KYBERNETIKA — VOLUME 15 (1979), NUMBER 1

Information Theoretical Optimization Techniques

FLEMMING TOPSØE

It is the object of this paper to show that a game theoretical viewpoint may be taken to underlie the maximum entropy principle as well as the minimum discrimination information principle, two principles of well known significance in theoretical statistics and in statistical thermodynamics. Our setting is very simple and certainly calls for future expansion.

INTRODUCTION OF ABSOLUTE AND RELATIVE GAMES

Let $(I, \mathscr{I})$ be a measurable space, referred to as the *state space*. The state space is *discrete* if I is countable and $\mathscr{I}$ consists of all subsets of I. By Π we denote the set of countable decompositions of I in measurable sets, directed in the usual way by refinement. Π_0 denotes the subset of Π consisting of finite decompositions.

By a *distribution* we mean a probability measure on $(I, \mathscr{I})$. The set of all distributions is denoted by M. For $\mu \in M$ and $\pi \in \Pi$, $\mu \mid \pi$ denotes the restriction of μ to the σ-algebra generated by π. For a point function φ on I, $\langle \varphi, \mu \rangle$ denotes the expectation of φ w.r.t. μ. For a set function φ on $\mathscr{I}$, $\langle \varphi, \mu \rangle_\pi$ denotes expectation of φ w.r.t. $\mu \mid \pi$, i.e.

$$\langle \varphi, \mu \rangle_\pi = \sum_{A \in \pi} \varphi(A)\, \mu(A)\,.$$

We employ the usual conventions regarding arithmetic involving $\pm\infty$, eg. $0 \,.\, a = 0$ if a is an extended real number or if a is some udefined quantity.

We need a very primitive concept of a code. To us, a *code* is a set function $\kappa : \mathscr{I} \to [0, \infty]$ such that

$$\sum_{A \in \pi} \mathrm{e}^{-\kappa(A)} = 1 \quad \text{for all} \quad \pi \in \Pi\,. \tag{1}$$

If the reader fixes his attention to a specific $\pi \in \Pi$, if he changes the base e to 2 and if he assumes that κ is integer valued, he will realize that (1) is transformed into the

well known Kraft inequality concerning binary codes; the reason why we insist on equality in (1) is because we only have codes without superfluous digits in mind. We see that our definition of a code does not reflect the structure of individual codewords, only the lengths of the codewords. The fact that we allow κ to take arbitrary real values corresponds to a convenient mathematical idealization and can to some extent be justified by the noiseless coding theorem.

We denote by K the set of all codes. There is a natural bijection between K and M by which $\kappa \in K$ and $\eta \in M$ correspond to each other by the formulas

$$\kappa(A) = \log\left(1/\eta(A)\right), \quad \eta(A) = \mathrm{e}^{-\kappa(A)}; \quad A \in \mathscr{I}. \tag{2}$$

We write $\kappa \leftrightarrow \eta$ and call κ the code *adapted* to η when (2) holds. Note that when $\kappa \leftrightarrow \eta$, then $\kappa(A) = \infty$ and $\eta(A) = 0$ are equivalent statements.

If (A_n) are pairwise disjoint measurable sets, and if $\kappa \in K$, then

$$\kappa(\bigcup A_n) = -\log\left(\sum \mathrm{e}^{-\kappa(A_n)}\right).$$

All our results concern the study of subsets of M. Consider one such subset, say C. We think of C as the set of "consistent" distribution, in more suggestive terms, C consists of those distributions which are consistent with the "preparation" of the "physical system" under study.

With C we shall associate various games. All games will be two-person-zero-sum games. Basicly, the idea is as follows. We think of the two participants as the "observer" of the physical system and as "nature". The observer chooses, or "plays", a code and nature chooses a distribution. Nature is bound to choose a consistent distribution whereas there are no limitations to the codes the observer can play. The cost of the game, seen from the point of view of the observer, is measured by the average codelength. We assume that nature acts in a way which is least favourable to the observer (hence we are led to zero-sum games).

We may consider it the objective of the observer to transmit the results of (independent) observations of the system. The average codeword length then really measures the cost of transmission per observation. Dually, one could think of the code as a device enabling the observer to carry out the observation.

Now, let us be more concrete. Assume that $(I, \mathscr{I})$ is discrete. Then, with each subset $C \subseteq M$ we associate a game denoted by the same letter C and called the *absolute game* associated with C. The game C is defined by the cost function

$$(\kappa, \mu) \to \langle \kappa, \mu \rangle; \quad (\kappa, \mu) \in K \times C.$$

This function is well defined since $(I, \mathscr{I})$ is discrete. The range of the cost function is $[0, \infty]$. Put

$$\alpha = \sup_{\mu \in C} \inf_{\kappa \in K} \langle \kappa, \mu \rangle,$$

$$\beta = \inf_{\kappa \in K} \sup_{\mu \in C} \langle \kappa, \mu \rangle .$$

We call μ an *optimal strategy for nature* if $\mu \in C$ and if

$$\inf_{\kappa \in K} \langle \kappa, \mu \rangle = \alpha . \tag{3}$$

And we call κ an *optimal strategy for the observer* if $\kappa \in K$ and if

$$\sup_{\mu \in C} \langle \kappa, \mu \rangle = \beta . \tag{4}$$

Clearly, $\alpha \leqq \beta$. If $\alpha = \beta$, we call this number the *value* of the game. A strategy $\kappa \in K$ for the observer is called a *cost-stable* strategy if $\langle \kappa, \mu \rangle$ is finite and independent of μ for $\mu \in C$.

Then consider an arbitrary state space $(I, \mathscr{I})$. If we try to define the absolute game associated with a subset $C \subseteq M$, we run into difficulties. Firstly, $\langle \kappa, \mu \rangle$ is not well defined. This is a minor difficulty since one could reasonably consider the limit of $\langle \kappa, \mu \rangle_\pi$ as π ranges over Π. But doing so, one sees that unless μ is discrete one will obtain the value ∞. The conclusion is that even though one could consider the absolute game, this game will be of no interest since the observer can not discriminate between the possible moves by means of the average codelength (analogous remark applies to nature).

The way out of the difficulty is to measure the performance of the observer relative to some fixed code. Thus we shall consider improvements in average codelength rather than average codelength itself. As there is a natural one-to-one correspondence between K and M, the fixation of a code implies the fixation of a distribution.

With the above remarks in mind, consider a fixed pair κ, η where $\kappa \leftrightarrow \eta$. We call κ the *reference code* and η the *reference distribution*. All codes will be related to κ. Therefore, we define a *codeimprovement* Δ (relative to κ) as a set function $\Delta : \mathscr{I} \to \to [-\infty, \infty)$ such that the set function κ_Δ defined by

$$\kappa_\Delta(A) = \kappa(A) - \Delta(A) ; \quad A \in \mathscr{I} \tag{5}$$

is a code. Note that the value of $\Delta(A)$ for sets with $\kappa(A) = \infty$ has no influence on the value of $\kappa_\Delta(A)$. Accordingly, two codeimprovements are considered to be identical if they only differ on sets with infinite codelength for κ, i.e. if they only differ on sets of η-measure 0.

