In addition to the scheduled program, the following two papers, by A. N. Kolmogorov and V. I. Siforov, were presented at the 1956 Symposium on Information Theory. However, the manuscripts were received too late for inclusion in the September (Symposium) issue of these TRANSACTIONS. The papers were submitted in response to our invitation to these distinguished Russian scientists, and the following translations were distributed to those attending the Symposium.-The Editor.

# On the Shannon Theory of Information Transmission in the Case of Continuous Signals\*

ANDREI N. KOLMOGOROV<sup>†</sup>

#### I. INTRODUCTION

THE ROLE of the entropy of a random object  $\xi$ , capable of taking the values  $x_1, x_2, \dots, x_n$  with the probabilities  $p_1, p_2, \cdots, p_n$ ,

$$H(\xi) = -\sum_{k} p_{k} \log p_{k}$$

in information theory and in the theory of information transmission using discrete signals, can be considered to have been explained sufficiently. Furthermore, I insist that the fundamental concept, which admits of generalization to perfectly arbitrary continuous information and signals, is not directly the entropy concept but the concept of the quantity of information  $I(\xi, \eta)$  in the random object  $\xi$  relative to the object  $\eta$ . In the discrete case this quantity is evaluated correctly according to the wellknown Shannon formula:<sup>1</sup>

$$I(\xi, \eta) = H(\eta) - MH(\eta/\xi).$$

For a finite-dimensional distribution, possessing density, the quantity  $I(\xi, \eta)$  is determined, according to Shannon, by the analogous formula

$$I(\xi, \eta) = h(\eta) - Mh(\eta/\xi),$$

where  $h(\eta)$  is the "differential entropy"

$$h(\eta) = -\int p(y) \log p(y) \, dy,$$

and  $h(\eta/\xi)$  is the conditional differential entropy defined in an analogous manner. It is well known that the quantity  $h(\xi)$  has no direct real interpretation and is not even invariant with respect to coordinate transformation in the space of the x's. For an infinitely-dimensional distribution, the analog of  $h(\xi)$  is nonexistent, in general.

According to the proper meaning of the word, the entropy of the object  $\xi$  with a continuous distribution is always infinite. If the continuous signals can, nevertheless, serve to transmit finitely great information, then it is only because they are always observed with bounded accuracy. Consequently, it is natural to define the appropriate " $\epsilon$ -entropy"  $H_{\epsilon}(\xi)$  of the object  $\xi$  by giving the accuracy of observation  $\epsilon$ . Shannon did thus under the designation "rate of creating information with respect to a fidelity criteron." Although choosing a new name for this quantity does not alter the situation, I decided to call your attention to that proposition by underlining the more widespread interest in the concept and its deep analogy to the ordinary exact entropy. I imply, beforehand, that, as remarked in Section IV, the theorem on the extremal role of the normal distribution (both in the finite-dimensional and the infinite-dimensional cases) is retained for the  $\epsilon$  entropy. Furthermore, I give in Sections II and III, without pretending to its unconditional newness, an abstract formulation of the definition and fundamental properties of  $I(\xi, \eta)$  and a survey of the fundamental problems of the Shannon theory of information transmission. Certain specific results obtained recently by Soviet investigators are explained in Sections IV to VI. I wish to emphasize, especially, the very significant interest, as it appears to me, in the investigations of the asymptotic behavior of the  $\epsilon$  entropy as  $\epsilon \rightarrow 0$ . The cases investigated earlier

$$H_{\epsilon}(\xi) \sim n \log \frac{1}{\epsilon}; \qquad \widetilde{H}_{\epsilon}(\xi) = 2w$$

where n is the number of measurements and w is the bandwidth of the spectrum, are only very particular cases of the rules which can be encountered here. In order to understand the perspectives disclosed here, my note,<sup>2</sup> explained in another terminology, might be of interest: hence, I am placing a certain number of reprints at the disposal of the participants of the symposium.

To a considerable degree, my report reproduces the contents of a report presented jointly with Iaglom and Gel'fand at the Third All-Union Mathematics Conference.

<sup>\*</sup> Presented at 1956 Symposium on Information Theory at Mass. Inst. Tech., Cambridge, Mass., September 10-12, 1956. Translated by Morris D. Friedman.

