KYBERNETIKA — VOLUME 43 (2007), NUMBER 5, PAGES 675 – 696

ON GENERALIZED ENTROPIES, BAYESIAN DECISIONS AND STATISTICAL DIVERSITY

Igor Vajda and Jana Zvárová

The paper summarizes and extends the theory of generalized ϕ-entropies $H_\phi(X)$ of random variables X obtained as ϕ-informations $I_\phi(X;Y)$ about X maximized over random variables Y. Among the new results is the proof of the fact that these entropies need not be concave functions of distributions p_X. An extended class of power entropies $H_\alpha(X)$ is introduced, parametrized by $\alpha \in \mathbb{R}$, where $H_\alpha(X)$ are concave in p_X for $\alpha \geq 0$ and convex for $\alpha < 0$. It is proved that all power entropies with $\alpha \leq 2$ are maximal ϕ-informations $I_\phi(X;X)$ for appropriate ϕ depending on α. Prominent members of this subclass of power entropies are the Shannon entropy $H_1(X)$ and the quadratic entropy $H_2(X)$. The paper investigates also the tightness of practically important previously established relations between these two entropies and errors $e(X)$ of Bayesian decisions about possible realizations of X. The quadratic entropy is shown to provide estimates which are in average more than 100 % tighter than those based on the Shannon entropy, and this tightness is shown to increase even further when α increases beyond $\alpha = 2$. Finally, the paper studies various measures of statistical diversity and introduces a general measure of anisotony between them. This measure is numerically evaluated for the entropic measures of diversity $H_1(X)$ and $H_2(X)$.

Keywords: ϕ-divergences, ϕ-informations, power divergences, power entropies, Shannon entropy, quadratic entropy, Bayes error, Simpson diversity, Emlen diversity

AMS Subject Classification: 94A17, 62C10

1. INTRODUCTION AND BASIC CONCEPTS

The models of data and variables in the digital world are usually discrete. Therefore we are interested in random variables with discrete true distributions $p = (p(i) : i \in \mathcal{I})$ and discrete hypothetical distributions $q = (q(i) : i \in \mathcal{I})$ with finite $\mathcal{I}$. We drop the indices $i \in \mathcal{I}$ irrelevant for both p and q in the sense $p(i) + q(i) = 0$, i. e. we suppose $p(i) + q(i) > 0$ for all $i \in \mathcal{I}$.

The divergence of distributions p, q is often expressed by the ϕ-divergence for ϕ from the class $\mathbf{\Phi}$ of real valued functions convex on the interval $(0, \infty)$ and strictly convex at $t = 1$ with $\phi(1) = 0$. Following Csiszár [4, 5] or Liese and Vajda [17, 18],

if $\phi \in \Phi$ then the ϕ-divergence of distributions p, q can be defined by formula

$$D_\phi(p\|q) = \sum_{i:q(i)>p(i)} q(i)\,\phi\left(\frac{p(i)}{q(i)}\right) + \sum_{i:q(i)<p(i)} p(i)\,\phi^*\left(\frac{q(i)}{p(i)}\right) \tag{1.1}$$

where $\phi^* \in \mathbf{\Phi}$ is adjoined to ϕ in the sense that for all $t \in (0,\infty)$

$$\phi^*(t) = t\phi\,(1/t)\,, \tag{1.2}$$

and the (possibly infinite) values $\phi(0)$, $\phi^*(0)$ needed in (1.1) are obtained as limits of $\phi(t)$, $\phi^*(t)$ for $t \downarrow 0$. It is clear from (1.1) that $D_{\phi^*}(p\|q) = D_\phi(q\|p)$. As well known, the ϕ-divergences take on values between 0 and $\phi(0) + \phi^*(0)$ where $D_\phi(p,q) = 0$ if and only if $p = q$. For this and further properties see [17] or [18].

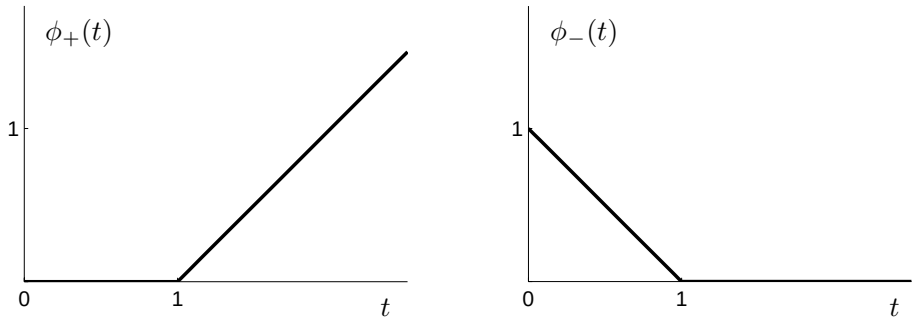

Fig. 1.1. Functions $\phi_+(t) = (t-1)^+ = \max(t-1,0)$ and $\phi_-(t) = (t-1)^- = -\min(t-1,0)$.

The most simple functions $\phi \in \mathbf{\Phi}$ are $\phi_+(t)$ and $\phi_-(t) = \phi_+^*(t)$ given in Figure 1.1 leading to the *upper variation*

$$D_{\phi_+}(p\|q) = V_+(p\|q) = \sum_{i:p(i)>q(i)} (p(i) - q(i)) \tag{1.3}$$

and *lower variation*

$$D_{\phi_-}(p\|q) = V_-(p\|q) = \sum_{i:p(i)<q(i)} (q(i) - p(i)), \tag{1.4}$$

and their sum $\phi(t) = \phi_+(t) + \phi_-(t) = |t-1|$ which is self-adjoined in the sense $\phi^*(t) = \phi(t)$ and leads to the *total variation*

$$D_\phi(p\|q) = V(p\|q) = \sum_i |p(i) - q(i)|. \tag{1.5}$$

Well known class of ϕ-divergences parametrized by $\alpha \in \mathbb{R}$ is obtained from the power functions

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \text{for } \alpha(\alpha-1) \neq 0 \tag{1.6}$$

and their limits

$$\phi_1(t) = t\ln t - t + 1, \quad \phi_0(t) = -\ln t + t - 1. \tag{1.7}$$

In what follows we use the simplified notation $D_\alpha(p\|q) = D_{\phi_\alpha}(p\|q)$ and call these expressions *power divergences*. The adjoining rule (1.2) for the power functions (1.6), (1.7) is $\phi^*_\alpha(t) = \phi_{1-\alpha}(t)$ and leads to the skew symmetry

$$D_\alpha(p\|q) = D_\alpha(q\|p), \quad \alpha \in \mathbb{R} \tag{1.8}$$

of the corresponding power divergences.

The best known power divergence is perhaps the statistical *Pearson divergence*

$$D_2(p\|q) = \sum_i \frac{p(i)^2}{q(i)} - 1 = \sum_i \frac{(p(i) - q(i))^2}{q(i)} \tag{1.9}$$

where (and also in the sequel) the summands with $q(i) = 0$ in the denominator are assumed to be infinite. Another important examples are the *double-Pearson divergence*

$$D_4(p\|q) = \sum_i \frac{p(i)^4}{q(i)^3} - 1, \tag{1.10}$$

the classical information-theoretic divergence often called *Kullback divergence*

$$D_1(p\|q) = \sum_i p(i)\ln\frac{p(i)}{q(i)} \tag{1.11}$$

and the *Hellinger divergence* (squared Hellinger distance)

$$D_{1/2}(p\|q) = 4\sum_i \left(\sqrt{p(i)} - \sqrt{q(i)}\right)^2. \tag{1.12}$$

The power divergences defined by the class of functions (1.6), (1.7) are usually attributed to Cressie and Read [3]. We would like to emphasize here that the power divergences $\bar{D}_\alpha(P\|Q)$ obtained for arbitrary probability measures P, Q and all $\alpha > 0$ from the functions

$$\bar{\phi}_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha - 1} \quad \alpha > 0, \alpha \neq 1$$

and their limit $\bar{\phi}_1(t) = t\ln t - t + 1 \equiv \phi_1(t)$ were introduced much earlier in the formulas (1.8), (1.9) of Perez [22]. This means that, in particular, the divergences $D_\alpha(p\|q)$ of Cressie and Read are obtained from $\bar{D}_\alpha(p\|q)$ of Perez in the domain $\alpha > 0$ simply by dividing by α. Further, the skew symmetry (1.8) implies that the Cressie-Read divergences follow from the version of Perez by the similar division rule

$$D_\alpha(p\|q) = \bar{D}_{1-\alpha}(q\|p)/(1-\alpha)$$

also in the domain $\alpha \leq 0$. For $\alpha = 1$ both the Perez and Cressie-Read versions coincide with the Kullback divergence.