The measure μ obtained from κ_Δ by the one-to-one correspondence between K and M is given by

$$\mu(A) = \mathrm{e}^{\Delta(A)} \eta(A) ; \quad A \in \mathscr{I} . \tag{6}$$

Note that μ is absolutely continuous with respect to η and that, conversely, any probability measure μ which is absolutely continuous with respect to η induces

a codeimprovement by means of the formula

$$\Delta(A) = \log\left[\mu(A)/\eta(A)\right]; \quad A \in \mathscr{S}, \quad \eta(A) > 0. \tag{7}$$

The value of $\Delta(A)$ for sets with $\eta(A) = 0$ may be defined arbitrarily in $[-\infty, \infty)$.

The set of distributions absolutely continuous with respect to η is denoted by M_η and the set of codeimprovements relative to κ (where $\kappa \leftrightarrow \eta$) is denoted by K_η. We have seen that (6) and (7) define a one-to-one correspondance between M_η and K_η.

Let $\Delta \in K_\eta$. Then Δ is a set function. But we may also consider Δ as a point function. As no confusion seems likely to arise, we shall use the same letter for both functions. The point function $\Delta : X \to [-\infty, \infty)$ is defined by

$$\Delta = \log \frac{\mathrm{d}\mu}{\mathrm{d}\eta},$$

where μ is the measure given by (6) and where $\mathrm{d}\mu/\mathrm{d}\eta$ denotes a finite and non-negative valued version of the Radon-Nikodym derivative of μ with respect to η. We can recapture the set function Δ from the point function Δ by the formula

$$\Delta(A) = \log\left(\frac{1}{\eta(A)}\int_A \mathrm{e}^{\Delta}\,\mathrm{d}\eta\right); \quad \eta(A) > 0. \tag{8}$$

The set K_η may thus be identified with the set of measurable functions $\Delta : I \to [-\infty, \infty)$ for which $\int \mathrm{e}^{\Delta}\,\mathrm{d}\eta = 1$ with the understanding that functions which only differ on an η-null set are considered as the same function. In terms of this identification, the correspondance between K_η and M_η is given by the formulas

$$\frac{\mathrm{d}\mu}{\mathrm{d}\eta} = \mathrm{e}^{\Delta}, \quad \Delta = \log \frac{\mathrm{d}\mu}{\mathrm{d}\eta}. \tag{9}$$

The commutative diagram

$$\begin{array}{ccc} K & \xleftarrow{\quad (2) \quad} & M \\ (5)\ \uparrow & & \uparrow\ \mathrm{id} \\ K_\eta & \xleftarrow[\text{or (9)}]{\text{(6) and (7)}} & M_\eta \end{array}$$

summarizes our discussion.

Lemma 1. Let $\Delta \in K_\eta$ and let $\mu \in M_\eta$. Then

$$\lim_{\pi \in \Pi} \langle \Delta, \mu \rangle_\pi = \langle \Delta, \mu \rangle \tag{10}$$

in the sense that when one of the sides exists as an extended real number, then so does the other and equality holds.

Notice that Δ on the left hand side of (10) refers to the set function Δ whereas the Δ on the right hand side refers to the point function Δ. Clearly. we have to assume $\mu \in M_\eta$ – otherwise neither the left hand nor the right hand member of (10) could determine well defined numbers.

Proof. With a "+" disignating "positive part", and a "−" designating "negative part" we claim that

$$\lim_{\pi \in \Pi} \langle \Delta^+, \mu \rangle_\pi = \langle \Delta^+, \mu \rangle \,, \tag{11}$$

$$\lim_{\pi \in \Pi} \langle \Delta^-, \mu \rangle = \langle \Delta^-, \mu \rangle \,. \tag{12}$$

Once this is proved, the result follows. We only prove (11) as (12) may be proved analogously.

Let $s > 1$. For $i \in \mathbf{Z}$ put

$$A_i = \{s^i \leqq \Delta < s^{i+1}\}$$

and let π denote the decomposition of X into the sets A_i and the set $\{\Delta \leqq 0\}$.

For every $i \in \mathbf{Z}$, every $x \in A_i$ and every measurable subset B of A_i with $\eta(B) > 0$ we have

$$s^{-1} \,.\, \Delta(x) \leqq \Delta(B) \leqq s \,.\, \Delta(x) \,.$$

Let $\sigma \geqq \pi$. Consider a fixed A_i and let B_{ij} denote the sets in σ contained in A_i and of positive η-measure. Then

$$\left| \int_{A_i} \Delta \, \mathrm{d}\mu - \sum_j \mu(B_{ij}) \,.\, \Delta(B_{ij}) \right| < (s-1) \int_{A_i} \Delta \, \mathrm{d}\mu \,.$$

In case $\int \Delta^+ \, \mathrm{d}\mu < \infty$, it follows that

$$\left| \int \Delta^+ \, \mathrm{d}\mu - \sum_{B \in \sigma} \mu(B) \,.\, \Delta^+(B) \right| \leqq (s-1) \int \Delta^+ \, \mathrm{d}\mu$$

so that (11) holds. If $\int \Delta^+ \, \mathrm{d}\mu = \infty$, it follows by putting $s = 3/2$ that

$$\sum_j \mu(B_{ij}) \,.\, \Delta(B_{ij}) \geqq \frac{1}{2} \int_{A_i} \Delta \, \mathrm{d}\mu \quad \text{for all } i$$

so that

$$\sum_{B \in \sigma} \mu(B) \,.\, \Delta^+(B) = \infty \,,$$

hence (11) also holds in this case. □

It will follow from a later result that, under a slight restriction (which is perhaps

entirely superfluous), we also have

$$\lim_{\pi \in \Pi_0} \langle \Delta, \mu \rangle_\pi = \langle \Delta, \mu \rangle \,. \tag{13}$$

This will follow in a rather roundabout way. We would have preferred to be able to give a direct proof.

Justified by Lemma 1, $\langle \Delta, \mu \rangle$ is to be interpreted as the *average codeimprovement.*

Let $\eta \in M$ and let C be a subset of M_η. With C we associate a game denoted by $C \parallel \eta$ and called the *relative game* associated with C. The game $C \parallel \eta$ is defined by the payoff function

$$(\Delta, \mu) \to \langle \Delta, \mu \rangle \,; \quad (\Delta, \mu) \in K_\eta \times C \,.$$

The range of the payoff function is $[-\infty, \infty]$ supplied with an extra element "not defined". Put

$$\alpha = \inf_{\mu \in C} \sup_{\Delta \in K_\eta} \langle \Delta, \mu \rangle \,,$$

$$\beta = \sup_{\Delta \in K_\eta} \inf_{\mu \in C} \langle \Delta, \mu \rangle \,.$$

Here it is understood that only well defined values of $\langle \Delta, \mu \rangle$ should be considered.

