g}$ is $\widetilde{\mathcal{A}}$-measurable if and only if for every real number t there is a set $A_t\in T^{-1}(\mathcal{B})$ such that

$$\mu(\{g<t\}\triangle A_t) = 0$$

where $\triangle$ denotes the symmetric difference of sets. Furthermore, if g takes on values in $[0,1]$ then it suffices to consider this condition only for all t from a subset D dense in $[0,1]$. Indeed, if $t_m\uparrow t$ and there are $A_m\in T^{-1}(\mathcal{B})$ with $\mu(\{g<t_m\}\triangle A_m) = 0$ then $\{g<t\} = \cup_{m=1}^{\infty}\{g<t_m\}$ implies

$$\begin{aligned}\mu(\{g<t\}\triangle(\cup_{n=1}^{\infty}A_n)) &\le \mu(\{g\ge t\}\cap(\cup_{n=1}^{\infty}A_n)) \\ &\quad + \mu((\cup_{m=1}^{\infty}\{g<t_m\})\cap(\cap_{n=1}^{\infty}\overline{A}_n)) \\ &\le \sum_{n=1}^{\infty}\mu(\{g\ge t_n\}\cap A_n) + \sum_{m=1}^{\infty}\mu(\{g<t_m\}\cap\overline{A}_m) = 0\,.\end{aligned}$$

A similar procedure can be used for $t_m\downarrow t$. Now the rest of the proof is easy. Indeed, put $\mu = P+Q$ and denote by p the density of P with respect to μ. Then $q := dQ/d\mu = 1-p$ and p takes on values in $[0,1]$. Since μ is finite measure, the set of all $t\in[0,1]$ for which $\mu(\{x: p(x)=t\})>0$ is at most countable. Hence,

$$D = \{t: 0<t<1,\ \mu(\{x:(1-t)p(x)=tq(x)\})=0\}$$

is dense in $[0,1]$. Denote by φ_{opt} the Bayes test for P,Q and by $\widetilde{\varphi}_{\text{opt}}$ the Bayes test for $\widetilde{P}=KP=P\circ T^{-1}$, $\widetilde{Q}=KQ=Q\circ$

T^{-1}. The densities $\widetilde{p} = d\widetilde{P}/d(\widetilde{P}+\widetilde{Q})$ and $\widetilde{q} = d\widetilde{Q}/d(\widetilde{P}+\widetilde{Q})$ are $T^{-1}(\mathcal{B})$-measurable. If $\mathcal{I}_\pi(P,Q) = \mathcal{I}_\pi(KP,KQ)$ then (68) implies that for all $\pi = 1-t$, $t \in D$ the symmetric difference of the sets

$$\{x : \pi p(x) < (1-\pi)q(x)\} \quad \text{and} \quad \{x : \pi\widetilde{p}(x) < (1-\pi)\widetilde{q}(x)\}$$

has μ measure zero. Therefore, $p = dP/d(P+Q)$ is $\widetilde{\mathcal{A}}$-measurable. □

Sufficiency is one of the basic concepts of mathematical statistics. The next theorem gives a complete characterization of the sufficiency in terms of f-divergences.

Theorem 14: For every $f \in \mathcal{F}$ and every stochastic kernel $K : \mathcal{X} \times \mathcal{B} \mapsto [0,1]$

$$D_f(KP,KQ) \le D_f(P,Q) \tag{105}$$

and the equality holds if the kernel is sufficient for $\{P,Q\}$. Conversely, if $f \in \mathcal{F}$ is strictly convex and $D_f(P,Q) < \infty$ then the equality in (105) implies that the kernel is sufficient for $\{P,Q\}$.