In his classical papers on information theory, Shannon introduced probability distributions $p_{X;Y}(i)$, $i \in \mathcal{I} = \mathcal{X} \times \mathcal{Y}$ as models for the situations where an $\mathcal{Y}$-valued observation Y informs about an $\mathcal{X}$-valued message X. As a measure of information he proposed a nonnegative quantity $I(X;Y)$ which is nothing but the Kullback divergence $D_1(p_{X,Y}\|p_X p_Y)$ between the joint distribution $p_{X,Y}$ of X, Y and the product $p_X p_Y$ of the marginal distributions

$$p_X(x) = \sum_y p_{X,Y}(x,y), \quad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

of X and Y on $\mathcal{X}$ and $\mathcal{Y}$. In other words, the *Shannon information* is

$$I_1(X;Y) = D_1(p_{X,Y}\|p_X p_Y) = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)\,p(y)} \tag{1.13}$$

(cf. (1.11)). Here and in the sequel we use the conventions

$$p(x,y) = p_{X,Y}(x,y), \quad p(x) = p_X(x) \quad \text{and} \quad p(y) = p_Y(y) \tag{1.14}$$

which are common in the literature on information theory. The product of marginals $p_X p_Y$ is a hypothetic distribution $q_{X,Y}$ which is true only if X, Y are independent. We see from (1.13) that the Shannon information is a nonnegative measure of association of the random variables X, Y which is equal zero if and only if X, Y are independent.

A similar measure of association was proposed much earlier by Pearson [21], namely the *mean square contingency*

$$I_2(X;Y) = D_2\left(p_{X,Y}\|p_X p_Y\right) = \sum_{x,y} \frac{(p(x,y) - p(x)\,p(y))^2}{p(x)\,p(y)} \tag{1.15}$$

(cf. (1.9) and (1.14)), used later as a basis in various criteria of statistical association (Cramér [2], Tschuprow [27], Höffding [14]). Later Höffding [15] proposed postulates for measures of association of random pairs (X,Y) based on the measure

$$V(X;Y) = V\left(p_{X,Y}\|p_X p_Y\right) = \sum_{x,y} |p(x,y) - p(x)\,p(y)| \tag{1.16}$$

(cf. (1.5) and (1.14)) called *Höffding coefficient* in Zvárová [32].

Motivated by these proposals and also by earlier papers of Rényi [23, 24], Csiszár [6] and Zvárová [32] introduced the general *ϕ-information*

$$I_\phi(X;Y) = D_\phi\left(p_{X,Y}\|p_X p_Y\right) = \sum_{x,y} p(x)\,p(y)\,\phi\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \tag{1.17}$$

(cf. (1.14)) where we used a simplified form of (1.1) since $p(x)\,p(y) > 0$ follows for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ from the assumption $p(x)\,p(y) + p(x,y) > 0$. In other words, the ϕ-information is nothing but the ϕ-divergence of true distribution $p_{X,Y}$ and the hypothetic distribution $p_X p_Y$ which is true only if X and Y are independent.

As observed already by Shannon (see e. g. Cover and Thomas [1]), the information $I_1(X;Y)$ is maximal if $Y = X$, i. e. if the observed variable is the message X itself. The amount of information $I_1(X;X)$ in the message X is the *Shannon entropy*

$$H_1(X) = -\sum_x p(x) \ln p(x) \quad (\text{in } \textit{nats}). \tag{1.18}$$

This is one element from the family of *power entropies* defined for arbitrary distributions $p = p_X$ and all $\alpha \in \mathbb{R}$ by the formula

$$H_\alpha(p) \equiv H_\alpha(X) = \sum_x p(x)\, \psi_\alpha(p(x)) \tag{1.19}$$

using the power functions given by

$$\psi_\alpha(\pi) = \frac{1 - \pi^{\alpha-1}}{\alpha - 1}, \quad \pi \in (0,1], \quad \psi_\alpha(0) = \lim_{\pi \downarrow 0} \psi_\alpha(\pi) \tag{1.20}$$

if $\alpha \neq 1$ and otherwise by the corresponding limit

$$\psi_1(\pi) = -\ln \pi, \quad \pi \in (0,1], \quad \psi_1(0) = \infty \tag{1.21}$$

with the additional rule $0\psi_\alpha(0) = 0$. The functions $\psi_\alpha(\pi)$ may be viewed as decreasing measures of information in an event of probability $\pi \in [0,1]$. Therefore the entropies $H_\alpha(X)$ are expected amounts of information in the individual events $X = x$. If we normalize in (1.20) by $(\alpha - 1)\ln 2$ then the limit entropy $H_1(X)$ for $\alpha \to 1$ will differ from (1.18) by $\log_2$ at the place of $\ln = \log_e$, i. e. the information $H_1(X)$ will be in *bits* instead of *nats*.

The subclass of the power entropies (1.19) with parameters $\alpha > 0$ was first introduced in a slightly differently normalized form by Havrda and Charvát [13]. As we shall see, interesting examples of power entropies, in addition to the Shannon entropy, are

$$H_2(X) = 1 - \sum_x p(x)^2 \tag{1.22}$$

called *quadratic entropy* by Vajda [28] and

$$H_0(X) = n - 1 \quad \text{for} \quad n = \sum_{x:p(x)>0} 1 \tag{1.23}$$

indicating the number of possible messages which differ from the delivered one, which may be called *Hartley entropy* (it is one-one related to what is commonly called the Hartley entropy, cf. e. g. [20]).

The power entropies (1.19) belong to the even wider class of *ψ-entropies*

$$H_\psi(p) \equiv H_\psi(X) = \sum_x p(x)\, \psi(p(x)) \tag{1.24}$$

for decreasing continuous functions $\psi(\pi)$ of variable $\pi \in (0,1]$ with $\psi(1) = 0$. These general entropies were studied in Vajda [29] who proved that they exhibit standard

desirable properties of information measures when $\pi\psi(\pi)$ is concave on $(0,1)$. The power entropies (1.19) are concave in this sense only for $\alpha \geq 0$.

Section 2 summarizes the results of the recent lecture [31]. It shows that the definition of generalized entropies of information sources as generalized informations in direct observations of these sources leads to some nonconcave entropies, in particular to infinitely many nonconcave power entropies. Section 3 studies relations between the entropies $H_\alpha(X)$ for $\alpha \geq 0$ and the errors $e(X)$ of Bayes decisions about X. Section 4 investigates mutual relation between two important particular power entropies, namely $H_1(X)$ and $H_2(X)$.

2. INFORMATIONS AND ENTROPIES

The ϕ-informations $I_\phi(X;Y)$ characterize statistical association between the observed random variable Y and an unknown random state of nature X. Motivated by statistical decision problems, in this paper we are interested in the situations when these informations achieve maximal values. However, there are many situations when the interest is concentrated on small values of these informations. As an example let us mention the situation when Y encrypts a message X or when $I_\phi(X;Y)$ serve as a mixing coefficients for weakly dependent stochastic processes.

Let us start this section with the formula of Zvárová [32]

$$\sup_Y I_\phi(X;Y) = I_\phi(X;X) = H_{\tilde{\phi}}(X) \quad \text{(cf. (1.24))} \tag{2.1}$$

where $H_{\tilde{\phi}}(X)$ is defined by (1.24) with $\psi(t)$ replaced by $\tilde{\phi}(t) = \phi^*(t) + \phi(0)\,(1-t)$ for $\phi^*(t)$ given by (1.2). It says that the ϕ-information about X distributed by $p(x) = p_X(x)$ obtained by observing an associated random variable Y is maximized by directly observing X, and that the maximal value of this information is given by the $\tilde{\phi}$-entropy

$$H_{\tilde{\phi}}(p) \equiv H_{\tilde{\phi}}(X) = \sum_x p(x)\,\phi^*(p(x)) + \phi(0)\,H_2(X) \quad \text{(cf. (1.24) and (1.20))}. \tag{2.2}$$

This result jointly with the next assertion emphasize the prominent role of the quadratic entropy $H_2(X)$.

Proposition 2.1. The ϕ-informations $I_\phi(X;Y)$ corresponding to the simple functions $\phi = \phi_+$ and $\phi = \phi_-$ from $\mathbf{\Phi}$ defined in Figure 1.1 achieve maxima given by the quadratic entropy, i. e.

$$H_{\tilde{\phi}_+}(X) = H_{\tilde{\phi}_-}(X) = H_2(X) \quad \text{(cf. (1.22))}. \tag{2.3}$$

Therefore the entropy $H_{\tilde{\phi}}(X)$ which maximizes the Höffding measure of information $V(X;Y)$ given in (1.16) is $2H_2(X)$.