<sup>†</sup> Academician, Academy of Science, USSR. <sup>1</sup> It seems expedient to me that the notation  $H(\eta/x)$  is the conditional entropy of  $\eta$  for  $\xi = x$  and  $MH(\eta/\xi)$  is the matical expectation of this conditional entropy for the variable  $\xi$ .

<sup>&</sup>lt;sup>2</sup> A. N. Kolmogorov, Doklady, AN USSR, vol. 108, no. 3, pp 385-388; 1956.

However, since the present symposium is of a more engineering character, I omitted a number of mathematical details. The work of Khinchin on the logical foundations of the theory remains beyond the limits of my survey.

As regards the work of Soviet radio engineers, you will hear about some of them from the other speakers. In the note itself, I will have occasion to note only the interest, in principle, of certain early work of Kotel'nikov, circa 1933 (see Section VI, further).

# II. QUANTITY OF INFORMATION IN ONE RANDOM Object Relative to Another

Let  $\xi$  and  $\eta$  be random objects with regions of possible values X and Y,

$$P_{\xi}(A) = P(\xi \varepsilon A); \qquad P_{\eta}(B) = P(\eta \varepsilon B)$$

the appropriate probability distributions, and

$$P_{\xi\eta}(C) = P((\xi, \eta) \varepsilon C)$$

the joint probability distribution of the objects  $\xi$  and  $\eta$ . By definition, the quantity of information in the random object  $\xi$  relative to the random object  $\eta$  is given by the formula

$$I(\xi, \eta) = \int_{X} \int_{Y} P_{\xi\eta}(dx \, dy) \log \frac{P_{\xi\eta}(dx \, dy)}{P_{\xi}(dx)P_{\eta}(dy)}.$$
 (1)

The exact meaning of this formula requires certain elucidation and the general properties of  $I(\xi, \eta)$ , given later, are correct only for certain limitations of a set-theoretical character on the distributions  $P_{\xi}$ ,  $P_{\eta}$  and  $P_{\xi\eta}$ , but I will not dwell on this here. In every case, the general theory can be explained, without great difficulty, in such a way that it will be applicable to random objects  $\xi$  and  $\eta$  of very general nature (vectors, functions, generalized functions, etc.).

Eq. (1) can be considered to be due to Shannon although he was limited to the case

$$P_{\xi}(A) = \int_{A} p_{\xi}(x) \, dx; \qquad P_{\eta}(B) = \int_{B} p_{\eta}(y) \, dy$$
$$P_{\xi\eta}(C) = \iint_{C} p_{\xi\eta}(x, y) \, dx \, dy$$

when (1) transforms into

$$I(\xi, \eta) = \int_{X} \int_{Y} P_{\xi\eta}(x, y) \log \frac{P_{\xi\eta}(x, y)}{P_{\xi}(x)P_{\eta}(y)} \, dx \, dy.$$

Sometimes, it is useful to represent the distribution P as

$$P_{\xi\eta}(C) = \iint_{C} a(x, y) P_{\xi}(dx) P_{\eta}(dy) + S(C)$$
(2)

where the function S(C) is singular relative to the product

$$P_{\xi} \times P_{\eta}.$$

If the singular component of S is lacking, then the formula

$$\alpha_{\xi\eta} = a(\xi, \eta) \tag{3}$$

determines the random quantity  $\alpha_{i\eta}$  uniquely to the accuracy of probability zero. Sometimes, the following theorem formulated by Gel'fand and Iaglom<sup>3</sup> is useful.

Theorem: If  $S(X \times Y) > 0$ , then  $I(\xi, \eta) = \infty$ . If  $S(X \times Y) = 0$ , then

$$\begin{split} I(\xi, \eta) &= \int_{X} \int_{Y} a(x, y) \log a(x, y) P_{\xi}(dx) P_{\eta}(dy) \\ &= \int_{X} \int_{Y} \log a(x, y) P_{\xi\eta}(dx \, dy) \\ &= M \log \alpha_{\xi\eta}. \end{split}$$
(4)

Let us enumerate certain fundamental properties of  $I(\xi, \eta)$ .

