ON MEASURES OF ENTROPY AND INFORMATION

ALFRÉD RÉNYI
MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES

1. Characterization of Shannon's measure of entropy

Let $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ be a finite discrete probability distribution, that is, suppose $p_k \geqq 0 (k = 1, 2, \cdots, n)$ and $\sum_{k=1}^{n} p_k = 1$. The amount of uncertainty of the distribution $\mathcal{P}$, that is, the amount of uncertainty concerning the outcome of an experiment, the possible results of which have the probabilities $p_1, p_2, \cdots, p_n$, is called the *entropy* of the distribution $\mathcal{P}$ and is usually measured by the quantity $H[\mathcal{P}] = H(p_1, p_2, \cdots, p_n)$, introduced by Shannon [1] and defined by

$$H(p_1, p_2, \cdots, p_n) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{1.1}$$

Different sets of postulates have been given, which characterize the quantity (1.1). The simplest such set of postulates is that given by Fadeev [2] (see also Feinstein [3]). Fadeev's postulates are as follows.

(a) *$H(p_1, p_2, \cdots, p_n)$ is a symmetric function of its variables for* $n = 2, 3, \cdots$.

(b) *$H(p, 1 - p)$ is a continuous function of p for* $0 \leqq p \leqq 1$.

(c) $H(1/2, 1/2) = 1$.

(d) $H[tp_1, (1 - t)p_1, p_2, \cdots, p_n] = H(p_1, p_2, \cdots, p_n) + p_1 H(t, 1 - t)$ *for any distribution* $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ *and for* $0 \leqq t \leqq 1$.

The proof that the postulates (a), (b), (c), and (d) characterize the quantity (1.1) uniquely is easy except for the following lemma, whose proofs up to now are rather intricate.

LEMMA. *Let $f(n)$ be an additive number-theoretical function, that is, let $f(n)$ be defined for* $n = 1, 2, \cdots$ *and suppose*

$$f(nm) = f(n) + f(m), \qquad n, m = 1, 2, \cdots. \tag{1.2}$$

Let us suppose further that

$$\lim_{n \to +\infty} [f(n + 1) - f(n)] = 0. \tag{1.3}$$

Then we have

$$f(n) = c \log n, \tag{1.4}$$

where c is a constant.

This lemma was first proved by Erdös [4]. In fact Erdös proved somewhat more, namely he supposed, as is usual in number theory, the validity of (1.2) only for n and m being relatively prime. Later the lemma was rediscovered by Fadeev. The proofs of both Erdös and Fadeev are rather complicated. In this section we give a new proof of the lemma, which is much simpler.

PROOF. Let $N > 1$ be an arbitrary integer and let us put

$$g(n) = f(n) - \frac{f(N) \log n}{\log N}, \qquad n = 1, 2, \cdots . \tag{1.5}$$

It follows evidently from (1.2) and (1.3) that

$$g(nm) = g(n) + g(m), \qquad n, m = 1, 2, \cdots \tag{1.6}$$

and that

$$\lim_{n \to +\infty} [g(n + 1) - g(n)] = 0. \tag{1.7}$$

We have further

$$g(N) = 0. \tag{1.8}$$

Let us now put $G(-1) = 0$ and

$$G(k) = \max_{N^k \leqq n < N^{k+1}} |g(n)|, \qquad k = 0, 1, \cdots , \tag{1.9}$$

and further,

$$\delta_k = \max_{N^k \leqq n < N^{k+1}} |g(n + 1) - g(n)|, \qquad k = 0, 1, \cdots . \tag{1.10}$$

Clearly we have

$$\lim_{k \to +\infty} \delta_k = 0. \tag{1.11}$$

Now we shall prove that

$$\lim_{n \to +\infty} \frac{g(n)}{\log n} = 0. \tag{1.12}$$

Since for $N^k \leqq n < N^{k+1}$, we have $|g(n)|/\log n \leqq G(k)/k \log N$, in order to prove (1.12) it is clearly sufficient to prove that

$$\lim_{k \to +\infty} \frac{G(k)}{k} = 0. \tag{1.13}$$

Now let n be an arbitrary integer and let k be defined by the inequalities $N^k \leqq n < N^{k+1}$. Let us put $n' = N[n/N]$ where $[x]$ denotes the integral part of x; thus n' is the greatest multiple of N not exceeding n. Then we have evidently $0 \leqq n - n' < N$ and thus

$$|g(n)| \leqq |g(n')| + \sum_{l=n'}^{n-1} |g(l + 1) - g(l)| \leqq |g(n')| + N\delta_k. \tag{1.14}$$

By (1.6) and (1.8) we have

$$g(n') = g\left(\left[\frac{n}{N}\right]\right) + g(N) = g\left(\left[\frac{n}{N}\right]\right), \tag{1.15}$$

and hence the inequalities $N^{k-1} \leqq [n/N] < N^k$, together with (1.14), imply that

$$G(k) \leqq G(k-1) + N\delta_k, \qquad k = 0, 1, \cdots. \tag{1.16}$$

Adding the inequalities (1.16) for $k = 0, 1, \cdots, m$, it follows that

$$\frac{G(m)}{m} \leqq N\left(\frac{\delta_0 + \delta_1 + \cdots + \delta_m}{m}\right). \tag{1.17}$$

Taking (1.11) into account, we obtain (1.13) and so (1.12). But clearly (1.12) implies

$$\lim_{n \to +\infty} \frac{f(n)}{\log n} = \frac{f(N)}{\log N}. \tag{1.18}$$

As N was an arbitrary integer greater than 1 and the left side of (1.18) does not depend on N, it follows that, denoting by c the value of the limit on the left side of (1.18), we have

$$f(N) = c \log N, \qquad N = 2, 3, \cdots. \tag{1.19}$$

By (1.2) we have evidently $f(1) = 0$. Thus the lemma is proved.

With a slight modification the above proof applies also in the case when the validity of (1.2) is supposed only for relatively prime m and n. A previous version of the above proof has been given by the author in [5]. The version given above is somewhat simpler than in [5].