P r o o f . For $\phi_+(t)$ we get $\phi_+(0) = 0$ and $\phi_+^*(t) = \phi_-(t) = 1 - t$ for all $t \in [0,1]$. Therefore $\tilde{\phi}_+(t) = \phi_+^*(t) + \phi_+(0)\,(1-t)$ (cf. the definition of general $\tilde{\phi}(t)$ above)

coincides with $\psi_2(t)$ from (1.20) so that $H_{\tilde{\phi}_+}(X) = H_2(X)$. For ϕ_- we get $\phi_-(0) = 1$ and $\phi_-^*(t) = \phi_+(t) = 0$ for all $t \in [0,1]$ so that again $\tilde{\phi}_-(t) = \psi_2(t)$ and the rest is as above. The last statement follows from the fact that if $\phi(t) = |t-1|$ then $\tilde{\phi}(t) = \tilde{\phi}_+(t) + \tilde{\phi}_-(t) = 2\psi_2(t)$. □

In the next proposition we are interested in the *modified power entropies*

$$\tilde{H}_\alpha(p) \equiv \tilde{H}_\alpha(X) = H_{\tilde{\phi}_\alpha}(X), \quad \alpha \in \mathbb{R} \tag{2.4}$$

given by (2.2) when $\phi = \phi_\alpha$, i. e. when

$$\tilde{\phi}_\alpha(t) = \phi_{1-\alpha}(t) + \phi_\alpha(0)\,(1-t)$$

(cf. the adjoining rule $\phi_\alpha^*(t) = \phi_{1-\alpha}(t)$ below (1.7)). These entropies maximize the general *power informations*

$$I_\alpha(X;Y) = D_\alpha\left(p_{X,Y} \| p_X p_Y\right), \quad \alpha \in \mathbb{R} \tag{2.5}$$

(cf. (1.17) and (1.6)). The trivial case when the Hartley entropy $H_0(X)$ is zero is excluded, i. e. the number n in (1.23) is supposed to be at least 2.

Proposition 2.2. The entropies $\tilde{H}_\alpha(X)$ are infinite for $\alpha \le 0$ and finite, given by

$$\tilde{H}_\alpha(X) = \frac{1}{\alpha} H_{2-\alpha}(X) \tag{2.6}$$

for $\alpha > 0$. This confirms the well known fact that the maximal Shannon information $\tilde{H}_1(X)$ is the Shannon entropy $H_1(X)$. However, this implies also that the maximal Pearson information $\tilde{H}_2(X)$ is half of the Hartley entropy $H_0(X)$ given in (1.23), the maximal Hellinger information $\tilde{H}_{1/2}(X)$ is the entropy

$$H_{3/2}(X) = 4 \sum_x p(x) \left(1 - \sqrt{p(x)}\right) \tag{2.7}$$

and the maximal double-Pearson information $\tilde{H}_4(X)$ is the nonconcave entropy

$$\frac{1}{2} H_{-2}(X) = \frac{1}{12} \left(\sum_x \frac{1}{p(x)^2} - 1 \right). \tag{2.8}$$

Proof. As mentioned above, the assumption $p(x)\,p(y) - p(x,y) > 0$ implies $p(x) > 0$ for all $x \in \mathcal{X}$. Therefore the sum in (2.2) is finite and $H_2(X)$ is by assumptions positive. Therefore $H_{\tilde{\phi}}(X) = \infty$ if and only if

$$\phi(0) = \phi_\alpha(0) = \infty.$$

From (1.6), (1.7) we see that this takes place for $\alpha \le 0$ and that $\phi_\alpha(0) = 1/\alpha$ for $\alpha > 0$. Now, assuming $\alpha > 0$ and substituting $\phi(0) = 1/\alpha$ and

$$\phi^*(t) = \phi_{1-\alpha}(t) = \frac{1 - t^{1-\alpha}}{\alpha(1-\alpha)} - \frac{1}{\alpha}(1-t) \quad \text{for } \alpha \neq 1$$

and

$$\phi^*(t) = \phi_0(t) = -\ln t + t - 1 \quad \text{for } \alpha = 1$$

in (2.2), we find the desired form (2.6) for $\tilde{H}_\alpha(X)$. The concrete expressions $\tilde{H}_2(X) = H_0(X)/2$ as well as the expressions (2.7) and (2.8) follow from (2.6) and from the definition of α-entropies in (1.19), (1.20). □

The last proposition shows that the maximal power informations with $\alpha > 2$ lead to modified power entropies $\tilde{H}_\alpha(p)$ which are convex in p. As an example, consider for $p = (\pi, 1-\pi)$ the nonconcave entropy

$$\tilde{H}_{-2}(p) = h(\pi) = \frac{1}{12}\left(\frac{1}{\pi^2} + \frac{1}{(1-\pi)^2} - 1\right) \tag{2.9}$$

obtained from (2.8). Since $\varphi(\pi) = 1/\pi^2$ is convex on $(0,1)$, we get

$$h(\pi) > \frac{1}{12}\left(\frac{2}{(1/2)^2} - 1\right) = h(1/2) \tag{2.10}$$

for $\pi \neq 1/2$. This as well as the discontinuity $h(\pi) \to \infty$ for $\pi \to 0$ contradicts what is observed in the case of concave entropies like $H_2(p) = 1 - \pi^2 - (1-\pi)^2$. But nevertheless the information measure $h(\pi)$ of (2.9) is justified. Namely, by (2.1) and Proposition 2.2, $h(\pi)$ given by (2.9) is the double-Pearson information $\tilde{H}_4(X) = I_4(X;X)$, i. e. it is the double-Pearson divergence of the 2×2 contingency tables for $p_{X,X}$ and $p_X\, p_X$ that follow.

π	0
0	$1-\pi$

π^2	$\pi(1-\pi)$
$\pi(1-\pi)$	$(1-\pi)^2$

We see from these tables that the absolute deviations $|(p(x,y)/p(x)\,p(y)) - 1|$ for $\pi \neq 1/2$ or $\pi = 1/2$ are

$$\frac{1-\pi}{\pi}, \quad 1, \quad 1, \quad \frac{\pi}{1-\pi} \qquad \text{or} \qquad 1,\ 1,\ 1,\ 1$$

respectively, so that the sum of positive powers of the left-hand deviations may be arbitrarily larger than the similar sum on the right-hand side. This helps to understand that if the information in X is measured on the double-Pearson scale by $h(\pi)$ then $h(\pi)$ with π close to zero may be considerably larger than $h(1/2)$. Explicitly one can calculate the double-Pearson divergence of the contingency tables for $\pi = 1/2$ which is smaller than for $\pi = 1/4$ while the Kullback divergence for $\pi = 1/2$ is larger than for $\pi = 1/4$ (and the standard Pearson divergence is in both cases the same). Therefore $h(\pi)$ of (2.9) satisfies (2.10) while the Shannon information

$$h(\pi) = -\pi \ln \pi - (1-\pi)\ln(1-\pi) \tag{2.11}$$

is for $\pi = 1/2$ larger than for $\pi = 1/4$ and the Pearson information $h(\pi) = 1/2$ is constant for all $\pi \in (0, 1)$.

Thus we can summarize that the form of the entropy measuring the information in a message X from a given source $(\mathcal{X}, p(x))$ depends on the ϕ-divergence used to quantify the information in Y about X. Our nontrivial observation is that some well known ϕ-divergences legitimize in this manner nonconcave entropies.