We call μ an *optimal strategy for nature* if $\mu \in C$ and if

$$\sup_{\Delta \in K_\eta} \langle \Delta, \mu \rangle = \alpha \,.$$

And we call Δ an *optimal strategy for the observer* if $\Delta \in K_\eta$ and if

$$\inf_{\mu \in C} \langle \Delta, \mu \rangle = \beta \,.$$

Clearly, $\alpha \geqq \beta$. If $\alpha = \beta$, we call this number the *value* of the game. A strategy $\Delta \in K_\eta$ is a *payoff-stable* strategy if $\langle \Delta, \mu \rangle$ is finite and independant of μ for $\mu \in C$.

THE ABSOLUTE GAME

In this section we assume that $(I, \mathcal{S})$ is discrete. The following trivial identity plays a key role for the discussion of the absolute game.

Lemma 2. Let $\eta \in M$ and denote by κ the code adapted to η. Then, for any $\mu \in M$,

$$\langle \kappa, \mu \rangle = D(\mu \parallel \eta) + H(\mu) \,.$$

Here, $D(. \parallel .)$ denotes information gain, cf. e.g. [10], [12] or [4], and $H(.)$ denotes entropy. The properties of these functions which we shall need, can all be found in the sources just mentioned. The letter "D" instead of the familiar "I" is suggested

by Csiszár and Körner in a forthcoming book. D may be taken to stand for "distance", "divergence" or "discrimination".

Lemma 2 implies:

Lemma 3. For any $\mu \in M$,

$$\min_{\kappa \in K} \langle \kappa, \mu \rangle = H(\mu)$$

and the minimum is achieved for the code adapted to μ. If $H(\mu) < \infty$, the minimum is not achieved for any other code.

Let $C \subseteq M$ and put

$$H_{\max}(C) = \sup_{\mu \in C} H(\mu) .$$

Lemma 3 shows that an optimal strategy for nature in the absolute game C is the same as a distribution μ with $\mu \in C$ and $H(\mu) = H_{\max}(C)$. Such a distribution is also called a *canonical distribution* for C. For discussions of the "Maximum Entropy Principle", which we are thus led to, see [8], [7], [14] and [2] just to mention some references.

Another useful consequence of Lemma 2 is:

Lemma 4. Let $\mu_1, \mu_2, \ldots, \mu_n$ be distributions and let $(p_1, p_2, \ldots, p_n)$ be a probability vector. Then

$$H(\textstyle\sum p_\nu \mu_\nu) = \sum p_\nu H(p_\nu) + \sum p_\nu D(\mu_\nu \,\|\, \sum p_j \mu_j) .$$

Proof. Put $\eta = \sum p_\nu \mu_\nu$. Denote by κ the code adapted to η. Then

$$H(\eta) = \langle \kappa, \eta \rangle = \langle \kappa, \textstyle\sum p_\nu \mu_\nu \rangle = \sum p_\nu \langle \kappa, \mu_\nu \rangle = \sum p_\nu [D(\mu_\nu \,\|\, \eta) + H(\mu_\nu)] =$$

$$= \textstyle\sum p_\nu D(\mu_\nu \,\|\, \eta) + \sum p_\nu H(\mu_\nu) . \qquad \square$$

We shall now search for canonical distributions. Topological considerations come into the picture. Note that as $(I, \mathcal{S})$ is discrete, there is only one sensible topology on M. Besides the natural condition $H_{\max}(C) < \infty$, we shall also assume that C is convex. We remark that in view of the basic applications, we do not wish to assume that C is closed.

Theorem 1. Assume that $C \subseteq M$ is convex and that $H_{\max}(C) < \infty$. Then there exists a unique distribution μ_C such that $\mu_n \to \mu_C$ for every sequence $(\mu_n) \subseteq C$ such that $H(\mu_n) \to H_{\max}(C)$.

In the terminology of convex analysis, the theorem says that the entropy function strongly exposes every convex set C with $H_{\max}(C) < \infty$.

Proof. Recall Pinsker's inequality $D(\mu \parallel \eta) \geqq \frac{1}{4} \parallel \mu - \eta \parallel^2$ where $\parallel . \parallel$ denotes total variation. From Lemma 4 we then get, for every n and m:

$$H_{\max}(C) \geqq H(\tfrac{1}{2}\mu_n + \tfrac{1}{2}\mu_m) = \tfrac{1}{2} H(\mu_n) + \tfrac{1}{2} H(\mu_m) +$$
$$+ \tfrac{1}{2} D(\mu_n \parallel \tfrac{1}{2}\mu_n + \tfrac{1}{2}\mu_m) + \tfrac{1}{2} D(\mu_m \parallel \tfrac{1}{2}\mu_n + \tfrac{1}{2}\mu_m) \geqq$$
$$\geqq \tfrac{1}{2} H(\mu_n) + \tfrac{1}{2} H(\mu_m) + \tfrac{1}{16}\|\mu_n - \mu_m\|^2 .$$

It follows that $\|\mu_n - \mu_m\| \to 0$ for $n, m \to \infty$. Hence there exists $\mu_C \in M$ such that $\mu_n \to \mu_C$. It is easy to see that μ_C is independent of the particular sequence (μ_n). □

The distribution μ_C of Theorem 1, we call the *center of attraction* for C. Clearly, if the canonical distribution exists, it must be the center of attraction. But μ_C may fail to be the canonical distribution since neither $\mu_C \in C$ nor $H(\mu_C) = H_{\max}(C)$ need hold in general – we can only assert that $\mu_C \in \bar{C}$, the closure of C and, by lower semi-continuity of H, that $H(\mu_C) \leqq H_{\max}(C)$.

It is an important observation that the trivial inequality $H(\mu) \leqq H_{\max}(C)$ for $\mu \in C$ can be strengthened.

Theorem 2. Assume, that C is convex and that $H_{\max}(C) < \infty$. Then, for every $\mu \in C$, we have

$$H(\mu) + D(\mu \parallel \mu_C) \leqq H_{\max}(C) .$$

Proof. Choose $(\mu_n) \subseteq C$ such that

$$n[H_{\max}(C) - H(\mu_n)] \to 0 .$$

With the given $\mu \in C$ we associate the distributions

$$\mu_n^* = \left(1 - \frac{1}{n}\right)\mu_n + \frac{1}{n}\mu; \; n \geqq 1 .$$

Then $\mu_n^* \in C$, hence $H(\mu_n^*) \leqq H_{\max}(C)$. By Lemma 4 we have:

$$H(\mu_n^*) = \left(1 - \frac{1}{n}\right) H(\mu_n) + \frac{1}{n} H(\mu) + \left(1 - \frac{1}{n}\right) D(\mu_n \parallel \mu_n^*) + \frac{1}{n} D(\mu \parallel \mu_n^*) \geqq$$
$$\geqq \left(1 - \frac{1}{n}\right) H(\mu_n) + \frac{1}{n} H(\mu) + \frac{1}{n} D(\mu \parallel \mu_n^*)$$

and

$$H(\mu) + D(\mu \parallel \mu_n^*) \leqq n[H_{\max}(C) - H(\mu_n)] + H(\mu_n)$$

follows. As $\mu_n^* \to \mu_C$, and as $D(. \parallel .)$ is jointly lower semi-continuous,

$$\liminf_{n \to \infty} D(\mu \parallel \mu_n^*) \geqq D(\mu \parallel \mu_C) ,$$

and combining with the previous inequality we get

$$H(\mu) + D(\mu \parallel \mu_C) \leqq H_{\max}(C)\,,$$

as desired. □

We remark that the inequality of Theorem 2 also holds for $\mu \in \bar{C}$. More important than this remark is the observation that the inequality characterizes μ_C:

Proposition 1. Let $C \subseteq M$ be convex with $H_{\max}(C) < \infty$. If $\mu^* \in M$ has the property that

$$H(\mu) + D(\mu \parallel \mu^*) \leqq H_{\max}(C)$$

for all $\mu \in C$, then $\mu^* = \mu_C$.

Proof. Let $(\mu_n) \subseteq C$ satisfy $H(\mu_n) \to H_{\max}(C)$. It follows that $D(\mu_n \parallel \mu^*) \to 0$, hence $\mu_n \to \mu^*$. As $\mu_n \to \mu_C$ too, $\mu^* = \mu_C$. □

So far, our results have been analogous to results of Csiszár [4].

We now establish the main facts concerning our information-theoretical game.

Theorem 3. Let $C \subseteq M$ be convex with $H_{\max}(C) < \infty$. Then the value of the absolute game C exists and is $H_{\max}(C)$, and the observer has an optimal strategy, viz. the code κ_C adapted to the center of attraction μ_C. Furthermore, this optimal strategy is unique, indeed, for any $\kappa \in K$ and $\eta \in M$ with $\kappa \leftrightarrow \eta$, we have

$$\sup_{\mu \in C} \langle \kappa, \mu \rangle \geqq H_{\max}(C) + D(\mu_C \parallel \eta)\,. \tag{14}$$

We could also add that nature has an optimal strategy if and only if $\mu_C \in C$ and $H(\mu_C) = H_{\max}(C)$. This has been observed previously.

Proof. Combining Lemma 2 and Theorem 2, we have

$$\inf_{\kappa \in K} \sup_{\mu \in C} \langle \kappa, \mu \rangle \leqq \sup_{\mu \in C} \langle \kappa_C, \mu \rangle = \sup_{\mu \in C} [H(\mu) + D(\mu \parallel \mu_C)] \leqq$$

$$\leqq H_{\max}(C)\,.$$

Since, as already noticed, $H_{\max}(C)$ is bounded above by the inf sup we started with, there must be equality throughout. This shows that the value of the game C exists and is $H_{\max}(C)$ and that κ_C is an optimal strategy for the observer.

To prove uniqueness of an optimal strategy for the observer, let $\kappa \in K$ and denote by η the distribution associated with κ. Let (μ_n) be a sequence in C for which $H(\mu_n) \to$ $\to H_{\max}(C)$. By Lemma 2, Theorem 1 and lower semi-continuity of $D(\,.\parallel .)$, we have

$$\sup_{\mu \in C} \langle \kappa, \mu \rangle \geqq \liminf_{n \to \infty} \langle \kappa, \mu_n \rangle = \liminf_{n \to \infty} [H(\mu_n) + D(\mu_n \parallel \eta)] \geqq$$

$$\geqq H_{\max}(C) + D(\mu_C \parallel \eta)\,.$$

This proves (14). As

$$\kappa = \kappa_C \Leftrightarrow \eta = \mu_C \Leftrightarrow D(\mu_C \parallel \eta) = 0\,,$$

(14) implies that

$$\sup_{\mu \in C} \langle \kappa, \mu \rangle > H_{\max}(C)\,,$$

unless $\kappa = \kappa_C$. Hence κ_C is the only optimal strategy. □

With a special assumption, taken from [4], the optimal strategy of the observer has an extra stability property:

Theorem 4. Let $C \subseteq M$ be convex with $H_{\max}(C) < \infty$. Assume that μ_C is the canonical distribution and that μ_C is an algebraic inner point of C, i.e. that to any $\mu \in C$ there exists $\mu' \in C$ and $0 < \alpha < 1$ such that $\mu_C = \alpha\mu + (1 - \alpha)\mu'$. Then, for every $\mu \in C$, we have

$$\langle \kappa_C, \mu \rangle = H_{\max}(C)\,.$$

Proof. As μ_C is assumed to be canonical, $\langle \kappa_C, \mu_C \rangle = H(\mu_C) = H_{\max}(C)$. Now consider any $\mu \in C$ and determine $\mu' \in C$ and $0 < \alpha < 1$ such that $\mu_C = \alpha\mu + (1 - \alpha)\mu'$. As κ_C is an optimal strategy for the observer,

$$\langle \kappa_C, \mu \rangle \leqq H_{\max}(C) \quad \text{and} \quad \langle \kappa_C, \mu' \rangle \leqq H_{\max}(C)\,. \tag{15}$$

But a proper convex combination of these inequalities yields $\langle \kappa_C, \mu_C \rangle \leqq H_{\max}(C)$ which is known to hold with equality. Hence equality also holds in (15). □

It is probably not true that μ_C algebraic inner implies that μ_C is canonical.

The conclusion of Theorem 4 says that the strategy κ_C is cost-stable.

Sometimes, it is possible directly to determine a cost-stable code κ^*. Then it lies nearby to ask if $\kappa^* = \kappa_C$ or, equivalently, if $\mu^* = \mu_C$ where $\kappa^* \leftrightarrow \mu^*$. We shall see that adding a condition not quite as strong as $\mu^* \in C$ but somewhat stronger than $\mu^* \in \overline{C}$, an afirmative answer can be obtained.

Theorem 5. Let $C \subseteq M$ be convex. Let $\mu^* \in M$ and let κ^* be the code adapted to μ^*. Assume that κ^* is cost-stable. Then $H_{\max}(C) < \infty$. If furthermore,

$$\inf_{\mu \in C} D(\mu \parallel \mu^*) = 0\,,$$

then $\langle \kappa^*, \mu \rangle = H_{\max}(C)$ for $\mu \in C$ and $\kappa^* = \kappa_C$, $\mu^* = \mu_C$.

Proof. Let h denote the common value of $\langle \kappa^*, \mu \rangle$; $\mu \in C$. For $\mu \in C$ we have

$$H(\mu) \leqq H(\mu) + D(\mu \parallel \mu^*) = \langle \kappa^*, \mu \rangle = h\,,$$

hence $H_{\max}(C) \leqq h$.