Let us add some remarks on the set of postulates (a) to (d). Let us denote $\mathcal{P} = (p_1, p_2, \cdots, p_m)$ and $\mathcal{Q} = (q_1, q_2, \cdots, q_n)$ as two probability distributions. Let us denote by $\mathcal{P} * \mathcal{Q}$ the direct product of the distributions $\mathcal{P}$ and $\mathcal{Q}$, that is, the distribution consisting of the numbers $p_j q_k$ with $j = 1, 2, \cdots, m$; $k = 1, 2, \cdots, n$. Then we have from (1.1)

$$H[\mathcal{P} * \mathcal{Q}] = H[\mathcal{P}] + H[\mathcal{Q}], \tag{1.20}$$

which expresses one of the most important properties of entropy, namely, its *additivity*: the entropy of a combined experiment consisting of the performance of two independent experiments is equal to the sum of the entropies of these two experiments. It is easy to see that one cannot replace the postulate (d) by (1.20) because (1.20) is much weaker. As a matter of fact there are many quantities other than (1.1) which satisfy the postulates (a), (b), (c), and (1.20). For instance, all the quantities

$$H_\alpha(p_1, p_2, \cdots, p_n) = \frac{1}{1-\alpha} \log_2\left(\sum_{k=1}^{n} p_k^\alpha\right), \tag{1.21}$$

where $\alpha > 0$ and $\alpha \neq 1$ have these properties. The quantity $H_\alpha(p_1, p_2, \cdots, p_n)$ defined by (1.21) can also be regarded as a measure of the entropy of the distribution $\mathcal{P} = (p_1, \cdots, p_n)$. In what follows we shall call

$$H_\alpha(p_1, p_2, \cdots, p_n) = H_\alpha[\mathcal{P}]$$

the *entropy of order* α of the distribution $\mathcal{P}$. We shall deal with these quantities in the next sections. Here we mention only that, as is easily seen,

$$\lim_{\alpha \to 1} H_\alpha(p_1, p_2, \cdots, p_n) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{1.22}$$

Thus Shannon's measure of entropy is the limiting case for $\alpha \to 1$ of the measure of entropy $H_\alpha[\mathcal{P}]$. In view of (1.22) we shall denote in what follows Shannon's measure of entropy (1.1) by $H_1(p_1, \cdots, p_n)$ and call it the measure of entropy of order 1 of the distribution. Thus we put

$$H_1[\mathcal{P}] = H_1(p_1, p_2, \cdots, p_n) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{1.23}$$

There are besides the quantities (1.22) still others which satisfy the postulates (a), (b), (c), and (1.20). For instance, applying a linear operation on $H_\alpha[\mathcal{P}]$ as a function of α we get again such a quantity. In the next section we shall show what additional postulate is needed besides (a), (b), (c), and (1.20) to characterize the entropy of order 1. We shall see that in order to get such a characterization of Shannon's entropy, it is advantageous to extend the notion of a probability distribution, and define entropy for these generalized distributions.

2. Characterization of Shannon's measure of entropy of generalized probability distributions

The characterization of measures of entropy (and information) becomes much simpler if we consider these quantities as defined on the set of *generalized probability distributions.* Let $[\Omega, \mathcal{B}, P]$ be a probability space, that is, Ω an arbitrary nonempty set, called the set of elementary events; $\mathcal{B}$ a σ-algebra of subsets of Ω, containing Ω itself, the elements of $\mathcal{B}$ being called events; and P a probability measure, that is, a nonnegative and additive set function for which $P(\Omega) = 1$, defined on $\mathcal{B}$. Let us call a function $\xi = \xi(\omega)$ which is defined for $\omega \in \Omega_1$ where $\Omega_1 \in \mathcal{B}$ and $P(\Omega_1) > 0$, and which is measurable with respect to $\mathcal{B}$, a *generalized random variable*. If $P(\Omega_1) = 1$ we call ξ an *ordinary* (*or complete*) *random variable,* while if $0 < P(\Omega_1) < 1$ we call ξ an *incomplete random variable.* Clearly, an incomplete random variable can be interpreted as a quantity describing the result of an experiment depending on chance which is not always observable, only with probability $P(\Omega_1) < 1$. The distribution of a generalized random variable will be called a generalized probability distribution. In particular, in the case when ξ takes on only a finite number of different values $x_1, x_2, \cdots, x_n$, the distribution of ξ consists of the set of numbers $p_k = P\{\xi = x_k\}$ for $k = 1, 2, \cdots, n$. Thus a finite discrete generalized probability distribution is simply a sequence $p_1, p_2, \cdots, p_n$ of nonnegative numbers such that putting $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ and

$$W(\mathcal{P}) = \sum_{k=1}^{n} p_k, \tag{2.1}$$

we have

$$0 < W(\mathcal{P}) \leqq 1. \tag{2.2}$$

We shall call $W(\mathcal{P})$ the *weight* of the distribution. Thus the weight of an ordinary distribution is equal to 1. A distribution which has a weight less than 1 will be called an *incomplete distribution.*

Let Δ denote the set of all finite discrete generalized probability distributions, that is, Δ is the set of all sequences $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ of nonnegative numbers such that $0 < \sum_{k=1}^{n} p_k \leqq 1$. We shall characterize the entropy $H[\mathcal{P}]$ (of order 1) of a generalized probability distribution $\mathcal{P} = (p_1, \cdots, p_n)$ by the following five postulates.

Postulate 1. *$H[\mathcal{P}]$ is a symmetric function of the elements of $\mathcal{P}$.*

Postulate 2. *If $\{p\}$ denotes the generalized probability distribution consisting of the single probability p then $H[\{p\}]$ is a continuous function of p in the interval* $0 < p \leqq 1$. Note that the continuity of $H[\{p\}]$ is supposed only for $p > 0$, but not for $p = 0$.

Postulate 3. $H[\{1/2\}] = 1$.

Postulate 4. *For $\mathcal{P} \in \Delta$ and $\mathcal{Q} \in \Delta$ we have $H[\mathcal{P} * \mathcal{Q}] = H[\mathcal{P}] + H[\mathcal{Q}]$.*

Before stating our last postulate we introduce some notation. If we denote $\mathcal{P} = (p_1, p_2, \cdots, p_m)$ and $\mathcal{Q} = (q_1, q_2, \cdots, q_n)$ as two generalized distributions such that $W(\mathcal{P}) + W(\mathcal{Q}) \leqq 1$, we put

$$\mathcal{P} \cup \mathcal{Q} = (p_1, p_2, \cdots, p_m, q_1, q_2, \cdots, q_n). \tag{2.3}$$

If $W(\mathcal{P}) + W(\mathcal{Q}) > 1$ then $\mathcal{P} \cup \mathcal{Q}$ is not defined. Now we can state our last postulate.

