MATHEMATICAL PROBLEMS IN THE SHANNON THEORY OF OPTIMAL CODING OF INFORMATION

R. L. DOBRUSHIN
MOSCOW UNIVERSITY

1. Introduction

1.1. The term "information theory" which has become fashionable during the last decade has as yet no unique interpretation. In technical and cybernetics literature, information theory is usually taken to mean the totality of the applications of mathematical methods to the problem of input, processing, storage, and transmission of information. From the mathematical point of view these applications have little in common since they are based on methods belonging to very diverse branches of mathematics. Such an interpretation of information theory covers the whole of mathematical statistics, much of the theory of random processes, the recently developed investigations into the power of the ϵ-net in functional spaces [54], which is regarded as an estimate of the quantity of information given by an element of this space, estimates of the "complexity" of the algorithms of mathematical analysis [2] and [96], and so on.

But within information theory, in the wide sense of this term, an important place is occupied by a young discipline which is also (particularly in mathematical literature) often called information theory. To be explicit, we shall call it the Shannon theory of optimal coding of information. The reason is that everything in this discipline is the direct development of ideas contained in the remarkable and fundamental work of C. E. Shannon [78], and that the Shannon theory investigates means of transmitting information based on an optimal choice of methods of coding and decoding information. Moreover, the special characteristic of this theory is the possibility of greatly varying the method of coding information. In those cases where the coding method is rigidly fixed, the Shannon theory is not best suited to the problem, but rather the more usual methods of contemporary mathematical statistics should be employed. This is the case, for example, in the majority of statistical problems, when it is not within the power of the statistician to alter the procedure by which the relevant information was selected, and the only question which arises is the choice of an optimal decoding method ("the decision rule").

1.2. It has recently become clear that the Shannon theory of coding information may be of interest to mathematicians not only because of the importance

in applications, but also because it gives rise to many original mathematical problems which are interesting quite apart from their technical applicability. Unexpected points of contact between information theory and other mathematical sciences have been discovered. Thus, the methods of group theory have proved useful in constructing codes (see section 5), and a connection between the general problems of information theory and of inference in stochastic processes has been clarified (see the work of M. Pinsker [72]).

Work under the leadership of A. N. Kolmogorov has shown that the ideas of the Shannon information theory are also useful in the classical domain of mathematical analysis—in the theory of dynamical systems [52], [53], and the theory of functions [54]. As is often the case in young branches of mathematics, some of the problems of Shannon information theory are completely elementary in their formulation, but their solution proves to be nontrivial (and sometimes extremely difficult). On the other hand, attempts to carry out a precise examination of the Shannon information theory under sufficiently general conditions result in a complicated and rather unwieldy structure which is based, as is always the case in probability theory, on abstract methods of set theory and measure theory. The unwieldiness of these structures even led some mathematicians to wonder, as in [22], whether it would not be wiser to abandon the axiomatization of information theory. It seems that the necessity of a precise foundation for the ideas of the Shannon information theory is no less unavoidable than, for example, its comparable necessity in the theory of stochastic processes; and that even the most general constructions of information theory are no more complicated and abstract than, say, the constructions of the contemporary theory of Markov processes (see in particular [23]). The doubts in this respect may perhaps be explained by the fact that, while the original theory of Markov chains has taken about fifty years to reach the present state of development of the theory of Markov processes, the Shannon information theory has covered a comparable distance in at most ten years. This rapid development is a reflection of the general tempo of contemporary science.

1.3. The aim of this paper is to give a survey of the basic directions of the Shannon theory of optimal coding of information from the common point of view. Naturally the author cannot hope to clarify all the important questions to the same extent. Most attention will be paid to questions connected with the author's work, including his most recent research. At the same time, some no less important questions will remain on the second plane. Particular attention will be devoted to publicizing questions of the Shannon information theory which are awaiting solution and which vary greatly in their difficulty and in the concreteness of their formulation. In order to single them out in the text, without encumbering the exposition, Roman numerals are used at the left margin.

The bibliography at the end of the article is fairly complete as far as articles published in mathematical journals are concerned. From the enormous tech-

nical literature only that which bears directly on the content of this paper has been selected.

2. The fundamental Shannon problem

2.1. We shall start with a formulation of the fundamental problem solved by C. E. Shannon, using as a basis the presentation proposed by A. N. Kolmogorov [51], and developed in the author's papers, [19], and preliminary publications [15], [18]. We feel that Kolmogorov's concept differs from previous work on the subject (A. Feinstein [29]; A. I. Khinchin [49]; M. Rosenblatt [76]; J. Nedoma [62]; J. Wolfowitz [97], [98], [100]; D. Blackwell, L. Breiman, A. J. Thomasian [8]) in that it is sufficiently general to encompass all the cases of practical interest and at the same time simpler and physically more natural than, for example, the concept proposed by A. Perez [65].

2.2. The following terminology and notation will be used. We shall denote by $X, Y, Z, \cdots$ sets of elements, and by $S_X, S_Y, S_Z, \cdots$ σ-fields of subsets of $X, Y, Z, \cdots$ respectively. A measurable space consists of a couple (X, S_X); and a measurable space (X, S_X) will be called a real line if X is the real line R^1 and S_X is the Borel σ-field in X. A random variable ξ *taking on values in a measurable space* (X, S_X) is a function defined on some probability space (Z, S_Z, P) *whose range is in* X and is S_X measurable, that is, for any set $A \in S_X$, the set $\xi^{-1}(A)$ is in S_Z.

2.3. *The transmitting system or transmitter* (Q, V) will be described by specifying the following three elements.

1) Two measurable spaces (Y, S_Y), $(\tilde{Y}, S_{\tilde{Y}})$. We shall call them the *spaces of input and output signals* of the transmitter.

2) A function $Q(y, A)$, defined for all $y \in Y$ and $A \in S_{\tilde{Y}}$ which is S_Y measurable for every fixed $A \in S_{\tilde{Y}}$ and is a probability measure on $S_{\tilde{Y}}$ for every fixed $y \in Y$. We shall call $Q(y, A)$ *the transition function.*

3) A subset V of the set of all probability measures on the product σ-field $S_Y \times S_{\tilde{Y}}$. This subset will be called the *restriction on the distribution of the signals.*

(Sometimes, for example, in the author's paper [15], what is defined above as a transmitting system is referred to as a channel. We prefer to reserve the term channel for a somewhat different concept to be encountered later.) We shall say that two random variables η and $\tilde{\eta}$ are *connected by the transmitter* if η and $\tilde{\eta}$ take on values in (Y, S_Y), and $(\tilde{Y}, S_{\tilde{Y}})$ respectively, the joint distribution of η and $\tilde{\eta}$ is in V, and for any $A \in S_{\tilde{Y}}$, the conditional probability $P\{\tilde{\eta} \in A | \eta\}$ is given by

$$P\{\tilde{\eta} \in A | \eta\} = Q(\eta, A) \tag{1}$$

with probability one.

From the intuitive point of view, Y is the totality of what is transmitted by the transmitter and $\tilde{Y}$ is the totality of what is received by the signal receiver.

(In applications, the spaces Y and $\tilde{Y}$ often coincide.) If the random value η of the input signal is given, equation (1) enables us to find the conditional distribution of the output signal $\tilde{\eta}$. Thus $Q(y, \cdot)$ is the distribution of the output signal if the signal y is the input. Finally, the introduction of the restriction V is related to the fact that in many applications it is impossible to consider the distribution of the input and output signals as arbitrary. Typical of this is the case where it is assumed that the mean square value (average power) of the input signal does not exceed a given constant. If, as occurred in most earlier work, the researcher does not wish to introduce such a restriction, then V must be taken to mean the totality of all the probability measures over $(Y \times \tilde{Y}, S_Y \times S_{\tilde{Y}})$.

In what follows the restriction V is not given arbitrarily, but it is assumed that it is defined as follows: Given N real-valued, $S_Y \times S_{\tilde{Y}}$ measurable function $\pi_i(y, \tilde{y})$ where $y \in Y$ and $\tilde{y} \in \tilde{Y}$, and a set $\overline{V}$ in the N-dimensional Euclidean space, the distribution of $(\eta, \tilde{\eta})$ belongs to V only if the vector of the mathematical expectations satisfies the condition:

$$(E\pi_1(\eta, \tilde{\eta}), E\pi_2(\eta, \tilde{\eta}), \cdots, E\pi_N(\eta, \tilde{\eta}) \in \overline{V}). \tag{2}$$

The assumption that the restriction V is given by equation (2) appears sufficient for all particular cases which are interesting from the practical point of view. The possible exceptions are of the following type. Let (Y, S_Y) be the space of real-valued functions $y(t)$ where $t \in [a, b]$ with the ordinary σ-algebra of measurable sets. Then the variable η with values in the space of the input signals turns out to be the random process $\eta(t)$ where $t \in [a, b]$. It is assumed that condition V is reduced to the requirement $E\{\eta^2(t)\} \leqq p_s$ for all $t \in [a, b]$, where p_s is a given constant. A restriction of this kind corresponds to a continuum of functions $\pi(\cdot, \cdot)$ and cannot be given in the form (2).

I The proof of Shannon's theorems, stated in section 3, applicable to this and similar cases, still is an open problem.

Some important particular classes of transmitting systems are enumerated below.

2.4. A transmitting system is called a *segment of length n of a homogeneous memoryless channel*, if there exist measurable spaces (Y_0, S_{Y_0}), $(\tilde{Y}_0, S_{\tilde{Y}_0})$ and a transition function $Q_0(y, A)$ defined for $y \in Y_0$ and $A \in S_{\tilde{Y}_0}$ such that the spaces of the input and output signals of the transmitter are the nth powers $(Y_0^n, S_{Y_0}^n)$, $(\tilde{Y}_0^n, S_{\tilde{Y}^0}^n)$ of the spaces (Y_0, S_{Y_0}), $(\tilde{Y}_0, S_{\tilde{Y}_0})$, that is to say, products of n copies of these spaces, and for any $y = (y_1, \cdots, y_n) \in Y_0^n$ the transition function $Q(y, \cdot)$ is given by

$$Q(y, \cdot) = Q_0(y_1, \cdot) \times Q_0(y_2, \cdot) \times \cdots \times Q_0(y_n, \cdot), \tag{3}$$

that is, it is the Cartesian product of the corresponding measures. Finally, the set of distributions V_0 must be such that the distribution of the variables $\eta = (\eta_1, \cdots, \eta_n)$ and $\tilde{\eta} = (\tilde{\eta}_1, \cdots, \tilde{\eta}_n)$ belongs to V if and only if the pairs $(\eta_i, \tilde{\eta}_i)$ have distributions belonging to V_0, for all $i = 1, \cdots, n$. Intuitively this definition means that the ith component $\tilde{\eta}_i$ of the output signal depends only on the

ith component η_i of the input signal, and that the restriction is reduced to restrictions imposed on each pair of components. If the spaces Y_0, $\tilde{Y}_0$ consist of a finite or denumerable set of elements, then the transition function $Q_0(\cdot\,,\cdot)$ can be defined by the matrix of the transition probabilities (as happens in the theory of Markov processes). In that case the restriction V is usually absent. Then the memoryless channel is called *finite* (or, correspondingly, *denumerable*).

The particular case of the memoryless channel, which is by far the most worked out and the most important in applications, is the *memoryless channel with additive noise.* In this case (Y_0, S_{Y_0}) and $(\tilde{Y}_0, S_{\tilde{Y}_0})$ are real lines, and $Q_0(y, \cdot\,)$ coincides with the distribution of the variable $\zeta + y$, where the noise ζ has a given distribution and where $E\{\zeta^2\} = p_n$ is called the *noise power.* In this case the condition $E\{\eta_i^2\} \leqq p_s$ is usually taken for V_0 where p_s is called the *signal power.* Such a channel is said to be *Gaussian*, if ζ has a Gaussian distribution.

2.5. A considerably more general class of transmitters is to be found in the concept of a *homogeneous channel with discrete time.* For the study of such a channel it is necessary to specify the measurable spaces (Y_0, S_{Y_0}) and $(\tilde{Y}_0, S_{\tilde{Y}_0})$ of the input and output signals at any time, as well as the measurable space (F, S_F), called the *state space* of the channel, and the *transition function* $Q_0(y_0, f, A)$ of the channel defined for $y_0 \in Y_0, f \in F$, and $A \in S_{\tilde{Y}} \times S_F$, which is measurable with respect to $S_{Y_0} \times S_F$ for each fixed A and is a probability measure on $\tilde{Y}_0 \times F$ for fixed y_0 and f. Such a channel can be regarded from an intuitive point of view as a pair of sequences of variables $\cdots, \eta_{-1}, \eta_0, \eta_1, \cdots$, and $\cdots, \tilde{\eta}_{-1}, \tilde{\eta}_0, \tilde{\eta}_1, \cdots$, taking on values in the spaces (Y_0, S_{Y_0}) and $(\tilde{Y}_0, S_{\tilde{Y}_0})$ respectively, together with a sequence of variables $\cdots, \phi_{-1}, \phi_0, \phi_1, \cdots$ with values in (F, S_F). Moreover given $\eta_i = y$ and $\phi_{i-1} = f$, the conditional distribution for the pairs $\tilde{\eta}_i, \phi_i$ is given by $Q_0(y, f, \cdot\,)$ and does not depend on further information about $\phi_k, \eta_k, \tilde{\eta}_k$ at earlier times. Thus the states of the channel describe its memory of the past.

More precisely, a transmitter, called a *segment of length n of the homogeneous channel with discrete time,* is associated with each *initial probability distribution* $p_0(\cdot\,)$ on (F, S_F) and each integer n. Here the spaces of the input and output signals of the transmitter are the powers $(Y_0^n, S_{Y_0}^n)$ and $(\tilde{Y}_0^n, S_{Y_0}^n)$. To avoid complicating the exposition, the transition function $Q(y, A)$ for $y \in Y_0^n$ and $A \in S_{\tilde{Y}_0}^n$ will be constructed only in the particular case of a transmitter for which measures $\mu_{\tilde{Y}_0}(\tilde{A})$ on $(\tilde{Y}_0, S_{\tilde{Y}_0})$ and $\mu_F(B)$ on (F, S_F) exist, such that the transition function $Q_0(y_0, f_0, A)$ is given by the formula

$$Q_0(y_0, f_0, A) = \int_A q_0(y_0, f, \tilde{y}_0, f')\mu_{\tilde{Y}_0}(d\tilde{y}_0)\mu_F(df') \tag{4}$$

that is to say, Q_0 is determined by the density $q_0(y_0, f, \tilde{y}_0, f')$. Then,

$$Q(y, \tilde{A}) = \int_{\tilde{A}} q(y, \tilde{y})\mu_{\tilde{Y}_0}^n(d\tilde{y}), \qquad y \in Y_0^n, \tilde{y} \in \tilde{Y}_0^n \tag{5}$$

where

$$y = (y_1, \cdots, y_n), \qquad \tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n) \tag{6}$$

and

$$q(y_1, \cdots, y_n; \tilde{y}_1, \cdots, \tilde{y}_n) = \int_F \cdots \int_F q_0(y_1, f_0, \tilde{y}_1, f_1) \cdots q_0(y_n, f_{n-1}, \tilde{y}_n, f_n) p_0(df_0) \mu_F(df_1) \cdots \mu_F(df_n). \tag{7}$$

From the intuitive point of view, representation (6) can be elucidated by saying that the integrand in (6) is the density of the joint probability distribution of the sequence of states of the channel and the sequence of output signals of the channel, for a given sequence of input signals to the channel. This density is averaged over all possible sequences of states of the channel. Furthermore, it is not difficult to give a more general definition of $Q(\cdot\,, \cdot)$ for the case when $Q_0(\cdot\,, \cdot)$ is not determined by a density. Finally, in a manner analogous to that used in the theory of Markov processes, it is easy to extend the definition of a segment of the channel to the case of channels which are not homogeneous in time and to channels with continuous time (compare similar constructions in [19], section 1.8).

