ON THE RATE OF INFORMATION GAIN IN EXPERIMENTS WITH A FINITE PARAMETER SET[1]

J. Krob and H. v. Weizsäcker

Received: Revised version:

Dedicated to the memory of Erik N. Torgersen[2]

Abstract

Let $(\mathcal{E}_k)$ be a sequence of experiments with the same finite parameter set. Suppose only that identification of the parameter is possible asymptotically. For large classes of information functionals we show that their exponential rates of convergence towards complete information coincide. As a special case we obtain the rate of the Shannon capacity of product experiments.

1 Introduction

There are various measures of the information content of a statistical experiment $\mathcal{E}$, among others the decision theoretic deficiency distance to the most informative experiment and the Shannon capacity (which was introduced in the statistical context by D.V. Lindley [9] and J.M. Bernardo [1]). These numbers are not easily computed. Therefore it is desirable to describe at least their asymptotic behaviour when the experiments get more and more informative. To our knowledge the asymptotics of the Shannon capacity has not been studied for finite parameter situations. We do this using the framework of f-(dis)similarities in the sense of Györfi and Nemetz [5].
The main message of our paper is the following observation: Let $\mathcal{E}_k$ $(k \in \mathbb{N})$ be a sequence of experiments with a common finite parameter set. The only assumption about these experiments is that they allow asymptotically the identification of the parameter, i.e. that they converge to an experiment with complete information. Then the exponential rate of the information gain does not depend on the special choice of the information functional,

[1]The paper is based on parts of the first author's doctorial thesis.
AMS 1980 subject classification 62B10, 94A17, 62C05
Keywords: information, deficiency, Shannon capacity, statistical experiment, f-dissimilarity, exponential rate, large deviations
[2]This work was stimulated by very instructive discussions with Erik Torgersen. We deeply regret his early death.

at least in a very wide class of functions f. In particular the capacity and the deficiency always have the same rate.
For product experiments E. Torgersen ([12]) extended the large deviation theorem of H. Chernoff ([2]) in order to identify the deficiency rate. Thus our result gives in particular the asymptotic rate of the Shannon capacity for finite parameter product experiments. For another model where these rates can be computed explicitly see the forthcoming paper [10] which treats finite Markov chain experiments.

The paper is organized as follows: Section 2 introduces f-similarities, discusses some examples and extends the concepts of equivocation, information gain and capacity to the more general setting. Section 3 collects a few useful facts about concave functions on the simplex of n-ary probability vectors. Section 4 discusses the convergence of the f-similarities if the identification of the parameters is possible. Section 5 contains the main results. Roughly speaking they say: Exponentially equivalent f have the same exponential rate for the similarity.

Notations. Let $\mathcal{E} = \{P_1, ..., P_n\}$ be a statistical experiment with parameter set $\{1, ..., n\}$, i.e. a finite family of probability distributions on the common measurable space X. The set of all probability vectors of length n is denoted by $Prob_n$. Let $\lambda \in Prob_n$ be considered as a prior distribution on the parameter space. We write $P_\lambda = \sum_{i=1}^n \lambda_i P_i$ and for every $x \in X$ the associated posterior distribution $(\frac{d\lambda_1 P_1}{dP_\lambda}(x), ..., \frac{d\lambda_n P_n}{dP_\lambda}(x)) \in Prob_n$ is denoted by $\hat{\lambda}(x)$. The uniform prior $(\frac{1}{n}, \cdots, \frac{1}{n})$ is denoted λ_{unif}. For $1 \le i \le n$ the symbol e_i denotes the i-th unit vector which corresponds to the point mass at i.

2 f-Similarities

Extending an idea of I. Csiszár [4], L. Györfi and T. Nemetz introduced (see the references in [5]) the concept of f-dissimilarities as a way to measure the degree of mutual singularity of a family of n probability distributions $P_1, ..., P_n$ on the same observation space. Here f is a convex function on $Prob_n$. In order to include the entropy functional we change signs and consider "f-similarities" for concave functions f. Actually for our main results we shall weaken the concavity requirement considerably. However, for us it is essential that the function f is bounded which excludes some cases studied by Csiszár. It turns out to be helpful to introduce a prior distribution as an additional parameter.

Definition 2.1 *Let $\mathcal{E} = \{P_1, ..., P_n\}$ be a finite family of probability distributions on the measurable space X. Let $f : Prob_n \longrightarrow \mathbb{R}$ be a continuous function. Let $\lambda \in Prob_n$ be a prior distribution on $\{1, ..., n\}$. The* f **-Similarity of $\mathcal{E}$ w.r.t.** λ *is given by*

$$H_f(\mathcal{E}, \lambda) = \int_X f(\frac{d\lambda_1 P_1}{dP_\lambda}(x), ..., \frac{d\lambda_n P_n}{dP_\lambda}(x))\ P_\lambda(dx).$$

Remark 2.2 Since f is evaluated at the posterior distribution $\hat{\lambda}(x)$ corresponding to the prior λ and the observation x the number $H_f(\mathcal{E}, \lambda)$ can be considered as the expected posterior value of f. The value $H_f(\mathcal{E}, \lambda_{unif})$ is denoted by $f(\mathcal{E})$ in decision theoretic literature, see e.g. [12], p.640.