Proof: We can assume without loss of generality $f(1)=0$, i.e., $f \in \mathcal{F}_1$, so that the representation of f-divergences $D_f(P,Q)$ by means of the statistical informations $\mathcal{I}_\pi(P,Q)$ given by Theorem 11 is applicable. Using this representation, we see that the inequality (105) follows from (102). If K is sufficient then we have the equality in (102) and therefore the equality in (105). Let us now suppose that $f \in \mathcal{F}_1$ is strictly convex. By Definition 1 and Theorem 1, this means that $f'_+(t)$ is strictly increasing on $(0,\infty)$ and, therefore, $\Gamma_f(B) > 0$ for open intervals $B \subseteq (0,1)$. If $D_f(P,Q)$ is finite and the equality in (105) holds then

$$\int \left[\mathcal{I}_\pi(KP,KQ) - \mathcal{I}_\pi(P,Q)\right] d\Gamma_f(\pi) = 0$$

so that $\mathcal{I}_\pi(KP,KQ) = \mathcal{I}_\pi(P,Q)$ for almost all $\pi \in (0,1)$. The rest is clear from the easily verifiable continuity of the mapping $\pi \mapsto \mathcal{I}_\pi(\tilde{P},\tilde{Q})$ for any pair $\tilde{P},\tilde{Q}$. □

Remark 6: Theorem 14, or its restriction to the information divergence $D_f(P,Q) = I(P,Q)$, is sometimes called an information processing theorem or a data processing lemma (see, e.g., Csiszár [12], Csiszár and Körner [13], Cover and Thomas [10]). For the information divergence it was first formulated by Kullback and Leibler [30]. For the f-divergences with strictly convex f it was first established by Csiszár [11], [12] and later extended by Mussmann [39]. A general version was proved in Theorem 14 of Liese and Vajda [33]. In all these papers, the sufficiency in the classical sense considered in (104) was used. The concept proposed in Definition 5 is not only simpler, but also more intuitive. At the same time, the proof of Theorem 14 is incomparably shorter and more transparent than the proof of Theorem 1.24 in [33].

Remark 7: Theorem 14 says that f-divergences are invariants of sufficiency of stochastic kernels and finite f-divergences with f strictly convex on $(0,\infty)$ are complete invariants of this sufficiency. It is interesting to observe that the functions $f_{\pi,0}$ given in (73), which define the statistical informations $\mathcal{I}_\pi(P,Q)$ as $f_{\pi,0}$-divergences, are not strictly convex on $(0,\infty)$. Thus, the statistical informations are invariants of sufficiency but none of them is a complete invariant. However, their collection $\{\mathcal{I}_\pi(P,Q) : \pi \in (0,1)\}$ is a complete invariant, as well as arbitrary integrals $\int \mathcal{I}_\pi(P,Q)\, d\Gamma(\pi)$ for Γ absolutely continuous with respect to the Lebesgue measure.

If $\mathcal{Y} = \mathcal{X}$, $\mathcal{B} \subseteq \mathcal{A}$, and $K(\cdot|x) = \delta_x$ then KP, KQ are the restrictions $P^{\mathcal{B}}, Q^{\mathcal{B}}$ of P,Q on the sub-σ-algebra $\mathcal{B} \subseteq \mathcal{A}$. From Theorem 14 follows the next corollary useful in the next section.

Corollary 2: If $f \in \mathcal{F}$ and $P^{\mathcal{B}}, Q^{\mathcal{B}}$ are the restrictions of P,Q on a sub-σ-algebra $\mathcal{B} \subseteq \mathcal{A}$ then $D_f(P^{\mathcal{B}},Q^{\mathcal{B}}) \le D_f(P,Q)$.

For example, the divergences $D_f(\boldsymbol{p},\boldsymbol{q})$ given in (99) and the informations $\mathcal{I}_\pi(\boldsymbol{p},\boldsymbol{q}) = D_{f_\pi}(\boldsymbol{p},\boldsymbol{q})$ (see (47)) satisfy the inequalities

$$D_f(\boldsymbol{p},\boldsymbol{q}) \le D_f(P,Q) \quad \text{and} \quad \mathcal{I}_\pi(\boldsymbol{p},\boldsymbol{q}) \le \mathcal{I}_\pi(P,Q)$$

i.e., quantizations cannot increase the f-divergences or statistical informations.