If the state space F of the channel is finite, then the channel is said to have *a finite number of states.* A memoryless channel (see section 2.4) can be interpreted as a channel in which the state space consists of a single element.

The definition given above was prompted by ideas from the contemporary theory of automata, [86], and it is of sufficient generality to include a large proportion of the physically interesting cases. For channels with a finite memory and finite signal spaces a very similar definition was studied by D. Blackwell, L. Breiman, and A. J. Thomasian [8]. An analogous definition was formulated in 1957 by A. N. Kolmogorov in his seminar at Moscow University.

Note that our definition would not be any more general if we allowed the output signal to depend also on the input and output signals at earlier instances of time. In fact this dependence can always be avoided by extending the state space of the channel, by defining as the state of the channel the totality consisting of the previous state together with the values of the input and output signals at all earlier times. (This idea is elaborated in [8], where a similar method shows that channels with a finite memory in the Khinchin-Feinstein sense [49], [31] are a special case of channels with a finite number of states.) This method enables one to convert into a channel a wide class of transmitters which operate in time and possess the property of *lack of anticipation.* (The corresponding definition is given by analogy with the definitions in [49] and [31].) However, in this case the property of homogeneity in time may be violated, though this does not occur if we introduce the analogous concepts not for finite segments of the channel but for channels operating over an infinite time interval.

2.6. In applications it is sometimes natural to assume that the channel depends on slowly changing parameters. As a limiting case, it may be supposed

that during the time of transmission the parameters do not change, but that their exact value is unknown and is a random variable with a given probability distribution. Thus the concept of a *channel with a random parameter* is reached. Such a channel has the following general definition: it is assumed (in the notation of section 2.5) that we are given a measurable space (B, S_B) of *values of the parameter* and a measurable function $\beta(f)$, $f \in F$, with values in (B, S_B) such that the transition density $q_0(y_0, f, \tilde{y}_0, f')$ is zero whenever $\beta(f) \neq \beta(f')$. (It is easy to give an analogous definition in the case when the density does not exist.) Then, for a given initial distribution $p_0(\cdot)$, it is natural to call *a random parameter of the channel* the random variable $\beta(\phi)$, where ϕ has the distribution $p_0(\cdot)$. From the intuitive point of view, $\beta(\phi)$ is a random parameter which does not change with time. It is then natural to introduce the concept of *a channel conditional on the value* of $b \in B$. The state space of this channel is the inverse image of the point b under the mapping $\beta(f)$ of F into B, and the transition density coincides with $q_0(y_0, f, \tilde{y}_0, f')$ with $\beta(f) = \beta(f') = b$.

2.7. The transmitter is called *Gaussian* if the spaces of input and output signals are spaces of real-valued functions, and the transition function $Q(y_i, \cdot)$ yields, for any fixed y a conditional Gaussian distribution (an infinite set of random variables is said to possess a conditional Gaussian distribution if any finite subset possesses, under an arbitrary condition, a finite-dimensional Gaussian distribution whose second moments are independent of the condition and whose first moments depend linearly on this condition); and finally the restriction V is imposed only on the first and second moments of the random variables which are linked by the transmitter. Such a transmitter assigns to each Gaussian input process η another Gaussian output process $\tilde{\eta}$.

An important role is played, particularly in applications, by receiving-transmitting systems with a finite bandwidth. An attempt to formulate the appropriate concept mathematically, and to prove Shannon's well known formula (46) in the case of a channel operating for an infinite time, meets with serious difficulties related to the fact that stationary processes with bounded spectra are singular (deterministic) and therefore yield no information. These difficulties may be avoided by the introduction of the following model of *a transmitter with finite bandwidth* $[\lambda, \lambda + W]$ *and additive noise for transmission over time* T. This model is perhaps a little clumsy from the point of view of the general theory of stochastic processes, but it reflects the real physical situation sufficiently well (compare [41]).

Here the spaces (Y, S_Y) and $(\tilde{Y}, S_{\tilde{Y}})$ are the spaces of the functions $y(t)$ for $t \in [0, T]$ with σ-fields of measurable sets introduced in the usual way. Further, the noise is given as the random process $\zeta(t)$ for $t \in [0, T]$. The operator A takes the function $v(t)$ for $t \in [0, T]$into the function

$$Av(t) = \sum_{\lambda_0 \leq k/T \leq \lambda_0 + W} \left(c_k \cos \frac{2\pi kt}{T} + d_k \sin \frac{2\pi kt}{T} \right) \tag{8}$$

where the c_k and d_k are Fourier coefficients of $v(t)$. Then, the measure $Q[y(t), \cdot]$

is the distribution of the random process $A[y(t) + \zeta(t)]$. The restriction V consists of the boundedness of the average power of the input signal; that is, it consists of the fact that

$$E\left\{\int_0^T \eta^2(t)dt\right\} \leqq p_s T \tag{9}$$

where p_s is a given constant, called the *average signal power*, The *average noise power* p_n is defined by

$$\frac{1}{T} E\left\{\int_0^T \zeta^2(t)dt\right\}. \tag{10}$$

We shall say that the noise is *white Gaussian* if

$$\zeta(t) = \sum_{\lambda_0 \leqq k/T \leqq \lambda_0 + W} \left(c_k \cos \frac{2\pi kt}{T} + d_k \sin \frac{2\pi kt}{T}\right), \tag{11}$$

where the c_k and d_k are independent Gaussian random variables with zero means and equal variances.

2.8. By the *message* (p, W) we shall mean the aggregate consisting of two measurable spaces (X, S_X) and $(\tilde{X}, S_{\tilde{X}})$, a probability distribution $p(\cdot)$ on the σ-field S_X and a subset W of the set of all the probability measures in the product σ-field $S_X \times S_{\tilde{X}}$. The spaces (X, S_X) and $(\tilde{X}, S_{\tilde{X}})$ will be called the spaces of *input and output values* of the message. The distribution $p(\cdot)$ will be called the *input distribution of the message*, and the subset W will be called the *condition of accuracy of reproduction*. The two random variables ξ and $\tilde{\xi}$ generate the message (p, W) if they take values in the spaces (X, S_X) and $(\tilde{X}, S_{\tilde{X}})$ respectively, if the distribution of the variable ξ coincides with $p(\cdot)$, and if the joint distribution of the pair belongs to W.

Thus, as is usual in information theory, the message being transmitted is regarded as random, with a given distribution $p(\cdot)$. Moreover, it is considered that the message obtained after transmission need not coincide exactly with the input message. (According to the definition used herein, even the spaces of the input and output message values need not in general coincide, although in the majority of applications they do coincide.) However, some bounds are indicated within which the received message may vary, depending on the transmitted message. This is the condition of the accuracy W.

It is assumed that the condition W is defined with the help of M functions $\rho_i(x, \tilde{x})$, for $x \in X$ and $\tilde{x} \in \tilde{X}$, and by a set $\overline{W}$ in a manner comparable with that by which equation (2) defines restriction V. The same remarks may be made here on the question of generality as were made on the generality of condition (2). In particular there remains an open question analogous to that mentioned in section 2.3.

2.9. An important particular case is that of a message with the *condition of perfect reproduction*. This is how we shall describe a message for which the spaces (X, S_X) and $(\tilde{X}, S_{\tilde{X}})$ coincide, and for which W is such that the pair $(\xi, \tilde{\xi})$ generate the message (p, W) if and only if $\xi = \tilde{\xi}$ with probability one.

In fact the research in a long series of papers started by Feinstein [29], and Khinchin [49] reduces to this very case under our treatment of Shannon's theorem. A more general treatment of communications, also introduced by Shannon [79] was made mathematically precise by Kolmogorov [51] and developed in the author's work. A recent paper of Shannon [84] investigates independently a particular class of messages which, in the terminology employed here, are called messages with a component-wise condition of accuracy of reproduction.

2.10. We shall assume that we are given measurable spaces (X_0, S_{X_0}) and $(\tilde{X}_0, S_{\tilde{X}_0})$ such that $(X, S_X) = (X_0^n, S_{X_0}^n)$ and $(\tilde{X}, S_{\tilde{X}}) = (\tilde{X}_0^n, S_{\tilde{X}_0}^n)$. Then the random variables ξ and $\tilde{\xi}$, taking on values in X and $\tilde{X}$ can be regarded as the sets $\xi = (\xi_1, \cdots, \xi_n)$, $\tilde{\xi} = (\tilde{\xi}_1, \cdots, \tilde{\xi}_n)$, where the ξ_i take on values in X_0, and $\tilde{\xi}_i$ in $\tilde{X}_0$. Then the variables ξ_i are called the *input components* of the message, and the $\tilde{\xi}_i$ the *output components* of the message. The condition W of accuracy of reproduction is *component-wise* if it is fulfillment or nonfulfillment for the pair $(\xi, \tilde{\xi})$ depends only on the pair-wise distribution of the pairs $(\xi_i, \tilde{\xi}_i)$, where $i = 1, \cdots, n$, but does not depend on the joint distributions of these pairs. The concept of a message *homogeneous in time* is introduced in a natural way. The component-wise condition of accuracy of reproduction will be called a *homogeneous condition of boundedness in mean square error*, if (X_0, S_{X_0}) and $(\tilde{X}_0, S_{X_0})$ are real lines and the condition consists of the fact that

$$E\{(\xi_i - \tilde{\xi}_i)^2\} \leqq p_d, \tag{12}$$

where p_d is a given constant called the *mean square error*. The *condition W of accuracy of reproduction is additive*, if each of the functions $\rho_i(x, \tilde{x})$ has the form

$$\rho_i(x, \tilde{x}) = \sum_{i=1}^{n} \rho_i^0(x_i^0, \tilde{x}_i^0), \tag{13}$$

where

$$x = (x_1^0, \cdots, x_n^0), \qquad \tilde{x} = (\tilde{x}_1^0, \cdots, \tilde{x}_n^0). \tag{14}$$

The input message has *independent components* if for the variable ξ with distribution $p(\cdot)$, the components $\xi_1, \cdots, \xi_n$ are independent variables.

2.11. Another important special case is that of a Gaussian message. Here it is assumed that the spaces of input and output message values are spaces of real-valued functions, so that the variables ξ and $\tilde{\xi}$ can be treated as collections of real variables.

The message is called *Gaussian* if the distribution $p(\cdot)$ gives a set of variables possessing joint Gaussian distributions, and condition W is imposed only on the first and second moments of the random variables under examination.

2.12. Now let the transmitter (Q, V) and the message (p, W) be given simultaneously. The name *encoding function* will be given to a function $P(x, A)$, for $x \in X$ and $A \in S_Y$, which for a fixed value x of the input message is a probability measure in the space Y of output signals and for fixed A is measurable with respect to S_X. The name *decoding function* will be given to a function $\tilde{P}(\tilde{y}, \tilde{A})$,

for $\tilde{y} \in \tilde{Y}$ and $\tilde{A} \in S_{\tilde{Y}}$, which for a fixed output signal $\tilde{y}$ is a measure on the space $\tilde{X}$ of output values of messages, and for a fixed $\tilde{A}$ is measurable with respect to $S_{\tilde{Y}}$. We shall say that *the message* (p, W) *can be transmitted by the transmitter* (Q, V) *with the help of the encoding function* $P(\cdot\ ,\ \cdot)$ *and the decoding function* $\tilde{P}(\cdot\ ,\ \cdot)$ if random variables ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$ can be constructed forming a Markov chain such that the pair ξ, $\tilde{\xi}$ generate the message (p, W), and the pair $(\eta, \tilde{\eta})$ are connected by the transmitter (Q, V), and, with probability one,

$$P\{\eta \in A|\xi\} = P(\xi, A), \qquad P\{\tilde{\xi} \in \tilde{A}|\tilde{\eta}\} = \tilde{P}(\tilde{\eta}, \tilde{A}). \tag{15}$$

In this case we shall say that the variables ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$ give the *method of transmission of the message* (p, W) *by the transmitter* (Q, V). Notice that the encoding function P and the decoding function $\tilde{P}$ uniquely determine the joint distribution of ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$.

We make here a few remarks on this definition. The use of the encoding function $P(\cdot\ ,\ \cdot)$ means from the intuitive point of view that if the message takes on the value x, then we transmit an input signal chosen with the distribution $P(x, \cdot)$. In most earlier papers only *nonrandomized encoding* was used, given by the function $f(x)$ for $x \in X$ with values in Y, such that the measure $P(x, \cdot)$ is concentrated at the point $f(x)$. *Nonrandomized decoding* is defined analogously. In practice, naturally, nonrandomized encoding and decoding is almost always used, but the introduction of randomization obviously enters in principle into the possibilities open to the communication system builder. From the mathematical point of view it simplifies the formulation of theorems. Let us suppose that the restriction V is omitted, the definition of condition W contains only the single function $\rho_1(\cdot\ ,\ \cdot)$, with $M = 1$, and $\overline{W} = [0, a]$. Then, *if the message* (p, W) *can be transmitted at all by the transmitter* (Q, V), *it can be transmitted with the help of nonrandomized encoding and decoding functions.* It is also not difficult to give examples showing that this assertion may be false in case $M = 2$. It can still be expected, however, that under very broad conditions (perhaps in some asymptotic sense, see section 3.10), some theorem may hold on the possibility of substituting nonrandomized for randomized encoding and decoding functions. III

The assumption that ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$ form a Markov chain is completely natural. Intuitively it means that in transmission the output signal depends only on the input signal and not on the value of the message encoded by it, and that in decoding only the output signal is used, and not the inaccessible input signal and message.

2.13. The *basic Shannon problem* can now be formulated. *For which pairs of the message* (p, W) *and the transmitter* (Q, V) *is it possible, and for which pairs is it impossible, to select coding and decoding functions such that* (p, W) *can be transmitted by the transmitter* (Q, V)? The solutions of these problems for different assumptions will be called *Shannon theorems.*

There also arises naturally a second problem of constructing, when transmission is possible, the encoding and decoding functions which realize this trans-

mission in the simplest and most effective way possible. The examination of these two problems, and also of their direct generalizations, forms at present the subject of the Shannon theory of the optimal coding of information.

3. Shannon's theorems

3.1. In his fundamental work, Shannon introduced quantities which enabled him to formulate an answer to the problem he raised. The principal one of these is a quantity called information.