3. ENTROPIES AND BAYESIAN DECISIONS

In the statistical decision theory we are interested in the expected loss $\mathsf{E}\mathcal{L}(d, X)$ resulting from a decision $d \in \mathcal{X}$ under the information about the state of nature $x \in \mathcal{X}$ represented by a random variable X distributed by p_X on $\mathcal{X}$. Note that in practical applications the unconditional *a priori* distribution p_X is usually replaced by conditional *a posteriori* distributions $p_{X|Y=y}, y \in \mathcal{Y}$ resulting from observations of a random variable Y statistically associated with X by a joint distribution $p_{X,Y}$ on $\mathcal{X} \otimes \mathcal{Y}$. For the indicator loss function $\mathcal{L}(d, x) = \boldsymbol{I}(x \neq d)$ the minimal *Bayes loss*

$$e(p_X) = \arg\min_{d \in \mathcal{X}} \mathsf{E}\mathcal{L}(d, X)$$

is achieved at $d = \arg\max p_X(x)$ and has the meaning of minimal decision error called Bayes error. In what follows we write simply $e(X)$ instead of $e(p_X)$, i. e. we deal with the *Bayes error*

$$e(X) = 1 - \max_x p(x) \quad \text{for} \quad p(x) = p_X(x). \tag{3.1}$$

It is often desirable to estimate this error by means of measures of information from the class (1.24), in particular by means of its most prominent members $H_1(X)$ and $H_2(X)$. R. M. Fano was the first who found for $e = e(X)$ an upper bound $H_1^+(e)$ achieved by the Shannon entropy $H_1 = H_1(X)$ and Kovalevskij [16] was probably the first who found the corresponding lower bound $H_1^-(e)$. For $h(\pi)$ given by (2.11) and n denoting the number of messages in $\mathcal{X}$, these Fano–Kovalevskij bounds satisfy the relations

$$H_1^-(e) = h(k(1-e)) + k(1-e)\ln k \leq H_1 \leq h(e) + e\ln(n-1) = H_1^+(e), \tag{3.2}$$

where the right hand equality holds in the whole range $0 \leq e \leq (n-1)/n$ while the left hand equality holds piecewise on the subranges

$$\frac{k-1}{k} \leq e \leq \frac{k}{k+1} \quad \text{for} \quad k = 1, \ldots, n-1. \tag{3.3}$$

Note that the n of (3.2) was defined as the number of messages $x \in \mathcal{X}$ independently of whether $p(x) > 0$ or $p(x) = 0$. Therefore it coincides with the n of (1.23) only if $p(x) > 0$ for all $x \in \mathcal{X}$ which is not assumed in this section.

With the help of derivatives one can verify that both the bounds $H_1^-(e)$ and $H_1^+(e)$ continuously increase in the variable $e \in [0, (n-1)/n]$ from the lowest common value $H_1^-(0) = H_1^+(0) = 0$ to the largest common value $H_1^-((n-1)/n) = H_1^+((n-1)/n) = \ln n$. An illustration is given in Figure 3.1.

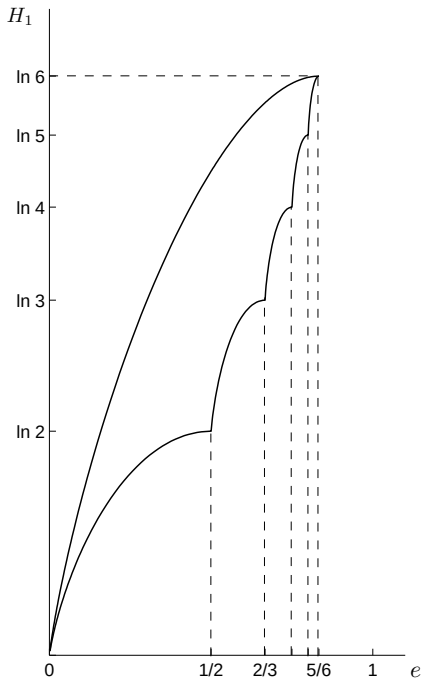

Fig. 3.1. Upper and lower bounds $H_1^+(e)$, $H_1^-(e)$ for $n = 6$.

Proposition 3.1. Upper bound $H_2^+(e)$ and lower bound $H_2^-(e)$ achieved by the quadratic entropy $H_2 = H_2(X)$ under the condition $e(X) = e$ satisfy the relations

$$H_2^-(e) = k(1-e)\,(1+e-(1-e)\,k) \le H_2 \le e\left(2 - \frac{n\,e}{n-1}\right) = H_2^+(e) \qquad (3.4)$$

where the right hand equality holds on the whole range $e \in [0, (n-1)/n]$ while the left hand equality holds piecewise on the subranges given in (3.3).

P r o o f. A general Theorem 1 of Vajda and Vašek [30] implies that if $H(p)$ is any Schur-concave function of probability distributions $p = (p_1, \ldots, p_n)$ then among all p with $e = 1 - \max p_i$ from the semiclosed interval $((k-1)/k,\ k/(k+1)]$, the function $H(p)$ is maximized at $p^+ = (1-e,\ e/(n-1), \ldots, e/(n-1))$ and minimized at $p^- = (1-e, \ldots, 1-e,\ 1-k(1-e), 0, \ldots, 0)$. It is easy to see that the quadratic entropy $H_2(p)$ is Schur-concave in the sense of [30] and that $H_2(p^+)$ and $H_2(p^-)$ are the bounds given in (3.4). □

The bounds of Proposition 3.1 are illustrated in Figure 3.2. With the help of derivatives one can verify that both these bounds are continuously increasing in the variable $e \in [0, (n-1)/n]$ from the lowest common value $H_1^-(0) = H_1^+(0) = 0$ to the largest common value $H_1^-((n-1)/n) = H_1^+((n-1)/n) = (n-1)/n$.

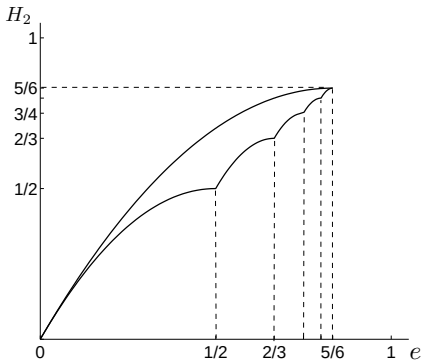

Fig. 3.2. Upper and lower bounds $H_2^+(e)$, $H_2^-(e)$ for $n = 6$.

Vajda [28] conjectured that the quadratic entropy provides tighter bounds for the Bayes error than the Shannon entropy. Up to now, this conjecture was neither rejected nor confirmed. It can be rigorously studied using the differences $e_\alpha^{\max}(H) - e_\alpha^{\min}(H)$ between maximal and minimal Bayes errors under the α-entropy $H_\alpha = H$ in the domain $0 \le H \le H_\alpha^{\max}$ and comparing the *average inaccuracies*

$$A_{\alpha,n} = \frac{1}{H_\alpha^{\max}} \int_0^{H_\alpha^{\max}} \left[e_\alpha^{\max}(H) - e_\alpha^{\min}(H) \right] \mathrm{d}H \tag{3.5}$$

of the best possible estimates of Bayes errors $e(X)$ on the basis of entropies $H_\alpha(X)$ continuously varying between $H_\alpha^{\min} = 0$ and $H_\alpha^{\max} < \infty$.

Such a rigorous study was carried out recently by Vajda and Zvárová [31] where the formula (3.5) was applied to the nonnegative Shannon entropy H_1 with $H_1^{\max} = \ln n$ and nonnegative quadratic entropy H_2 with $H_2^{\max} = (n-1)/n$ to demonstrate that the inaccuracy $A_{1,n}$ of the estimates based on the Shannon entropy is at least 100 % above the inaccuracy $A_{2,n}$ of the estimates based on the quadratic entropy.

Here we are interested in the wider class of inaccuracies $A_{\alpha,n}$ of the estimates of $e(X)$ based on the power entropies $H_\alpha(X)$ of orders $\alpha \in \mathbb{R}$. If $\alpha < 0$ then the condition $H_\alpha^{\max} < \infty$ of our definition (3.5) is violated whenever $n > 2$ because then the Bayes errors $0 < e(X) \le 1/2$ can be achieved at the distributions p_X with one component zero, for which $H_\alpha(X) = \infty$. Therefore we restrict attention to $\alpha \ge 0$. The following generalization of (3.2) – (3.4) and modification (3.5) will be useful. It uses the indicator function $\boldsymbol{I}$ and the notation

$$e_k = \frac{k-1}{k} \quad \text{for} \quad 1 \le k \le n\,. \tag{3.6}$$

Proposition 3.2. If $\alpha \ge 0$ then the functions

$$H_\alpha^+(e) = \frac{1 - (1-e)^\alpha - (n-1)^{1-\alpha} e^\alpha}{\alpha - 1}\, \boldsymbol{I}(0 < e \le e_n)\,, \tag{3.7}$$

$$H_\alpha^-(e) = \sum_{k=1}^{n-1} \frac{1-k(1-e)^\alpha-(1-k(1-e))^\alpha}{\alpha-1}\, \boldsymbol{I}(e_k < e \le e_{k+1}) \tag{3.8}$$

together with their limits (cf. (3.2), (3.3))

$$H_1^+(e) = h(e) + e\ln(n-1)\,, \tag{3.9}$$

$$H_1^- = \sum_{k=1}^{n-1} k(1-e)(1+e-(1-e)k)\, \boldsymbol{I}(e_k < e \le e_{k+1}) \tag{3.10}$$

are attainable upper and lower bounds for $H_\alpha(X)$ in the class of all X with fixed Bayes errors $e \in [0, e_n]$. If $\alpha > 0$ then the inaccuracies (3.5) satisfy the relation

$$A_{\alpha,n} = \frac{1}{H_\alpha^{\max}} \int_0^{e_n} \left[H_\alpha^+(e) - H_\alpha^-(e)\right] \mathrm{d}e \tag{3.11}$$

for H_α^+, H_α^- defined by (3.7) – (3.10).