Postulate 5. *If $\mathcal{P} \in \Delta$, $\mathcal{Q} \in \Delta$, and $W(\mathcal{P}) + W(\mathcal{Q}) \leqq 1$, we have*

$$H[\mathcal{P} \cup \mathcal{Q}] = \frac{W(\mathcal{P})H[\mathcal{P}] + W(\mathcal{Q})H[\mathcal{Q}]}{W(\mathcal{P}) + W(\mathcal{Q})}. \tag{2.4}$$

Postulate 5 may be called the *mean-value property* of entropy; the entropy of the union of two incomplete distributions is the weighted mean value of the entropies of the two distributions, where the entropy of each component is weighted with its own weight. One of the advantages of defining the entropy for generalized distributions, and not merely for ordinary (complete) distributions, is that this mean-value property is much simpler in the general case.

We now prove

Theorem 1. *If $H[\mathcal{P}]$ is defined for all $\mathcal{P} \in \Delta$ and satisfies the postulates 1, 2, 3, 4, and 5, then $H[\mathcal{P}] = H_1[\mathcal{P}]$, where*

$$H_1[\mathcal{P}] = \frac{\sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}}{\sum_{k=1}^{n} p_k}. \tag{2.5}$$

Proof. The proof is very simple. Let us put

$$h(p) = H[\{p\}], \qquad 0 < p \leqq 1, \tag{2.6}$$

where $\{p\}$ again denotes the generalized distribution consisting of the single probability p. We have by postulate 4

$$h(pq) = h(p) + h(q) \qquad \text{for} \quad 0 < p \leqq 1; \qquad 0 < q \leqq 1. \tag{2.7}$$

By postulate 2, $h(p)$ is continuous for $0 < p \leqq 1$ and by postulate 3 we have $h(1/2) = 1$. Thus it follows that

$$h(p) = H[\{p\}] = \log_2 \frac{1}{p}. \tag{2.8}$$

Now it follows from postulate 5 by induction that if $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_n$ are incomplete distributions such that $\sum_{k=1}^{n} w(\mathcal{P}_k) \leqq 1$, then

$$H[\mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_n] = \frac{w(\mathcal{P}_1)H[\mathcal{P}_1] + w(\mathcal{P}_2)H[\mathcal{P}_2] + \cdots + w(\mathcal{P}_n)H[\mathcal{P}_n]}{w(\mathcal{P}_1) + w(\mathcal{P}_2) + \cdots + w(\mathcal{P}_n)}. \tag{2.9}$$

As any generalized distribution $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ can be written in the form

$$\mathcal{P} = \{p_1\} \cup \{p_2\} \cup \cdots \cup \{p_n\}, \tag{2.10}$$

the assertion of theorem 1 follows from (2.9) and (2.10).

An advantage of the above introduction of the notion of entropy is that the term $\log_2 (1/p_k)$ in Shannon's formula is interpreted as the entropy of the generalized distribution consisting of the single probability p_k and thus it becomes evident that (1.1) is really a mean value. This point of view was emphasized previously by some authors, especially by G. A. Barnard [6].

The question arises of what other quantity is obtained if we replace in postulate 5 the arithmetic mean by some other mean value. The general form of a mean value of the numbers $x_1, x_2, \cdots, x_n$ taken with the weights $w_1, w_2, \cdots, w_n$, where $w_k > 0$ and $\sum_{k=1}^{n} w_k = 1$, is usually written in the form (for example, see [7])

$$g^{-1}\left[\sum_{k=1}^{n} w_k g(x_k)\right], \tag{2.11}$$

where $y = g(x)$ is an arbitrary strictly monotonic and continuous function and $x = g^{-1}(y)$ denotes the inverse function of $y = g(x)$. The function $g(x)$ is called the Kolmogorov-Nagumo function corresponding to the mean value (2.10). Thus we are led to replace postulate 5 by

POSTULATE 5′. *There exists a strictly monotonic and continuous function $y = g(x)$ such that if $\mathcal{P} \in \Delta$, $\mathcal{Q} \in \Delta$, and $w(\mathcal{P}) + w(\mathcal{Q}) \leqq 1$, we have*

$$H[\mathcal{P} \cup \mathcal{Q}] = g^{-1}\left[\frac{w(\mathcal{P})g(H[\mathcal{P}]) + w(\mathcal{Q})g(H[\mathcal{Q}])}{w(\mathcal{P}) + w(\mathcal{Q})}\right]. \tag{2.12}$$

It is an open question which choices of the function $g(x)$ are admissible, that is, are such that postulate 5′ is compatible with 4. Clearly, if $g(x) = ax + b$ with $a \neq 0$, then postulate 5′ reduces to 5. Another choice of $g(x)$ which is admissible is to choose $g(x)$ to be an exponential function. If $g(x) = g_\alpha(x)$ where $\alpha > 0$, $\alpha \neq 1$, and

$$g_\alpha(x) = 2^{(\alpha-1)x}, \tag{2.13}$$

then postulates 1, 2, 3, 4, and 5′ characterize the entropy of order α. In other words the following theorem is valid.

THEOREM 2. *If $H[\mathcal{P}]$ is defined for all $\mathcal{P} \in \Delta$ and satisfies postulates* 1, 2, 3, 4, *and* 5′ *with $g(x) = g_\alpha(x)$, where $g_\alpha(x)$ is defined by* (2.13), *$\alpha > 0$, and $\alpha \neq 1$, then $H[\mathcal{P}] = H_\alpha[\mathcal{P}]$, where, putting $\mathcal{P} = (p_1, p_2, \cdots, p_n)$, we have*

$$H_\alpha[\mathcal{P}] = \frac{1}{1-\alpha}\log_2\left[\frac{\sum_{k=1}^{n} p_k^\alpha}{\sum_{k=1}^{n} p_k}\right]. \tag{2.14}$$

The quantity (2.14) will be called the *entropy of order α* of the generalized distribution $\mathcal{P}$. Clearly if $\mathcal{P}$ is an ordinary distribution, (2.14) reduces to (1.21). It is also easily seen that

$$\lim_{\alpha\to1} H_\alpha[\mathcal{P}] = H_1[\mathcal{P}], \tag{2.15}$$

where $H_1[\mathcal{P}]$ is defined by (2.5).