Let the two random variables ζ and $\tilde{\zeta}$, taking on values in the measurable spaces (Z, S_Z) and $(\tilde{Z}, S_{\tilde{Z}})$ respectively, be given. Let $p_{\zeta\tilde{\zeta}}(D)$ for $D \in S_Z \times S_{\tilde{Z}}$, together with $p_\zeta(C)$ for $C \in S_Z$, and $p_{\tilde{\zeta}}(\tilde{C})$ for $\tilde{C} \in S_{\tilde{\zeta}}$ be their joint and marginal distributions. If the measure $p_{\zeta\tilde{\zeta}}(\cdot)$ is not absolutely continuous with respect to the product of the measures $p_\zeta \times p_{\tilde{\zeta}}(\cdot)$, then we define *information* by

$$I(\zeta, \tilde{\zeta}) = +\infty. \tag{16}$$

But if $p_{\zeta\tilde{\zeta}}(\cdot)$ is absolutely continuous with respect to $p_\zeta \times p_{\tilde{\zeta}}(\cdot)$, then $a_{\zeta\tilde{\zeta}}(\cdot)$ denotes the Radon-Nikodym density of the measure $p_{\zeta\tilde{\zeta}}(\cdot)$ with respect to $p_\zeta \times p_{\tilde{\zeta}}(\cdot)$ and the name *information density* is given to the function

$$i_{\zeta\tilde{\zeta}}(\cdot) = \log a_{\zeta\tilde{\zeta}}(\cdot). \tag{17}$$

We shall give the name *information* to

$$\begin{aligned} I(\zeta, \tilde{\zeta}) &= \int_{Z\times\tilde{Z}} i_{\zeta\tilde{\zeta}}(z, \tilde{z}) p_{\zeta\tilde{\zeta}}(dz, d\tilde{z}) \\ &= \int_{Z\times\tilde{Z}} a_{\zeta\tilde{\zeta}}(z, \tilde{z}) \log a_{\zeta\tilde{\zeta}}(z, \tilde{z}) p_\zeta(dz) p_{\tilde{\zeta}}(d\tilde{z}). \end{aligned} \tag{18}$$

If (Z, S_Z) and $(\tilde{Z}, S_{\tilde{Z}})$ are real lines, and the distributions $p_{\zeta\tilde{\zeta}}(\cdot)p_\zeta(\cdot)$, and $p_{\tilde{\zeta}}(\cdot)$ are given by the densities $q_{\zeta\tilde{\zeta}}(z, \tilde{z})$, $q_\zeta(z)$, and $q_{\tilde{\zeta}}(\tilde{z})$, then definition (18) takes the more classical form

$$I(\zeta, \tilde{\zeta}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} q_{\zeta\tilde{\zeta}}(z, \tilde{z}) \log \frac{q_{\zeta\tilde{\zeta}}(z, \tilde{z})}{q_\zeta(z) q_{\tilde{\zeta}}(\tilde{z})} \, dz \, d\tilde{z}. \tag{19}$$

We shall give the name *entropy* of the random variable ζ to its information with respect to itself, namely

$$I(\zeta, \zeta) = H(\zeta). \tag{20}$$

Frequently instead of expressing the definition of information as given in equation (18), it is expressed as the limit (or upper bound) of a certain sequence of integral sums. The relationship between the different possible definitions of information is investigated in detail in a series of papers by Perez [64]; Gelfand, Kolmogorov and Yaglom [36]-[38]; Tzian-Tze-Peia [94]; and Dobrushin [19]. (The somewhat belated paper of Fleischer [34] reflects the insufficient contact between Soviet and American investigators in this field.)

3.2. We feel that the main evidence for the scientific importance of the concepts of information and entropy is the fact that with their help a solution to Shannon's problem can be obtained. There exist separate and, up to now, unconnected indications that the concepts of information and entropy may be of value, quite apart from the circumstances described in section 2 (see section 8). However, on the whole it must be said that the numerous attempts to employ the mathematical concepts of information theory and entropy in places where the word information is applied in its nonspecialized treatment only compromise these concepts. From this point of view the very popular axiomatic definition of entropy, which was introduced at first by Shannon in [79], and later simplified by other writers (Khinchin [48], and Faddeev [28]), played even a negative role. The point is that not all the axioms which enter into the definition of entropy are equally natural (for example, the formula of conditional entropy), and also that, as Kolmogorov notes, it is not a priori clear whether there exists a unique one-dimensional numerical characteristic which adequately reflects the properties of such a complicated phenomenon that information turns out to be in a diverse real situation. Nevertheless, it would be interesting to give also an axiomatic definition of information. Obviously the basic axioms here must be the nonnegative properties of information, equation (21); the formula of conditional information (22); and, possibly, the formula connecting information and entropy (20).

IV

3.3. Let us note some of the important properties of information. Information is nonnegative

$$I(\zeta, \tilde{}) \geqq 0, \text{ moreover } I(\zeta, \tilde{\zeta}) = 0 \tag{21}$$

if and only if ζ and $\tilde{\zeta}$ are independent. Its dependence on the distribution $p_{\zeta\tilde{\zeta}}(\cdot)$ is not continuous, but only semicontinuous from below (see Gelfand and Yaglom [38]). This is explained (see [21]) by the fact that information density (in the probability sense of convergence) depends continuously on the distribution (with the metric of convergence in quadratic mean); but from the convergence in probability of a sequence of integrands it does not follow that the sequence of their integrals converges, unless these integrals are uniformly convergent. If there exists a conditional probability distribution for the pair ζ and $\tilde{\zeta}$ subject to the condition that a third variable γ is given, then the conditional information $I(\zeta, \tilde{\zeta}|\gamma)$ can be defined in a natural way. (The definition of conditional information can also be given without the assumption of the existence of the conditional distributions (see [19] and [52]).) It is, of course, a random variable. Then the following important *formula for conditional information* holds

$$I((\zeta, \gamma)\tilde{\zeta}) = I(\gamma, \tilde{\zeta}) + EI(\zeta, \tilde{\zeta}|\gamma). \tag{22}$$

It can be deduced from equations (21) and (22) that if $\phi(\cdot)$ is some measurable function, then

$$I(\phi(\zeta), \tilde{\zeta}) \leqq I(\zeta, \tilde{\zeta}). \tag{23}$$

It also follows from equation (22) that if the variables ζ, γ, and $\tilde{\xi}$ form a Markov chain, then

$$I((\zeta, \gamma), \tilde{\xi}) = I(\gamma, \tilde{\xi}). \tag{24}$$

3.4. Now these results can be applied to Shannon's problem. Suppose the variables ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$ give the method of transmission. Then, by applying first equation (23) and then equation (24), it is found that

$$I(\xi, \tilde{\xi}) \leqq I((\xi, \eta), (\tilde{\eta}, \tilde{\xi})) \leqq I(\eta, \tilde{\eta}). \tag{25}$$

That is, *the information between the input and output messages cannot be greater than the information between the input and output signals.*

The name *capacity* of the transmitter (Q, V) will be given to the number

$$C(Q, V) = \sup I(\eta, \tilde{\eta}), \tag{26}$$

where the upper bound is taken over all the pairs of variables η, $\tilde{\eta}$ connected by the transmitter (Q, V). The name *W-entropy of the message* is given to the number

$$H(p, W = \inf I(\xi, \tilde{\xi}), \tag{27}$$

where the lower bound is taken over all pairs ξ, $\tilde{\xi}$ which form the message (p, W). [The name entropy is justified by the fact that, for a message with the condition of perfect reproduction, $H(p, W) = H(\xi)$.] Then equation (25) gives the following *Shannon theorem: If the message (p, W) can be transmitted by the transmitter (Q, V), then*

$$H(p, W) \leqq C(Q, V). \tag{28}$$

It will be called the *converse Shannon theorem.*

This general and simple proof of the assertion of the converse Shannon theorem, based only on the "algebraic" properties of information, was apparently first pointed out by Kolmogorov [51]. It seems that it is not yet sufficiently widely used in information theory literature, since in some papers facts which are special cases of this result are still proved by more unwieldy *ad hoc* methods.

3.5. As for the statement of the Shannon theorem about sufficient conditions for the possibility of transmission, the subject becomes more complicated. It is easy to give examples which show that the condition of equation (28) (even if $\leqq$ is replaced by $<$ there) is insufficient for the possibility of transmission. Condition (28) is sufficient for the possibility of transmission only in a certain asymptotic sense, and in this respect the assumption that $H(p, W) \to \infty$ is essential, since condition (28) is necessary and sufficient only when applied to the problem of transmitting a sufficiently large quantity of information. The numerous other assumptions are of the nature of regularity assumptions, and it may be boldly assumed that they are fulfilled in applications.

The most vulnerable aspect of Shannon's theorem, from the point of view of practical applicability, is the fact that the coding and decoding methods, whose existence is guaranteed by the fulfillment of condition (28), are obviously unavoidably connected with the coding of a large quantity of information

as a single unit, and are therefore inevitably complex. (Furthermore, their complexity increases when $H/C \to 1$.) In this connection the possibility of obtaining quantitative estimates of the minimal complexity for the algorithms of optimal coding and decoding is tempting, though still remote (see [2], where similar estimates are obtained for the algorithms of computational analysis). From this point of view [102] and [26] are very interesting, since they give upper bounds for the average number of operations for decoding algorithms in simple situations.

V

3.6. The concepts used in the mathematical formulation of the direct assertion of Shannon's theorem will be introduced below. The sequence of pairs of random variables $(\zeta^t, \tilde{\zeta}^t)$, where $t = 1, 2, \cdots$, will be called *information-stable* if $0 < I(\zeta^t, \tilde{\zeta}^t) < \infty$ and

$$\lim_{t\to\infty} \frac{i_{\zeta^t,\tilde{\zeta}^t}(\zeta^t, \tilde{\zeta}^t)}{I(\zeta^t, \tilde{\zeta}^t)} = 1 \tag{29}$$

in the sense of convergence in probability. (The case of infinite information, for which a theory can also be developed [19] is not considered here.)

In many applications, it is natural to take ζ^t and $\tilde{\zeta}^t$ as the sequences of segments $[0, t]$ of certain processes $\{\alpha_s\}$ and $\{\tilde{\alpha}_s\}$ respectively, that is,

$$\zeta^t = \{\alpha_s, s \in [0, t]\}, \qquad \tilde{\zeta}^t = \{\tilde{\alpha}_s, s \in [0, t]\}. \tag{30}$$

If the processes α_s and $\tilde{\alpha}_s$ have discrete time, and if the pairs $(\alpha_s, \tilde{\alpha}_s)$ for $s = \cdots, -1, 0, 1, \cdots$ are completely independent, then the assertion of information stability of the segments ζ^t and $\tilde{\zeta}^t$ reduces to the assertion of the law of large numbers for a sequence of sums of independent random variables. Therefore, the necessary and sufficient conditions for information stability can easily be obtained from the theory of limit theorems for the sums of independent random quantities. Similarly, if the pairs $(\alpha_s, \tilde{\alpha}_s)$ form a Markov chain, then the assertion of information stability reduces to the assertion of the law of large numbers for functions of variables connected in a Markov chain. This was first pointed out by Rosenblatt-Rot [75]. If α_s and $\tilde{\alpha}_s$ form a joint stationary ergodic process with discrete time and state-space, then the assertion of information stability is easily obtained from the well known theorem of McMillan [58] and this theorem itself signifies information stability of the sequence of pairs $(\zeta^t, \tilde{\zeta}^t)$. The earliest results on conditions for information stability for stationary processes with an arbitrary set of states, and for processes with continuous time were obtained by Perez [65], [67]; and were established under rather wide conditions by Pinsker [72].

The next task here seems to be the generalization of these results to the case of a "sequential scheme" (compare [40]), that is, to the case when $\zeta^t = \{\alpha_s^t, s \in [0, t]\}$, $\tilde{\zeta}^t = \{\tilde{\alpha}_s^t, s \in [0, t]\}$. Recently, Jacobs [43] introduced an interesting class of almost periodic processes and proved McMillan's theorem for them, which also makes it possible to establish their information stability. It seems important at present to carry over Jacob's results to almost periodic processes with continuous time and state spaces,

VI

since this is exactly the statistical nature of many of the processes
VII arising in radio-engineering in connection with modulations of stationary processes.

Pinsker's result [72] is interesting and conclusive, since it shows that, if ζ^t and $\tilde{\zeta}^t$ are sets of variables possessing joint Gaussian distributions, then for information stability it is necessary and sufficient that $I(\zeta^t, \tilde{\zeta}^t) \to \infty$.

It is clear from the above that the property of information stability holds under very wide conditions. Its generality can be compared with that of the law of large numbers, which is related to it.

3.7. Examine further the sequence of transmitters (Q^t, V^t) with $C(Q^t, V^t) < \infty$. It will be called *information stable* if there exists an information stable sequence of pairs $(\eta^t, \tilde{\eta}^t)$, connected by the transmitters (Q^t, V^t) such that

$$\lim_{t\to\infty} \frac{I(\eta^t, \tilde{\eta}^t)}{C(Q^t, V^t)} = 1. \tag{31}$$

On many grounds one may hope that the property of information stability for a sequence of transmitters when $C(Q^t, V^t) \to \infty$ will turn out to be as general as is the property of information stability for sequences of random variables. The mathematical results so far obtained, however, are much less wide here. It was proved by the author [19] that sequences of increasing segments of a homogeneous memoryless channel are information stable. By using the
VIII theory of sums of independent variables it would obviously not be difficult to obtain general (and perhaps necessary and sufficient) conditions for information stability for memoryless channels in the nonhomogeneous case, and also for a sequential scheme.

Numerous investigations [90], [32], [33], [12] devoted to the solution of the problem posed by Khinchin on the coincidence of ergodic and ordinary capacity of a channel may be regarded as the proof of information stability for the segments of a channel with finite memory in the Feinstein-Khinchin sense. From the results of [8], it is possible, by using the theorem of Khu-Go-Din, which is given below (see section 3.10), to deduce indirectly the information stability of segments of homogeneous channels with finite memory and discrete state spaces, when certain restrictions are imposed on the transition function q_0. It seems interesting to give a direct proof of information stability for this case and to establish (in the fashion of the ergodic theory of homogeneous finite
IX Markov chains) a condition which is necessary and sufficient for information stability for any initial state.

For the nondiscrete case the only general result is that of Pinsker [72] which proves that in order that a sequence of Gaussian transmitters should be information stable, it is necessary and sufficient that $C(Q^t, V^t) \to \infty$. The problem of obtaining wide sufficient conditions for information stability of
X segments of channels, with generalizations to the case of continuous time and the case of nonhomogeneity in time (perhaps almost periodicity), is of great importance here.

3.8. The sequence of messages (p^t, W^t) with $H(p^t, W^t) < \infty$ will be called *information stable* if there exists an information stable sequence of pairs $(\xi^t, \tilde{\xi}^t)$ forming the message (p^t, W^t) such that

$$\lim_{t\to\infty} \frac{I(\xi^t, \tilde{\xi}^t)}{H(p^t, W^t)} = 1. \tag{32}$$

Less is known about the conditions for information stability for a sequence of messages than for a transmitter.

In the case of messages with the condition of perfect reproduction, the problem reduces to the information stability of pairs of variables $(\xi^t, \tilde{\xi}^t)$, where ξ^t has the distribution p^t. It can usually be solved on the basis of the results given earlier. In another paper, [19], we gave proof of the information stability of a sequence of communications homogeneous in time, with a component-wise condition of accuracy of reproduction, and independent components as the number n of components tends to infinity. Here the problem of immediate interest is comparable to problem VIII for transmitters. Pinsker proved that for information stability of a sequence of Gaussian messages it is necessary and sufficient that $H(p^t, W^t) \to \infty$. German, a student of Moscow University, proved in his thesis that a sequence of messages with component-wise conditions of accuracy of reproduction is information stable when the input process is stationary and ergodic and takes on a finite number of values. In this connection, the generalization to the case of a continuous set of message values, and of continuous time, is of immediate importance.

XI

XII

3.9. The following definition is introduced. Let $\mathfrak{U} \subset R^n$, where R^n is n-dimensional Euclidean space with points $\bar{x} = (x_1, \cdots, x_n)$. It is convenient to take the number $r(\bar{x}', \bar{x}'') = \max_i |x_i' - x_i''|$ as the distance between the points $\bar{x}' = (x_1', \cdots, x_n')$ and $\bar{x}'' = (x_1'', \cdots, x_n'')$. By $[\mathfrak{U}]_\epsilon$, where $\epsilon \geqq 0$, we denote the set of all points $\bar{x} \in R^n$ such that for some $\bar{\bar{x}} \in \mathfrak{U}$, the distance $r(\bar{x}, \bar{\bar{x}}) \leqq \epsilon$. We denote by (Q, V_ϵ) a transmitter for which, in definition (2), the set $\overline{V}$ is replaced by $[\overline{V}]_\epsilon$; by (p, W_ϵ) we denote a message for which the set $\overline{W}$ is replaced by $[\overline{W}]_\epsilon$.