P r o o f. If $\alpha \ge 0$, then the α-entropies $H_\alpha(p)$ are Schur-concave in the sense of [30]. It is easy to verify that $H_\alpha(p^+), H_\alpha(p^-)$ for p^+, p^- from the proof of Proposition 3 are the functions given in (3.7) – (3.10). Therefore it remains to prove (3.11). If $\alpha > 0$ then the bounds (3.7) – (3.10) continuously increase in the domain $e \in [0, e_n]$, and $e_\alpha^{\max}(H), e_\alpha^{\min}(H)$ are inverse to $H_\alpha^+(e), H_\alpha^-(e)$. Therefore the integrals in (3.5) and (3.11) coincide which completes the proof. □

Example 3.1. The bounds (3.9), (3.10) coincide with those given in (3.2), (3.3) and putting $\alpha = 2$ in (3.7), (3.8) we obtain the bounds given in (3.4), (3.3). For $\alpha = 0$ we get from (3.7), (3.8)

$$H_0^+(e) = (n-1)\, \boldsymbol{I}(0 < e \le e_n), \quad H_0^-(e) = \sum_{k=1}^{n-1} k\, \boldsymbol{I}(e_k < e \le e_{k+1})\,. \tag{3.12}$$

It is evident from this example that if $\alpha = 0$ then the upper bound $H_0^+(e)$ of (3.7) is discontinuous at $e = 0$ and the lower bound $H_0^-(e)$ of (3.8) is discontinuous at all $e \in \{e_1, e_2, \ldots, e_{n-1}\}$. Here $H_0(X)$ is the Hartley entropy achieving only n possible integer values between $H_0^{\min} = 0$ and $H_0^{\max} = n-1$ so that the average inaccuracy $A_{0,n}$ can be obtained by the formal extension of (3.11) to $\alpha = 0$ and application of (3.12). However, it cannot be obtained by observing that if $H_0(X) = k \in \{0, 1, \ldots, n-1\}$ then the Bayes error $e(X)$ takes on values between

$$e_0^{\min}(k) = 0 \quad \text{and} \quad e_0^{\max}(k) = k/(k+1)\,.$$

Therefore we get the individual inaccuracies

$$e_0^{\max}(k) - e_0^{\min}(k) = k/(k+1) \quad \text{for} \quad 0 \le k \le n-1$$

leading to the average inaccuracy

$$A_{0,n} = \frac{1}{n}\sum_{k=1}^{n-1} \frac{k}{k+1} = 1 - \frac{1}{n}\sum_{k=1}^{n} \frac{1}{k}\,. \tag{3.13}$$

The following assertion uses this formula. It is based on Proposition 3.2.

Proposition 3.3. For $\alpha > 0$ the average inaccuracy $A_{\alpha,n}$ is given by the formula

$$A_{\alpha,n} = \frac{1}{(1-n^{1-\alpha})(\alpha+1)} \left(\sum_{k=1}^{n-1} \frac{1}{k(k+1)^{\alpha}} - \frac{n-1}{n^{\alpha}} \right) \tag{3.14}$$

if $\alpha \neq 1$ and by

$$A_{1,n} = \lim_{\alpha \to 1} A_{\alpha,n} = \frac{n-1}{2n} - \frac{1}{2\ln n} \sum_{k=1}^{n-1} \frac{\ln(k+1)}{k(k+1)} \tag{3.15}$$

if $\alpha = 1$. At $\alpha = 0$ the average inaccuracy (3.13) is larger than the corresponding limit for $\alpha \downarrow 0$, namely it holds

$$A_{0,n} = \lim_{\alpha \downarrow 0} A_{\alpha,n+1}. \tag{3.16}$$

Proof. Let $\alpha > 0$ be different from 1. It suffices to prove (3.14) because the functions (3.7), (3.8) applied in (3.11) are bounded and continuous in the neighborhoods of $\alpha = 0$ and $\alpha = 1$ with the limits (3.9), (3.10) for $\alpha \to 1$ and (3.12) for $\alpha \downarrow 0$. By a routine integration we get

$$\int_0^{e_n} H_{\alpha}^{+}(e)\,\mathrm{d}e = \frac{1}{\alpha-1} \left[e_n - \frac{n^{\alpha}+n-2}{(\alpha+1)n^{\alpha}} \right]$$

and

$$\int_0^{e_n} H_{\alpha}^{-}(e)\,\mathrm{d}e = \frac{1}{\alpha-1} \left[e_n - \frac{1}{\alpha+1} \sum_{k=1}^{n-1} \frac{(k+1)^{\alpha} - (k-1)k^{\alpha-1}}{[k(k+1)]^{\alpha}} \right].$$

If we use the fact that

$$H_{\alpha}^{\max} = \frac{1-n^{1-\alpha}}{\alpha-1}$$

and that the last sum equals

$$\sum_{k=1}^{n-1} \frac{1}{k^{\alpha}} - \sum_{k=1}^{n-1} \frac{1}{(k+1)^{\alpha}} + \sum_{k=1}^{n-1} \frac{1}{k(k+1)^{\alpha}} = \frac{n^{\alpha}-1}{n^{\alpha}} + \sum_{k=1}^{n-1} \frac{1}{k(k+1)^{\alpha}}$$

and substitute these expressions in (3.11), we get the desired result (3.14). The remaining assertions (3.15), (3.16) follow by taking limits for $\alpha \to 1$ and $\alpha \downarrow 0$ in (3.14). The equality of (3.16) and (3.13) is easily seen. □

Example 3.2. By putting $\alpha = 2$ in (3.14) we get the average inaccuracy

$$A_{2,n} = \frac{n}{3(n-1)} \sum_{k=1}^{n-1} \frac{1}{k(k+1)^2} - \frac{1}{3n}.$$

Using

$$\sum_{k=1}^{n-1} \frac{1}{k(k+1)^2} = \sum_{k=1}^{n-1} \frac{1}{k(k+1)} - \sum_{k=1}^{n-1} \frac{1}{(k+1)^2} = 2 - \frac{1}{n} - \sum_{k=1}^{n} \frac{1}{k^2}$$

we get

$$A_{2,n} = \frac{2n}{3n-1} - \frac{2n-1}{3n(n-1)} - \frac{1}{3} \sum_{k=1}^{n} \frac{1}{k^2}. \tag{3.17}$$

Proposition 3.4. The average inaccuracies (3.5) for the power entropy estimates of the Bayes error satisfy the limit laws

$$A_{\alpha,\infty} = \lim_{n\to\infty} A_{\alpha,n} = \begin{cases} \dfrac{1}{\alpha+1} & \text{if} \quad 0 \le \alpha \le 1 \\ \dfrac{1}{\alpha+1} \displaystyle\sum_{k=1}^{\infty} \frac{1}{k(k+1)^{\alpha}} & \text{if} \quad \alpha > 1\,. \end{cases} \tag{3.18}$$

P r o o f. For $\alpha = 0$ this follows from (3.16) where

$$\sum_{k=1}^{n-1} \frac{1}{k} = \ln(n-1) + C + o(1) \quad \text{as} \quad n \to \infty \tag{3.19}$$

for the Euler constant C. For $0 < \alpha < 1$ this follows from (3.14) and (3.19) because

$$\sum_{k=1}^{n-1} \frac{1}{k(k+1)^{\alpha}} \le \sum_{k=1}^{n-1} \frac{1}{k}\,.$$

For $\alpha = 1$ it suffices to use (3.15) and the expansion

$$\frac{\ln(k+1)}{k(k+1)} = \frac{\ln((k+1)/k)}{k} + \frac{\ln k}{k} - \frac{\ln(k+1)}{k+1}$$

where $\ln((k+1)/k)$ is bounded above by $1/k$. Indeed, this expansion implies

$$0 \le \sum_{k=1}^{n-1} \frac{\ln(k+1)}{k(k+1)} \le \sum_{k=1}^{\infty} \frac{1}{k^2} - \frac{\ln n}{n}$$

so that, by the Euler formula, the constant

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \tag{3.20}$$

uniformly upper bounds the sum. For $\alpha > 1$ the desired result follows immediately from (3.14). □