The next corollary is useful when the concept of sufficiency introduced in Definition 5 is compared with the traditional concepts based on likelihood ratios, as mentioned in Remark 6. It is obtained from Theorem 14 by taking into account that if $f \in \mathcal{F}$ then $\widehat{f}(t) = f(2t/(t+1)),\ (t+1)/2$ belongs to $\mathcal{F}$, and

$$D_{\widehat{f}}(P,Q) = D_f(P,(P+Q)/2)\,.$$

Moreover, $\widehat{f}$ is strictly convex when f is strictly convex.

Corollary 3: Stochastic kernel $K : \mathcal{X} \times \mathcal{B} \mapsto [0,1]$ is sufficient for $\{P,Q\}$ if and only if it is sufficient for $\{P,(P+Q)/2\}$ and the latter condition holds if and only if the kernel is sufficient for $\{Q,(P+Q)/2\}$.

VII. Continuity of Divergences

This section presents a new approach to the continuity of f-divergences (finite approximations, convergence on nested sub-σ-algebras). This approach is based on the integral representation of f-divergences by statistical informations in Theorem 11. The results of this section are not new but the proofs are more self-contained and much simpler than those known from the earlier literature.

Denote by $P^{\mathcal{B}}, Q^{\mathcal{B}}$ the restrictions of probability measures P,Q on a sub-σ-algebra $\mathcal{B} \subseteq \mathcal{A}$ and let us return to the formula (24) for f-divergence $D_f(P,Q)$. Let p,q be the densities of P,Q with respect to a measure μ considered in this formula, and let μ be a probability measure. Finally, let $\mu^{\mathcal{B}}$ be the restriction of μ on $\mathcal{B}$.

Theorem 15: If $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \cdots$ is a sequence of sub-σ-algebras of $\mathcal{A}$ the union of which generates $\mathcal{A}$, then for every $f \in \mathcal{F}$

$$\lim_{n\to\infty} D_f\left(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n}\right) = D_f(P,Q)\,. \tag{106}$$

Proof: We can assume that $f \in \mathcal{F}_1$. Let

$$p_n = \mathsf{E}(p|\mathcal{B}_n), \quad q_n = \mathsf{E}(q|\mathcal{B}_n).$$

Then by the Lévy theorem (see, e.g., Kallenberg [29, Theorem 6.23])

$$\mathsf{E}|p-p_n| = \int |p-p_n|\,d\mu \to 0, \qquad \text{as } n\to\infty.$$

Similarly, $\mathsf{E}|q-q_n| \to 0$. Combining these convergences with the elementary inequality $|a\wedge b - c\wedge d| \le |a-c|+|b-d|$ we get the convergence of the Bayes loss

$$B_\pi(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n}) = \int (\pi p_n)\wedge((1-\pi)q_n)\,d\mu$$

to

$$B_\pi(P,Q) = \int (\pi p)\wedge((1-\pi)q)d\mu\,.$$

In other words, we obtain for all $\pi\in(0,1)$ the convergence

$$\lim_{n\to\infty} \mathcal{I}_\pi(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n}) = \mathcal{I}_\pi(P,Q) \tag{107}$$

of the statistical informations. By Corollary 2, this convergence is monotone, from below. By Theorem 11

$$D_f(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n}) = \int \mathcal{I}_\pi(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n})\,d\Gamma_f(\pi)$$

and

$$D_f(P,Q) = \int \mathcal{I}_\pi(P,Q)\,d\Gamma_f(\pi)$$

so that (106) follows from (107) and from the monotone convergence theorem for integrals. □

Let now $\mathcal{B}\subseteq\mathcal{A}$ be the subalgebra generated by a finite quantization T considered in previous section, i.e., let $\mathcal{B}$ be generated by a finite partition $\mathcal{P}=\{A_1,\dots,A_n\}\subseteq\mathcal{A}$ of the observation space $\mathcal{X}$. Further, let $\boldsymbol{p}^{\mathcal{P}}, \boldsymbol{q}^{\mathcal{P}}$ denote the discrete distributions $\boldsymbol{p},\boldsymbol{q}$ defined by (98) as functions of finite partitions $\mathcal{P}$. By (98) and (99), $D_f(P^{\mathcal{B}},Q^{\mathcal{B}}) = D_f(\boldsymbol{p}^{\mathcal{P}},\boldsymbol{q}^{\mathcal{P}})$.