Remark 2.3 If f is affine then one has $H_f(\mathcal{E}, \lambda) = f(\lambda) = \sum_{i=1}^n \lambda_i f(e_i)$. Thus for affine f this number contains no information about the experiment.
We are interested mainly in the asymptotic behaviour of

$$H_f(\mathcal{E}_k, \lambda) - H_f(\mathcal{M}_a, \lambda) \tag{1}$$

for sequences of experiments $\mathcal{E}_k = \{P_1^{(k)}, \cdots, P_n^{(k)}\}$ where the measures $P_i^{(k)}$ become more and more singular to each other and the experiment $\mathcal{M}_a$ is one with complete information, i.e. with pairwise singular measures P_i. For $\mathcal{M}_a$ one has $P_i\{\hat{\lambda}_i = e_i\} = 1$ for each i and hence $H_f(\mathcal{M}_a, \lambda) = \sum_{i=1}^n \lambda_i f(e_i)$ is just the affine part of the function f. Thus the difference (1) will not be changed if we change the affine part of f. Therefore we shall assume in many cases $f(e_i) = 0$ or equivalently $H_f(\mathcal{M}_a, \lambda) = 0$. As $\mathcal{E}$ approaches $\mathcal{M}_a$ the posterior distributions are close to the extreme points e_i and, thus, what really matters for the asymptotics of the above difference is the asymptotic behaviour of $f(\lambda)$ as λ approaches these extreme points. In particular we shall see that changing f by, say, a factor of logarithmic order does not change the exponential rate of $H_f(\mathcal{E}, \lambda)$. Therefore this rate depends very little on the explicit 'parametric' form of f.

If f is concave then Jensen's inequality implies the inequality

$$H_f(\mathcal{E}, \lambda) \leq f(\int \hat{\lambda}_1 dP_\lambda, ..., \int \hat{\lambda}_n dP_\lambda) = f(\lambda_1, ..., \lambda_n).$$

Hence for concave f the number $H_f(\mathcal{E}, \lambda)$ as a function of $\mathcal{E}$ has the maximal value $f(\lambda)$ which is attained if $P_1 = ... = P_n$, i.e. if $\mathcal{E}$ is the least informative experiment.

Like the dissimilarity (cf. [5]) the similarity includes for particular choices of f several known functionals.

Example 2.4 Let $f : Prob_n \longrightarrow \mathbb{R}_+$ be the entropy functional given by

$$f(z) = \sum_{i=1}^n -z_i \log z_i.$$

Then $H_f(\mathcal{E}, \lambda)$ is the conditional entropy of the parameter given the observation where the parameter and the observation are considered as random variables on the probability space $(\{1, ..., n\} \times X, \sum_{i=1}^n \lambda_i \varepsilon_i \otimes P_i)$. In information theory the number $H_f(\mathcal{E}, \lambda)$ in this case is sometimes called **equivocation**.

Example 2.5 Let $b : Prob_n \longrightarrow \mathbb{R}_+$ be defined by $b(z_1, ..., z_n) = 1 - (\max_j z_j)$. Then $H_b(\mathcal{E}, \lambda)$ is the Bayes-risk for the prior λ (and $0-1$ loss function). As is shown e.g. in [13],p.255, the **deficiency** $\delta(\mathcal{E}, \mathcal{M}_a)$ of the experiment $\mathcal{E}$ to the most informative experiment $\mathcal{M}_a$ in which all P_i are mutually singular, is equal to $2\max_\lambda H_b(\mathcal{E}, \lambda)$. If Ω is the observation space of the experiment $\mathcal{E}$, the number $\delta(\mathcal{E}, \mathcal{M}_a)$ can be defined as

$$\inf_K \max_{1\leq i\leq n} \frac{1}{2} \parallel P_i K - e_i \parallel_1,$$

where K varies over all transition kernels $K : \Omega \to Prob_n$, cf. [8], p. 8.

Example 2.6 For $\alpha \in Prob_n$ consider the (concave) function f_α given by $f_\alpha(z) = \prod_{i=1}^n z_i^{\alpha_i}$. Then the number $H_{f_\alpha}(\mathcal{E}, \lambda_{unif})$ is the **Hellinger transform** of the experiment at the point α.

Example 2.7 In [12] the rate of (1) is determined for product experiments where $f(z) = -\psi(z)$ and ψ is a sublinear map on the whole space $\mathbb{R}^n$ (i.e. ψ satisfies $\psi(x+y) \leq \psi(x) + \psi(y)$ and $\psi(ax) = a\psi(x)$ for all $a > 0$ and $x, y \in \mathbb{R}^n$). Clearly every sublinear ψ is convex and hence the induced f on $Prob_n$ is concave. The function b in example 2.5 is of this type: Take $\psi(z) = \parallel z \parallel_\infty - \sum_{i=1}^n z_i$. However, not every concave function f on $Prob_n$ is of this form. Neither the entropy functional of example 2.4 nor the functions f_α in example 2.6 can be extended to concave functions on $\mathbb{R}^n$ because on any line through an extreme point e_i their slope near e_i becomes infinite.