Example 3.3. We see from Proposition 3.4 that $A_{0,\infty} = 1$, $A_{1/2,\infty} = 2/3$ and $A_{1,\infty} = 1/2$. Applying the Euler formula (3.20) in (3.17) we get

$$A_{2,\infty} = \frac{2}{3} - \frac{\pi^2}{18} \doteq 0.118\,. \tag{3.21}$$

Next follows a table presenting exact values of some of the average inaccuracies $A_{\alpha,n}$ specified in Propositions 3.3, 3.4 and Examples 3.2, 4.2. We see from this table

that the inaccuracy $A_{1,n}$ of the Shannon-entropy based estimate exceeds the inaccuracy $A_{2,n}$ of the quadratic-entropy-based estimate at least twice which confirms the above mentioned hypothesis of Vajda [28]. The average inaccuracy of the quadratic entropy is at most 12 % while for the Shannon entropy it can be close to 50 %. But at the same time it is seen more, namely that the average inaccuracies $A_{\alpha,n}$ of the power-entropy-based estimates decrease with α increasing in the domain $[0, \infty)$ and already for $\alpha = 4$ they remain close to 1 % uniformly for all observations space sizes $n > 1$.

Table 3.1. Average inaccuracies $A_{\alpha,n}$ for selected α and n.

n	2	3	4	5	6	7	8	9	10	...	∞
$A_{0,n}$	.250	.389	.479	.543	.592	.630	.660	.686	.707	...	1.000
$A_{1/2,n}$	0	.145	.225	.278	.315	.344	.367	.385	.401	...	.667
$A_{1,n}$	0	.092	.142	.175	.198	.215	.229	.240	.249	...	.500
$A_{3/2,n}$	0	.061	.093	.113	.127	.137	.144	.151	.156	...	.229
$A_{2,n}$	0	.042	.062	.074	.081	.087	.091	.094	.096	...	.118
$A_{3,n}$	0	.020	.027	.031	.033	.034	.035	.036	.036	...	.038
$A_{4,n}$	0	.009	.012	.013	.013	.014	.014	.014	.014	...	.014

4. ENTROPIES AND MEASURES OF DIVERSITY

The quadratic entropy $H_2(X)$ proposed as a measure of quality of statistical decisions based on X in Vajda [28] was quite frequently used in this role, see e. g. Devijver and Kittler [7] or Devroye et al. [8] and further references there. But $H_2(X)$ was proposed much earlier as a measure of equality of an income distributed by p_X among given social groups $x \in \mathcal{X}$ by Dalton [6] (cf. e. g. Sen [22]). It is also called *Simpson's measure of diversity* in biological literature (or Simpson–Gini index), with a reference to Simpson [23] and Gini [10] (cf. e. g. Emlen [9] or Marshal and Olkin [17]).

The quadratic entropy is included in a wider class of diversity measures for distributions $p(x) = p_X(x)$

$$\mathcal{D}_U(X) = \sum_x U(p(x)) \tag{4.1}$$

introduced for all concave "utility functions" $U(\pi)$ of variable $\pi \in [0, 1]$ by Dalton [6]. Thus $\mathcal{D}_U(X)$ are concave functions of distributions $(p(x) : x \in \mathcal{X})$, among them $H_2(X)$ obtained for $U(\pi) = \pi(1 - \pi)$. Such functions were systematically studied as uncertainty measures in Vajda [26] and Morales et al. [18]. In spite of that this class is very general, many diversity measures considered in the literature are $\widetilde{\phi}$-entropies $H_{\widetilde{\phi}}(X)$ in the sense of 2.1 which are nonconcave, and thus not belonging to the Dalton class (4.1).

Example 4.1. On p. 412 of Marshal and Olkin [17] study the *Fishlow diversity measure* of the form

$$H_{\widetilde{\phi}}(X) = \sum_x p(x)\widetilde{\phi}(p(x)) = \sum_x (1/n - p(x))^+ \quad \text{(cf. (1.24) and Figure 1.1)}$$

where $(f(x))^+$ denotes the positive part of any function $f(x)$. Here $\widetilde{\phi}(t) = \phi_+(1/(nt))$ $= (1/(nt) - 1)^+$ corresponds in the sense

$$\widetilde{\phi}(t) = \phi^*(t) + \phi(0)(1-t) \tag{4.2}$$

considered in (2.1) to the convex function

$$\phi(t) = t\left(\frac{t}{n} - 1\right)^+ = \left(\frac{t^2}{n} - t\right)^+ \tag{4.3}$$

of variable $t > 0$. Thus $H_{\widetilde{\phi}}(X)$ is an average amount of information obtained by observing realizations $x \in \mathcal{X}$ of X where the individual amounts of information $\widetilde{\phi}(p(x)) = (1/(np(x)) - 1)$ are nonzero only if the probabilities $p(x)$ of messages $x \in \mathcal{X}$ are significant in the sense that they are less than average, i. e. if

$$p(x) < \frac{1}{n} = \frac{1}{n}\sum_x p(x)\,.$$

In other words, the events with average and more than average probabilities are considered to be nonsignificant and thus noninformative. The present diversity measure $H_{\widetilde{\phi}}(X)$ has the form (4.1) for the utility function $U(\pi) = (1/n - \pi)^+$ which is convex on $[0, 1]$. Therefore this diversity measure does not belong to the Dalton class. Moreover this diversity measure extends the class of entropies introduced in Section 2 which are nonconcave and at the same time are ϕ-information for a convex ϕ with $\phi(1) = 0$. The present particular function ϕ is given in (4.3).

Example 4.2. Let us now consider the *Emlen diversity measure*

$$\mathcal{D}(X) = \sum_x p(x)\, e^{-p(x)} - c^{-1}, \quad c = 2.718...$$

introduced to the biometry in [10]. Here we subtracted from the original Emlen's proposal the constant $c^{-1} = (2.718...)^{-1}$ in order to shift the range of values to the interval $[0, \infty)$. Obviously,

$$\mathcal{D}(X) = H_{\widetilde{\phi}}(X) = \sum_x p(x)\,\widetilde{\phi}(p(x))$$

where $\widetilde{\phi}(t) = (c^{1-t} - 1)/c$ corresponds in the sense (4.2) to the convex function

$$\phi(t) = t(c^{1-1/t} - 1)/c, \quad t > 0 \tag{4.4}$$

from the class $\mathbf{\Phi}$ considered in Sections 1 and 2. Thus the Emlen's $\mathcal{D}(X)$ is the ϕ-information $I_\phi(X;X)$ in the sense of (2.1) for $\phi \in \mathbf{\Phi}$ given in (4.4). At the same time it holds

$$\mathcal{D}(X) = \sum_x U(p(x))$$

where $U(\pi) = \pi(c^{1-\pi} - 1)/c$ is not concave on $[0,1]$. Hence the Emlen diversity does not belong to the Dalton class (4.1) and is thus another example of nonconcave entropy which is at the same time a maximal $\phi-$ information. However, contrary to the previous example where the measure of diversity was a convex function of distribution $p = p_X$, here $\mathcal{D}(X)$ is a strictly Schur-concave function of p. Consequently the biometry can serve as a new source of motivation for the Schur-concave entropies systematically studied in Morales et al. [18] as a natural extension of the concave entropies studied previously in [26].

An important conclusion from what has been said above is that the diversities of random variables X can be measured by concave, Schur-concave or convex entropies. Zvárová [31], Zvárová and Mazura [32] and recently Zvárová and Vajda [33] studied diversity of the genes X taking on on various alleles $x \in \mathcal{X}$ with relative frequencies $p_X(x)$ which do not remain the same if we go from one population to other. They proposed and more deeply investigated the class of measures of genetic diversity

$$\left\{ H^{\sim}_{\phi}(X) = I_\phi(X;X) : \phi \in \mathbf{\Phi} \right\}$$

containing as a particular cases the *Shannon measure* $H_1(X)$ given in (1.18) and the Simpson measure given in (1.22) as well as some nonconcave entropies discussed in Section 2. We have seen that the properties of these diversity measures may be quite different.