Theorem 16: For every $f\in\mathcal{F}$

$$D_f(P,Q) = \sup_{\mathcal{P}} D_f(\boldsymbol{p}^{\mathcal{P}},\boldsymbol{q}^{\mathcal{P}}) \tag{108}$$

where the supremum is taken over all finite partitions $\mathcal{P}\subseteq\mathcal{A}$ of the observation space $\mathcal{X}$.

Proof: Denote by $\mathcal{B}_0\subseteq\mathcal{A}$ the sub-σ-algebra generated by densities p and q, set $\mathcal{Y}=\mathcal{X}$, and let $\mathcal{B}_0$ be the sub-σ-algebra of subsets of $\mathcal{Y}$. Then for the identity mapping T, Theorems 14 and 13 imply the equality

$$D_f(P^{\mathcal{B}_0}, Q^{\mathcal{B}_0}) = D_f(P,Q).$$

As $\mathcal{B}_0$ is countably generated, there exist finite partitions $\mathcal{P}_1\subseteq\mathcal{P}_2\subseteq,\dots$ generating algebras $\mathcal{B}_1\subseteq\mathcal{B}_2\subseteq\cdots$ the union of which generates $\mathcal{B}_0$. Therefore, by Theorem 15

$$\lim_{n\to\infty} D_f(P^{\mathcal{B}_n}, Q^{\mathcal{B}_n}) = D_f(P^{\mathcal{B}_0}, Q^{\mathcal{B}_0})$$

which completes the proof. □

Example 11: From Theorem 16 we obtain, e.g., the formula

$$I(P,Q) = \sup_{\mathcal{P}} \sum_{A\in\mathcal{P}} P(A)\ln\frac{P(A)}{Q(A)}$$

for the I-divergence (30). Similar formulas are valid also for the remaining f-divergences considered in the examples of previous sections.

The formula of Example 11 was established by Gel'fand *et al.* [21]. Its extension (108) was proved in [51].

VIII. Divergences in Statistical Testing and Estimation

In this section, we consider the standard model $\{(\mathcal{X},\mathcal{A}),\mathbb{P}\}$ of mathematical statistics where $\mathbb{P}$ is a class of mutually equivalent probability measures on $(\mathcal{X},\mathcal{A})$. The mutual equivalence means that the distributions in $\mathbb{P}$ have a common support to which can be reduced without loss of generality the observation space $\mathcal{X}$. Due to the equivalence, the likelihood ratio $\ell_{P,Q}$ of any $P\in\mathbb{P}$ with respect to any $Q\in\mathbb{P}$ reduces to the Radon–Nikodym density of P with respect to Q, i.e.,

$$\ell_{P,Q} = \frac{dP}{dQ}, \qquad P,Q\in\mathbb{P}. \tag{109}$$

Moreover, all likelihood ratios (109) can be assumed without loss of generality positive and finite everywhere on $\mathcal{X}$. Then Definition 2 implies for every $f\in\mathcal{F}$

$$D_f(P,Q) = \int f(\ell_{P,Q})\,dQ, \qquad P,Q\in\mathbb{P}. \tag{110}$$

We are interested in the applicability of the f-divergences (110) in testing statistical hypothesis

$$\boldsymbol{H}: P_0\in\mathbb{P}_0 \text{ against } \boldsymbol{A}: P_0\in\mathbb{P}-\mathbb{P}_0,\ \emptyset\ne\mathbb{P}_0\subseteq\mathbb{P} \tag{111}$$

on the basis of independent observations $X_1,\dots,X_n$ from $(\mathcal{X},\mathcal{A},P_0)$, and in estimation of the true distribution $P_0\in\mathbb{P}$ on the basis of these observations.