The fact that $H_\alpha[\mathcal{P}]$ is characterized by the same properties as $H_1[\mathcal{P}]$, with only the difference that instead of the arithmetic mean value in postulate 5 we have an exponential mean value in 5′, and the fact that $H_1[\mathcal{P}]$ is a limiting case of $H_\alpha[\mathcal{P}]$ for $\alpha \to 1$, both indicate that it is appropriate to consider $H_\alpha[\mathcal{P}]$ also as a measure of entropy of the distribution $\mathcal{P}$. In the next section we shall show that if we formulate the problem in a more general form, the only admissible choices of the function $g(x)$ are those considered above. That is, that $g(x)$ has to be either a linear or an exponential function.

3. Characterization of the amount of information $I(\mathcal{Q}|\mathcal{P})$

The entropy of a probability distribution can be interpreted not only as a measure of uncertainty but also as a measure of information. As a matter of fact, the amount of information which we get when we observe the result of an experiment (depending on chance) can be taken numerically equal to the amount of uncertainty concerning the outcome of the experiment before carrying it out.

There are however also other amounts of information which are often considered. For instance we may ask what is the amount of information concerning a random variable ξ obtained from observing an event E, which is in some way connected with the random variable ξ. If $\mathcal{P}$ denotes the original (unconditional) distribution of the random variable ξ and $\mathcal{Q}$ the conditional distribution of ξ under the condition that the event E has taken place, we shall denote a measure of the amount of information concerning the random variable ξ contained in the observation of the event E by $I(\mathcal{Q}|\mathcal{P})$. Clearly $\mathcal{Q}$ is always absolutely continuous with respect to $\mathcal{P}$; thus the quantity $I(\mathcal{Q}|\mathcal{P})$ will be defined only if $\mathcal{Q}$ is absolutely continuous with respect to $\mathcal{P}$. Denoting by $h = d\mathcal{Q}/d\mathcal{P}$ the Radon-Nikodym derivative of $\mathcal{Q}$ with respect to $\mathcal{P}$, a possible measure of the amount of information in question is

$$I_1(\mathcal{Q}|\mathcal{P}) = \int_\Omega \log_2 h \, d\mathcal{Q} = \int_\Omega h \log_2 h \, d\mathcal{P}. \tag{3.1}$$

In the case when the random variable ξ takes on only a finite number of different values $x_1, x_2, \cdots, x_n$ and we put $P\{\xi = x_k\} = p_k$ and $P\{\xi = x_k|E\} = q_k$ for $k = 1, 2, \cdots, n$, then (3.1) reduces to

$$I_1(\mathcal{Q}|\mathcal{P}) = \sum_{k=1}^{n} q_k \log_2 \frac{q_k}{p_k}. \tag{3.2}$$

It should however be added that other interpretations of the quantity (3.1) or of (3.2) have also been given (see Kullback [8], where further literature is also indicated). Notice that the quantity (3.2) is defined for two finite discrete probability distributions $\mathcal{P} = (p_1, \cdots, p_n)$ and $\mathcal{Q} = (q_1, \cdots, q_n)$ only if $p_k > 0$ for $k = 1, 2, \cdots, n$ (among the q_k there may be zeros) and if there is given a one-to-one correspondence between the elements of the distribution $\mathcal{P}$ and $\mathcal{Q}$, which must therefore consist of an equal number of terms. It follows easily from Jensen's inequality (see, for example, [7]) that the quantities (3.1) or (3.2) are always nonnegative, and they are equal to 0 if and only if the distributions $\mathcal{P}$ and $\mathcal{Q}$ are identical.

While many systems of postulates have been given to characterize the entropy, it seems that a similar characterization of the quantity (3.2) has not been attempted. In this section we shall characterize the quantity (3.2) by certain intuitively evident postulates. At the same time we shall consider also other possible measures of the amount of information in question. It turns out that the only alternative quantities are the quantities

$$I_\alpha(\mathcal{Q}|\mathcal{P}) = \frac{1}{\alpha - 1} \log_2 \left(\sum_{k=1}^{n} \frac{q_k^\alpha}{p_k^{\alpha-1}} \right), \tag{3.3}$$

where $\alpha \neq 1$. Evidently we have

$$\lim_{\alpha \to 1} I_\alpha(\mathcal{Q}|\mathcal{P}) = I_1(\mathcal{Q}|\mathcal{P}). \tag{3.4}$$

We shall call the quantity (3.3) the information of order α contained in the observation of the event E with respect to the random variable ξ or, for the sake of brevity, *the information of order α obtained if the distribution $\mathcal{P}$ is replaced by the distribution $\mathcal{Q}$*. We shall give a system of postulates, analogous to the postulates for entropy considered in section 2, which characterize the quantities $I_\alpha(\mathcal{Q}|\mathcal{P})$, including the case $\alpha = 1$.

As in the case of entropy, it is advantageous to consider the quantity $I(\mathcal{Q}|\mathcal{P})$ for generalized probability distributions, not only for complete distributions. We suppose that, associated with any generalized probability distribution $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ such that $p_k > 0$ for $k = 1, 2, \cdots, n$, and any generalized probability distribution $\mathcal{Q} = (q_1, q_2, \cdots, q_n)$ whose terms are given in a one-to-one correspondence with those of $\mathcal{P}$ (as determined by their indices), there corresponds a real number $I(\mathcal{Q}|\mathcal{P})$ which satisfies the following postulates.

POSTULATE 6. *$I(\mathcal{Q}|\mathcal{P})$ is unchanged if the elements of $\mathcal{P}$ and $\mathcal{Q}$ are rearranged in the same way so that the one-to-one correspondence between them is not changed.*

POSTULATE 7. *If $\mathcal{P} = (p_1, p_2, \cdots, p_n)$ and $\mathcal{Q} = (q_1, q_2, \cdots, q_n)$, and $p_k \leqq q_k$ for $k = 1, 2, \cdots, n$ then $I(\mathcal{Q}|\mathcal{P}) \geqq 0$; while if $p_k \geqq q_k$ for $k = 1, 2, \cdots, n$ then $I(\mathcal{Q}|\mathcal{P}) \leqq 0$.*

POSTULATE 8. $I(\{1\}|\{1/2\}) = 1$.