3.10. The following Shannon theorem can now be formulated: *Let the information stable sequences of transmitters (Q^t, V^t) and messages (p^t, W^t) be such that the functions $\pi_i^t(\cdot\,, \cdot)$ and $\rho_j^t(\cdot\,, \cdot)$, where $i = 1, \cdots, N^t$, and $j = 1, \cdots, M^t$, are uniformly bounded in i, j, and t and for any $a > 0$, $N^t = o\{\exp [aC(Q^t, V^t)]\}$, $M^t = o\{\exp [aH(p^t, W^t)]\}$. Let $H(p^t, W^t) \to \infty$, and let the lower limit*

$$\lim_{t\to\infty} \frac{C(Q^t, V^t)}{H(p^t, W^t)} > 1. \tag{33}$$

Then for any $\epsilon > 0$ there exists a t_0 sufficiently large so that for all $t \geqq t_0$ the message (p^t, W_ϵ^t) can be transmitted by the transmitter (Q^t, V_ϵ^t).

This theorem will be called the direct Shannon theorem.

Let us consider this result. First, in formulating the theorem we note that it is not always possible to replace W_ϵ^t by W^t, and V_ϵ^t by V^t. This is clearly evident,

for example, in the case of messages with the condition of perfect reproduction, since the error arising in transmission (with however small a probability) in any noisy channel leads here to the incomplete congruency of the input and output messages. But the converse Shannon theorem about the impossibility of transmission shows that the message (p^t, W_ϵ^t) can be transmitted by the transmitter (Q^t, V_ϵ^t) only if $[C(Q^t, V_\epsilon^t)]/[H(p^t, W_\epsilon^t)] \geqq 1$. Since this ratio usually depends continuously on ϵ, it follows from the converse Shannon theorem that for a wide class of cases condition (33) (with the sign $>$ replaced by $\geqq$), is also necessary for fulfilling the assertion of the formulated theorem. Indeed, the converse Shannon theorem is given in this very form in most of the previous papers. In [19] conditions are given under which it is then possible to replace W_ϵ^t by W^t and V_ϵ^t by V^t in the formulated theorem.

The restriction on the number of functions $\pi_i^t(\cdot\,,\cdot)$ and $\rho_j^t(\cdot\,,\cdot)$ turns out to be rather weak and is fulfilled in interesting particular cases. Conversely, the requirement of the boundedness of the functions $\pi_i^t(\cdot\,,\cdot)$ and $\rho_j^t(\cdot\,,\cdot)$ is extremely stringent. In [19] the supplementary conditions necessary to replace the requirement of boundedness are investigated in detail. They are closely related to the requirements for information stability and recall Liapunov's conditions for the central limit theorem. It would be interesting to see how they are fulfilled in concrete situations (see sections 3.7 and 3.8). In [19] this question is partially examined for the cases of memoryless channels and a message with a component-wise condition of accuracy and independent components.

XIII

Information stability conditions for sequences of transmitters and messages are of fundamental importance. It was recently proved by Khu-Go-Din that a stipulation very close to the assertion of information stability of sequences of transmitters is necessary in order that, for sufficiently large t, it should be possible to transmit by the transmitter any information stable message with a condition of perfect reproduction, for which equation (33) held. Khu-Go-Din's result shows that the introduction of a condition of information stability for transmitters is unavoidable. It seems interesting to attempt to obtain a similar result for information stability requirement for sequences of messages.

XIV

This form of the theorem contains as a special case (if inessential differences of formulation are ignored) all earlier published theorems of this type (see [49], [34], [65], and [62]). However, these papers examine, besides the Shannon theorem itself, also the question of conditions for the information stability of transmitters and messages studied in them, a matter which in our presentation is treated as a separate problem.

3.11. In the case where the sequence of transmitters is a sequence of segments of a homogeneous channel, and the sequence of messages is a sequence of homogeneous messages with an increasing number of components, a somewhat different formulation of the basic Shannon theorem is often preferred.

In this connection the following concepts will be introduced. Let $(Q_{p_0}^n, V^n)$ be

the transmitter given by a segment of length n of a homogeneous channel with an initial distribution p_0. Then,

$$\overline{C} = \lim_{n\to\infty} \frac{1}{n} C(Q^n_{p_0}, V^n) \tag{34}$$

will be called the *mean capacity of the channel* if this limit exists for any initial distribution p_0, and does not depend on this distribution. Further let $\{p^n, W^n\}$ be a message homogeneous in time, with n components. The name *rate of creation of the message* will be given to the limit:

$$\overline{H} = \lim_{n\to\infty} \frac{1}{n} H(p^n, W^n). \tag{35}$$

These definitions can easily be generalized to the case of continuous time. It follows directly from the Shannon theorems formulated above, if $\overline{H} > \overline{C}$, that for sufficiently large n the message $\{p_n, W^n\}$ cannot be transmitted by the transmitter $(Q^n_{p_0}, V^n)$. If, on the other hand, $\overline{C} > \overline{H}$, and the requirements of information stability are fulfilled, then, for sufficiently large n, the message (p^n, W^n) can be transmitted by the transmitter $(Q^n_{p_0}, V^n)$.

Sometimes a slight variation of these definitions is preferable. Let $\mathcal{Z} = \{\zeta_n, \tilde{\zeta}_n\}$ be a joint stationary processes. Then

$$\bar{I}(\mathcal{Z}) = \lim_{n\to\infty} \frac{1}{n} I\{(\zeta_1, \cdots, \zeta_n), (\tilde{\zeta}_1, \cdots, \tilde{\zeta}_n)\} \tag{36}$$

will be called the *rate of transmission of information*. (The conditions for existence of the limit in equation (36) as well as other variations of this definition were studied by Pinsker [72]). It is then natural to assume that

$$\overline{C} = \sup I(\mathcal{K}) \tag{37}$$

where the upper bound is taken over all processes $\mathcal{K} = \{\eta_n, \tilde{\eta}_n\}$, such that for any n, the variables $(\eta_1, \cdots, \eta_n)$ and $(\tilde{\eta}_1, \cdots, \tilde{\eta}_n)$ are connected by a segment of length n of the channel with some initial distribution for the state of the channel. It is also assumed that

$$\bar{I} = \inf I(\Xi) \tag{38}$$

where the lower bound is taken over all processes $\Xi = \{\xi_n, \tilde{\xi}_n\}$ such that for any n, the variables $(\xi_1, \cdots, \xi_n)$ and $(\tilde{\xi}_1, \cdots, \tilde{\xi}_n)$ form a message with n components. The deduction of equation (37) for channels with finite memory in the Khinchin-Feinstein sense is contained in [90] and [32]. The problem of establishing general conditions for its fulfillment, and also the almost completely uninvestigated problem of establishing conditions for equation (38) are obviously closely linked with problems IX, X, and XII of obtaining conditions for the information stability of the corresponding channel and message.

XV

3.12. In this paper the formulation of Shannon's theorem reflects mathematically the real picture of transmission of information, lasting for a certain long but finite period of time. But in most of the earlier papers, a different mathe-

matical idealization was introduced in which the transmission of information was thought of as lasting for an infinite time. The corresponding definitions, as they are given for simple cases in well known papers, will not be given here. For the general case they are unwieldy and can be found in [19], section 1.8.

For simple cases, which have already been studied in earlier literature, the transition from the Shannon theorem in the formulation given above to the corresponding theorem for infinite transmission is almost trivial. However, this is not so for the general case. An example will be described intuitively to clarify the difficulties which arise here. Let the message under transmission be $\xi^t \equiv \xi$, where $-\infty < t < \infty$, ξ is a random variable with a continuous distribution, and where the condition of accuracy of reproduction is the condition of perfect reproduction. The rate of creation of such a message is infinite. However, there exist coding and decoding methods which enable it to be transmitted by a transmitter with a nonzero mean capacity which is as small as desired.

To do this, having chosen the monotone sequence of times $T_1 > T_2 > \cdots$, $T_n \to -\infty$, for the segment $[T_{n+1}, T_n]$, we transmit information about ξ_t with accuracy of up to 2^{-n}. By choosing $T_{n+1} - T_n$ sufficiently large, this can be done with as small a probability of error as desired. For any time t, information about ξ_t will be obtained at the output right up to that very time with as great an accuracy and as small a probability of error as desired. Thus, although $\overline{H} > \overline{C}$ here, transmission over infinite time is nevertheless possible. An example of a similar (but essentially more complicated) situation for which $\overline{H} < \infty$ can also be introduced.

It would be interesting to clear up this cycle of problems completely. In order to do this it may be necessary to reformulate the definition of entropy and of capacity (compare with the different definitions of the rate of transmission of information in [51]). XVI

4. The calculation of capacity, of W-entropy, and of optimal distributions

4.1. Since the solution to Shannon's problem is given by the capacity of the transmitter and the W-entropy of the message, it is important to learn how to calculate explicitly these quantities in concrete cases. Also of interest, and closely related to the above, is the question of the calculation of optimal input distributions to the transmitter, that is, of the distributions p in the space of input signals, such that if η has the distribution p and the pair $(\eta, \tilde{\eta})$ is connected by the transmitter (Q, V), then

$$I(\eta, \tilde{\eta}) = C(Q, V), \tag{39}$$

as well as the question of the calculation of optimal message distributions, that is, of the joint distributions $r(\cdot\,, \cdot)$ in the product of the spaces of input and output message values, such that the pairs $(\xi, \tilde{\xi})$ with distribution $r(\cdot\,, \cdot)$ form the message (p, W), and

$$I(\xi, \tilde{\xi}) = H(p, W). \tag{40}$$

The importance of optimal distributions is explained by the fact that, as the inequality of relation (25) shows, if the variables ξ, η, $\tilde{\eta}$, and $\tilde{\xi}$ provide the means of transmitting the message (p, W) through the transmitter (Q, V), then

$$I(\xi, \tilde{\xi}) \leqq C(Q, V), \qquad I(\eta, \tilde{\eta}) \geqq H(p, W). \tag{41}$$

It follows from this that when the entropy of the message is close to the capacity of the transmitter (that is, when this capacity is used almost completely) the coding and decoding methods must be chosen so that the differences $I(\xi, \tilde{\xi}) - H(p, W)$ and $C(Q, V) - I(\eta, \tilde{\eta})$ are small. It is natural to deduce from this that the distributions of the input signal η and of the message values ξ and $\tilde{\xi}$ are almost optimal, which means that the value of the optimal distributions suggests how to choose the transmission method. From the mathematical point of view, it is far from obvious that the smallness of the differences $I(\xi, \tilde{\xi}) - H(p, W)$ and $C(Q, V) - I(\eta, \tilde{\eta})$ implies the nearness of the corresponding distributions, since the space of the distributions is not compact and information is not a continuous functional. Nevertheless it appears that this is true under sufficiently broad assumptions.

XVII

4.2. In accordance with the definition of capacity, in equation (26), the problem of calculating it is the problem of calculating the maximum of a functional (and in the case of discrete space, of calculating of a function of several variables). The character of the problem is related to the special analytical form of the functional under examination, which enables the solution to be simplified; and also to the fact that restrictions in terms of inequalities are imposed on the region of variation of its argument (the restriction V and the requirement of nonnegativity of probability distributions). The general theory of variational problems with such restrictions has not yet been adequately worked out (see [5]).

An explicit expression for the capacity can be attained only in exceptional cases, but in papers by Shannon, and especially Muroga [79], [60], and [61], an algorithm has been worked out which reduces the problem to the solution of a transcendental equation. These investigations were carried out at a level of "physical accuracy" and it seems important (especially for the more complicated nondiscrete case) to complete them. This could be done by analyzing in detail questions of existence, examining special cases and so on. In [81] Shannon gave a geometrical treatment of the problem for the discrete case and showed, in particular, that the set of optimal distributions is convex. It would be interesting to find methods of calculating the extreme points of this set and also to carry over the investigation to the general case.

XVIII

XIX

Analogous problems exist in connection with calculating the entropy of messages and optimal distributions for messages. However, almost nothing has been done here (apart from the cursory remarks in [86]).

XX

4.3. In the case where the transmitter is a segment of a memoryless channel, it is not difficult to show (see [19] or [31]) that among the optimal input distributions there exists a distribution with independent components. It is easily deduced from this that, if C_n is the capacity of a segment of length n, then

$$C_n = nC_1. \tag{42}$$

However, the calculation of C_1 usually turns out to be an equally difficult problem. For a Gaussian memoryless channel with additive noise *Shannon's formula*

$$C_n = n \log \left(1 + \frac{p_s}{p_n}\right), \tag{43}$$

holds, and the optimal distribution is Gaussian.

Similarly, for a message with independent components and with a component-wise condition of accuracy, there are among the optimal distributions (see [19]) some for which the pairs $(\xi_i, \tilde{\xi}_i)$ are independent. It follows again from this that, if H_n is the entropy of the message with n components, then

$$H_n = nH_1. \tag{44}$$

And particularly simple is the following formula for a homogeneous Gaussian message with a condition of boundedness on the mean square error,

$$H_n = n \log \left(1 + \frac{p_e}{p_d}\right), \tag{45}$$

where p_d is the mean square error, and p_e is the variance of a component of the input signal.

4.4. In more general situations explicit formulas for the capacities and W-entropies cannot be obtained. In this connection of great importance is a somewhat simpler problem, namely to calculate the mean capacily $\overline{C}$ and the rate of creation $\overline{H}$ of the message (compare section 3.11).

The problem of calculating $\overline{C}$ and $\overline{H}$ has already been adequately solved for Gaussian transmitters and messages which are especially important from the point of view of applications. In fact it can be shown fairly easily (see Pinsker [69], [72]) that in these cases there are always Gaussian distributions among the optimal distributions. It is possible to give for Gaussian distributions (see [80] and [38]), a simple expression for information in the finite-dimensional case. This can be generalized to the infinite-dimensional case. (In particular, a method of calculating information for certain classes of Gaussian processes is developed in [38].) General formulas for the rate of transmission of information, for the mean capacity, and for the rate of creation of a message in terms of the corresponding spectral densities were obtained by Pinsker in [68], [69], and [72]. His results were partly repeated in [73]. The methods of solving such problems are closely linked with the analytical methods of second order stationary processes.

The following Shannon formula is the most popular in radio engineering applications. If C_T is the capacity of a transmitter with a finite bandwidth and

with additive white Gaussian noise for transmission during the time T, then as $T \to \infty$

$$C_T \sim 2WT \log\left(1 + \frac{p_s}{p_n}\right). \tag{46}$$

The situation is worse for other classes of processes. Thus, for example for Rayleigh processes, which are frequently used in radio engineering, the problem immediately leads to serious difficulties. (Rayleigh processes arise in passing narrowband Gaussian processes through envelope detectors. See for example [14]. Something was done on this subject in [61].)

XXI

4.5. It is easy to deduce a simple formula for the rate of message creation with the condition of perfect reproduction, if the input process is a homogeneous finite Markov chain (see [79] and [48]). However, if we even simply replace the Markov chain by a process which is a function of a Markov chain, the problem becomes immeasurably more difficult. An ingenious solution to this problem was proposed by Blackwell [6]. Most important here is the question of whether it is impossible to generalize this solution to the case where the condition of accuracy is an arbitrary component-wise condition or an additive condition. It is possible that analogous methods can be applied to the wholly unexplored problem of finding the mean channel capacity having finite memory. In this case there may be (see [8]) among the optimal input processes, that is processes $\mathfrak{K}$ for which the upper bound in equation (37) is attained, ones which are functions of Markov processes. It is of interest to give a method for calculating the parameters of such processes.

XXII

XXIII

XXIV

4.6. The difficulty of finding explicit expressions for the W-entropy and capacities naturally leads one to consider making use of the smallness of some of the channel and message parameters.