Basic applications of genetic diversity measures are comparisons of diversities of two different genes X and Y in the same population or of the same gene in two different populations (this can also be characterized by two random variables X and Y). The problem is whether or to what extent the comparison of genetic or ecological or any other diversities depends on the used diversity measure. In the rest of this section we study this problem. The role of different diversity measures $\mathcal{D}_1(X)$ and $\mathcal{D}_2(X)$ will be played mainly by the Shannon and Simpson measures $H_1(X)$ and $H_2(X)$.

Example 4.3. Let X be geometric random variable with $p_X(i) = (1-\pi)\pi^i$ for $i = 0, 1, 2, \ldots$. Then

$$H_1(X) = \frac{h(\pi)}{1-\pi} \quad \text{and} \quad H_2(X) = \frac{2\pi}{1+\pi}$$

for $\pi \in [0,1)$ and $h(\pi)$ given in (2.11). Since both these functions are increasing in $\pi \in [0,1)$, the diversity measures H_1 and H_2 are isotone in the family $\mathcal{P}$ of geometric models in the sense that

$$H_1(X) \lesseqgtr H_1(Y) \quad \text{iff} \quad H_2(X) \lesseqgtr H_2(Y) \tag{4.5}$$

for two geometric random variables X and Y. If however $\mathcal{P}_3$ is the class of all discrete distributions of size $n = 3$ and

$$p_X = \left(\frac{1}{2}, \frac{1}{2}, 0\right), \qquad p_Y = \left(\frac{3}{4}, \frac{1}{8}, \frac{1}{8}\right)$$

are distributions of X, Y from $\mathcal{P}_3$ then

$$H_1(X) = \ln 2 < H_1(Y) = \frac{3}{4} \ln \frac{8}{3}$$

while

$$H_2(X) = \frac{1}{2} > H_2(Y) = \frac{3}{8}.$$

This means that (4.5) fails to hold, i. e. that H_1 and H_2 are not isotone in the family $\mathcal{P}_3$.

Let $\mathcal{D}_1(X)$ and $\mathcal{D}_2(X)$ be diversity measures defined for all observations X distributed by p_X for a given family $\mathcal{P}$ of discrete distributions. These measures are said to be *isotone* on $\mathcal{P}$ if for all X and Y with distributions from $\mathcal{P}$

$$\mathcal{D}_1(X) \lesseqgtr \mathcal{D}_1(Y) \quad \text{iff} \quad \mathcal{D}_2(X) \lesseqgtr \mathcal{D}_2(Y). \tag{4.6}$$

If moreover both $\mathcal{D}_j(X)$ take on $\mathcal{P}$ all values between $\mathcal{D}_j^{\min} < \mathcal{D}_j^{\max}$ then the set of all $(d_1, d_2) \in (\mathcal{D}_1^{\min}, \mathcal{D}_1^{\max}) \times (\mathcal{D}_2^{\min}, \mathcal{D}_2^{\max})$ such that

$$\inf_{\mathcal{D}_2(X)=d_2} \mathcal{D}_1(X) < d_1 < \sup_{\mathcal{D}_2(X)=d_2} \mathcal{D}_1(X) \tag{4.7}$$

is called *anisotony domain* and denoted $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$.

Proposition 4.1. If an anisotony domain $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P}) \subset \mathbb{R}^2$ has non-void interior then the diversity measures $\mathcal{D}_1(X)$ and $\mathcal{D}_2(X)$ are not isotone on $\mathcal{P}$.

P r o o f. If the assumption holds then there exists a non-void sphere, and consequently a non-void square $(d_1, d_1 + \varepsilon) \times (d_2, d_2 + \varepsilon)$ contained in $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$. By assumptions and (4.7), this means that there exist X, Y with distributions in $\mathcal{P}$ satisfying the relations

$$\mathcal{D}_2(X) = d_2, \quad D_2(Y) = d_2 + \varepsilon$$

and

$$\mathcal{D}_1(X) > d_1 + \varepsilon, \quad D_1(Y) < d_1.$$

These relations contradict (4.6) which completes the proof. □

The anisotony domains $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$ are usually open in $\mathbb{R}^2$ with non-void interior as in the case of $\mathcal{A}(H_1, H_2|\mathcal{P}_3)$ from Example 4.3, but they may be also empty as in the case of $\mathcal{A}(H_1, H_2|\mathcal{P})$ from the same example. If $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$ is measurable with the Lebesgue measure $\mu(\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P}))$ then the relative size

$$\alpha(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P}) = \frac{\mu(\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P}))}{(\mathcal{D}_1^{\max} - \mathcal{D}_1^{\min})(\mathcal{D}_2^{\max} - \mathcal{D}_2^{\min})} \tag{4.8}$$

of the uncertainty domain $\mathcal{A}(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$ in the rectangle $(\mathcal{D}_1^{\min}, \mathcal{D}_1^{\max}) \otimes (\mathcal{D}_2^{\min}, \mathcal{D}_2^{\max})$ can serve as a *measure of anisotony* of the diversity measures $\mathcal{D}_1(X)$ and $\mathcal{D}_2(X)$ on the family $\mathcal{P}$ (a more sophisticated motivation of this measure can be found in [37]).

It is clear from the definition that if $\mathcal{D}_1(X)$ and $\mathcal{D}_2(X)$ are isotone on $\mathcal{P}$ then the measure $\alpha(\mathcal{D}_1, \mathcal{D}_2|\mathcal{P})$ of their anisotony on $\mathcal{P}$ is well defined and equal zero. An example for this is $\alpha(H_1, H_2|\mathcal{P}) = 0$ for the set $\mathcal{P}$ of geometric distributions from Example 4.3. The following general result enables to evaluate as a particular case the measure of anisotony $\alpha(H_1, H_2|\mathcal{P}_3)$ for the set $\mathcal{P}_3$ of all discrete distribution of size $n = 3$ from Example 4.3.

Proposition 4.2. Let $\mathcal{P}_n$ be the set of all discrete distributions of size $n \geq 2$ and let $e_1, \ldots, e_n$ be the numbers from $(0, 1)$ defined by (3.6). Then the anisotony domain for the diversity measures $H_1(X), H_2(X)$ on $\mathcal{P}_n$ is given by

$$\mathcal{A}(H_1, H_2|\mathcal{P}_n) = \big\{(d_1, e) : 0 < e < e_n, H_1^-(e) < d_1 < H_1^+(e)\big\} \tag{4.9}$$

where

$$H_1^+(e) = h(s_n(e)) + s_n(e)\, \ln(n-1) \tag{4.10}$$

and

$$H_1^-(e) = \sum_{k=1}^{n-1} [h(t_{k+1}(e)) + t_{k+1}(e)\, \ln k]\, \boldsymbol{I}(e_k < e \leq e_{k+1}) \tag{4.11}$$

for

$$s_n(e) = e_n - \sqrt{e_n(e_n - e)}, \qquad t_k(e) = e_k + \sqrt{e_k(e_k - e)}\,. \tag{4.12}$$

P r o o f . This proof is based on Theorem II.1 of Harremöes and Topsøe [11]. One can deduce from there that if $e \in (e_k, e_{k+1}]$ for $1 \leq k \leq n-1$ then

$$\max_{p:H_2(p)=e} H_1(p) = H_1(p^+(e)) \tag{4.13}$$

and

$$\min_{p:H_2(p)=e} H_1(p) = H_1(p^-(e)) \tag{4.14}$$

where

$$p^+(e) = \left(s, \frac{1-s}{n-1}, \ldots, \frac{1-s}{n-1}\right), \qquad p^-(e) = \left(\frac{1-t}{k}, \ldots, \frac{1-t}{k}, t, 0, \ldots, 0\right)$$

are distributions from $\mathcal{P}_n$ with $s \geq (1-s)/(n-1)$ and $t \leq (1-t)/k$ depending on e by means of the condition

$$H_2(p^+(e)) = H_2(p^-(e)) = e\,.$$

This condition represents two different quadratic equations in the variable e. Their unique solutions $s = s_n(e)$ and $t = t_k(e)$ satisfying the conditions $s \geq (1-s)/(n-1)$ and $t \leq (1-t)/k$ are presented in (4.12). The remaining steps leading to the formulas (4.10), (4.11) for the extremal values of the diversity H_1 are clear. Further, $H_2^{\min} = 0, H_2^{\max} = e_n$ and for every $d_2 \in (e_k, e_{k+1}] \subset (0, e_n]$ the values $H_1^-(d_2), H_1^+(d_2)$ obtained from (4.10), (4.11) represent the infima and suprema from (4.7). Hence the formula (4.9) with e replaced by d_2 represents exactly what is prescribed by the definition (4.7). □

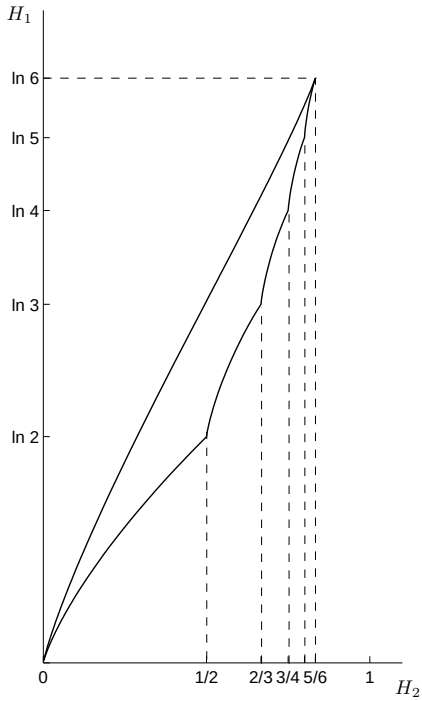

Fig. 4.1. Upper and lower bounds $H_1^+ = H_1^+(H_2)$ and $H_1^- = H_1^-(H_2)$ for $n = 6$.