Let $(\mu_n) \subseteq C$ satisfy $D(\mu_n \parallel \mu^*) \to 0$. Then, going to the limit in the equation

$$h = \langle \kappa^*, \mu_n \rangle = H(\mu_n) + D(\mu_n \parallel \mu^*),$$

we realize that $H(\mu_n) \to h$. Thus $H_{\max}(C) = h$. From Theorem 1, $\mu_n \to \mu_C$ follows. As $\mu_n \to \mu^*$ too, $\mu^* = \mu_C$ follows. □

We end this section with some results of a combined information-theoretical and topological nature giving details about sets with $H_{\max}(C) < \infty$. We assume that I is infinite. For topological details, the reader is referred to [5].

Put

$$M_0 = \{\mu \in M \mid H(\mu) < \infty\},$$

$$M_\infty = \{\mu \in M \mid H(\mu) = \infty\}.$$

M_0 and M_∞ are both convex dense subsets of M. Furthermore, M_∞ is a G_δ-set, indeed, it is a countable intersection of open and convex sets:

$$M_\infty = \bigcap_1^\infty \{H(\mu) > n\}.$$

In particular, M_∞ is Polish (separable and metrizable with a complete metric). The set M_0 is not Polish, it is not even a Baire space; to see this notice that the sets

$$M_0 \cap \{H(\mu) > n\}$$

are open and dense subsets of M_0 with empty intersection.

The position of M_0 and M_∞ as subsets of M resembles that of the rationals and the irrationals as subsets of the reals.

For a subset $C \subseteq M$, we denote by $\bar{C}$, $\mathrm{co}(C)$ and by $\overline{\mathrm{co}}(C)$ the closure, the convex hull and the closed convex hull of C, respectively.

Theorem 6. (i) If $C \subseteq M_0$ and if $\mathrm{co}(C)$ is a Baire space, then $H_{\max}(C) < \infty$.

(ii) Let $g : M \to]-\infty, \infty]$ be an affine lower semicontinuous function, let x be real and put

$$C = \{\mu \in M \mid \langle g, \mu \rangle = x\}.$$

If $C \subseteq M_0$, then $H_{\max}(C) < \infty$.

(iii) If $C \subseteq M$ is convex, then $H_{\max}(C) < \infty$ if and only if $\bar{C} \subseteq M_0$.

Proof. (i): Let us prove a slightly more general result and assume that there exists a Baire space A such that

$$\mathrm{co}(C) \subseteq A \subseteq \overline{\mathrm{co}}(C) \cap M_0. \tag{3}$$

As A is a Baire space which is covered by the closed sets $\{H(\mu) \leqq B\}$, we may apply the Baire category argument to conclude that there exist $B < \infty$, $\mu_0 \in A$ and $\varepsilon > 0$ such that, for any $\mu \in A$ with $\|\mu - \mu_0\| < \varepsilon$, we have $H(\mu) \leqq B$.

Now consider any $\mu \in \mathrm{co}(C)$ and put

$$\eta = \left(1 - \frac{\varepsilon}{3}\right)\mu_0 + \frac{\varepsilon}{3}\mu .$$

Then $\|\eta - \mu_0\| < \varepsilon$. As $\mu_0 \in \overline{\mathrm{co}}(C)$ and as $\mathrm{co}(C) \subseteq A$, there exists a sequence (η_n) of distributions in A such that $\eta_n \to \eta$ and such that $\|\eta_n - \mu_0\| < \varepsilon$ for all n. Then $H(\eta_n) \leqq B$ for all n and $H(\eta) \leqq B$ follows. Since

$$H(\eta) \geqq \left(1 - \frac{\varepsilon}{3}\right) H(\mu_0) + \frac{\varepsilon}{3} H(\mu) \geqq \frac{\varepsilon}{3} H(\mu) ,$$

we can now conclude that $H(\mu) \leqq 3B\varepsilon^{-1}$.

(ii): Clearly, $C = \{\langle g, \mu\rangle = x\}$ is convex. From

$$C = \{\langle g, \mu\rangle \leqq x\} \cap \bigcap_1^\infty \left\{\langle g, \mu\rangle > x - \frac{1}{n}\right\} ,$$

we see that C is a G_δ-subset of M, hence Polish, in particular a Baire space. The result now follows from (i).

(iii): Even without an assumption of convexity, $H_{\max}(C) < \infty$ implies $\overline{C} \subseteq M_0$. Now assume that C is convex and that $\overline{C} \subseteq M_0$. Then $A = \overline{C}$ is a Baire space and satisfies (3) hence, from the result proved above, $H_{\max}(C) < \infty$ follows. □

Naturally, the conclusion of (i) may be strengthened to $H_{\max}(\mathrm{co}(C)) < \infty$. We note that $H_{\max}(C) < \infty$ does not imply $H_{\max}(\mathrm{co}(C)) < \infty$ in general; to see this, consider the set of one-point masses. This example also illustrates the role of convexity in (iii) since it shows that $H_{\max}(C) < \infty$ does not imply $\overline{\mathrm{co}}(C) \subseteq M_0$ even though both $\mathrm{co}(C)$ and $\overline{C}$ are subsets of M_0; actually, we have the extreme situation that $H_{\max}(C) = 0$ but $\overline{\mathrm{co}}(C) = M$. Of course, for a general set C, we have $H_{\max}(C) < \infty$ if $\overline{\mathrm{co}}(C) \subseteq M_0$.

To illustrate the role of convexity in (i), we mention that there exist compact sets (hence also Baire sets) C, with $C \subseteq M_0$ and $H_{\max}(C) = \infty$; to see this, construct a sequence of distributions with finite support and with "large" entropies converging to a unit mass.

Then follow some observations related to compactness properties.

Theorem 7. (i) If C is convex and if $H_{\max}(C) < \infty$, then C is relatively compact.

(ii) A convex and closed subset of M_0 is compact.

(iii) A convex Baire subset of M_0 is relatively compact.

Proof. (i): For the purpose of an indirect proof assume that C is not relatively compact. Then there exists $\varepsilon > 0$ such that, for every finite subset $J \subseteq I$, there exists $\mu \in C$ with $\mu(I \setminus J) > \varepsilon$. This implies that there exists a sequence (I_n) of finite and pairwise disjoint subsets of I and a sequence (μ_n) of distributions in C such that $\mu_n(I_n) \geqq \varepsilon$ for all $n \geqq 1$. For each $m \geqq 1$ put

$$\eta_m = \frac{1}{m} \sum_1^m \mu_n .$$

By convexity, $\eta_m \in C$.