In the case when the transmitter is such that the output signal differs very little from the input signal, and in the case when the message is such that its output meaning is close to its input meaning, the problem is closely linked with the nonprobabilistic theory of ϵ-entropies and ϵ-capacities (see [54], section 2). Another useful restriction is the assumption that the channel has a random parameter. It is natural to assume that in this case the mean capacity

$$\overline{C} = \sup_{\zeta} \int_B \overline{I}(Z_b)\bar{p}(db), \tag{47}$$

where $\bar{p}(\cdot)$ is the distribution of the parameter $\beta(\phi)$, and $\overline{I}(Z_b)$ is the rate of information transmission for a pair of processes $(\zeta_n, \bar{\zeta}_n^b)$ such that $(\zeta_1, \cdots, \zeta_n)$ and $(\bar{\zeta}_1^b, \cdots, \bar{\zeta}_n^b)$ are connected by a corresponding conditional channel with parameter b. The upper bound is taken over all input processes $\{\zeta_n\}$ for which there exists an input process $\{\bar{\zeta}_n\}$ such that the pairs $(\zeta_1, \cdots, \zeta_n)$ and $(\bar{\zeta}_1, \cdots, \bar{\zeta}_n)$ are connected by the channel under investigation. It can be deduced from papers

by Tsybakov [91], [92], and [93], that formula (47) enables us to calculate the capacity of a wide class of physically realizable channels for the propagation of radio waves, particularly when phase fluctuations exist. In [16], formula (47) was proved for a special case which could easily be extended to the general case if only the space B of parameter values were finite. But since, in the applications of formula (47) which we mentioned, one must deal with a variable β which must have a continuous distribution, it would be interesting to deduce formula (47) also for that case.

XXV

It should be noted however that a channel with a random parameter is nonergodic under any sensible interpretation of this term. Therefore the direct Shannon theorem does not hold for it. But this does not take away the physical significance of $\overline{C}$ since it can be treated as an approximate value for the capacity of a channel with a slowly changing parameter, for which the direct Shannon theorem is applicable.

4.7. From among the important domains of research which are not discussed here, let us note the papers on the calculation of the capacity of multiwire channels [63], and on the statistical evaluation of information-theoretic parameters of realizable channels and messages [4], [17], and [56].

5. The investigation of optimal codes

5.1. The question naturally arises as to the simplest way of constructing encoding and decoding methods whose existence is guaranteed by Shannon's theorem. This problem will be examined only for the simplest case where the message is one with the condition of perfect reproduction, and such that the space X of input message values consists of S elements $E_1, \cdots, E_S$, and $p_\xi(E_i) = 1/S$, for $i = 1, \cdots, S$. This message will be denoted by $(p, \mathfrak{M}^S)$. Such a case is particularly important, since the course of the proof of Shannon's theorem for an arbitrary message suggests, [19], that the solution of the problem for this special case enters as a direct component in the general solution. We shall not dwell here on an interesting series of investigations, [39], devoted to the so-called nonuniform code, which can be interpreted as investigations of effective coding methods for other classes of messages.

5.2. Let the transmitter (Q, V) be given. By $e(Q, V, S)$ we shall denote the smallest ϵ such that the message $(p, \mathfrak{M}_\epsilon^S)$ can be transmitted by the transmitter (Q, V). Since, for $S \to \infty$, the sequence of messages $\mathfrak{M}^S$ is information stable with entropy $\log S$, then if

$$\lim_{t\to\infty} \frac{\log S^t}{C(Q^t, V^t)} < 1, \tag{48}$$

and if the assumptions of the direct Shannon theorem are satisfied for the sequence $C(Q^t, V^t)$, the following relation is satisfied

$$\lim_{t\to\infty} e(Q^t, V^t, S^t) = 0. \tag{49}$$

It follows from the converse Shannon theorem that if, on the contrary,

$$\lim_{t\to\infty} \frac{\log S^t}{C(Q^t, V^t)} > 1, \tag{50}$$

then

$$\varlimsup_{t\to\infty} e(Q^t, V^t, S^t) > 0. \tag{51}$$

By using constructions which were applied in the proof of the direct Shannon theorem, it is not difficult to prove that, if the sequence (Q^t, V^t) is information stable, equation (51) can be replaced by the stronger equation

$$\lim_{t\to\infty} e(Q^t, V^t, S^t) = 1. \tag{52}$$

The difference between the formulation of equations (51) and (52) was emphasized by Wolfowitz [101].

5.3. If equation (52) describes in a sufficiently exhaustive fashion the asymptotic value of $e(\cdot\ ,\ \cdot\ ,\ \cdot)$ under condition (50), then conversely, in the case when equation (48) is satisfied, it is desirable to make relation (49) more precise, estimating the rate at which $e(\cdot\ ,\ \cdot\ ,\ \cdot)$ tends to zero. Such an estimate is particularly important if (Q^n, V^n) is a segment of length n of a homogeneous channel, since if $e(Q^{n_0}, V^{n_0}, S^{n_0})$ is sufficiently small, the encoding for a channel operating continuously can be carried out in separate blocks of length n_0 (compare with similar constructions in [31] and [49]). This means that the smaller n_0 is, the simpler will be such a coding method. Perhaps, under very broad assumptions it follows from equation (48) that, for some $a > 0$,

$$e(Q^t, V^t, S^t) = o[2^{-aC(Q^t, V^t)}]. \tag{53}$$

In any case, by following through the proof of the direct Shannon theorem [19], it is easy to see that this will be so if the assumption of information stability of the sequence (Q^t, V^t) is replaced by the stronger assumption that there exists a sequence of pairs $(\eta^t, \tilde{\eta}^t)$ connected by the transmitter (Q^t, V^t), such that for any $\epsilon > 0$ and some $d(\epsilon) > 0$,

$$\lim_{t\to\infty} P\left\{\left|\frac{i_{\eta^t\tilde{\eta}^t}(\eta^t, \tilde{\eta}^t)}{C(Q^t, V^t)} - 1\right| > \epsilon\right\} = o[2^{-d(\epsilon)C(Q^t, V^t)}]. \tag{54}$$

For a memoryless channel, condition (54) follows from the well known theorems on large deviations of sums of independent variables [13] if it is assumed that the optimal input distribution is "sufficiently good." By such means Feinstein [30] proved assertion (53) for homogeneous finite memoryless channels.

XXVI For channels with a finite number of states, assertion (54) can obviously be deduced from the well known estimates for large deviations for sums of variables connected in a Markov chain, but in a more general situation separate investigations are necessary.

5.4. Assume now that (Q^n, V^n) is a segment of length n of a homogeneous channel. Further, assume that $S^n = [2^{nH}]$. Condition (48) now means that

(55) $$H < \overline{C},$$

where $\overline{C}$ is the mean channel capacity. Investigations which were carried out in simplest cases allow us to hope that when $n \to \infty$ and for certain $a, b > 0$,

(56) $$e(Q^n, V^n, [2^{nH}]) \asymp n^a 2^{-bn}.$$

Here the notation $x_n \asymp y_n$ means that

(57) $$0 < \varliminf_{n\to\infty} \frac{x_n}{y_n} \leqq \varlimsup_{n\to\infty} \frac{x_n}{y_n} < \infty.$$

The more precise asymptotic expression

(58) $$e(Q^n, V^n, [2^{nH}]) \sim cn^a 2^{-bn},$$

where c is a constant, seems obviously impossible to obtain (al-
XXVII though this has not yet been proved even for one case). The point is that in order to investigate the probability $e(\cdot\,, \cdot\,, \cdot)$ with such a high degree of accuracy, the arithmetic properties of the number n are essential.

It would be interesting to prove the existence of the constants a and b, but naturally it is more important to learn how to calculate them. This turns out to be difficult even for the simplest finite memoryless channels. The author studied this problem for the case of memoryless channels where the matrix of transition probabilities $Q_0 = (q_{ij})$ possesses the following *property of symmetry*: each of its rows is a permutation of any other row and each of its columns a permutation of any other column. It appeared that for such channels

(59) $$b = \log R(\bar{h}) + (1 - \bar{h}) \log m(\bar{h}) + \log N$$
$$a = \frac{-1}{2\bar{h}},$$

where

(60) $$R(h) = \frac{N}{M} \sum_{i=1}^{M} (q_{i1})^h, \qquad m(h) = \frac{d \log R(h)}{dh}$$
$$\log R(\bar{h}) - \bar{h} m(\bar{h}) = -H,$$

and where M is the number of elements in the input signal space and N is the number in the output signal space, provided that

(61) $$H \geqq H_{\text{crit}} = \frac{1}{2} m\left(\frac{1}{2}\right) - \log R\left(\frac{1}{2}\right).$$

A particular case of the channels under investigation is the *symmetric binary channel* with matrix

(62) $$Q = \begin{pmatrix} q_1 & q_2 \\ q_2 & q_1 \end{pmatrix},$$

where $q_1 + q_2 = 1$. Elias [24] and [25] studied a problem in connection with this channel which is of interest here. (His expression for a, however, is erroneous.) When $H < H_{\text{crit}}$, it is only possible to obtain distinct upper and lower

bounds for b. The problem of calculating b exactly, even for the
XXVIII case of the simplest symmetric binary channel, is very difficult. Evidently, its solution can be made possible only with the aid of some new ideas. Elias also studied [24] the case of a *binary erasure channel*, that is, a channel with the matrix

$$\begin{pmatrix} q_1 & 0 & q_2 \\ 0 & q_1 & q_2 \end{pmatrix}, \qquad q_1 + q_2 = 1 \tag{63}$$

which is not symmetric in the sense used here. The results of Elias can apparently be generalized to a certain class of nonbinary channels. How-
XXIX ever, the problem of finding the parameters a and b even for a nonsymmetric binary channel again becomes very difficult, and the approach to it is as yet not clear. (Roughly speaking, H_{crit} coin-
XXX cides here with the mean capacity $\overline{C}$, so that the difficulties are the same as for the symmetric binary channel when $H < H_{\text{crit}}$.) In the general case of a finite memoryless channel, it is possible to obtain only certain estimates (see [82]).

In [85] Shannon examined the problem for an important class of Gaussian memoryless channels with additive noise, and once again he found a complete solution only for the case when $H \geqq H_{\text{crit}}$. It is not clear to what
XXXI extent his results can be generalized to other Gaussian channels. Apparently the results obtained for symmetric finite memoryless channels can be generalized to certain channels with a continuous set of signals, which generate the circle $|z| = 1$, and such that their transition
XXXII probabilities are invariant under rotations. (Such channels can well represent transmissions with phase modulation.)

5.5. In all the papers referred to on constructive coding methods with minimal error, the method of *random coding* is used. This consists of the following. We shall consider only nonrandomized coding. In accordance with the remarks made in section 2.11, this is no real restriction of generality. The coding can then be defined as a function $f(E_i)$. By the code $\mathcal{K}(S)$ we shall denote the collection of values $f(E_1), \cdots, f(E_S)$. By $e\{Q, V, \mathcal{K}(S)\}$ we shall denote the smallest ϵ such that the message $(p, \mathfrak{M}_\epsilon^S)$ can be transmitted by the transmitter (Q, V) using the code $\mathcal{K}(S)$ and some method of decoding. Now assume that there is given a certain probability distribution $r(\cdot)$ on the space (X, S_X) of input signals, and a system of independent random variables $\phi_1, \cdots, \phi_S$, having the distribution $r(\cdot)$. Then $\mathcal{K}(S) = (\phi_1, \cdots, \phi_S)$ will be called a *random code*, and the mathematical expectation

$$\bar{e}(Q, V, S) = E(e\{Q, V, \mathcal{K}(S)\}) \tag{64}$$

will be called the *mean probability of error for the distribution* $r(\cdot)$. Obviously $\bar{e}(Q, V, S) \geqq e(Q, V, S)$. All the well known asymptotic upper estimates (enumerated above) for the probability $e(Q, V, S)$ are based on this inequality; but

XXXIII the study of the mean probability of error is also independently of interest, in that it characterizes the probability of error for a typical code. There are grounds for hoping that, in a wide class of cases, the following law of large numbers is true, *as* $n \to \infty$

$$\frac{e\{Q^n, V^n, \mathcal{K}([2^{nH}])\}}{\bar{e}\{Q^n, V^n, [2^{nH}]\}} \to 1 \text{ (in probability).} \tag{65}$$

It would also be interesting to study in greater detail the asymptotic distribution of $e\{Q^n, V^n, \mathcal{K}([2^{nH}])\}$ as $n \to \infty$. Most commonly $r(\cdot)$ is taken to be the optimal distribution of the input signal. In this connection, it is clear, from the usual proof of Shannon's theorem [19] that condition (54) implies not only relation (53), but also the stronger assertion

$$\bar{e}(Q^t, V^t, S^t) = o(2^{-aC(Q^t, V^t)}). \tag{66}$$

XXXIV Apparently, in a wide class of cases,

$$\bar{e}(Q^n, V^n, [2^{nH}]) \asymp n^{\bar{a}} 2^{-\bar{b}n}. \tag{67}$$

The values of $\bar{a}$ and $\bar{b}$ have so far been calculated only for the same cases (see above) when the constants a and b were considered. It was shown that, when $H \geqq H_{\text{crit}}$, $a = \bar{a}$ and $b = \bar{b}$. This means that for sufficiently high rates of information transmission, the random code is (with accuracy up to a constant) asymptotically "good" as well as optimal, which (see [35]) permits one to construct "good" codes by the Monte Carlo method. It should be noted, however, that for a Gaussian channel Shannon [85] had to take for $r(\cdot)$ not the optimal distribution but a certain distribution which was only asymptotically optimal. (Instead of the distribution of n independent normal variables with zero mean and variance p_s, he had to take the uniform distribution over an n-dimensional sphere of radius $\sqrt{np_s}$.) It would be of interest to know whether it

XXXV is possible to reduce $\bar{a}$ and $\bar{b}$ by a similar method in other cases.

When $H < H_{\text{crit}}$, explicit expressions for $\bar{a}$ and $\bar{b}$ can similarly be found. For example, for the symmetric channels described above

$$\bar{a} = -\frac{1}{2}, \qquad \bar{b} = H + 2 \log R\left(\frac{1}{2}\right) + \log N. \tag{68}$$

It is interesting that, in all the cases studied, H_{crit} turned out to be the break point of the derived function $\bar{b}(H)$. It would appear that $b < \bar{b}$

XXXVI whenever $H < H_{\text{crit}}$, but this has not been proved even for a symmetric binary channel (compare problem XXVIII). The problem of calculating $\bar{a}$ and $\bar{b}$ is simpler than that of calculating the con-

XXXVII stants a and b, and it appears realistic to find $\bar{a}$ and $\bar{b}$ in a rather wide class of cases.

5.6. Because of the difficulty in the practical realization of transmission and in the theoretical investigation of arbitrary codes, a special class of group codes has been singled out [42], [88], [10], [95].

Let us postulate a finite memoryless channel such that the number of ele-

ments in the input signals space is $M = q^k$, where q is a prime number and k is an integer. Y_0 shall be identified with the direct product of k cyclic groups of order q, and Y^n will be taken to be the direct product of n groups Y_0. The code $\mathcal{K}(S)$ will be called a *group code* if it forms a subgroup of Y^n. We shall denote by $\bar{e}(Q, V, S)$ the smallest value of the probability $e\{Q, V, \mathcal{K}(S)\}$, where the minimum is taken over all group codes. It turns out, for a wide class of homogeneous memoryless channels whose transition matrix is symmetric, and also for a symmetric binary channel and for a binary erasure channel, that [25] in the limit as $n \to \infty$ and $H \geqq H_{\text{crit}}$,

$$e(Q^n, V^n, [2^{nH}]) \asymp \bar{e}(Q^m, V^n, [2^{nH}]), \tag{69}$$

and thus the best group code is asymptotically as good as the best among all codes. This follows from the fact that, under a certain natural definition of the concept of a random group code, the mean value of the probability of error over all group codes, for all H, coincides asymptotically with $\bar{e}(Q^n, V^n, [2^{nH}])$.