The observations $X_1,\dots,X_n$ define the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ on $(\mathcal{X},\mathcal{A})$. Assuming that $\mathcal{A}$ contains all singletons $\{x\}$, $x\in\mathcal{X}$, we can say that P_n is supported by the finite set $\{X_1,\dots,X_n\}\subseteq\mathcal{X}$. The distributions supported by finite subsets of $\mathcal{X}$ are usually called discrete. The distributions attributing zero probability to all singletons are called continuous. If $\mathbb{P}$ is a family of continuous distributions then P_n is singular with $P\in\mathbb{P}$ so that, by Theorem 5

$$D_f(P_n,P) = f(0)+f^*(0), \qquad \text{for all } P\in\mathbb{P} \tag{112}$$

i.e., $D_f(P_n,P)$ is not an appropriate measure of proximity of P_n and distributions P from continuous models.

Let us restrict ourselves to $f\in\mathcal{F}$ with $f(1)=0$ (i.e., $f\in\mathcal{F}_1$) strictly convex at 1. Since P_n tends to P_0 in the sense that $\lim_{n\to\infty}P_n(A) = P_0(A)$ a.s., the obvious property of f-divergences

$$\inf_{P\in\mathbb{P}_0} D_f(P_0,P) = \begin{cases} =0, & \text{under } \boldsymbol{H} \\ >0, & \text{under } \boldsymbol{A} \end{cases} \tag{113}$$

suggests the divergence test statistics

$$T_n = \inf_{P\in\mathbb{P}_0} D_f(P_n, P) \tag{114}$$

for various f under consideration. In these statistics, the unknown true distributions $P_0 \in \mathbb{P}_0$ are replaced by the known empirical distributions P_n tending to P_0 for $n \to \infty$. Therefore, these statistics are expected to be asymptotically zero under the hypothesis $\boldsymbol{H}$ and asymptotically positive under the alternative $\boldsymbol{A}$. Similarly, the property

$$P_0 = \arg\min_{P\in\mathbb{P}} D_f(P_0, P) \tag{115}$$

suggests the minimum divergence estimators

$$\hat{P}_n = \arg\min_{P\in\mathbb{P}} D_f(P_n, P) \tag{116}$$

for various f under consideration. These estimators are expected to tend to P_0 in a reasonable topology if $n \to \infty$. But (112) indicates that this direct approach does not yield universally applicable tests and estimators.

Various methods of bypassing the difficulty represented by (112), namely, that the f-divergences overemphasize the effect of the singularity $P_n \perp P$, were proposed in earlier literature, leading to interesting classes of minimum-distance tests, minimum-distance estimators, and M-estimators. In this paper, we propose an alternative approach based on the following theorem introducing a new "supremal" representation of f-divergences. In this theorem, $L_1(P)$ denotes the L_1-space of functions on $(\mathcal{X}, \mathcal{A}, P)$.

Theorem 17: If $\mathbb{P}$ is a class of mutually absolutely continuous distributions such that for any triplet P_0, P, Q from $\mathbb{P}$

$$f'_+(\ell_{P,Q}) \in L_1(P_0) \tag{117}$$

then for every $f \in \mathcal{F}$ the f-divergence $D_f(P_0, P)$ can be represented as the supremum

$$D_f(P_0, P) = \sup_{Q\in\mathbb{P}} \mathcal{D}_f(P_0, P \mid Q) \tag{118}$$

where

$$\mathcal{D}_f(P_0, P \mid Q) = D_f(Q, P) + \mathsf{E}_{P_0} f'_+(\ell_{Q,P}) - \mathsf{E}_Q f'_+(\ell_{Q,P}) \tag{119}$$

and this supremum is achieved at $Q = P$.