Example 2.8 Let us specialize to $n = 2$. Fix $1 \leq r < \infty$. Let $f(z_1, z_2) = 1 - |z_1^{\frac{1}{r}} - z_2^{\frac{1}{r}}|^r$. Because of $b(z_1, z_2) = 1 - \max(z_1, z_2) = \min(z_1, z_2)$ and the general relation $a + b - |a - b| = 2\min(a, b)$ one has $f(z) = 2b(z)$ in the special case $r = 1$. Moreover writing $Q = (P_1 + P_2)/2$ we get

$$H_f(\mathcal{E}, (\frac{1}{2}, \frac{1}{2})) = 1 - \frac{1}{2} \int |\frac{dP_1}{dQ}^{\frac{1}{r}} - \frac{dP_2}{dQ}^{\frac{1}{r}}|^r \, dQ$$

Here the integral is for $r = 1$ the total variation distance and for $r = 2$ the square of the Hellinger distance between P_1 and P_2. So in particular

$$2H_b(\mathcal{E}, (\frac{1}{2}, \frac{1}{2})) = 2 - \parallel P_1 - P_2 \parallel_1 .$$

Example 2.9 For $n = 2$ one can also consider the function $f(z_1, z_2) = z_1 \tilde{f}(\frac{z_1}{z_2})$ for a function $\tilde{f}$ on $\mathbb{R}_+$. This formally gives the $\tilde{f}$-divergence of Csiszár [4]. However our boundedness restriction (which is enforced by the continuity) excludes interesting convex examples like the Kullback-Leibler number which corresponds to the function $\tilde{f}(x) = \log x$. Of course one could symmetrize and rescale, taking

$$f(z_1, z_2) = \exp(-z_1 \tilde{f}(\frac{z_1}{z_2}) - z_2 \tilde{f}(\frac{z_2}{z_1})).$$

This function is not concave. The corresponding similarity then is related in a weak sense to the Kullback-Leibler distance and it fits into our framework.

For some purposes it is convenient to pass from $Prob_n$ to the full cone $\mathbb{R}^n_+$.

Definition 2.10 *Given a real function f on $Prob_n$ we extend it to $\overline{f}$ on the cone $\mathbb{R}^n_+$ by setting*

$$\overline{f}(0) = 0 \quad and \quad \overline{f}(z) = \| z \|_1 f(\frac{z}{\| z \|_1})$$

where $\| z \|_1 = \sum_{i=1}^n |z_i|$.

Remark 2.11 Clearly $\overline{f}$ is positively homogeneous, i.e. it satisfies $\overline{f}(az) = a\overline{f}(z)$ for all $a \geq 0$ and $z \in \mathbb{R}^n_+$. Moreover $\overline{f}$ is continuous if f is continuous. If we replace f by $\overline{f}$ in the definition of $H_f(\mathcal{E}, \lambda)$ the P_λ could be replaced by any measure dominating the $\lambda_i P_i$. In this connection we note (but shall not use) the fact that the number $H_f(\mathcal{E}, \lambda)$ is the value $m_\mathcal{E}(\overline{f}_\lambda)$ of the conical measure $m_\mathcal{E}$ associated with the experiment $\mathcal{E}$ at the positively homogeneous function $\overline{f}_\lambda : z \mapsto \overline{f}(\lambda_1 z_1, ..., \lambda_n z_n)$ (cf. [7],ch.3). So in this sense our paper deals with 'large deviations of conical measures'.

Extending the approach of D.V. Lindley [9] from entropy to general similarities one introduces the expected amount of information (measured in terms of the functional f) which one gains by passing from the prior to the posterior distribution. This corresponds to the transmission rate in Shannon theory. The maximal information gain corresponds to the Shannon capacity.

Definition 2.12 *The number*

$$I_f(\mathcal{E}, \lambda) = f(\lambda_1, ..., \lambda_n) - H_f(\mathcal{E}, \lambda)$$

*is called the f-**information gain** of the experiment $\mathcal{E}$ w.r.t. the prior λ. The symbol $C_f(\mathcal{E})$ denotes the maximal f-information gain*

$$C_f(\mathcal{E}) = \max_{\lambda \in Prob_n} I_f(\mathcal{E}, \lambda).$$

The following lemma implies in particular that indeed on the compact set $Prob_n$ the value $I_f(\mathcal{E}, \lambda)$ attains its maximum.

Lemma 2.13 *If $f : Prob_n \longrightarrow \mathbb{R}$ is continuous then $H_f(\mathcal{E}, \lambda)$ and $I_f(\mathcal{E}, \lambda)$ are continuous in $\lambda \in Prob_n$.*

Proof: Let $Q = \sum_{i=1}^n P_i$. We use the representation from remark 2.11

$$H_f(\mathcal{E}, \lambda) = \int \overline{f}(\lambda_1 \frac{dP_1}{dQ}(x), ..., \lambda_n \frac{dP_n}{dQ}(x))\, dQ(x).$$

In this integral the argument of $\overline{f}$ is a continuous function of λ for each x and by the special choice of Q it is uniformly bounded in each component as a function of x and λ. Since $\overline{f}$ is also continuous the continuity of $H_f(\mathcal{E}, \lambda)$ and hence also of $I_f(\mathcal{E}, \lambda)$ follows by dominated convergence. *q.e.d.*

Remark 2.14 It should be kept in mind that the prior λ_{opt} which attains $C_f(\mathcal{E})$ typically is different from the prior $\overline{\lambda}$ which maximizes f. For the entropy functional $f(z) = -\sum z_i \log z_i$ one has $\overline{\lambda} = \lambda_{unif}$ whereas this is not true for λ_{opt}. For example let $P_i = Bin(k, \frac{i}{n+1})$ for $1 \leq i \leq n$. Then one can compute numerically (e.g. by the algorithm of Arimoto-Blahut) the optimal prior and it turns out that it gives greater weight to the parameters near the boundary of $(0,1)$. For a theoretical explanation of this phenomenon see e.g. [3] and more explicitely the recent paper [11]. However a simple observation is the following

Lemma 2.15 *Assume that f attains its maximal value on $Prob_n$ at the prior $\overline{\lambda}$. Then the following estimates hold*

$$I_f(\mathcal{E}, \overline{\lambda}) \leq C_f(\mathcal{E}) \leq f(\overline{\lambda}).$$

3 Some properties of concave functions on $Prob_n$

In our main results we shall need less than concavity for the function f which defines the similarity. Nevertheless comparison with concave functions is important. The following proposition collects a couple of useful and presumably known facts about nonnegative concave functions on $Prob_n$. For the readers' convenience we give all proofs.