POSTULATE 9. *If $I(\mathcal{Q}_1|\mathcal{P}_1)$ and $I(\mathcal{Q}_2|\mathcal{P}_2)$ are defined, and if $\mathcal{P} = \mathcal{P}_1 * \mathcal{P}_2$ and $\mathcal{Q} = \mathcal{Q}_1 * \mathcal{Q}_2$ and the correspondence between the elements of $\mathcal{P}$ and $\mathcal{Q}$ is that induced by the correspondence between the elements of $\mathcal{P}_1$ and $\mathcal{Q}_1$, and those of $\mathcal{P}_2$ and $\mathcal{Q}_2$, then*

$$I(\mathcal{Q}|\mathcal{P}) = I(\mathcal{Q}_1|\mathcal{P}_1) + I(\mathcal{Q}_2|\mathcal{P}_2). \tag{3.5}$$

POSTULATE 10. *There exists a continuous and strictly increasing function $y = g(x)$ defined for all real x, such that denoting by $x = g^{-1}(y)$ its inverse function, if $I(\mathcal{Q}_1|\mathcal{P}_1)$ and $I(\mathcal{Q}_2|\mathcal{P}_2)$ are defined, and $0 < w(\mathcal{P}_1) + w(\mathcal{P}_2) \leqq 1$ and $0 < w(\mathcal{Q}_1) + w(\mathcal{Q}_2) \leqq 1$, and the correspondence between the elements of $\mathcal{P}_1 \cup \mathcal{P}_2$ and $\mathcal{Q}_1 \cup \mathcal{Q}_2$ is that induced by the correspondence between the elements of $\mathcal{P}_1$ and $\mathcal{Q}_1$ and those of $\mathcal{P}_2$ and $\mathcal{Q}_2$, then we have*

$$I(\mathcal{Q}_1 \cup \mathcal{Q}_2|\mathcal{P}_1 \cup \mathcal{P}_2) = g^{-1}\left\{\frac{w(\mathcal{Q}_1)g[I(\mathcal{Q}_1|\mathcal{P}_1)] + w(\mathcal{Q}_2)g[I(\mathcal{Q}_2|\mathcal{P}_2)]}{w(\mathcal{Q}_1) + w(\mathcal{Q}_2)}\right\}. \tag{3.6}$$

Let us mention that if $\bar{g}(x) = ag(x) + b$ where $a \neq 0$, then the right side of (3.6) remains unchanged if we replace $g(x)$ by $\bar{g}(x)$. Thus if postulate 10 holds with $g(x)$ it holds also for $\bar{g}(x)$ instead of $g(x)$. We now prove

THEOREM 3. *Suppose that the quantity $I(\mathcal{Q}|\mathcal{P})$ satisfies the postulates 6, 7, 8, 9, and 10. Then the function $g(x)$ in 10 is necessarily either a linear or an exponential function. In the first case $I(\mathcal{Q}|\mathcal{P}) = I_1(\mathcal{Q}|\mathcal{P})$, where*

$$I_1(\mathcal{Q}|\mathcal{P}) = \frac{\sum_{k=1}^{n} q_k \log_2 \frac{q_k}{p_k}}{\sum_{k=1}^{n} q_k}, \tag{3.7}$$

while in the second case $I(\mathcal{Q}|\mathcal{P}) = I_\alpha(\mathcal{Q}|\mathcal{P})$ with some $\alpha \neq 1$, where

$$I_\alpha(\mathcal{Q}|\mathcal{P}) = \frac{1}{\alpha - 1} \log_2 \frac{\sum_{k=1}^{n} \frac{q_k^\alpha}{p_k^{\alpha-1}}}{\sum_{k=1}^{n} q_k}. \tag{3.8}$$

REMARK. If $\mathcal{P}$ and $\mathcal{Q}$ are complete distributions then clearly the formulas (3.7) and (3.8) reduce respectively to the formulas (3.2) and (3.3).

PROOF. Let us put

$$f(q, p) = I(\{q\}|\{p\}), \qquad 0 < p \leqq 1, \qquad 0 < q \leqq 1. \tag{3.9}$$

It follows from postulate 9 that

$$f(q_1q_2, p_1p_2) = f(q_1, p_1) + f(q_2, p_2). \tag{3.10}$$

Putting $q_1 = q_2 = 1$ in (3.10), we get

$$f(1, p_1p_2) = f(1, p_1) + f(1, p_2), \tag{3.11}$$

while for $q_1 = p_2 = 1$, $p_1 = p$, $q_2 = q$ we get from (3.10)

$$f(q, p) = f(1, p) + f(q, 1). \tag{3.12}$$

On the other hand, it follows from postulate 7 that $I(\mathcal{P}|\mathcal{P}) = 0$ for any $\mathcal{P}$, and thus

$$f(1, p) + f(p, 1) = 0. \tag{3.13}$$

Hence we obtain from (3.12)

$$f(q, p) = f(1, p) - f(1, q). \tag{3.14}$$

Now, according to postulate 7, it follows from (3.14) that $f(1, p)$ is a decreasing function of p, and by taking postulate 8 into account it follows from (3.11) that

$$f(1, p) = \log_2 \frac{1}{p}. \tag{3.15}$$

Thus from (3.14) we obtain

$$f(q, p) = I(\{q\}|\{p\}) = \log_2 \frac{q}{p}, \qquad 0 < p \leqq 1, \qquad 0 < q \leqq 1. \tag{3.16}$$

Using now postulate 10, considering the decompositions $\mathcal{P} = \{p_1\} \cup \{p_2\} \cup \cdots \cup \{p_n\}$ and $\mathcal{Q} = \{q_1\} \cup \{q_2\} \cup \cdots \cup \{q_n\}$ and applying induction we obtain

$$I(\mathcal{Q}|\mathcal{P}) = g^{-1}\left[\frac{\sum_{k=1}^{n} q_k g\left(\log_2 \frac{q_k}{p_k}\right)}{\sum_{k=1}^{n} q_k}\right]. \tag{3.17}$$