Let $\sigma : I \to \{1, 2, \ldots\}$ be a map which assumes the value n on I_n. Denote by ξ_m the image distribution of η_m under σ. Since there are at most 2 points where ξ_m assumes a value exceeding e^{-1}, since the function $\varphi(x) = -x \log x$ is increasing in $[0, e^{-1}]$ and since $\xi_m(n) \geqq \varepsilon/m$ for $n = 1, 2, \ldots, m$, we find (for $m \geqq 2$):

$$H(\eta_m) \geqq H(\xi_m) = \sum_1^\infty \varphi(\xi_m(n)) \geqq \sum_1^m \varphi(\xi_m(n)) \geqq (m - 2)\, \varphi(\varepsilon/m) \geqq$$

$$\geqq \varepsilon(1 - 2/m) \log m .$$

Letting $m \to \infty$, we see that $H(\eta_m) \to \infty$ contradicting the assumption $H_{\max}(C) < \infty$. Thus C must be relatively compact.

(ii) and (iii) follow from (i) and from Theorem 6. □

THE RELATIVE GAME

We proceed along lines very much parallel to those given for the discrete game. For this reason we only give indications of a few proofs. The state space $(I, \mathcal{I})$ is now arbitrary. We fix a reference measure η and the corresponding reference code κ.

Lemma 5. Let $\mu^* \in M_\eta$ and let $\Delta^* \in K_\eta$ be the codeimprovement adapted to μ^*. Then, for any $\mu \in M_\eta$,

$$\langle \Delta^*, \mu \rangle = D(\mu \parallel \eta) - D(\mu \parallel \mu^*)$$

in the sense that if the right hand side exists as a well defined extended real number, then so does the left hand side and equality holds.

This is a restatement of [4], equation (2.6). The two next results are easy corollaries.

Lemma 6. For any $\mu \in M_\eta$,

$$\max_{\Delta \in K_\eta} \langle \Delta, \mu \rangle = D(\mu \parallel \eta)$$

and the maximum is achieved for the codeimprovement adapted to μ. If $D(\mu \parallel \eta) < \infty$, the maximum is not achieved for any other codeimprovement.

Let $C \subseteq M_\eta$ and put

$$D_{\min}(C \parallel \eta) = \inf_{\mu \in C} D(\mu \parallel \eta) .$$

Lemma 6 shows that an optimal strategy for nature in the relative game $C \parallel \eta$ is the same as a distribution μ with $\mu \in C$ and $D(\mu \parallel \eta) = D_{\min}(C \parallel \eta)$. We are thus led to the so called "minimum discrimination information principle". For a discussion of this principle with many results besides those we shall give, see [10], [3], [4] and [1], esp. Section 9.1.

Lemma 7. Let $\mu_1, \ldots, \mu_n$ be distributions in M_η and let $(p_1, \ldots, p_n)$ be a probability vector. Then

$$\sum p_\nu D(\mu_\nu \parallel \eta) = D(\sum p_\nu \mu_\nu \parallel \eta) + \sum p_\nu D(\mu_\nu \parallel \sum p_j \mu_j) .$$

This identity was perhaps first considered by the author in 1967. We refer to [13].

As a further consequence of Lemma 5, we prove that (13) holds provided $D(\mu \parallel \eta)$ and $D(\mu^* \parallel \eta)$ are not both infinite where μ^* is the distribution in M_η corresponding to Δ. This follows easily from the formula

$$\sup_{\pi \in \Pi_0} D(\mu \parallel \eta)_\pi = D(\mu \parallel \eta) \tag{16}$$

where $D(\mu \parallel \eta)_\pi$ stands for $D(\mu \mid \pi \parallel \eta \mid \pi)$. It is easy to establish (16) with Π_0 replaced by Π (cf. also Lemma 1), and the reduction from Π to Π_0 then only needs a few extra comments. The equation (16) one also finds in [9] and in [11].

We need two topologies on M. In the weak topology, $\mu_n \to \mu$ means that $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{A}$ and in the strong topology, $\mu_n \to \mu$ means that $\|\mu_n - \mu\| \to 0$.

The map $(\mu, \eta) \to D(\mu \parallel \eta)$ of $M \times M$ into $[0, \infty]$ is jointly lower semi-continuous in the weak topology (employ (16)). For fixed $\eta \in M$ and $a < \infty$, the set of $\mu \in M$ with $D(\mu \parallel \eta) \leqq a$ is a weakly compact and convex subset of M. As I. Csiszár pointed out to the author, this set need not be strongly compact.

Theorem 8. Let $C \subseteq M_\eta$ be convex and assume that $D_{\min}(C \parallel \eta) < \infty$. Then there exists a unique distribution $\mu_{C\|\eta} \in M_\eta$ such that $\mu_n \to \mu_{C\|\eta}$, strongly for every sequence $(\mu_n) \subseteq C$ such that $D(\mu_n \parallel \eta) \to D_{\min}(C \parallel \eta)$. Furthermore, for every $\mu \in C$, we have

$$D(\mu \parallel \eta) \geqq D_{\min}(C \parallel \eta) + D(\mu \parallel \mu_{C\|\eta}) . \tag{17}$$

The proof is analogous to the proofs of Theorems 1 and 2. The result is a slight improvement over [4] and the proof hinted at is partly a simplification.

The distribution $\mu_{C\|\eta}$ of Theorem 8, we call the *relative center of attraction*. Equation (17) can also be used to characterize $\mu_{C\|\eta}$ (compare with Proposition 1). Below, $\Delta_{C\|\eta}$ denotes the codeimprovement in K_η adapted to $\mu_{C\|\eta}$.

Theorem 9. Let $C \subseteq M_\eta$ be convex and assume that $D(\mu \,\|\, \eta) < \infty$ for all $\mu \in C$. Then the value of the relative game $C \,\|\, \eta$ exists and is $D_{\min}(C \,\|\, \eta)$, and the observer has an optimal strategy, viz. the codeimprovement $\Delta_{C\|\eta}$. Furthermore, this optimal strategy is unique, indeed, for any $\Delta^* \in K_\eta$ with corresponding distribution $\mu^* \in M_\eta$, we have

$$\inf_{\mu \in C} \langle \Delta^*, \mu \rangle \leqq D_{\min}(C \,\|\, \eta) - D(\mu_{C\|\eta} \,\|\, \mu^*) .$$

Clearly, $\mu_{C\|\eta}$ is an optimal strategy for nature if and only if $\mu_{C\|\eta} \in C$.

Probably, Theorem 9 holds with the condition $D(\mu \,\|\, \eta) < \infty$ for $\mu \in C$ replaced by the weaker assumption $D_{\min}(C \,\|\, \eta) < \infty$. What we have to prove is that if $\mu \in C$ satisfies $D(\mu \,\|\, \mu_{C\|\eta}) = \infty$, hence also $D(\mu \,\|\, \eta) = \infty$, and if $\langle \Delta_{C\|\eta}, \mu \rangle$ is a well defined extended real number, then $\langle \Delta_{C\|\eta}, \mu \rangle \geqq D_{\min}(C \,\|\, \eta)$ holds. This will follow if (13) holds generally.

The results analogous to Theorems 4 and 5 are left to the reader both to formulate and to prove.

When $(I, \mathcal{S})$ is discrete one may ask for which reference measures η, the relative game $C \,\|\, \eta$ leads to the same result as the absolute game in the sense that $\mu_{C\|\eta} = \mu_C$. A partial answer in a special case is as follows.