Proposition 3.1 *Let $f : Prob_n \longrightarrow \mathbb{R}_+$ be concave. Then*
a) the positively homogeneous extension $\overline{f}$ of f to $\mathbb{R}^n_+$ is concave and monotone for the coordinatewise ordering of $\mathbb{R}^n_+$.
b) Let $a > 0$ and $\lambda, \mu \in Prob_n$ be such that $a^{-1}\mu_i \leq \lambda_i \leq a\mu_i$ for all $i \in \{1, ..., n\}$. Then for every experiment with parameter set $\{1, \cdots, n\}$ and all observations x we have at the posterior distributions the inequality $f(\hat{\lambda}(x)) \leq a^2 f(\hat{\mu}(x))$.
c) For every $\lambda \in Prob_n$ and every experiment $\mathcal{E}$ one has $H_f(\mathcal{E}, \lambda) \leq n H_f(\mathcal{E}, \lambda_{unif})$ where $\lambda_{unif} = (\frac{1}{n}, \cdots, \frac{1}{n})$.
d) Let $b(z) = 1 - \max_{j=1}^n z_i$. If f vanishes precisely at the extreme points e_i then f satisfies $\gamma\ b \leq f$ for some positive constant γ.
e) If f can be extended to a finite concave function on the whole space $\mathbb{R}^n$ and if $f(e_i) = 0$ for all $i \in \{1, \cdots, n\}$ then there is some constant α such that $f \leq \alpha\ b$ on $Prob_n$.

Proof: a) To see the concavity of $\overline{f}$, let $\alpha \in (0,1)$ and $y, z \in \mathbb{R}^n_+$ be given. Then $y = a\lambda$ and $z = b\mu$ for some $a, b > 0$ and $\lambda, \mu \in Prob_n$. Write γ instead of $\alpha a + (1-\alpha)b$. Then

$$\begin{aligned}\overline{f}(\alpha y + (1-\alpha)z) = \gamma f(\frac{\alpha a}{\gamma}\lambda + \frac{(1-\alpha)b}{\gamma}\mu) &\geq \alpha a f(\lambda) + (1-\alpha) z f(\mu) \\ &= \alpha \overline{f}(y) + (1-\alpha)\overline{f}(z).\end{aligned}$$

Next we verify the monotonicity: Let $y, z \in \mathbb{R}^n_+$ be such that $y_i \leq z_i$ for every $i \in \{1, ..., n\}$. Then $z - y \in \mathbb{R}^n_+$ and hence

$$\overline{f}(z) = \overline{f}(\frac{2y + 2(z-y)}{2}) \geq \frac{\overline{f}(2y) + \overline{f}(2(z-y))}{2} \geq \frac{\overline{f}(2y)}{2} = \overline{f}(y).$$

b) It is easily verified that for all observations the posterior distributions satisfy the inequalities $\hat{\lambda}(x)_i \leq a^2\hat{\mu}(x)_i$ for $i \in \{1, ..., n\}$. Then the desired estimate follows from part a).
c) Let $Q = \sum P_i$. Then by the monotonicity of $\overline{f}$ and remark 2.11 we have

$$H_f(\mathcal{E}, \lambda) = \int \overline{f}(\frac{d\lambda_1 P_1}{dQ}, \cdots, \frac{d\lambda_n P_n}{dQ})dQ \leq \int \overline{f}(\frac{dP_1}{dQ}, \cdots, \frac{dP_n}{dQ})dQ = nH_f(\mathcal{E}, \lambda_{unif}).$$

d) (cf. p. 38 in [6].) For each i let $D_i = \{z \in Prob_n \ : \ z_i = \max_j z_j\}$. Then $Prob_n$ is the union of the polytopes D_i. The set D_i has only finitely many extreme points, one of them being the point e_i at which $f(e_i) = 0 = b(e_i)$. At the other extreme points f is strictly positive. Hence there is some $\gamma_i > 0$ such that $f \geq \gamma_i \ b$ on exD_i. But on D_i the function b is and hence the function $f - \gamma_i b$ is concave. Since a concave function attains its infimum at an extreme point of D_i we have $f - \gamma_i \ b \geq 0$ on D_i and thus $f \geq \gamma \ b$ on $Prob_n$ for $\gamma = \min_i \gamma_i$.
e) Let us denote the concave extension again by f. For each i there is a vector $y^i \in \mathbb{R}^n$ and a number β_i such that the affine function $g(x) = \langle y^i, x\rangle + \beta_i$ satisfies $f(x) \leq g(x)$ for all $x \in \mathbb{R}^n$ and $g(e_i) = f(e_i) = 0$. In particular $\beta_i = -\langle y^i, e_i\rangle = -y^i_i$. Thus we get for all $z \in Prob_n$

$$f(z) \leq \sum_{j=1}^{n} y^i_j z_j - y^i_i = \sum_{j=1}^{n}(y^i_j - y^i_i)z_j \leq \alpha_i \sum_{j\neq i} z_j = \alpha_i(1 - z_i)$$

where $\alpha_i = \max_{j\neq i} y^i_j - y^i_i$. Choosing $\alpha = \max_i \alpha_i$ we get $f(z) \leq \alpha \min_i(1 - z_i) = \alpha \ b(z)$. *q.e.d.*