XXXVIII Thus, almost all group codes, when $H \geqq H_{\text{crit}}$, are constructed as optimal codes. (In order to make this assertion more precise, it would be necessary to obtain an equation such as (65) for group codes.) When $H < H_{\text{crit}}$, the asymptotic behavior of $\bar{e}(Q^n, V^n, [2^{nH}])$ is unknown even for a binary symmetric channel. The results described above can apparently be extended to a somewhat wider class of memoryless chan-

XXXIX nels (compare problem XXIX). The question of whether group codes are asymptotically optimal is extremely interesting, even

XL only for the simplest classes of channels with memory.

The creation of optimal algebraic coding methods for $M \neq q^k$ is apparently impossible. Then, as far as the channels with a continuous set of states are concerned, generalizations are apparently possible to the case referred to in connection with problem XXXII, where Y can be identified with the multiplicative group of complex numbers $|z| = 1$. No method is at present known for generalizing algebraic coding methods in such a way as to make them applicable to Gaussian memoryless channels.

5.7. The problem of constructing optimal codes, that is, the problem of creating relatively simple algorithms yielding a code $\mathcal{K}(S)$ such that $e\{Q^n, V^n, \mathcal{K}(S)\}$ coincides with or at least is close to $e(Q^n, V^n, S)$, is very difficult. It can be solved only for rather small separate values of n and S. It becomes only a little easier if it is restricted to the examination of group codes, [88]. For this reason other nonprobabilistic methods for estimating the quality of a code have been introduced and profitably studied. Assume that (Q^n, V^n) is a finite memoryless channel. Then Y^n is the space of sequences $y = (y_1, \cdots, y_n)$ where y_i takes on the values $1, \cdots, M$. Defining $\rho\{(y_1, \cdots, y_n), (y'_1, \cdots, y'_n)\}$ as the number of indices $i = 1, \cdots, n$ for which $y_i \neq y'_i$, we transform Y^n into a metric space.

The *code distance* for the code $\mathcal{K}\{f(E_1), \cdots, f(E_S)\}$ will be defined as

$$d(\mathcal{K}) = \min_{i \neq j} \rho\{f(E_i), f(E_j)\}. \tag{70}$$

Intuition suggests that, generally speaking, for channels with symmetric matrices, and codes with large code distances the probability of error is smaller. It is not difficult, however, to give examples, even for a binary channel, of codes such that $d(\mathcal{K}) < d(\mathcal{K}')$, but $e(Q^n, V^n, \mathcal{K}) < e(Q^n, V^n, \mathcal{K}')$. The coincidence of codes which are optimal with respect to the probability of error and to code distance can only be proved for a very narrow class of cases [87] when there exist *densely packed codes*, that is when the whole space V^n can be represented as the sum of nonintersecting spheres with equal radii (the centers of these spheres form the densely packed code). Even the asymptotic coincidence of these two senses of optimality is not proved. However, one may conjecture that if

XLI

$$d(n, S) = \max d\{\mathcal{K}(S)\} \tag{71}$$

(where the maximum is taken over all codes $\mathcal{K}(S)$ of size S, for a transmission of length n) then for a sequence of codes $\mathcal{K}\{[2^{nH}]\}$ with $n = 1, 2, \cdots$

$$\frac{d\{\mathcal{K}([2^{nH}])\}}{d\{n, [2^{nH}]\}} \to 1, \qquad n \to \infty \tag{72}$$

it follows that

$$\frac{e\{Q^n, V^n, \mathcal{K}([2^{nH}])\}}{e\{Q^n, V^n, [2^{nH}]\}} \to 1, \qquad n \to \infty. \tag{73}$$

The converse is false. The function $d(n, S)$ is determined only for certain values of the arguments (see, for example, [44], [87]) and up to the present time no solution has been found for the most natural problem of calculating the constant α such that

XLII

$$d\{n, [2^{nH}]\} \sim \alpha n, \tag{74}$$

despite the elementary nature of its formulation. Problems similar to XLI and XLII arise and remain unsolved when these problems are restricted only to the study of group codes.

XLIII

XLIV

The concept of code distance can also be introduced in a natural way for the simplest channels with a continuous set of states. In this case one takes for $\rho(\cdot\,, \cdot)$ the usual Euclidean distance, and the investigation of $d(n, S)$ approaches the well known geometric investigations into the filling up of space by spheres.

5.8. The assumption that $S = [2^{nH}]$ made above is the most natural, since it corresponds to the assumption of constancy of the transmission rate. There is, however, some interest in investigating what occurs for another asymptotic value of S. In particular it has appeared that the problem becomes substantially simpler if it is assumed that S is constant. In this case solutions have been found for the analogues of problems XLII (see [3]; it is interesting that here for a binary channel $\alpha \geqq 1/2$ but $\alpha \to 1/2$ when $S \to \infty$), XLI, XLIII and XLIV, and for the question of calculating the constants a and b. Here, of course, $\bar{b} > b$, so that the random code turns out to be worse than the optimal one. However, even

XLV for a nonsymmetric binary memoryless channel (quite apart from more complicated cases) the asymptotic value of $e(Q^n, V^n, S)$ remains uninvestigated.

The assumption that S is constant means that transmission at a very slow rate is being studied. Another extreme case is that of transmission at the maximum possible rate, which approaches the capacity. The following problem was pointed out by M. Pinsker. Let $N(n, \epsilon)$ be the largest possible value of S such that the message $(p, \mathfrak{M}_\epsilon^S)$ can be transmitted by the transmitter (Q^n, V^n). The asymptotic value of $N(n, \epsilon)$, when ϵ is constant and $n \to \infty$, must be studied. It is easy to see that $e\{Q^n, V^n, N(n, \epsilon)\} = \epsilon$, so that studying $N(n, \epsilon)$ is one way of studying the asymptotic value of the probability $e(Q^n, V^n, S)$. It is not difficult to show that, for a homogeneous memoryless channel with a symmetric matrix,

$$N(n, \epsilon) \asymp n^{-1/2} 2^{n\overline{C} + \sigma u_\epsilon \sqrt{n}} \tag{75}$$

where $\overline{C}$ is the channel capacity, u_ϵ is the solution of the equation $\Psi(u_\epsilon) = \epsilon$ where $\Psi(\cdot)$ is the normal distribution function, and $\sigma = [dm(0)/dh]^{1/2}$, see equation (60). Certain estimates for $N(n, \epsilon)$ can be extracted from the papers of Wolfowitz. It is interesting to investigate $N(n, \epsilon)$ for other channels. XLVI

All the statements introduced above on the asymptotic investigation of $e(Q^n, V^n, S)$ can be regarded as special cases of the following problem: find a relatively simple function $g(n, S)$ such that, when $n \to \infty$, XLVII

$$e(Q^n, V^n, S) \asymp g(n, S) \tag{76}$$

uniformly in S. (See similar problems in the limit theory for sums of random summands [50].) However, this problem will arise only after the solution of problem XXVIII.

5.9. Kolmogorov stated the problem of this section somewhat differently. To be precise, he changed the hypotheses of section 5.1 in that he considered $S = a^s$, where a and s are integers, and the message values are identified with the sets $(E_{i_1}, \cdots, E_{i_s})$ where $i_k = 1, \cdots, a$. Further, suppose that the space $\bar{X}$ of the output message values coincides with X, and that the accuracy conditions will be given by the set of s functions

$$\rho_k\{(E_{i_1}, \cdots, E_{i_s}), \quad (E_{j_1}, \cdots, E_{j_s})\} = \begin{cases} 0 & \text{if } E_{i_k} = E_{j_k}, \\ 1 & \text{if } E_{i_k} \neq E_{j_k}. \end{cases} \tag{77}$$

We shall take the input distribution to be uniform. Finally, the set $\overline{W}$ will consist of the single point $(0, \cdots, 0)$ (compare section 2.7). Such a message will be denoted by $(p, \mathfrak{M}^s)$. By $g(Q, V, S)$ (compare section 5.2) will be denoted the smallest ϵ, such that the message $(p, \mathfrak{M}_\epsilon^S)$ can be transmitted by the transmitter (Q, V). From the intuitive point of view, it can be said that here the input message is $\xi = (\xi_1, \cdots, \xi_s)$, the output message is $\bar{\xi} = (\bar{\xi}_1, \cdots, \bar{\xi}_s)$, and

$$g(Q, V, S) = \inf \max_k P\{\xi_k \neq \bar{\xi}_k\}, \tag{78}$$

where the lower bound is taken over all methods of transmission by (Q, V). The described criterion reflects better the technical requirements on the transmission accuracy. Evidently there is a formula for $g(Q^n, V^n, [2^{nH}])$ which is analogous to the formula (56) for $e(Q^n, V^n, [2^{nH}])$. We can prove that the value $\overline{\overline{b}}$ which describes the main exponential term of the asymptotic is the same for both formulas. The corresponding question about the value a is open.

XLVIII

6. Unique transmission methods

6.1. One of the basic assumptions of the theory developed thus far was that the distribution $p_\xi(\cdot)$ of the input message and the transition function $Q(\cdot\ ,\ \cdot)$ of the transmitter were regarded as given. However, this assumption is unreasonable in many real situations. Either because the statistical parameters of the channel change rapidly with time so that there is no way of obtaining the a priori distributions of these parameters, or because the same receiving–transmitting set must be adapted to operate under various conditions. Finally a game-theoretic situation may be represented, in which one of the players chooses the distributions p and Q, and the other chooses the transmission method.

6.2. The mathematical description of this situation reduces to the following. A certain set Γ of parameter values is given, and with each $\gamma \in \Gamma$ there is associated a transmitter (Q_γ, V_γ) such that the input and output signal spaces do not depend on γ. Further, there is given a set of messages (p_δ, W_δ) where $\delta \in \Delta$ such that the ranges of message values at the input and the output do not depend on δ. We shall say that there exists a *unique method of transmitting a system of messages* (p_δ, W_δ), *where* $\delta \in \Delta$, *by a system of transmitters* (Q_γ, W_γ), where $\gamma \in \Gamma$, if there exist a coding function $P(\cdot\ ,\ \cdot)$ and a decoding function $\tilde{P}(\cdot\ ,\ \cdot)$ (independent of γ and δ) such that, for all $\gamma \in \Gamma$ and $\delta \in \Delta$, the message (p_δ, W_δ) can be transmitted by (Q_γ, V_γ) with the help of the coding function $P(\cdot\ ,\ \cdot)$ and the decoding function $\tilde{P}(\cdot\ ,\ \cdot)$.

By the capacity of the system (Q_γ, V_γ), for $\gamma \in \Gamma$, we shall mean

$$C(\Gamma) = \inf_{\eta} \sup_{\gamma \in \Gamma} I(\eta, \eta_\gamma), \tag{79}$$

where (η, η_γ) are connected by (Q_γ, V_γ), and where the lower bound is taken over all the variables η for which there exists such a pair (η, η_γ), for all γ. Further, the quantity

$$H(\Delta) = \sup_{\delta \in \Delta} H(p_\delta, W_\delta). \tag{80}$$

will be called the *entropy of the system.*

It follows easily from the arguments of section 3.4 that *the existence of a unique transmission method implies that*

$$H(\Delta) \leqq C(\Gamma). \tag{81}$$

The concept of a unique transmission method was recently introduced in

papers by Blackwell, Breiman and Thomasian [9] and by the author [20]. In [9] the case of finite homogeneous memoryless channels was studied in detail. The general case, where Γ is arbitrary and Δ consists of a single element, was examined in [20], without giving proofs or detailed formulations.

Note that,

$$C(\Gamma) \leqq \inf_{\gamma \in \Gamma} C(Q_\gamma, V_\gamma), \tag{82}$$

and moreover, equality holds when the system of functions Q_γ is in some sense "convex." In the general case (see [20]), the symbol $<$ can be taken in equation (81).

6.3. There are grounds for believing that, in a fairly wide class of cases, the direct Shannon theorem is true for unique transmission methods, asserting that in the general case, under certain restrictions on the sequence of systems (Q^t_γ, V^t_γ) with $\gamma \in \Gamma^t$, and (p^t_δ, W^t_δ) with $\delta \in \Delta^t$, the inequality,

$$\varlimsup_{t \to \infty} \frac{C(\Gamma^t)}{H(\Delta^t)} > 1 \tag{83}$$

implies that for any $\epsilon > 0$ there exists a T so large that for all $t \geqq T$ there exists a unique method of transmitting the system of messages $(p^t_\delta, W^t_{\delta\epsilon})$ with $\gamma \in \Gamma^t$ by the system of transmitters $(Q^t_\gamma, V^t_{\gamma\epsilon})$. The question of establishing sufficiently general conditions (compare section 3.10) to satisfy this theorem is still an open one. However, those special cases for which it is not difficult to prove the validity of the theorem can be indicated. The theorem is true if the systems of messages and transmitters satisfy one of the following conditions:

XLIX

a) Γ^t consists of a single element (Q^t, V^t) and the sequence of transmitters (Q^t, V^t) is information stable;

b) (Q^t_γ, V^t_γ), where $\gamma \in \Gamma^t$, are segments of homogeneous channels with discrete time, finite sets of input and output signals, and finite memory; and moreover, the transition matrices of the channels are such that

$$\sum_{\tilde{y}_0 \in \tilde{Y}_0} q_0(y_0, f_0, \tilde{y}_0, f') \geqq \alpha > 0 \tag{84}$$

for all $y_0 \in Y_0$, $f_0 \in F$, and $f' \in F$;

c) (Q^t_γ, V^t_γ) where $\gamma \in \Gamma^t$, are segments of nonhomogeneous memoryless channels of length t with finite sets of signals; moreover, the conditions V^t_γ are absent, and the functions

$$Q^t_\gamma(y, \cdot) = Q_{a^1_\gamma}(y_1, \cdot) Q_{a^2_\gamma}(y_2, \cdot) \cdots Q_{a^t_\gamma}(y_t, \cdot) \tag{85}$$

where the set of $(a^1_\gamma, \cdots, a^t_\gamma)$ with $\gamma \in \Gamma^t$ consists of all possible sets of elements $a \in A$, and $\{Q_a, a \in A\}$ is a certain convex set of transition functions. (The set $\{Q_a, a \in A\}$ of transition functions is said to be *convex* if, for Q_a with $a \in A$ and $Q_{\bar{a}}$ with $\bar{a} \in A$ given by the matrices

$$(q^a_{ij}) \quad \text{and} \quad (q^{\bar{a}}_{ij}) \tag{86}$$

respectively, the function corresponding to the matrix

$$(\lambda_i q_{ij}^a + (1 - \lambda_i) q_{ij}^a), \qquad \text{with } 0 \leqq \lambda_i < 1, \tag{87}$$

also belongs to A.)

a′) Δ^t consists of a finite number r (independent of t) of messages (p_i^t, W_i^t) where $i = 1, \cdots, r$, and all the sequences of messages (p_i^t, W_i^t) where $t = 1, 2, \cdots$, are information stable.

b′) The conditions W_δ^t are independent of δ, and the distributions p_δ^t can be described as the distribution of a set $\{f_\delta(\zeta_1^\delta), \cdots, f_\delta(\zeta_t^\delta)\}$, where the ζ_t^δ form a finite homogeneous Markov chain with a fixed number of states. Moreover, all the elements of the transition probability matrix are larger than a positive constant α, uniformly in δ.

c′) The conditions W_δ^t are independent of δ and the (p_δ^t, W_δ^t) are messages which are nonhomogeneous in time with component-wise conditions of accuracy of reproduction, and with independent components. Moreover,

$$p_\delta^t = p^{a_\delta^1} \times \cdots \times p^{a_\delta^t}, \tag{88}$$

where $(a_\delta^1, \cdots, a_\delta^t)$, with $\delta \in \Delta$, are all the possible sets of elements $a \in A$, and $\{p_a, a \in A\}$ is a convex set of probability distributions on the finite space X_0.