Proof: By Theorem 1, for every $a, b > 0$

$$f(b) \geq f(a) + f'_+(a)(b-a)\,.$$

Substituting $b = \ell_{P_0,P}$, $a = \ell_{Q,P}$, and integrating with respect to P, we obtain from (109) and (110)

$$D_f(P_0, P) \geq \mathcal{D}_f(P_0, P \mid Q). \tag{120}$$

Now (118) follows directly from this inequality reducing to the equality when $Q = P_0$. □

Note that the inequality (120) with a different proof was presented for differentiable f already in Liese and Vajda [33, p. 172] where it was used to characterize the distributions minimizing f-divergences $D_f(P_0, P)$ on convex sets of distributions $\mathbb{P}$. Let us emphasize that for the empirical distribution the supremum representation

$$D_f(P_n, P) = \sup_{Q\in\mathbb{P}} \mathcal{D}_f(P_n, P \mid Q)$$

does not hold if P, Q are from a continuous family $\mathbb{P}$. The triplet P_n, P, Q satisfies the assumptions of Theorem 17, and thus also this representation, only when the observation space $\mathcal{X}$ is finite and P_n is supported by the whole $\mathcal{X}$. This situation will be considered in the last theorem below.

The main result of the present section are the following two definitions in which the test statistics (114) and estimators (116) are modified in the sense that the substitution $P_n \longmapsto P_0$ is applied to the representation $\sup_{Q\in\mathbb{P}} \mathcal{D}_f(P_0, P|Q)$ of the f–divergence $D_f(P_0, P)$ rather than to this f–divergence itself. In other words, $D_f(P_n, P)$ in the previous definitions (114) and (116) is replaced by

$$\sup_{Q\in\mathbb{P}} \mathcal{D}_f(P_n, P \mid Q) = \sup_{Q\in\mathbb{P}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f'_+(\ell_{Q,P}(X_i)) + \triangle_f(Q, P) \right\} \tag{121}$$

where

$$\triangle_f(Q, P) = D_f(Q, P) - \mathsf{E}_Q f'_+(\ell_{Q,P}).$$

Definition 6: The divergence test statistics for the hypothesis (111) about the models satisfying assumptions of Theorem 17 are defined by the formula

$$T_n = \inf_{P\in\mathbb{P}_0} \sup_{Q\in\mathbb{P}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f'_+(\ell_{Q,P}(X_i)) + \triangle_f(Q, P) \right\} \tag{122}$$

for $f \in \mathcal{F}_1$ strictly convex at 1 and $\triangle_f(Q, P)$ given above.

Definition 7: The minimum divergence estimators of distribution P_0 in the models satisfying assumptions of Theorem 17 are defined by the formula

$$\hat{P}_n = \arg\min_{P\in\mathbb{P}} \sup_{Q\in\mathbb{P}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f'_+(\ell_{Q,P}(X_i)) + \triangle_f(Q, P) \right\} \tag{123}$$

for $f \in \mathcal{F}_1$ strictly convex at 1 and $\triangle_f(Q, P)$ given above.

The next theorem implies that Definitions 6 and 7 are extensions of the definitions (114) and (116). It is well known that the definitions (114) and (116) are effective in discrete models (see, e.g., Read and Cressie [45], Morales *et al.* [37], and Menéndez *et al.* [38]) while the extensions (122) and (123) are applicable also in continuous and mixed models.

Theorem 18: If the distributions from $\mathbb{P}$ are mutually absolutely continuous and discrete, supported by a finite $\mathcal{X}$, then for all P_n with the same support $\mathcal{X}$

$$D_f(P_n, P) = \sup_{Q \in \mathbb{P}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f'_+(\ell_{Q,P}(X_i)) + \triangle_f(Q, P) \right\}. \tag{124}$$

Consequently, the test statistics (114) and (122) as well as the estimators (116) and (123) mutually coincide with a probability tending exponentially to 1 as $n \to \infty$.

Proof: If the family $\mathbb{P}$ satisfies the assumptions and P_n is supported by the whole observation space $\mathcal{X}$ then $P_n \in \mathbb{P}$ so that (124) follows from Theorem 17. If $\mathcal{X} = \{1, \ldots, k\}$ then the probability that a fixed $i \in \{1, \ldots, k\}$ is not in the support of P_n is

$$(1 - P_0(\{i\}))^n, \qquad \text{where } P_0(\{i\}) > 0.$$

Thus, the test statistics defined by (114) or (122) as well as the estimators defined by (116) or (123) differ with probability at most