Theorem 10. Let $(I, \mathcal{S})$ be discrete. Assume that $C \subseteq M$ is convex and that $H_{\max}(C) < \infty$. Consider a distribution $\eta \in M$ and let $\kappa \in K$ be the code adapted to η. If κ is cost-stable for the absolute game C, then $\mu_{C\|\eta} = \mu_C$.

Proof. By definition, there exists a finite constant h such that $\langle \kappa, \mu \rangle = h$ for all $\mu \in C$. By Lemma 2, we then see that

$$D(\mu \,\|\, \eta) = h - H(\mu) ; \quad \mu \in C .$$

Hence, maximizing $H(\mu)$ over C and minimizing $D(\mu \,\|\, \eta)$ over C amounts to the same thing. The result now follows from Theorems 1 and 8. □

One may remark that the assumption $H_{\max}(C) < \infty$ was not really necessary.

A SPECIAL CASE

We shall study a special discrete system first considered in detail by Ingarden and Urbanik in [7]. In essence, all results of this section are due to Ingarden and Urbanik, but our general results permit us to simplify the exposition. We mention that detailed

studies of some continuous systems have occurred in [3], [4] and in [1]. There also exists an interesting set of lecture notes in swedish from 1970 by Per Martin-Löf.

We assume that $(I, \mathscr{I})$ is discrete. Given is a function $E : I \to [0, \infty)$, the *energy function*. The sets of interest to us are the sets of the form

$$C(\bar{E}) = \{\mu \in M \mid \langle E, \mu \rangle = \bar{E}\} .$$

We put

$$H_{\max}(\bar{E}) = H_{\max}(C(\bar{E})) .$$

The *density function* $\Omega = \Omega(\bar{E})$ is the function

$$\Omega(\bar{E}) = \text{number of} \quad i \in I \quad \text{with} \quad E_i \leqq \bar{E} .$$

We assume that $\Omega(\bar{E}) < \infty$ for all $\bar{E}$ (otherwise $H_{\max}(\bar{E})$ could not be finite).

For our purposes we may and do assume that $I = \{1, 2, 3, \ldots\}$, that $E_1 \leqq E_2 \leqq \leqq \ldots$ and that $E_i \to \infty$. We write $E_{\min}$ in place of E_1.

Define the *partition function* $Z = Z(x)$ by

$$Z(x) = \sum_{i \in I} \mathrm{e}^{-E_i x} .$$

This series is a Dirichlet series. Let γ be the abscissa of convergence (ef. eg. [6]). Then

$$\gamma = \limsup_{i \to \infty} \frac{\log i}{E_i} = \limsup_{E \to \infty} \frac{\log \Omega(E)}{E} .$$

For all x with $Z(x) < \infty$ we define $\mu_x \in M$ by

$$\mu_x(i) = \mathrm{e}^{-E_i x}/Z(x) ; \quad i \in I .$$

The family (μ_x) where x ranges over all values with $Z(x) < \infty$ is an *exponential family*. The code adapted to μ_x is denoted κ_x. We have

$$\kappa_x(i) = \log Z(x) + xE_i ; \quad i \in I .$$

The reader should notice that the codes κ_x, hence also the exponential family and the partition function, appears quite naturally in the search for cost-stable codes.

Define a function Φ by

$$\Phi(x) = \langle E, \mu_x \rangle \quad \text{for all } x \text{ with } Z(x) < \infty .$$

For $x > \gamma$ we have

$$\Phi(x) = -Z'(x)/Z(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \log Z(x) .$$

Note that $-Z'(x)$ is a Dirichlet series with the same abscissa of convergence as (Zx).

For $n \geqq 1$ define approximations to $Z(x)$, μ_x, κ_x and $\Phi(x)$:

$$Z_n(x) = \sum_1^n e^{-E_i x},$$

$$\mu_{nx}(i) = e^{-E_i x}/Z_n(x) \qquad \text{for } i \leqq n \ (0 \text{ otherwise}),$$

$$\kappa_{nx}(i) = \log Z_n(x) + xE_i \quad \text{for } i \leqq n \ (\infty \text{ otherwise})$$

$$\Phi_n(x) = \langle E, \mu_{nx} \rangle .$$

These definitions make sense for all real x.
We leave the proof of the following result to the reader.

Proposition 2. Assume that $\gamma < \infty$. Then:

(a) $\Phi_1 \leqq \Phi_2 \leqq \ldots$,

(b) Φ_n is strictly decreasing on R (except if $E_n = E_{\min}$),

(c) $\lim_{x \to \infty} \Phi_n(x) = E_{\min}$, $\lim_{x \to -\infty} \Phi_n(x) = E_n$,

(d) Φ is strictly decreasing on (γ, ∞),

(e) $\lim_{x \to \infty} \Phi(x) = E_{\min}$,

(f) $\Phi(\gamma+) = \infty \Leftrightarrow -Z'(\gamma) = \infty$,

(g) $-Z'(x_0) < \infty \Rightarrow \Phi_n \to \Phi$, uniformly on $[x_0, \infty)$,

(h) $\lim_{n \to \infty} \Phi_n(x) = \infty$ for $x < \gamma$,

(i) $\lim_{n \to \infty} \Phi_n(\gamma) = \Phi(\gamma+)$.

We put

$$E_{\text{crit}} = \Phi(\gamma+).$$

Notice that $E_{\text{crit}} = \infty$ if $Z(\gamma) = \infty$, which will usually be the case in applications (in fact, $\gamma = 0$ will usually hold).

Theorem 11. Assume that $\gamma < \infty$. Then:

(a) $H_{\max}(\bar{E}) < \infty$ for all $E_{\min} < \bar{E} < \infty$,

(b) For $E_{\min} < \bar{E} < \infty$, all codes κ_x (with $Z(x) < \infty$) are cost-stable strategies for the game $C(\bar{E})$,

(c) If $E_{\min} < \bar{E} \leqq E_{\text{crit}}$, then the center of attraction for the game $C(\bar{E})$ is also the canonical distribution and is determined as the only distribution in the exponential family (μ_x) with mean energy $\bar{E}$,

(d) If $E_{\text{crit}} < \bar{E} < \infty$, then μ_γ is the center of attraction for the game $C(\bar{E})$. In this case, μ_γ is not canonical, in fact we both have $\langle E, \mu_\gamma \rangle < \bar{E}$ and $H(\mu_\gamma) < H_{\max}(\bar{E})$.

Proof. As

$$\langle \kappa_x, \mu \rangle = \log Z(x) + x\langle E, \mu \rangle; \; Z(x) < \infty ,$$

(b) follows. The existence of cost-stable strategies implies (a). If $E_{\min} < E_{\text{crit}}$, we can find x with $Z(x) < \infty$ such that $\mu_x \in C(\bar{E})$; this follows from Proposition 2. Then (c) follows from Theorem 5.