4 Convergence

We are interested in the asymptotic behaviour for large k of the quantities $H_f(\mathcal{E}_k, \lambda)$ and $C_f(\mathcal{E}_k)$ for various f and a sequence of experiments $\mathcal{E}_k = \{P^{(k)}_1, ..., P^{(k)}_n\}$ with the same parameter set $\{1, ..., n\}$. The classical example of course is the case of product experiments $\mathcal{E}_k = \mathcal{E}^k = \{P^k_1, ..., P^k_n\}$.
We assume that the measures $P^{(k)}_i$ are almost mutually singular for large k, with the idea in mind of measuring the speed of this process with our quantities.

Assumption A For all $i \neq j$ and $\varepsilon > 0$ we have $\lim_{k\to\infty} P^{(k)}_i\{\frac{dP^{(k)}_j}{dP^{(k)}_i} > \varepsilon\} = 0$.

There are many alternative ways to express this condition. It means that asymptotically the true parameter can be estimated with arbitrarily small error probabilities. Obviously it is implied by the condition that for $i \neq j$ one has for some $s > 0$

$$\lim_{k\to\infty} E^{(k)}_i[\frac{dP^{(k)}_j}{dP^{(k)}_i}]^s = 0.$$

In the terminology of the theory of comparison of experiments the assumption A says that (the standard measure of) $\mathcal{E}_k$ converges weakly to (the standard measure $\sum_{i=1}^n \delta_{e_i}$ of) the most informative experiment $\mathcal{M}_a$ which in turn is equivalent to the fact that the minimal Bayes risk $H_b(\mathcal{E}, \lambda)$ converges to 0 (cf. e.g. [13],p. 395f). For our quantities one gets

Proposition 4.1 *Suppose that A holds. Let $f : Prob_n \to \mathbb{R}_+$ be continuous with $f(e_i) = 0$ for all $i \in \{1, \ldots, n\}$. Then $\lim_{k\to\infty} H_f(\mathcal{E}_k, \lambda) = 0$ uniformly in $\lambda \in Prob_n$ and $\lim_{k\to\infty} C_f(\mathcal{E}_k) = M_f$ where $M_f = \max_{\lambda \in Prob_n} f(\lambda)$.*

Proof: Rather than deducing this from the abstract theory of standard measures of experiments quoted above we give a direct argument. Note that under assumption A for any $i \in \{1, ..., n\}$ the posterior distributions converge under the $i-$th hypothesis to the vector e_i in the following sense: For every $\delta > 0$ and every $\eta > 0$ there is some k_0 such that for $k \geq k_0$ we have

$$P_i^{(k)}\{||\hat{\lambda} - e_i|| > \delta\} < \eta$$

uniformly in the set $\{\lambda \in Prob_n : \lambda_i \geq \eta\}$. Given $\varepsilon > 0$ choose $\eta = \frac{\varepsilon}{4nM_f}$ and $\delta > 0$ such that $f(z) \leq \varepsilon/2$ whenever $\| z - e_i \| < \delta$ for some $i \in \{1, \cdots, n\}$. Choose k_0 as above. Then we get for all $\lambda \in Prob_n$ and $k \geq k_0$ the estimate

$$P_\lambda^{(k)}\{f(\hat{\lambda}) > \varepsilon/2\} \leq \sum_{i:\lambda_i<\eta} \lambda_i + \sum_{i:\lambda_i\geq\eta} \lambda_i P_i^{(k)}\{||\hat{\lambda} - e_i|| > \delta\} \leq 2n\eta$$

and hence

$$H_f(\mathcal{E}_k, \lambda) = \int f(\hat{\lambda})\, dP_\lambda^{(k)} \leq \varepsilon/2 + \int_{\{f(\hat{\lambda})>\varepsilon/2\}} M_f \; dP_\lambda^{(k)} < \varepsilon$$

which proves the first assertion. For the second part let f attain its maximal value M_f at $\overline{\lambda}$. Then we have by lemma 2.15

$$0 \leq M_f - C_f(\mathcal{E}_k) \leq f(\overline{\lambda}) - I_f(\mathcal{E}_k, \overline{\lambda}) = H_f(\mathcal{E}_k, \overline{\lambda}) \stackrel{k\to\infty}{\longrightarrow} 0.$$

q.e.d.

5 Comparing exponential rates

In many situations one can expect that the convergence in proposition 4.1 is exponentially fast. We want to show that in a wide class of functions f actually the exponential rate of convergence does not depend on the particular choice of f. For this consider the following definition. It is concerned with the comparison of the small values of two bounded nonnegative functions f, g.

Definition 5.1 *Let f and g be two nonnegative bounded real functions on a set Z. We say that f is* **exponentially dominated** *by g if for each $\varepsilon \in (0,1)$ there is a constant $C(\varepsilon) < \infty$ such that*

$$f(z) \leq C(\varepsilon) g(z)^{1-\varepsilon}$$

for all $z \in Z$. If f and g are exponentially dominated by each other we call f and g **exponentially equivalent**.