Now let us consider what possible choices of the function $g(x)$ are compatible with postulate 9. It follows from postulate 9 that for any $\lambda \geqq 0$ and $\mu \geqq 0$ we have

$$I(\mathcal{Q} * \{e^{-\lambda}\}|\mathcal{P} * \{e^{-\mu}\}) = I(\mathcal{Q}|\mathcal{P}) + \mu - \lambda. \tag{3.18}$$

Thus, putting $\mu - \lambda = y$, we see that for an arbitrary real y we have

$$g^{-1}\left[\frac{\sum_{k=1}^{n} q_k g\left(\log_2 \frac{q_k}{p_k} + y\right)}{\sum_{k=1}^{n} q_k}\right] = g^{-1}\left[\frac{\sum_{k=1}^{n} q_k g\left(\log_2 \frac{q_k}{p_k}\right)}{\sum_{k=1}^{n} q_k}\right] + y. \tag{3.19}$$

Now if $w_1, w_2, \cdots, w_n$ is any sequence of positive numbers such that $\sum_{k=1}^{n} w_k = 1$ and $x_1, x_2, \cdots, x_n$ is any sequence of real numbers, we may choose the generalized distributions $\mathcal{P}$ and $\mathcal{Q}$ in such a way that

$$(3.20)\qquad \frac{q_k}{\sum_{k=1}^{n} q_k} = w_k \quad\text{and}\quad \log_2 \frac{q_k}{p_k} = x_k, \qquad k = 1, 2, \cdots, n.$$

As a matter of fact, we can choose $q_k = \rho w_k$ and $p_k = \rho w_k 2^{-x_k}$ for $k = 1, 2, \cdots, n$, where $\rho > 0$ is so small that $\sum_{k=1}^{n} p_k \leqq 1$ and $\sum_{k=1}^{n} q_k \leqq 1$. Thus we obtain from (3.19) the result that for any such sequences w_k and x_k and for any real y we have

$$(3.21)\qquad g^{-1}\left[\sum_{k=1}^{n} w_k g(x_k + y)\right] = g^{-1}\left[\sum_{k=1}^{n} w_k g(x_k)\right] + y.$$

Now (3.21) can be expressed in the following form. If

$$(3.22)\qquad g_y(x) = g(x + y),$$

then we have

$$(3.23)\qquad g_y^{-1}\left[\sum_{k=1}^{n} w_k g_y(x_k)\right] = g^{-1}\left[\sum_{k=1}^{n} w_k g(x_k)\right].$$

That is, the functions $g(x)$ and $g_y(x)$ generate the same mean value. According to a theorem of the theory of mean values (see theorem 83 in [7]) this is possible only if $g_y(x)$ is a linear function of $g(x)$, that is, if there exist constants $a(y) \neq 0$ and $b(y)$ such that

$$(3.24)\qquad g_y(x) = g(x + y) = a(y)g(x) + b(y).$$

Without restricting the generality we may suppose $g(0) = 0$. Thus we obtain $b(y) = g(y)$, that is,

$$(3.25)\qquad g(x + y) = a(y)g(x) + g(y).$$

But (3.25) is true for any x and y. Thus we may interchange the roles of x and y and we get

$$(3.26)\qquad g(x + y) = a(x)g(y) + g(x).$$

Thus if $x \neq 0$ and $y \neq 0$ we obtain, comparing (3.25) and (3.26),

$$(3.27)\qquad \frac{a(y) - 1}{g(y)} = \frac{a(x) - 1}{g(x)}.$$

It follows from (3.27) that there exists a constant k such that

$$(3.28)\qquad a(x) - 1 = kg(x)$$

for all real x. Now we have to distinguish two cases. If $k = 0$ then $a(x) \equiv 1$ and thus by (3.25) we obtain for $g(x)$ the functional equation

$$(3.29)\qquad g(x + y) = g(x) + g(y)$$

for any real x and y. As $g(x)$ is by supposition monotonic it follows that $g(x) = cx$ where $c \neq 0$ is a constant. In this case we see from (3.17) that $I(\mathcal{Q}|\mathcal{P}) = I_1(\mathcal{Q}|\mathcal{P})$, where $I_1(\mathcal{Q}|\mathcal{P})$ is defined by (3.7). In the second case, when $k \neq 0$, the substitution of (3.28) into (3.25) yields

$$a(x + y) = a(x)a(y) \tag{3.30}$$

for any real x and y. Now (3.28) shows that $a(x)$ is monotonic and hence it follows that $a(x)$ is an exponential function, and so it can be written in the form

$$a(x) = c2^{(\alpha-1)x}, \tag{3.31}$$

where $\alpha \neq 1$ and $c \neq 0$ are constants. It follows from (3.28) that

$$g(x) = \frac{c2^{(\alpha-1)x} - 1}{k}. \tag{3.32}$$

Substituting (3.32) into (3.17) we obtain the result that $I(\mathcal{Q}|\mathcal{P}) = I_\alpha(\mathcal{Q}|\mathcal{P})$, where $I_\alpha(\mathcal{Q}|\mathcal{P})$ is defined by (3.8). Thus theorem 3 is proved. (The last part of the proof is essentially identical with the proof of theorem 84 of [7].)

Notice that our postulates do not demand that $I(\mathcal{Q}|\mathcal{P})$ should be a continuous function of the variables p_k, q_k for $k = 1, 2, \cdots, n$. Instead of continuity we have postulated a certain sort of monotony by means of postulate 7. This is the reason why the quantities $I_\alpha(\mathcal{Q}|\mathcal{P})$ with $\alpha \leqq 0$ are not excluded by the postulates. However $I_\alpha(\mathcal{Q}|\mathcal{P})$ can be considered to be a reasonable measure of information only if $\alpha > 0$. Thus to exclude the quantities $I_\alpha(\mathcal{Q}|\mathcal{P})$ with $\alpha \leqq 0$ we have to add a postulate of continuity. For instance, we may add

POSTULATE 11. $\lim_{\epsilon\to+0} [(p, \epsilon)|(p, p)] = 0$ *for some p with* $0 < p < 1/2$.

Clearly postulates 6 through 11 characterize the quantities $I_\alpha(\mathcal{Q}|\mathcal{P})$ with $\alpha > 0$.