It only remains to prove (d). So assume that $-Z'(\gamma) < \infty$ and that $\bar{E} > E_{\text{crit}}$. In order to prove that μ_γ is the center of attraction for $C(\bar{E})$ it suffices to show that

$$\langle \kappa_\gamma, \mu \rangle \leqq H_{\max}(\bar{E}) \quad \text{for} \quad \mu \in C(\bar{E}) ,$$

i.e., we have to show that

$$H_{\max}(\bar{E}) \geqq \log Z(\gamma) + \gamma \bar{E} . \tag{18}$$

This follows from Theorem 3 (or from Proposition 1).

Determine n_0 so that $E_{n_0} > \bar{E}$. For $n \geqq n_0$ determine x_n so that

$$\Phi_n(x_n) = \bar{E} .$$

By Proposition 2,

$$x_{n_0} \leqq x_{n_0+1} \leqq \ldots , \; \lim_{n \to \infty} x_n = \gamma .$$

It follows easily that

$$\limsup Z_n(x_n) \geqq Z(\gamma) . \tag{19}$$

Put $\mu_n = \mu_{n,x_n}$. Then $\mu_n \in C(\bar{E})$ and

$$H(\mu_n) = \log Z_n(x_n) + x_n \bar{E} .$$

From this and from (19), (18) follows. □

Probably, all cost-stable codes are of the form κ_x with $Z(x) < \infty$. If true, this together with (b) of Theorem 11 gives a game-theoretical description of the exponential family.

We remark that by Theorem 10, we may add to Theorem 11 that if we choose a reference measure from the exponential family (μ_x), then the relative center of attraction coincides with the (absolute) center of attraction.

If we are in the critical case (d) of Theorem 11 and we put $C = \{\mu \mid \langle E, \mu \rangle \leqq \bar{E}\}$, then C is a compact convex set with $H_{\max}(C) < \infty$ for which the canonical distribution does not exist.

That the critical case is theoretically possible may be seen by considering energy functions of the form

$$E_i = \log (i + 2) + K \log \log (i + 2) ; \quad i \geqq 1 .$$

Then, for all K, $\gamma = 1$. If $K \leqq 1$, $Z(\gamma) = -Z'(\gamma) = \infty$. If $1 < K \leqq 2$, $Z(\gamma) < \infty$ and $-Z'(\gamma) = \infty$. And if $K > 2$, $Z(\gamma)$ and $-Z'(\gamma)$ are both finite. E.g. if $K = 3$ there is a finite critical energy ($=2{\cdot}99$). We refer to the Figure 1. The example is due to Ingarden and Urbanik [7].

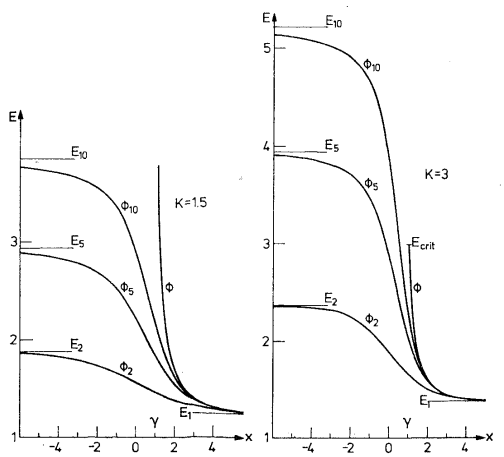

Fig. 1.

Theorem 12. A necessary and sufficient condition that every distribution with finite mean energy has finite entropy, is that $\gamma < \infty$.

Proof. Sufficiency has already been noticed. Let us prove necessity. We actually prove a stronger statement by only assuming that for some $E_{\min} < \bar{E} < \infty$, $H(\mu) < \infty$ for all $\mu \in C(\bar{E})$. By Theorem 6, $H_{\max}(\bar{E}) < \infty$. Determine x_n and μ_n (for $n \geqq n_0$) precisely as in the proof of (d) of Theorem 11. Since, for $n \geqq n_0$,

$$H_{\max}(\bar{E}) \geqq H(\mu_n) = \log Z_n(x_n) + x_n \bar{E} \geqq x_n(\bar{E} - E_{\min}),$$

$\lim x_n < \infty$. Choose $x \geqq \lim x_n$. Then, for $n \geqq n_0$,

$$\log Z_n(x) \leqq \log Z_n(x_n) \leqq H_{\max}(\bar{E}) - x_{n_0}\bar{E},$$

hence $Z(x) < \infty$ and thus $\gamma < \infty$. □

The present results were first developed without knowledge of some of the basic papers mentioned in the references. It seems now that the main novelty lies in the game-theoretical point of view. It would be interesting if this point of view could be extended to cover the modern needs of statistical thermodynamics.

I have had helpful discussions with J. P. R. Christensen and with I. Csiszár. Especially, my acquaintance with Csiszár's ideas let to substantial simplifications of some proofs since they permitted me to substitute a general and rather deep minimax inequality by an intrinsicly information theoretical argument.

(Received August 28, 1978.)

REFERENCES

[1] O. Barndorff-Nielsen: Information and exponential families in Statistical Theory. John Wiley, New York 1978.
[2] W. Bayer, W. Ochs: Quantum States with Maximum Information Entropy, I and II. Zeitschrift für Naturforschung *28a* (1973), 693—701 and 1571—1585.
[3] N. N. Čencov: Statistical Decision Rules and Optimal Decisions. (In russian.) Nauka, Moscow 1972.
[4] I. Csiszár: *I*-divergence geometry of probability distributions and minimization problems. Annals of Probability *3* (1975), 146—158.
[5] R. Engelking: General Topology. PWN, Warszawa 1977.
[6] G. H. Hardy, M. Riesz: The general Theory of Dirichlet's Series. Cambridge University Press, Cambridge 1915.
[7] R. S. Ingarden, K. Urbanik: Quantum Informational Thermodynamics. Acta Physica Polonica *21* (1962), 281—304.
[8] E. T. Jaynes: Information Theory and Statistical Mechanics, I and II. Physical Reviews *106* (1957), 620—630 and *108*, 171—190.
[9] G. Kallianpur: On the amount of information contained in a σ-field. In: Contributions to probability and statistics. Stanford 1960, 265—271.
[10] S. Kullback: Information theory and statistics, John Wiley, New York 1959.
[11] A. Perez: Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. In: Trans. of the first Prague Conference, Prague 1957, 183—208.
[12] A. Rényi: Wahrscheinlichkeitrechnung mit einem Anhang über Informationstheorie. VEB, Berlin 1962.
[13] F. Topsøe: Informationstheorie, eine Einführung. Teubner, Stuttgart 1974.
[14] I. Vincze: On the maximum probability principle in statistical physics. In: Ninth European meeting of statisticians. Budapest 1972, 869—893.

Dr. Flemming Topsøe, The University of Copenhagen, Institute of Mathematics, Universitetsparken 5, 2100 Copenhagen. Denmark.