Here is an alternative description of this concept.

Remark 5.2 *Let f, g be bounded nonnegative functions. Then f is exponentially dominated by g iff $\{g = 0\} \subset \{f = 0\}$ and*

$$\liminf_{g(z) \to 0} \frac{\log f(z)}{\log g(z)} \geq 1.$$

For us the most interesting examples are given by the following lemma.

Lemma 5.3 *For every $n \in \mathbb{N}$ the the entropy functional $f(z) = \sum_{i=1}^{n} -z_i \log z_i$ is exponentially equivalent on $Prob_n$ to the function $b(z) = 1 - \max_{j=1}^{n} z_j$. The same is true for every function f on $Prob_n$ which has a finite concave extension to $\mathbb{R}^n$ and which vanishes precisely at the extreme points e_i.*

The verification of the first part is straightforward calculus. The second statement follows from proposition 3.1, parts d) and e). In example 2.7 it is shown that the first part is not a special case of the second part. In the case $n = 2$ the functions f given in example 2.8 are exponentially equivalent to b even for $r > 1$. However, it is easily seen that the functions f_α from example 2.6 are not exponentially dominated by b whereas f_α dominates b exponentially by proposition 3.1 d).
Now we come to our main results.

Theorem 5.4 *Let f and g be nonnegative bounded functions on $Prob_n$ and let f be exponentially dominated by g. Let $(\mathcal{E}_k)_{k \in \mathbb{N}}$ be any sequence of experiments and let λ be a prior. Then*
a) We have

$$\limsup_{k \to \infty} \left(\sqrt[k]{H_f(\mathcal{E}_k, \lambda)} - \sqrt[k]{H_g(\mathcal{E}_k, \lambda)} \right) \leq 0. \tag{2}$$

b) If either f or g is exponentially equivalent to a concave function then for every number $a > 0$ the relation

$$\limsup_{k \to \infty} \left(\sqrt[k]{H_f(\mathcal{E}_k, \mu)} - \sqrt[k]{H_g(\mathcal{E}_k, \lambda)} \right) \leq 0 \tag{3}$$

holds uniformly in the set of all $\mu \in Prob_n$ which satisfy $a^{-1}\lambda \leq \mu \leq a\lambda$ in the coordinatewise ordering.

c) In particular, if $\lim_{k\to\infty} \sqrt[k]{H_f(\mathcal{E}_k,\lambda)}$ *exists then this limit still exists with the same value if we replace f by an exponentially equivalent function g. If f is exponentially equivalent to a concave function then the limit has the same value for all strictly positive priors.*

Proof: a) Jensen's inequality for the concave function $x \mapsto x^{1-\varepsilon}$ gives for every experiment $\mathcal{E}$

$$\begin{aligned} H_f(\mathcal{E},\lambda) &= \int f(\hat{\lambda})dP_\lambda \leq \int C(\varepsilon)g(\hat{\lambda})^{1-\varepsilon}dP_\lambda \\ \leq \quad & C(\varepsilon)\,(\int g(\hat{\lambda})dP_\lambda)^{1-\varepsilon} = C(\varepsilon)\ H_g(\mathcal{E},\lambda)^{1-\varepsilon}. \end{aligned}$$

Given the sequence $(\mathcal{E}_k)_{k\in\mathbb{N}}$ write $H_k = H_g(\mathcal{E}_k,\lambda)$ and $H'_k = H_f(\mathcal{E}_k,\lambda)$. We want to show that $\limsup_k(\sqrt[k]{H'_k} - \sqrt[k]{H_k}) \leq 0$. Since the sequences (H_k) and (H'_k) are bounded, we may pass to a suitable subsequence and assume that $H = \lim_k \sqrt[k]{H_k}$ and $H' = \lim_k \sqrt[k]{H'_k}$ exist. Since the k-th roots of the constant converge to 1 we get the estimate $H' \leq H^{1-\varepsilon}$ for every $\varepsilon \in (0,1)$ and hence $H' \leq H$. This proves part a).
b) If f is exponentially equivalent to a concave function h then we have by proposition 3.1 b) for all experiments and uniformly in $\{\mu \in Prob_n : a^{-1}\lambda \leq \mu \leq a\lambda\ \}$ the estimate

$$H_h(\mathcal{E},\mu) = \int h(\hat{\mu})dP_\mu \leq a\int h(\hat{\mu})dP_\lambda \leq a^3\int h(\hat{\lambda})dP_\lambda = a^3 H_h(\mathcal{E},\lambda)$$

and also by exponential equivalence for every $\varepsilon > 0$

$$H_f(\mathcal{E},\mu) \leq C_1(\varepsilon)H_h(\mathcal{E},\mu)^{1-\varepsilon}.$$

These two estimates give as in the proof of a) the inequality (3) with h instead of g, uniformly in μ. Since h is also exponentially dominated by g one then can apply a) to replace h by g. If g rather than f is exponentially dominated by a convex function one argues similarly.
c) is indeed a direct consequence of a) and b). *q.e.d.*

The next theorem shows that the rates given by the previous result also apply to the capacity under the assumption A. In order to motivate it let us recall the remark 2.14 that the prior which attains $C_f(\mathcal{E})$ is in general not close to the prior $\overline{\lambda}$ which maximizes f.