The above examples show that the direct Shannon theorem is true in a rather wide setting. On the other hand, its generality is less than that of the Shannon theorem of section 3.10. Thus, the conclusion of the theorem will generally be false if among the (p_δ, W_δ) there are messages with identical p_δ but different W_δ. (Here, everything is reduced to the necessity of satisfying a stronger accuracy condition, which turns out to be the intersection of several W_δ; a solution of this problem can be obtained by the methods of section 3.) Further, if in case (c) (and similarly in c′) the set $\{Q_a, a \in A\}$ is not convex, then the Shannon theorem holds only if in definition (80) [and similarly in equation (81)] the system Q^δ is replaced by its convex hull.

6.4. Note several important special cases of equations (80) and (81) for the capacity and entropy of systems. Let (Q_γ, V_γ) with $\gamma \in \Gamma$, be the system of all memoryless channels of length n with additive noise, with noise power p_n and signal power p_s. Then $C(\Gamma)$ is given by Shannon formula (43). Further, if (Q_γ, V_γ) with $\gamma \in \Gamma$ is the system of all transmitters with finite bandwidth W and additive noise for transmission during time T, then $C(\Gamma)$ is given asymptotically by Shannon formula (46). Finally, if (p_δ, W_δ), with $\delta \in \Delta$, is the system of all messages with component-wise accuracy conditions given by bounding the mean square error by the constant p_d, whose input message components have a dispersion equal to p_c, then expression (45) holds for $H(\Delta)$. Equations (43), (45), and particularly (46), which are usually derived for the Gaussian case, are in practice applied to a much wider class of cases, including the case where it is impossible to say anything definite about the character of the distributions. The new interpretation of these formulas given above can serve as a justification of such a practice.

6.5. Still open is the question of creating general methods for calculating the capacity $C(\Gamma)$ and the entropy $H(\Delta)$, and also the corresponding "optimal strategies," that is, the values of η, γ and δ at which the extremes in equations (79) and (80) are attained (compare problems XV, XIX and XX). Further, it seems interesting, after the fashion of the problems analyzed in section 5, to investigate the asymptotic error probability for the unique transmission methods. A possible first step might be an investigation of the error probability for the system of all memoryless channels having square matrices of fixed order, and having moreover diagonal elements $p_{ii} \geqq 1 - \epsilon$, where ϵ is a given constant.

L

LI

7. Coding with the use of supplementary information

7.1. It is sometimes natural to assume that at the input of a channel certain supplementary information about the state of the channel or about the output signals at earlier moments of time are known and can therefore be utilized in coding. (The channel may yield such information by feedback, or the information may be obtained by certain methods of investigating the medium through which the transmission is propagated.) The question arises of how to alter in this case the statement and the solution of Shannon's problem. This question has been studied in certain special cases by Shannon [80] and [83] and by the author [16].

7.2. For simplicity, the complete mathematical formulation of the problem will be given only for the case of a homogeneous channel with discrete time, although the reader will understand without difficulty how this definition may be extended to the general case. Assume that, in addition to the objects which enter into the definition of a segment of length n of a homogeneous channel with discrete time, we are given a sequence of measurable spaces (D_k, S_D^k) called the *spaces of supplementary information at time k*. Also given is the transition function

$$R_k(y_1, \cdots, y_{k-1}; \tilde{y}_1, \cdots, \tilde{y}_{k-1}; f_1, \cdots, f_k; A), \qquad k = 1, \cdots, n \tag{89}$$

where $y_i \in Y_0$, $\tilde{y}_i \in \tilde{Y}_0$, $f_i \in F$, and $A \in S_D^k$, which is a probability measure on S_D^k for fixed values of its first $3k - 2$ arguments. From the intuitive point of view, $R_k(\cdot\,, \cdot)$ gives the probability distribution of information present at the input of the channel at time k, given the past values of signals and channel states. Thus, the given formulation includes the case where the information may be random due, for example, to noise in the feedback channel. However, in this case it is assumed that the method of transmission through the feedback channel is fixed. An interesting problem would be that in which only the feedback channel itself is given, and the method of transmission through it is to be chosen optimally. There are no obvious approaches to the solution of such a problem.

LII

Let us note some very important cases. If $D_k = Y_0^{k-1} \times \tilde{Y}_0^{k-1} \times F^k$ and

$R_k(d, \cdot)$ with $d \in D_k$ is concentrated at the point d, then we shall say that we are given a *channel with perfect information about the past.* If $D_k = \tilde{Y}_0^{k-1}$, and $R_k(\cdot\ , \cdot)$ depends only on $\tilde{y}_1, \cdots, \tilde{y}_{k-1}$, and is concentrated at the corresponding point, then we shall say that we are given a *channel with perfect feedback* (sometimes inaccurately described as a feedback channel with infinite capacity). Finally, if $D_k = F$, and $R_k(\cdot\ , \cdot)$ depends only on f_k and is, as a measure, concentrated at the corresponding point, then we shall say that we are given a *channel with perfect information about its state.*

7.3. *A method of transmission for the message* (p, W) *with the use of supplementary information* is said to be given if there is given a system of variables $\xi, \tilde{\xi}, \eta_1, \cdots, \eta_n, \tilde{\eta}_1, \cdots, \tilde{\eta}_n, \phi_1, \cdots, \phi_n, \delta_1, \cdots, \delta_n$ with values $X, \tilde{X}, Y_0, \tilde{Y}_0, F$, and D respectively, such that

a) the pair $(\xi, \tilde{\xi})$ forms the message (p, W),

b) for all k and for all $A \in S_{\tilde{Y}} \times S_F$ there is the conditional probability

(90)

$$P\{(\tilde{\eta}_k, \phi_k) \in A | \eta_1, \cdots, \eta_k, \tilde{\eta}_1, \cdots, \tilde{\eta}_{k-1}, \phi_1, \cdots, \phi_{k-1}, \xi\} = Q_0(\eta_{k-1}, \phi_{k-1}, A),$$

c) the distribution of the variable ϕ_0 coincides with the initial distribution p_0 for the channel,

d) for any k, and any $A \in S_D^k$

(91)
$$P\{\delta_k \in A | \eta_1, \cdots, \eta_{k-1}, \tilde{\eta}_1, \cdots, \tilde{\eta}_{k-1}, \tilde{\phi}_1, \cdots, \tilde{\phi}_{k-1}, \delta_1, \cdots, \delta_{k-1}\}$$
$$= R_k(\eta_1, \cdots, \eta_{k-1}, \tilde{\eta}_1, \cdots, \tilde{\eta}_{k-1}, \tilde{\phi}_1, \cdots, \tilde{\phi}_{k-1}, A),$$

e) for any k and any $A \in S_Y$,

(92)
$$P\{\eta_k \in A | \eta_1, \cdots, \eta_{k-1}, \tilde{\eta}_1, \cdots, \tilde{\eta}_{k-1}, \tilde{\phi}_1, \cdots, \tilde{\phi}_k, \delta_1, \cdots, \delta_k, \xi\}$$
$$= P\{\eta_k \in A | \eta_1, \cdots, \eta_{k-1}, \delta_k, \xi\},$$

f) for any $A \in S_{\tilde{X}}$

(93)
$$P\{\tilde{\xi} \in A | \eta_1, \cdots, \eta_n, \tilde{\eta}_1, \cdots, \tilde{\eta}_n, \phi_1, \cdots, \phi_n, \delta_1, \cdots, \delta_n, \xi\}$$
$$= P\{\tilde{\xi} \in A | \tilde{\eta}_1, \cdots, \tilde{\eta}_n\}.$$

The intuitive meaning of conditions (a), (b), (c), and (d) is clear. Condition (e) means that in choosing the input signal at the kth moment, use is made only of the knowledge of the input message, of previous values of the input signal, and of the supplementary information δ_k. Condition (f) means that decoding is based only on knowledge of the output signal.

There is no supplementary information if the spaces of supplementary information D_k consist of one element each. It is not difficult to see that in this case the above definition is equivalent to that of section 2.12.

7.4. In [82], Shannon studies a channel having perfect information about its state and such that $q_0(y_0, f, \tilde{y}_0, f')$ depends only on f' (if there is no supplementary information, this channel becomes a memoryless channel). Shannon indicates a procedure which makes it possible to reduce the study of a channel with supplementary information to that of an ordinary channel. His arguments can easily

be extended to the most general case as follows. It is always possible to construct a transmitter such that the existence of a method for transmitting the message (p, W) by this transmitter is equivalent to the existence of a method for transmitting the same message through the original channel with supplementary information. This transmitter is constructed as follows: the space $\tilde{Y}$ coincides (as it did in the original channel) with Y_0^n. Further, the spaces $Y_0(D_k)$, for $k = 1, \cdots, n$, of all measurable mappings $y(d_k)$ from D_k into Y are introduced with the naturally induced σ-algebra of measurable subsets. Also, Y is taken as $Y = Y_0(D_1) \times \cdots \times Y_0(D_n)$. As for the transition function of the new transmitter, in the case of a channel with perfect information about its state, the transmitter must be regarded as a channel with the same state space as before, and a new transition density

$$\tilde{q}_0\{y(d_k), f_0, \tilde{y}_0, f'\} = q_0\{y(f_0), \tilde{y}_0, f'\}, \tag{94}$$

where $y(f_0)$ is the value of $y(d_k)$ when $d_k = f_0$. In the general case the definition of the transition function is unwieldy. Thus, we shall only mention that its basic idea remains the same. In those cases where $y(d_k) \in Y(D_k)$ is given at the transmitter input, and the information about the past is equal to d_k^0 (this can be mathematically defined for the new transmitter also, but now it must not influence the coding), it is necessary that transmission should occur, as it did in the previous channel, where the input signal $y(d_k^0)$ was given. Thus, at least formally, the Shannon problem for a channel with supplementary information reduces to the Shannon problem as formulated in section 2. However, since the resulting transmitter turns out to be very complicated, the criteria for information stability discussed in section 3 cannot be applied to it. Therefore, the question of deducing specific conditions for information stability in this case remains open.

LIII

7.5. The question of calculating the capacity of channels with supplementary information is extremely interesting. First, note the following fact. The capacity of a channel with perfect information about its past always coincides with the capacity of the same channel with perfect information about its state. This fact is a corollary of a more general result which will not be quoted here and which, from the intuitive point of view, shows that feedback is useful for increasing the capacity only to the extent to which it provides information about the state of the channel. Further, by using the method developed in [16], it is not difficult to show that the capacity of a channel with supplementary information coincides with the ordinary channel capacity if the following condition is fulfilled. Let η and η_f be random variables connected by a transmitter with the input and output signal spaces (Y_0, S_{Y_0}) and $(\tilde{Y}_0, S_{\tilde{Y}_0})$, and the transition function $Q_0(\cdot, f, \cdot)$. Then, $I(\eta, \eta_f)$ is independent of f and depends only on the distribution of η. The application of this assertion to a memoryless channel yields Shannon's result, [79], which states that feedback does not increase the capacity of a memoryless channel. (At the time of publication of [16], which proves the same result, the author was unfortunately not aware of Shannon's work.)

Another important example of a channel possessing the above property is given by Gaussian transmitters (if they are converted into channels by the method described in section 2). The point is that for such channels, $Q_0(\cdot\ , f, \cdot)$ gives a Gaussian transmitter and, moreover, only the mean value (and not the second moment) depends on f. The variation of the mean value of η_f does not affect the information $I(\eta, \eta_f)$. Thus feedback does not increase the capacity of a Gaussian transmitter. This was first pointed out by M. Pinsker. Shannon, [82], calculated the capacity for the above class of channels with perfect information about their state. Another interesting example is provided by channels with a random parameter and perfect information about the parameter value $b \in B$. Here (compare [47]) the mean capacity is clearly

$$\bar{C} = \int_B \bar{C}_b \tilde{P}(db), \tag{95}$$

where $\bar{C}_b$ is the mean capacity for a conditional channel with parameter b. This formula is easily proved (compare [16]) for the case where the space B is finite. Since it has interesting applications to the case of a continuous distribution of the parameter (Ovseevich, Pinsker, and Tsybakov used it to show
LVI that in certain physically realizable situations, feedback can increase the channel capacity by 50 per cent), it is interesting to prove it for the general case as well (compare section 4.6). It is important to learn how to calculate the capacity of channels with supplementary information in other cases also.

It is interesting to study too the question of whether, in the case of homogeneous memoryless channels, the use of feedback will diminish the optimal probability of error $e\{Q^n, V^n, [2^{nH}]\}$ (see section 5.2). It is not difficult to show that for channels with symmetric matrices, when $H \geqq H_{\text{crit}}$, the use of feedback does not alter the constants a and b in equation (56). In a paper of Elias, [25], there is a remark which can be interpreted as an assertion that, if $H < H_{\text{crit}}$, equation (56), with constants a and b again given by equation (59), holds for a binary symmetric channel with feedback. The author does not know the
LV proof of this theorem, nor how far it can be extended to an arbitrary channel with a symmetric matrix. As was shown in [25], similar facts can easily be proved for a symmetric binary erasure channel. Another immediate problem outstanding in the investigation of the proba-
LVI bility of error $e\{Q^n, V^n, [2^{nH}]\}$ for Gaussian memoryless channels with additive noise and feedback.

8. New applications of the concepts of the Shannon theory of information

8.1. In the preceding sections of this survey, only those questions have been studied which are contained within the framework of the basic Shannon problem of optimal transmission of information as formulated in section 2. Here, it

is desirable to discuss briefly other, as yet hardly noticed, directions in applying the concepts of entropy and information. It is possible that in the future all these directions will merge into some single theory, but so far not even the outlines of such a unified theory are visible.

8.2. The first such direction is the use of a generalized entropy for a pair of distributions as a measure of their difference, in problems of mathematical statistics. A survey of numerous works on this subject is contained in the recent book [55] by S. Kullback. In essence, the published papers contain an enumeration of the properties of generalized entropy, which confirms that it can conveniently be used as a measure of statistical difference. It is felt that the attempt to show that one can obtain an asymptotic answer to certain classical problems of mathematical statistics by means of generalized entropy is more important. This was done in the case of independent observations in papers by E. Mourier [59] and S. A. Aivazian [1]. In a new paper by Dobrushin, Pinsker, and Shiriaev it is shown that the asymptotic results noted can be extended to a very wide class of dependent trials.

8.3. The second promising direction has been presented so far only in the form of an investigation using the concept of entropy of the problem of the identification of false coins with the least possible number of weighings, well known from popular books of mathematical problems. Such an investigation was begun in [46], and was developed in detail in the book [103]. Here we can formulate the general problem of the minimal number of experiments necessary for obtaining certain information, and apparently the asymptotic solution of this problem can be given by means of the entropy concept under certain conditions.

8.4. The author does not understand the reasons for the appearance of the concepts of entropy and information in certain game-theoretic constructions (see [47]). The perspectives of this direction are not clear.

The author thanks A. N. Kolmogorov, M. S. Pinsker and A. M. Yaglom. Continual contact with them is mainly reflected in the content of this paper.

REFERENCES

[1] S. A. Aivazian, "A comparison of the optimal properties for the criteria of the Neyman-Pearson and Wald," *Teor. Veroyatnost. i Primenen.*, Vol. 4 (1959), pp. 86–93.

[2] N. S. Bakhvalov, "On the number of arithmetic operations in solving Poisson's equation for a square by the method of finite differences," *Dokl. Akad. Nauk SSSR*, Vol. 113 (1957), pp. 252–254.

[3] L. A. Bakut, "On the theory of error-correcting codes with an arbitrary basis," *Nauchn. Dokl. Vissh. Shkoli, Radiotek. i Elektron.*, No. 1, 1959.

[4] G. P. Basharin, "On the statistical estimate of the entropy of a sequence of independent random variables," *Teor. Veroyatnost. i Primenen.*, Vol. 4 (1959), pp. 361–364.

[5] R. Bellman, *Dynamic Programming*, Princeton, Princeton University Press, 1957.

[6] D. Blackwell, "The entropy of function of finite state Markov chains," *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, 1957, pp. 13–20.