$$\sum_{i=1}^{k} (1 - P_0(\{i\}))^n$$

which vanishes exponentially for $n \to \infty$. □

Example 12: If $f(t) = t \ln t$ then $f'(t) = \ln t + 1$ so that we obtain from Definition 6 and (121) the information divergence test statistic

$$T_n = \inf_{P \in \mathbb{P}_0} \sup_{Q \in \mathbb{P}} \frac{1}{n} \sum_{i=1}^{n} \ln(\ell_{Q,P}(X_i)) = \frac{1}{n} \ln \frac{\sup_{P \in \mathbb{P}} \prod_{i=1}^{n} p(X_i)}{\sup_{P \in \mathbb{P}_0} \prod_{i=1}^{n} p(X_i)}. \tag{125}$$

This is a general form of the generalized likelihood ratio test statistics. If $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ and $\mathbb{P}_0 = \{P_\theta : \theta \in \Theta_0\}$ then we obtain this statistics in the well-known textbook form

$$T_n = \frac{1}{n} \ln \frac{\sup_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^{n} p_\theta(X_i)}. \tag{126}$$

Example 13: For the same $f(t) = t \ln t$ as above we obtain from Definition 7 and (121) the minimum information divergence estimator

$$\begin{aligned} \hat{P}_n &= \arg\min_{P \in \mathbb{P}} \sup_{Q \in \mathbb{P}} \frac{1}{n} \sum_{i=1}^{n} \ln(\ell_{Q,P}(X_i)) \\ &= \arg\min_{P \in \mathbb{P}} \left[\sup_{Q \in \mathbb{P}} \frac{1}{n} \sum_{i=1}^{n} \ln q(X_i) - \frac{1}{n} \sum_{i=1}^{n} \ln p(X_i) \right] \\ &= \arg\max_{P \in \mathbb{P}} \frac{1}{n} \sum_{i=1}^{n} \ln p(X_i) \end{aligned}$$

which is a general form of the MLE. If $\mathbb{P} = \{P_\theta : \theta \in \Theta\}$ then $\hat{P}_n = P_{\hat{\theta}_n}$ where

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ln p_\theta(X_i)$$

is the well known MLE point estimator.

An interesting open problem is whether for some other functions $f \in \mathcal{F}_1$ the double optimizations in (122) and (123) can be reduced to two or more simple optimizations observed for $f(t) = t \ln t$ in Examples 12 and 13. Another interesting task is the general asymptotic theory of the divergence statistics T_n and the minimum divergence estimators $\hat{P}_n$ and $\hat{\theta}_n$ encompassing the maximum-likelihood theories as special cases.

ACKNOWLEDGMENT

The authors wish to thank Dr. T. Hobza for valuable comments and help with preparation of this paper.

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, no. 28, pp. 131–140, 1966.
[2] S. Arimoto, "Information-theoretical considerations on estimation problems," *Info. Contr.*, vol. 19, pp. 181–194, 1971.
[3] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimates consistent in total variation and two types of information divergence," *IEEE Trans. Inf. Theory*, vol. 38, no. 5, pp. 1437–1454, Sep. 1990.
[4] A. Berlinet, I. Vajda, and E. C. van der Meulen, "About the asymptotic accuracy of Barron density estimates," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 999–1009, May 1990.
[5] A. Bhattacharyya, "On some analogues to the amount of information and their uses in statistical estimation," *Sankhya*, vol. 8, pp. 1–14, 1946.
[6] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Adison-Wesley, 1987.
[7] A. Buzo, A. H. Gray Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–28, no. 5, pp. 562–574, Oct. 1980.
[8] H. Chernoff, "A measure of asymptotic efficiency for test of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, pp. 493–507, 1952.
[9] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.
[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 on Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, ser. A, vol. 8, pp. 84–108, 1963.
[12] ——, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[13] I. Csiszár and J. Körner, *Information Theory. Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Akademiai Kaidó, 1981.
[14] I. Csiszar, "Generalized cutoff rates and Renyi information measures," *IEEE Trans. Infm. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.
[15] M. H. De Groot, "Uncertainty, information and sequential experiments," *Ann. Math. Statist.*, vol. 33, pp. 404–419, 1962.
[16] ——, *Optimal Statistical Decisions*. New York: McGraw Hill, 1970.
[17] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.
[18] A. A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1491–1498, Jun. 2003.
[19] A. Feinstein, *Information Theory*. New York: McGraw-Hill, 1958.
[20] D. Feldman and F. Österreicher, "A note on f-divergences," *Studia Sci. Math. Hungar.*, vol. 24, pp. 191–200, 1989.