It remains to characterize $I_1(\mathcal{Q}|\mathcal{P})$ instead of all $I_\alpha(\mathcal{Q}|\mathcal{P})$. Of course this can be done by replacing postulate 10 by another postulate which demands that $I(\mathcal{Q}_1 \cup \mathcal{Q}_2|\mathcal{P}_1 \cup \mathcal{P}_2)$ be the weighted *arithmetic* mean of $I(\mathcal{Q}_1|\mathcal{P}_1)$ and $I(\mathcal{Q}_2|\mathcal{P}_2)$, that is, by

POSTULATE 10′. *If $I(\mathcal{Q}_1|\mathcal{P}_1)$ and $I(\mathcal{Q}_2|\mathcal{P}_2)$ are defined, and $w(\mathcal{P}_1) + w(\mathcal{P}_2) \leqq 1$ and $w(\mathcal{Q}_1) + w(\mathcal{Q}_2) \leqq 1$, and if the correspondence between the elements of $\mathcal{P}_1 \cup \mathcal{P}_2$ and $\mathcal{Q}_1 \cup \mathcal{Q}_2$ is that induced by the correspondence between the elements of $\mathcal{P}_1[P_2]$ and $\mathcal{Q}_1[Q_2]$, then we have*

$$I(\mathcal{Q}_1 \cup \mathcal{Q}_2|\mathcal{P}_1 \cup \mathcal{P}_2) = \frac{w(\mathcal{Q}_1)I(\mathcal{Q}_1|\mathcal{P}_1) + w(\mathcal{Q}_2)I(\mathcal{Q}_2|\mathcal{P}_2)}{w(\mathcal{Q}_1) + w(\mathcal{Q}_2)}. \tag{3.33}$$

The proof of theorem 3 contains the proof of

THEOREM 4. *If $I(\mathcal{Q}|\mathcal{P})$ satisfies postulates* 6, 7, 8, 9, and 10′, *then* $I(\mathcal{Q}|\mathcal{P}) = I_1(\mathcal{Q}|\mathcal{P})$, *where $I_1(\mathcal{Q}|\mathcal{P})$ is defined by* (3.7).

Another way of characterizing $I_1(\mathcal{Q}|\mathcal{P})$ is to retain postulate 10 but add

POSTULATE 12. *If* $\mathcal{P} = (p_1, p_2, \cdots, p_n)$, $\mathcal{Q} = (q_1, q_2, \cdots, q_n)$, *and* $\mathcal{R} = (r_1, r_2, \cdots, r_n)$ *are generalized distributions such that*

$$\frac{r_k}{q_k} = \frac{q_k}{p_k}, \qquad k = 1, 2, \cdots, n, \tag{3.34}$$

then we have

$$I(\mathcal{Q}|\mathcal{P}) + I(\mathcal{Q}|R) = 0. \tag{3.35}$$

It is easy to see that only $I(\mathcal{Q}|\mathcal{P}) = I_1(\mathcal{Q}|\mathcal{P})$ satisfies postulates 6, 7, 8, 9, 10, and 12.

4. Information-theoretical proof of a limit theorem on Markov chains

The idea of using measures of information to prove limit theorems of probability theory is due to Linnik [9]. In this section we shall show how this method works in a very simple case.

Let us consider a stationary Markov chain with a finite number of states. Let p_{jk} for $j, k = 1, 2, \cdots, N$ denote the transition probability in one step and $p_{jk}^{(n)}$ the transition probability in n steps from state j to state k. We restrict ourselves to the simplest case when *all transition probabilities* p_{jk} *are positive.* In this case, as is well known, we have

$$(4.1) \qquad \lim_{n \to +\infty} p_{jk}^{(n)} = p_k, \qquad j, k = 1, 2, \cdots, N,$$

where the limits p_k are all positive and satisfy the equations

$$(4.2) \qquad \sum_{j=1}^{N} p_j p_{jk} = p_k, \qquad k = 1, 2, \cdots, N,$$

and

$$(4.3) \qquad \sum_{k=1}^{N} p_k = 1.$$

Our aim is to give a new proof of (4.1) by the use of the measure of information $I_1(\mathcal{Q}|\mathcal{P})$. The fact that the system of equations (4.2) and (4.3) has a solution $(p_1, p_2, \cdots, p_N)$ consisting of positive numbers can be deduced by a well-known theorem of matrix theory. In proving (4.1) we shall take it for granted that such numbers p_k exist. Let us put $\mathcal{P} = (p_1, p_2, \cdots, p_N)$ and $\mathcal{P}_j^{(n)} = (p_{j1}^{(n)}, p_{j2}^{(n)}, \cdots, p_{jN}^{(n)})$ and consider the amounts of information

$$(4.4) \qquad I_1(\mathcal{P}_j^{(n)}|\mathcal{P}) = \sum_{k=1}^{N} p_{jk}^{(n)} \log_2 \frac{p_{jk}^{(n)}}{p_k}.$$

According to the definition of transition probabilities, we have

$$(4.5) \qquad p_{jk}^{(n+1)} = \sum_{l=1}^{N} p_{jl}^{(n)} p_{lk}.$$

Now let us introduce the notation

$$(4.6) \qquad \pi_{lk} = \frac{p_l p_{lk}}{p_k}.$$

The probabilistic meaning of the numbers π_{lk} is clear: π_{lk} is the conditional probability for the chain's being in state l, under the condition that at the next step it will be in state k, provided that the initial distribution is the stationary distribution given by the numbers $p_1, p_2, \cdots, p_N$. The conditional probabilities π_{lk} are often called the "backward" transition probabilities of the Markov chain. Now we have clearly $\sum_{l=1}^{N} \pi_{lk} = 1$ for $k = 1, 2, \cdots, N$ and by (4.5)

$$(4.7) \qquad I_1(\mathcal{P}_j^{(n+1)}|\mathcal{P}) = \sum_{k=1}^{N} p_k \left[\sum_{l=1}^{N} \pi_{lk} \left(\frac{p_{jl}^{(n)}}{p_l}\right)\right] \log_2 \left[\sum_{l=1}^{N} \pi_{lk} \left(\frac{p_{jl}^{(n)}}{p_l}\right)\right].$$

Applying Jensen's inequality [7] to the convex function $x \log_2 x$, for each value of k, we obtain from (4.7)

$$I_1(\mathcal{P}_j^{(n+1)}|\mathcal{P}) \leqq \sum_{k=1}^{N} p_k \sum_{l=1}^{N} \pi_{lk} \frac{p_{jl}^{(n)}}{p_l} \log_2 \frac{p_{jl}^{(n)}}{p_l}. \tag{4.8}$$