[7] ———, "Infinite codes for memoryless channels," *Ann. Math. Statist.*, Vol. 30 (1959), pp. 1242–1244.
[8] D. BLACKWELL, L. BREIMAN, and A. J. THOMASIAN, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 1209–1220.
[9] ———, "The capacity of a class of channels," *Ann. Math. Statist.*, Vol. 30 (1959), pp. 1223–1241.
[10] L. F. BORODIN, "Some questions in the theory of constructing error-correcting codes," The A. S. Popov Scientific Technical Society of Radio Engineering and Electronics, *Transactions No.* 2, Moscow (1958), pp. 110–151.
[11] L. BREIMAN, "The individual ergodic theorem of information theory," *Ann. Math. Statist.*, Vol. 29 (1957), pp. 809–811.
[12] L. CARLESON, "Two remarks on the basic theorems of information theory," *Scand. Math.*, Vol. 6 (1958), pp. 175–180.
[13] H. CRAMÉR, "Sur un nouveau théorème—limite de la théorie des probabilités," *Les Sommes et les Fonctions de Variables Aléatoires*, Paris, Hermann, 1938.
[14] W. B. DAVENPORT and W. L. ROOT, *An Introduction to the Theory of Random Signals and Noise*, New York, McGraw-Hill, 1958.
[15] R. L. Dobrushin, "On the formulation of the fundamental Shannon theorem" (a summary of an address delivered to a meeting of a seminar in probability theory, March 19, 1957), *Teor. Veroyatnost. i Primenen.*, Vol. 2 (1957), pp. 480–482.
[16] ———, "The transmission of information along a channel with feedback," *Teor. Veroyatnost. i Primenen.*, Vol. 3 (1958), pp. 395–412.
[17] ———, "A simplified method for an experimental estimate of the entropy of a stationary sequence," *Teor. Veroyatnost. i Primenen.*, Vol. 3 (1958), pp. 462–464.
[18] ———, "A general formulation of the fundamental Shannon theorem in information theory," *Dokl. Akad. Nauk SSSR*, Vol. 126 (1959), pp. 474–477.
[19] ———, "A general formulation of the fundamental Shannon theorem in information theory," *Uspehi Mat. Nauk*, Vol. 14 (1959), pp. 3–104.
[20] ———, "The optimal transmission of information along a channel with unknown parameters," *Radiot. i Elektron.*, Vol. 4 (1959), pp. 1951–1956.
[21] ———, "A limiting transition under the symbol of information and entropy," *Teor. Veroyatnost. i Primenen.*, Vol. 5 (1960), pp. 28–37.
[22] J. L. DOOB, "Editorial," *IRE Trans. Inf. Theory*, IT-5, No. 1 (1959), p. 3.
[23] E. B. DYNKIN, *The Foundations of the Theory of Markov Processes*, Moscow, Fizmatgiz, 1959.
[24] P. ELIAS, "Coding for noisy channels," *IRE Convention Record*, No. 4 (1955), pp. 37–46.
[25] ———, "Coding for two noisy channels," *Information Theory, Third London Symposium*, 1955 and 1956, pp. 61–74.
[26] M. A. EPSTEIN, "Algebraic decoding for a binary erasure channel," *IRE Convention Record*, No. 4 (1958), pp. 56–59.
[27] V. EROKHIN, "The ϵ-entropy of discrete random object," *Teor. Veroyatnost. i Primenen.*, Vol. 3 (1958), pp. 103–107.
[28] D. K. FADDEEV, "On the concept of the entropy for a finite probability model," *Uspehi Mat. Nauk*, Vol. 11 (1958), pp. 227–231.
[29] A FEINSTEIN, "A new basic theorem of information theory," *Trans. IRE* (1954), PG-IT-4, pp. 2–22.
[30] ———, "Error bounds in noisy channels without memory," *IRE Trans. Inf. Theory*, Vol. 1 (1955), pp. 13–14.
[31] ———, *Foundations of Information Theory*, New York, McGraw-Hill, 1958.
[32] ———, "On the coding theorem, and its converse for finite-memory channels," *Information and Control*, Vol. 2 (1959), pp. 25–44.

[33] ———, "On the coding theorem, and its converse for finite-memory channels," *Nuovo Cimento*, Suppl. 13, No. 2 (1959), pp. 345–352.

[34] J. FLEISCHER, "The central concepts of communication theory for infinite alphabets," *J. Math. Phys.*, Vol. 37 (1958), pp. 223–228.

[35] A. B. FONTAINE and W. W. PETERSON, "On coding for binary symmetric channels," *Trans. Amer. Inst. Elec. Eng.*, Vol. 77 (1958), pp. 638–647.

[36] I. M. GELFAND, A. N. KOLMOGOROV, and I. M. YAGLOM, "Towards a general definition of the quantity of information," *Dokl. Akad. Nauk SSSR*, Vol. 111 (1956), pp. 745–748.

[37] ———, "The quantity of information and entropy for continuous distributions," *Proceedings of the Third Mathematical Congress, Moscow*, Iz-vo Akad. Nauk SSSR, 1958, Vol. 3, pp. 300–320.

[38] I. M. GELFAND and A. M. YAGLOM, "On the calculation of the quantity of information about a random function contained in another such function," *Uspehi Mat. Nauk*, Vol. 12 (1957), pp. 3–52.

[39] E. N. GILBERT and E. F. MOORE, "A variable-length binary encoding," *Bell System Tech. J.*, Vol. 38 (1959), pp. 933–967.

[40] B. V. GNEDENKO and A. N. KOLMOGOROV, *Limit Distributions for the Sums of Independent Variables*, Moscow and Leningrad, Gostekhizdat, 1958.

[41] I. J. GOOD and K. C. DOOG, "A paradox concerning the rate of information," *Information and Control*, Vol. 1 (1958), pp. 91–112.

[42] R. W. HAMMING "Error detecting and error correcting codes," *Bell System Tech. J.*, Vol. 29 (1950), pp. 147–160.

[43] K. JACOBS, "Die Ubertragung diskreter Informationen Durch Periodische und Fastperiodische Kanöle," *Math. Ann.*, Vol. 137 (1959), pp. 125–135.

[44] A. D. JOSHI, "A note on upper bounds for minimum distance codes," *Information and Control*, Vol. 1 (1958), pp. 289–295.

[45] A. A. JUSHKEVICH, "On limit theorems connected with the concept of the entropy of Markov chains," *Uspehi Mat. Nauk*, Vol. 8 (1953), pp. 177–180.

[46] P. J. KELLOG and D. J. KELLOG, "Entropy of information and the odd ball problem," *J. Appl. Phys.*, Vol. 25 (1954), pp. 1438–1439.

[47] J. L. KELLY, JR., "A new interpretation of information rate," *Bell System Tech. J.*, Vol. 35 (1956), pp. 917–926.

[48] A. I. KHINCHIN, "The concept of entropy in probability theory," *Uspehi Mat. Nauk*, Vol. 8 (1953), pp. 3–20.

[49] ———, "On the basic theorems of information theory," *Uspehi Mat. Nauk*, Vol. 11 (1956), pp. 17–75.

[50] A. N. KOLMOGOROV, "Some work of recent years on limit theorems in probability theory," *Vestnik Moskov. Univ. Ser. Fiz.-Mat. Estest Nauk*, Vol. 8 (1953), pp. 29–38.

[51] ———, "The theory of the transmission of information," 1956, Plenary session of the Academy of Sciences of the USSR on the automatization of production, Moscow, Izd. Akad. Nauk SSSR, 1957, pp. 66–99.

[52] ———, "A new metric invariant of dynamic systems and of automorphisms in Lebesgue spaces," *Dokl. Akad. Nauk SSSR*, Vol. 119 (1958), pp. 861–864.

[53] ———, "On entropy per unit time as a metric invariant of automorphisms," *Dokl. Akad. Nauk SSSR*, Vol. 124 (1959), pp. 754–755.

[54] A. N. KOLMOGOROV and V. N. TIKHOMIROV, "The ϵ-entropy and the ϵ-capacity of sets in metric spaces," *Uspehi Mat. Nauk*, Vol. 14 (1959), pp. 3–86.

[55] S. KULLBACK, *Information Theory and Statistics*, New York, Wiley, 1959.

[56] Z. A. LOMNIZKY and S. K. ZAREMBA, "The asymptotic distributions of the amount of transmitted information," *Information and Control*, Vol. 2 (1959), pp. 266–284.

[57] E. J. MCCLUSKEY, JR., "Error correcting codes—a linear programming approach," *Bell System Tech. J.*, Vol. 38 (1959), pp. 1485–1512.

[58] B. McMillan, "The basic theorem of information theory," *Ann. Math. Statist.*, Vol. 24 (1953), pp. 196–219.

[59] E. Mourier, "Étude du choix entre-deux lois de probabilité," *C. R. Acad. Sci. Paris*, Vol. 223 (1946), pp. 712–714.

[60] S. Muroga, "On the capacity of a discrete channel. 1. Mathematical expression of capacity of a channel which is disturbed by noise in its very one symbol and expressible in one state diagram," *J. Phys. Soc. Japan*, Vol. 8 (1953), pp. 484–494.

[61] ———, "On the capacity of a noisy continuous channel," *IRE Trans. Inf. Theory*, IT-3 (1957), pp. 44–51.

[62] J. Nedoma, "The capacity of a discrete channel," *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, 1957, pp. 143–182.

[63] I. A. Ovseevich and M. S. Pinsker, "The speed of transmission of information, the capacity of a multi-channel system, and a procedure by a transformation method by a linear operator," *Radiotek.*, No. 3 (1959), pp. 9–21.

[64] A. Perez, "Notions generalisées d'uncertitude, d'entropie, et d'information du point de vue de la théorie de martingales," *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, 1957, pp. 183–208.

[65] ———, "Sur la théorie de l'information dans le cas d'un alphabet abstrait," *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, 1957, pp. 209–244.

[66] ———, "Sur la convergence des uncertitudes, entropies et information enchantillon vers leurs valeurs vraises," *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Prague, 1957, pp. 245–252.

[67] ———, "Information theory with an abstract alphabet. Generalized aspects of McMillan's limit theorem for the case of discrete and continuous time," *Teor. Veroyatnost. i Primenen.*, Vol. 4 (1959), pp. 105–109.

[68] M. S. Pinsker, "The quantity of information about a Gaussian random process contained in a second process which is stationary with respect to it," *Dokl. Akad. Nauk SSSR*, Vol. 99 (1954), pp. 213–216.

[69] ———, "The evaluation of the rate of creation of messages by a stationary random process and the capacity of a stationary channel," *Dokl. Akad. Nauk SSSR*, Vol. 111 (1956), pp. 753–756.

[70] ———, "The extrapolation of homogeneous random fields and the quantity of information about a Gaussian random field contained in another Gaussian random field," *Dokl. Akad. Nauk SSSR*, Vol. 112 (1957), pp. 815–818.

[71] ———, "The extrapolation of random vector processes and the quantity of information contained in one vector stationary random process with respect to another one, related to it in a stationary manner," *Dokl. Akad. Nauk SSSR*, Vol. 121 (1958), pp. 49–51.

[72] ———, "Information and information stability of random variables and processes," Moscow, Izd. Akad. Nauk SSSR, 1960.

[73] K. Powers, "A prediction theory approach to information rates," *IRE Convention Record*, Vol. 4 (1956), pp. 132–139.

[74] A. Rényi and J. Balatoni, "Uber den Begriff der Entropie," *Math. Forsch.*, Vol. 4 (1957), pp. 117–134.

[75] M. Rosenblatt-Rot, "The entropy of stochastic processes," *Dokl. Akad. Nauk SSSR*, Vol. 112 (1957), pp. 16–19.

[76] ———, "The theory of the transmission of information through statistical communication channels," *Dokl. Akad. Nauk SSSR*, Vol. 112 (1957), pp. 202–205.

[77] ———, "The normalized ϵ-entropy of sets and the transmission of information from continuous sources through continuous communication channels," *Dokl. Akad. Nauk SSSR*, Vol. 130 (1960), pp. 265–268.

[78] G. S. RUBINSTEIN and K. URBANIK, "The solution of an extremal problem," *Teor. Veroyatnost. i Primenen.*, Vol. 2 (1957), pp. 375–377.
[79] C. E. SHANNON, "A mathematical theory of communi cation," *Bell System Tech. J.*, Vol. 27 (1948), pp. 379–423 and 623–656.
[80] ———, "The zero capacity of a noisy channel," *IRE Trans. Inf. Theory*, IT-2 (1956), pp. 8–19.
[81] ———, "Some geometric results in channel capacity," *Nachr. Tech. Fachber.*, Vol. 6 (1956), pp. 13–15.
[82] ———, "Certain results in coding theory for noisy channels," *Information and Control*, Vol. 1 (1957), pp. 6–25.
[83] ———, "Channels with side information at the transmitter," *IBM J. Res. Devel.*, Vol. 2 (1958), pp. 289–293.
[84] ———, "Coding theorems for a discrete source with a fidelity criterion," *IRE Convention Record*, Vol. 7 (1959), pp. 142–163.
[85] ———, "Probability of error for optimal codes in a Gaussian channel," *Bell System Tech. J.*, Vol. 38 (1959), pp. 611–655.
[86] C. E. SHANNON and J. MCCARTHY, *Automata Studies*, Princeton, Princeton University Press, 1956.
[87] H. S. SHAPIRO and D. Ł. SLOTNICK, "On the mathematical theory of error-correcting codes," *IBM J. Res. Devel.*, Vol. 3 (1959), pp. 25–34.
[88] D. SLEPIAN, "A class of binary signalling alphabets," *Bell System Tech. J.*, Vol. 35 (1956), pp. 203–234.
[89] K. TAKANO, "On the basic theorem of information theory," *Ann. Inst. Statist. Math., Tokyo*, Vol. 9 (1958), pp. 53–77.
[90] I. P. TSAREGRADSKII, "A note on the capacity of a stationary channel with finite memory," *Teor. Veroyatnost. i Primenen.*, Vol. 3 (1958), pp. 84–96.
[91] B. S. TSYBAKOV, "On the capacity of two-path communication channels," *Radiotek. i Elektron.*, Vol. 4 (1959), pp. 1117–1123.
[92] ———, "On the capacity of channels with a large number of paths," *Radiotek. i Elektron.*, Vol. 4 (1959), pp. 1427–1433.
[93] ———, "The capacity of certain multi-path channels," *Radiotek. i Elektron.*, Vol. 4 (1959), pp. 1602–1608.
[94] TZIAN-TZE-PEI, "A note on the definition of quantity of information," *Teor. Veroyatnost. i Primenen.*, Vol. 3 (1958), pp. 99–103.
[95] W. ULRICH, "Non-binary error-correcting codes," *Bell System Tech. J.*, Vol. 36 (1957), pp. 53–77.
[96] A. G. VITUSHKIN, *An Estimate of the Complexity of Tabulation Problems*, Moscow, Fizmatgiz, 1959.
[97] J. WOLFOWITZ, "The coding of messages subject to chance error," *Illinois J. Math.*, Vol. 1 (1957), pp. 591–606.
[98] ———, "An upper bound of the rate of transmission of messages," *Illinois J. Math.*, Vol. 2 (1958), pp. 137–141.
[99] ———, "The maximum achievable length of an error-correcting code," *Illinois J. Math.*, Vol. 2 (1958), pp. 454–458.
[100] ———, "Information theory for mathematics," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 351–356.
[101] ———, "Strong converse of the coding theorem for semicontinuous channels," *Illinois J. Math.*, Vol. 3 (1959), pp. 477–489.
[102] J. M. WOZENCRAFT, "Sequential decoding for reliable communication," *IRE Convention Record*, Vol. 5 (1957), pp. 11–25.
[103] A. M. YAGLOM and I. M. YAGLOM, *Probability and Information*, Moscow, Gostekhizdat, 1960.