Theorem 5.5 *Let the continuous nonnegative function f on $Prob_n$ be exponentially equivalent to a concave function. Suppose that f vanishes at e_i for all $i \in \{1,...,n\}$ and that f does not attain its maximum $M_f = \max_{z\in Prob_n} f(z)$ on the geometric boundary of the simplex $Prob_n$. In addition assume that the sequence $(\mathcal{E}_k)$ satisfies assumption A. Then*

$$\lim_{k\to\infty}\left(\sqrt[k]{M_f - C_f(\mathcal{E}_k)} - \sqrt[k]{H_f(\mathcal{E}_k,\lambda)}\right) = 0$$

for every strictly positive prior λ.

Proof: Let $\overline{\lambda}$ be a point at which f attains its maximum M_f. Let $\lambda^{(k)}$ be the optimal prior at which the transmission rate $I_f(\mathcal{E}_k, \lambda)$ attains its maximum, i.e. $f(\lambda^{(k)}) - H_f(\mathcal{E}_k, \lambda^{(k)}) = C_f(\mathcal{E}_k)$. We can choose $\eta > 0$ according to the assumption on f such that every prior λ with $f(\lambda) > M_f - \eta$ satisfies $\lambda_i > \eta$ for all i. Since the sequence $(\mathcal{E}_k)$ fulfils assumption A we can choose by proposition 4.1 an index k_0 such that $H_f(\mathcal{E}_k, \overline{\lambda}) < \eta$ for all $k \geq k_0$. Therefore for $k \geq k_0$,

$$f(\lambda^{(k)}) \geq f(\lambda^{(k)}) - H_f(\mathcal{E}_k, \lambda^{(k)}) \geq f(\overline{\lambda}) - H_f(\mathcal{E}_k, \overline{\lambda}) \geq M_f - \eta$$

and hence, by the choice of η, $\lambda_i^{(k)} > \eta$ for all i. Similarly $\overline{\lambda}_i > \eta$ for all i. Letting $a = \eta^{-1}$ this gives $a^{-1}\overline{\lambda} \leq \lambda^{(k)} \leq a\overline{\lambda}$ for all $k \geq k_0$. Since f is exponentially equivalent to itself, we have by the preceding theorem

$$\lim_{k \to \infty} \left(\sqrt[k]{H_f(\mathcal{E}_k, \lambda^{(k)})} - \sqrt[k]{H_f(\mathcal{E}_k, \overline{\lambda})} \right) = 0. \tag{4}$$

In this equation $H_f(\mathcal{E}_k, \lambda^{(k)})$ can be replaced by $M_f - C_f(\mathcal{E}_k)$ since

$$\begin{aligned} H_f(\mathcal{E}_k, \lambda^{(k)}) \leq \quad & f(\overline{\lambda}) - f(\lambda^{(k)}) + H_f(\mathcal{E}_k, \lambda^{(k)}) \\ = \quad & M_f - C_f(\mathcal{E}_k) \leq M_f - f(\overline{\lambda}) + H_f(\mathcal{E}_k, \overline{\lambda}) \\ = \quad & H_f(\mathcal{E}_k, \overline{\lambda}). \end{aligned}$$

Finally again by the preceding theorem we may pass in (4) from the prior $\overline{\lambda}$ to any other strictly positive prior λ. This completes the proof. *q.e.d.*

In [6], theorem 3.37 a similar result is shown for functionals which are allowed to attain their maximum on the boundary of the simplex. However in that result the sequence of experiments is assumed to be of the product type.

The following result shows in particular that the rate of convergence of the Shannon capacity is the same as the rate of the deficiency $\delta(\mathcal{E}, \mathcal{M}_a)$ to the most informative experiment. It extends to general experiments the fact which is known for product experiments that this rate is determined by the worst pair of parameters. By lemma 5.3 it even extends to general experiments the result of [12] that the deficiency rate equals the rate of the difference (1) for sublinear (resp. superlinear) functionals.

Corollary 5.6 *Let f be the entropy functional or any other exponentially equivalent function on $Prob_n$. Let the sequence $(\mathcal{E}_k)_{k\in\mathbb{N}}$ of experiments satisfy assumption A. Let $\lambda \in Prob_n$ be a strictly positive prior. Consider for each k the following five numbers:*

$$\sqrt[k]{M_f - C_f(\mathcal{E}_k)}; \;\; \sqrt[k]{\delta(\mathcal{E}_k, \mathcal{M}_a)}; \;\; \sqrt[k]{H_f(\mathcal{E}_k, \lambda)}; \;\; \sqrt[k]{H_b(\mathcal{E}_k, \lambda)};$$
$$\max_{1 \leq i < j \leq n} \sqrt[k]{2 - \| P_i^{(k)} - P_j^{(k)} \|_1}.$$

If one of these expressions converges as $k \to \infty$ then the others converge as well and all have the same limit.