Taking into account the fact that

$$\sum_{k=1}^{N} p_k \pi_{lk} = p_l, \tag{4.9}$$

it follows from (4.8) that

$$I_1(\mathcal{P}_j^{(n+1)}|\mathcal{P}) \leqq I_1(\mathcal{P}_j^{(n)}|\mathcal{P}). \tag{4.10}$$

Thus the sequence $I_1(\mathcal{P}_j^{(n)}|\mathcal{P})$ is decreasing, and as $I_1(\mathcal{P}_j^{(n)}|\mathcal{P}) \geqq 0$, the limit

$$L = \lim_{n \to +\infty} I_1(\mathcal{P}_j^{(n)}|\mathcal{P}) \tag{4.11}$$

exists. We shall show now that $L = 0$ and simultaneously that (4.1) holds. As the number of states is finite, we can find a sequence $n_1 < n_2 < \cdots < n_s < \cdots$ of positive integers, such that the limits

$$\lim_{s \to +\infty} p_{jk}^{(n_s)} = q_{jk}, \qquad k = 1, 2, \cdots, N, \tag{4.12}$$

exist. As $\sum_{k=1}^{N} p_{jk}^{(n)} = 1$, we have evidently

$$\sum_{k=1}^{N} q_{jk} = 1. \tag{4.13}$$

Let us put further

$$q'_{jk} = \sum_{l=1}^{N} q_{jl} p_{lk}, \qquad k = 1, 2, \cdots, N, \tag{4.14}$$

and put for the sake of brevity $\mathcal{Q}_j = (q_{j1}, q_{j2}, \cdots, q_{jN})$ and $\mathcal{Q}'_j = (q'_{j1}, q'_{j2}, \cdots, q'_{jN})$. Clearly we have

$$\lim_{s \to +\infty} I_1(\mathcal{P}_j^{(n_s)}|\mathcal{P}) = I_1(\mathcal{Q}_j|\mathcal{P}) = L \tag{4.15}$$

and

$$\lim_{s \to +\infty} I_1(\mathcal{P}_j^{(n_s+1)}|\mathcal{P}) = I_1(\mathcal{Q}'_j|\mathcal{P}) = L. \tag{4.16}$$

Again using Jensen's inequality, exactly as in proving (4.10), we have

$$I_1(\mathcal{Q}'_j|\mathcal{P}) = \sum_{l=1}^{N} p_k \left[\sum_{l=1}^{N} \pi_{lk} \left(\frac{q_{jl}}{p_l}\right)\right] \log_2 \left[\sum_{l=1}^{N} \pi_{lk} \left(\frac{q_{jl}}{p_l}\right)\right] \leqq I_1(\mathcal{Q}_j|\mathcal{P}) \tag{4.17}$$

with equality holding in (4.17) only if $q_{jl}/p_l = c$ for $l = 1, 2, \cdots, N$, where c is a constant. But by (4.16) it follows that there is equality in (4.17), and thus we have

$$q_{jl} = cp_l, \qquad l = 1, 2, \cdots, N. \tag{4.18}$$

Notice that here we have made essential use of the supposition that all p_{jk} and thus all π_{lk} are positive. In view of (4.3) and (4.13) the constant c in (4.18) is equal to 1, and therefore $\mathcal{Q}_j = \mathcal{P}$. It follows from (4.15) that

$$L = I_1(\mathcal{P}|\mathcal{P}) = 0. \tag{4.19}$$

We have incidentally proved that (4.1) holds, as we have shown that if for an arbitrary subsequence n_s we have (4.12) then necessarily $q_{jl} = p_l$ for $l = 1, 2, \cdots, N$. But if (4.1) were false, we could find a subsequence n_s of integers such that (4.12) holds with $q_{jl} \not\equiv p_l$.

It is clear from the above proof that instead of the quantities (4.4) we could have used the analogous sums

$$\sum p_{jk}^{(n)} f\left(\frac{p_{jk}^{(n)}}{p_k}\right), \tag{4.20}$$

where $f(x)$ is any function such that $xf(x)$ is strictly convex. Thus for instance we could have taken $f(x) = x^{\alpha-1}$ with $\alpha > 1$ or $f(x) = -x^{\alpha-1}$ with $0 < \alpha < 1$. This means that instead of the measure of information of the first order, we could have used the measure of information of any order $\alpha > 0$, and deduce (4.1) from the fact that $\lim_{n\to+\infty} I_\alpha(\mathcal{P}_j^{(n)}|\mathcal{P}) = 0$.

In proving limit theorems of probability theory by considering measures of information, it is usually an advantage that one can choose between different measures. In the above simple case each measure $I_\alpha(\mathcal{Q}|\mathcal{P})$ was equally suitable, but in other cases the quantity $I_2(\mathcal{Q}|\mathcal{P})$, for example, is more easily dealt with than the quantity $I_1(\mathcal{Q}|\mathcal{P})$. The author intends to return to this question, by giving a simplified version of Linnik's information-theoretical proof of the central limit theorem, in another paper.

REFERENCES

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication,* Urbana, University of Illinois Press, 1949.

[2] D. K. Fadeev, "Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas," *Arbeiten zur Informationstheorie I,* Berlin, Deutscher Verlag der Wissenschaften, 1957, pp. 85–90.

[3] A. Feinstein, *Foundations of Information Theory,* New York, McGraw-Hill, 1958.

[4] P. Erdös, "On the distribution function of additive functions," *Ann. of Math.,* Vol. 47 (1946), pp. 1–20.

[5] A. Rényi, "On a theorem of Erdös and its application in information theory," *Mathematica,* Vol. 1 (1959), pp. 341–344.

[6] G. A. Barnard, "The theory of information," *J. Roy. Statist. Soc., Ser. B,* Vol. 13 (1951), pp. 46–64.

[7] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities,* Cambridge, Cambridge University Press, 1934.

[8] S. Kullback, *Information Theory and Statistics,* New York, Wiley, 1959.

[9] Yu. V. Linnik, "An information theoretical proof of the central limit theorem on Lindeberg conditions," *Teor. Veroyatnost. i Primenen.,* Vol. 4 (1959), pp. 311–321. (In Russian.)