[21] I. M. Gel'fand, A. N. Kolmogorov, and A. M. Yaglom, "On the general definition ofthe amount of information," *Dokl. Akad. Nauk. SSSR*, vol. 11, pp. 745–748, 1956.

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

[23] L. Györfi, G. Morvai, and I. Vajda, "Information-theoretic methods in testing the goodness of fit," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 28.

[24] P. Harremoës and F. Topsøe, "Inequalities between entropy and the index of coincidence derived from information diagrams," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2944–2960, Nov. 2001.

[25] H. Hewitt and K. Stromberg, *Real and Abstract Analysis*. Berlin, Germany: Springer, 1965.

[26] L. K. Jones and C. L. Byrne, "Generalized entropy criteria for inverse problems, with applications to data comprassion, pattern classification and cluster analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 1, pp. 23–30, Jan. 1990.

[27] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[28] S. Kakutani, "On equivalence of infinite product measures," *Ann. Math.*, vol. 49, pp. 214–224, 1948.

[29] O. Kallenberg, *Foundations of Modern Probability*. New York: Spinger, 1997.

[30] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

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

[32] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.

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

[34] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[35] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints," *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 241–255, Mar. 1991.

[36] D. Morales, L. Pardo, and I. Vajda, "Uncertainty of discrete stochastic systems: General theory and statistical inference," *IEEE Trans. Syst. Man Cybern., Part A*, vol. 26, no. 6, pp. 681–697, Nov. 1996.

[37] ——, "Minimum divergence estimators based on grouped data," *Ann. Inst. Statist. Math.*, vol. 53, pp. 277–288, 2001.

[38] M. L. Menéndez, D. Morales, L. Pardo, and I. Vajda, "Some new statistics for testing hypotheses in parametric models," *J. Multivar. Analysis*, vol. 62, pp. 137–168.

[39] D. Mussmann, "Sufficiency and f-divergences," *Studia Sci. Math. Hungar.*, vol. 14, pp. 37–41, 1979.

[40] F. Österreicher, "On a class of perimeter-type distances of probability distributions," *Kybernetika*, vol. 32, pp. 389–393, 1996.

[41] F. Österreicher and I. Vajda, "Statistical information and discrimination," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 1036–1039, May 1993.

[42] ——, "A new class of metric divergences on probability spaces and its applicability in statistics," *Ann. Inst. Statist. Math.*, vol. 55, no. 3, pp. 639–653, 2003.

[43] H. V. Poor, "Robust decision design using a distance criterion," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 575–587, Sep. 1980.

[44] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Probability Theory and Mathematical Statist.*, Berkeley, CA, 1961, pp. 547–561.

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

[46] C. E. Shanon, "A mathematical theory of communication," *Bell. Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[47] H. Strasser, *Mathematical Theory of Statistics*. Berlin, Germany: De Gruyter, 1985.

[48] F. Topsøe, "Information-theoretical optimization techniques," *Kybernetika*, vol. 15, pp. 7–17, 1979.

[49] ——, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 1602–1609, Apr. 2000.

[50] E. Torgersen, *Comparison of Statistical Experiments*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[51] I. Vajda, "On the f-divergence and singularity of probability measures," *Periodica Math. Hungar.*, vol. 2, pp. 223–234, 1972.

[52] ——, "χ^{α}–divergence and generalized Fisher's information," in *Trans. 6th Prague Conf. Information Theory*, Prague, Czechoslovakia, 1973, pp. 873–886.

[53] ——, *Theory of Statistical Inference and Information*. Boston, MA: Kluwer, 1989.

[54] ——, "On convergence of information contained in quantized observations," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2163–2172, Aug. 2002.