Proof: By the result mentioned in example 2.5 and lemma 3.1c) one has

$$2H_b(\mathcal{E}, \lambda_{unif}) \leq 2\max_{\lambda} H_b(\mathcal{E}, \lambda) = \delta(\mathcal{E}, \mathcal{M}_a) \leq nH_b(\mathcal{E}, \lambda_{unif}).$$

Since by the above theorems the rate does not depend on the choice of the prior the second and the fourth expression have the same rate. Since f and b are exponentially equivalent according to lemma 5.3 we can include the third and then by the preceding result also the first expression.
In order to include the last expression we introduce the function

$$s(z) = \sum_{1\leq i<j\leq n} \min(z_i, z_j)$$

on $Prob_n$. This sum defines a concave function on $\mathbb{R}^n$ and the extreme points e_i are precisely the points in $Prob_n$ at which all terms in the sum vanish. So the function s is, according to lemma 5.3, exponentially equivalent to b. Thus in the following sequence of quantities each has the same exponential rate as its neighbours. The symbol $\mathcal{E}_k^{ij}$ denotes the two parameter experiment ('dichotomy') $\{P_i^{(k)}, P_j^{(k)}\}$.

$$\begin{aligned}
H_b(\mathcal{E}, \lambda_{unif});\\
H_s(\mathcal{E}, \lambda_{unif});\\
\int \sum_{1\leq i<j\leq n} \min(\frac{dP_i^{(k)}}{dQ}, \frac{dP_j^{(k)}}{dQ})\ dQ;\\
\sum_{1\leq i<j\leq n} H_b(\mathcal{E}_k^{ij}, (\frac{1}{2}, \frac{1}{2}));\\
\max_{1\leq i<j\leq n} H_b(\mathcal{E}_k^{ij}, (\frac{1}{2}, \frac{1}{2}));
\end{aligned}$$

By the representation given in example 2.5 this completes the proof. *q.e.d.*

Combining this result with either [12], theorem 4.2 (see also [14]) or directly the classical result of Chernoff [2] we get the following explicit result. As mentioned before in [10] a similar result is proved for Markov chains.

Corollary 5.7 *Let $\mathcal{E} = \{P_1, ..., P_n\}$ be a fixed experiment. Then the Shannon capacities $C_f(\mathcal{E}^k)$ of the k-fold product $\mathcal{E}^k = \{P_1^k, ..., P_n^k\}$ satisfy*

$$\lim_{k\to\infty} \sqrt[k]{\log n - C_f(\mathcal{E}^k)} = \max_{i\neq j} \inf_{0<t<1} \int dP_i^t dP_j^{1-t}.$$

Example. At least in the following numerical example for moderate size of k the numbers $\sqrt[k]{H_f(\mathcal{E}^k,\lambda)}$ and $\sqrt[k]{H_b(\mathcal{E}^k,\lambda)}$ are actually much nearer to each other than to the limit. So it would be interesting to prove that the deficiency and the Shannon capacity are close even in the sense of a more refined asymptotic analysis. In this example $\mathcal{E}$ is the Bernoulli experiment with the three parameters 0.3, 0.5, 0.7, the prior λ is (0.3, 0.3, 0.4) and the limit rate equals approximatively 0.9789.

k	10	50	100	500	1000
$\sqrt[k]{H_f(\mathcal{E}^k,\lambda)}$	0.8229	0.8428	0.8444	0.9122	0.954536
$\sqrt[k]{H_b(\mathcal{E}^k,\lambda)}$	0.7295	0.8113	0.8239	0.9122	0.954536

Clearly in the continuous paramater situation the capacity converges to infinity but with much slower speed because of the overlap of parameter which are very close to each other. This question requires more subtle arguments, see [3],[11].

Acknowledgement. As already mentioned we owe much to the helpful discussions with Erik Torgersen. Also we are indebted to the referees for a couple of clarifying comments.

References

[1] J.M. Bernardo. Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc., Ser. B*, 41:113–147, 1979.

[2] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[3] B.S. Clarke and A.R. Barron. Jeffreys prior is asymptotically least favorable under entropy risk. *J. of Stat. Planning and Inference*, 41:37–60, 1994.

[4] I. Csiszár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[5] L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Ann. Inst. Statist. Math.*, 30, Part A:105–113, 1978.

[6] J. Krob. *Kapazität statistischer Experimente.* PhD thesis, Univ. Kaiserslautern, 1992.

[7] L. LeCam. *Asymptotic Methods in Statistical Decision Theory.* Springer, Berlin - Heidelberg - New York, 1986.

[8] L. LeCam and G.L. Yang. *Asymptotics in Statistics, Some Basic Concepts.* Springer Series in Statistics, Berlin - Heidelberg - New York, 1990.

[9] D.V. Lindley. On a measure of the information provided by an experiment. *Ann. Math. Stat.*, 27:986–1005, 1956.

[10] P. Scheffel and H.v. Weizsäcker. Risk rates in finite parameter Markov chain experiments. *Preprint*, 1995. University of Kaiserslautern.

[11] H.R. Scholl. Shannon optimal priors on iid experiments converge weakly to jeffreys' prior. *Preprint, University of Kaiserslautern*, 1996.

[12] E.N. Torgersen. Measures of information based on comparison with total information and with total ignorance. *Ann. of Statistics*, 9:638–657, 1981.

[13] E.N. Torgersen. *Comparison of Statistical Experiments.* Cambridge University Press, Cambridge - New York, 1991.

[14] I. Vajda. On the amount of information contained in a sequence of independent obsevations. *Kybernetica*, 6:306–324, 1970.

Jürgen Krob
Bertelsmann Club GmbH
Postfach 1109
33339 Rheda-Wiedenbrüeck
Germany
e-mail: Kro16@deziv011.bertelsmann.de

Heinrich v. Weizsäcker
Fachbereich Mathematik der Universität
D 67663 Kaiserslautern
Germany
e-mail: weizsaecker@mathematik.uni-kl.de