Figure 4.1 given below illustrates the bounds $H_1^+(d), H_1^-(d)$ of Proposition 4.2 and the area between them is an example of the anisotony domain $\mathcal{A}(H_1, H_2|\mathcal{P}_n)$ for $n = 6$. Analytic evaluation of the integral in the formula

$$\alpha_n(H_1, H_2|\mathcal{P}_n) = \frac{n}{(n-1)\ln n} \int_0^{e_n} \left[H_1^+(e) - H_1^-(e)\right] \mathrm{d}e \tag{4.15}$$

obtained from (4.8) for the measure of anisotony of the Shannon and Simpson diversities $H_1(X)$ and $H_2(X)$ on $\mathcal{P}_n$ with general $n \geq 2$ is too complicated to be given here. Instead we present in Table 4.1 the values of $\alpha_n = \alpha_n(H_1, H_2|\mathcal{P}_n)$ computed for selected n by means of numerical integration in (4.15) with the guaranteed accuracy in the first 3 decimal places. By [37],

Table 4.1. Measures of anisotony $\alpha_n = \alpha_n(H_1, H_2|\mathcal{P}_n)$ for selected values of n.

n	2	3	4	5	6	7	8	9	10	...	∞
α_n	0	.070	.097	.113	.125	.133	.139	.145	.150	...	.333

$$\lim_{n\to\infty} \alpha_n = \frac{1}{3} + O\left(\frac{1}{\ln n}\right) \quad \text{as } n \to \infty. \tag{4.16}$$

This justifies the value $\alpha_\infty = 0.333$ in the last column of the table and at the same time indicates that the rate of convergence in (4.16) is slow, of the logarithmic order. For example for $n = 10^6$ we obtain $\alpha_n = 0.292$ which is still far away from $\alpha_\infty = 0.333$.

We see from Table 4.1 that the anisotony between the Shannon and Simpson diversities is not negligible. It slowly increases from 0 to roughly 33 % when the size of observation space increases in the interval $2 \leq n \leq \infty$ but remains to be moderate, below 15 %, for the sizes $2 \leq n \leq 10$.

ACKNOWLEDGEMENT

Both authors welcome this opportunity to express their thanks to Albert Perez for introducing them to the area of generalized information measures. They acknowledge also the support of the Ministry of Education, Youth and Sports of the Czech Republic under project 1M06014 and the Czech Science Foundation under grant 102/07/1131.

(Received June 27, 2006.)

REFERENCES

[1] T. Cover and J. Thomas: Elements of Information Theory. Wiley, New York 1991.
[2] B. H. Cramér. Remarks on correlation. Skand. Akt. Tidskr. *7* (1924), 230–231.
[3] N. Cressie and T. R. C. Read: Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B *46* (1984), 440–464.
[4] I. Csiszár: Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. Ser. A *8* (1963), 85–108.
[5] I. Csiszár: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. *2* (1967), 299–318.
[6] I. Csiszár: A class of measures of informativity of observation channels. Period. Math. Hungar. *2* (1972), 191–213.
[7] H. Dalton: The Inequality of Incomes. Ruthledge & Keagan Paul, London 1925.
[8] P. Devijver and J. Kittler: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, NJ 1982.
[9] L. Devroy, L. Györfi, and G. Lugosi: A Probabilistic Theory of Pattern Recognition. Springer, Berlin 1996.
[10] J. M. Emlen: Ecology: An Evolutionary Approach. Adison-Wesley, Reading 1973.
[11] C. Gini: Variabilitá e Mutabilitá. Studi Economico-Giuridici della R. Univ. di Cagliari. *3* (1912), Part 2, p. 80.
[12] P. Harremöes and F. Topsøe: Inequalities between etropy and index of concidence. IEEE Trans. Inform. Theory *47* (2001), 2944–2960.
[13] J. Havrda and F. Charvát: Concept of structural a-entropy. Kybernetika *3* (1967), 30–35.
[14] W. Höffding: Masstabinvariante Korrelationstheorie. Teubner, Leipzig 1940.
[15] W. Höffding: Stochastische Abhängigkeit und funktionaler Zusammenhang. Skand. Aktuar. Tidskr. *25* (1942), 200–207.
[16] V. A. Kovalevskij: The problem of character recognition from the point of view of mathematical statistics. Character Readers and Pattern Recognition, 3–30. Spartan Books, New York 1967.
[17] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[18] F. Liese and I. Vajda: On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory *52* (2006), 4394–4412.
[19] A. W. Marshall and I. Olkin: Inequalities: Theory of Majorization and its Applications. Academic Press, New York 1979.
[20] D. Morales, L. Pardo, and I. Vajda: Uncertainty of discrete stochastic systems. IEEE Trans. Systems, Man Cybernet. Part A *26* (1996), 681–697.
[21] K. Pearson: On the theory of contingency and its relation to association and normal correlation. Drapers Company Research Memoirs, Biometric Ser. 1, London 1904.
[22] A. Perez: Information-theoretic risk estimates in statistical decision. Kybernetika *3* (1967), 1–21.
[23] A. Rényi: On measures of dependence. Acta Math. Acad. Sci. Hungar. *10* (1959), 441–451.
[24] A. Rényi: On measures of entropy and information. In: Proc. Fourth Berkeley Symposium on Probab. Statist., Volume 1, Univ. Calif. Press, Berkeley 1961, pp. 547–561.
[25] A. Sen: On Economic Inequality. Oxford Univ. Press, London 1973.
[26] E. H. Simpson: Measurement of diversity. Nature *163* (1949), 688.
[27] A. Tschuprow: Grundbegriffe und Grundprobleme der Korrelationstheorie. Berlin 1925.
[28] I. Vajda: Bounds on the minimal error probability and checking a finite or countable number of hypotheses. Information Transmission Problems *4* (1968), 9–17.
[29] I. Vajda: Theory of Statistical Inference and Information. Kluwer, Boston 1989.
[30] I. Vajda and K. Vašek: Majorization concave entropies and comparison of experiments. Problems Control Inform. Theory *14* (1985), 105–115.
[31] I. Vajda and J. Zvárová: On relations between informations, entropies and Bayesian decisions. In: Prague Stochastics 2006 (M. Hušková and M. Janžura, eds.), Matfyzpress, Prague 2006, pp. 709–718.
[32] J. Zvárová: On measures of statistical dependence. Čas. pěst. matemat. *99* (1974), 15–29.
[33] J. Zvárová: On medical informatics structure. Internat. J. Medical Informatics *44* (1997), 75–81.
[34] J. Zvárová: Information Measures of Stochastic Dependence and Diversity: Theory and Medical Informatics Applications. Doctor of Sciences Dissertation, Academy of Sciences of the Czech Republic, Institute of Informatics, Prague 1998.
[35] J. Zvárová and I. Mazura: Stochastic Genetics (in Czech). Charles University, Karolinum, Prague 2001.
[36] J. Zvárová and I. Vajda: On genetic information, diversity and distance. Methods of Inform. in Medicine *2* (2006), 173–179.
[37] J. Zvárová and I. Vajda: On isotony and anisotony between Gini and Shannon measures of diversity. Preprint, Prague 2006.

Igor Vajda, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: vajda@utia.cas.cz

Jana Zvárová, Centre of Biomedical Informatics, Institute of Computer Science – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8. Czech Republic.
e-mail: zvarova@cs.cas.cz