- 1)  $I(\xi, \eta) = I(\eta, \xi)$ .
- 2)  $I(\xi, \eta) \ge 0$ ;  $I(\xi, \eta) = 0$  only if  $\xi$  and  $\eta$  are independent.
- 3) If the pair  $(\xi_1, \eta_1)$  and  $(\xi_2, \eta_2)$  are independent, then

$$I[(\xi_1, \xi_2), (\eta_1, \eta_2)] = I(\xi_1, \eta_1) + I(\xi_2, \eta_2).$$

- 4)  $I[(\xi, \eta), \zeta] \ge I(\xi, \zeta).$
- 5)  $I[(\xi, \eta), \zeta] = I(\eta, \zeta)$ , if and only if  $\xi, \eta, \zeta$  is a Markov sequence, *i.e.*, if the conditional distribution of  $\zeta$  depends only on  $\eta$  for fixed  $\xi$  and  $\eta$ .

Apropos property 4, it is useful to note the following. In the case of the entropy

$$H(\xi) = I(\xi, \xi),$$

there is the bound on the entropy of the  $(\xi, \eta)$  pair from above:

$$H(\xi, \eta) \leq H(\xi) + H(\eta)$$

as well as the bound from below which results from 1 and 4.

$$H(\xi, \eta) \ge H(\xi); \qquad H(\xi, \eta) \ge H(\eta).$$

A similar estimate for the quantity of information in  $\zeta$ relative to the  $(\xi, \eta)$  pair does not exist. From

$$I(\xi, \zeta) = 0; \qquad I(\eta, \zeta) = 0$$

there still does not result the equality

$$I[(\xi, \eta), \zeta] = 0,$$

as can be shown by elementary examples.

For later use, let us note the special case when  $\xi$  and  $\eta$  are the random vectors:

$$\begin{aligned} \boldsymbol{\xi} &= (\xi_1, \cdots, \xi_m) \\ \boldsymbol{\eta} &= (\eta_1, \cdots, \eta_n) = (\xi_{m+1}, \cdots, \xi_{m+n}), \end{aligned}$$

and the quantities

$$\xi_1, \xi_2, \cdots, \xi_{m+n}$$

are distributed normally with the second central moments

$$s_{ij} = M[(\xi_i - M\xi_i)(\xi_j - M\xi_j)].$$

<sup>3</sup> A. N. Kolmogorov, A. M. Iaglom, and I. M. Gel'fand, "Quantity of Information and Entropy for Continuous Distributions," Report at Third All-Union Math. Conf., 1956. If the determinant

$$C = |s_{ij}|_{1 \le i,j \le m+n}$$

is not zero, then, as was calculated by Gel'fand and Iaglom

$$I(\xi, \eta) = \frac{1}{2} \log \frac{AB}{C}$$
(5)

where

$$A = [s_{ij}|_{1 \leq i,j \leq m}; \qquad B = [s_{ij}|_{m < i,j \leq m+n}]$$

It is often more expedient, however, to use another approach without the C > 0 limitation. As is known,<sup>4</sup> all the second moments  $s_{ij}$  except those for which i = j or j = m + i go to zero after a suitable linear coordinate transformation in the X and Y spaces. For such a choice of coordinates

$$I(\xi, \eta) = -\frac{1}{2} \sum [1 - r^2(\xi_k, \eta_k)]$$
(6)

where the summation is taken over those

 $k \leq \min(m, n)$ 

for which the denominator in the expression of the correlation coefficient

$$r(\xi_k, \eta_k) = \frac{s_{k,m+k}}{\sqrt{s_{kk} \cdot s_{m+k,m+k}}}$$

is not zero.

## III. Abstract Explanation of the Principles of the Shannon Theory

Shannon considers the transmission of information according to the scheme

 $\xi \to \eta \to \eta' \to \xi'$ 

where the "transmitting apparatus"

$$\eta \rightarrow \eta'$$

is characterized by the conditional distribution

$$P_{\eta'/\eta}(B'/y) = P(\eta' \varepsilon B'/\eta = y)$$

of the "output signal"  $\eta'$  for a given "input signal"  $\eta$ and a certain limitation

#### $P_{\eta} \epsilon V$

of the input signal distribution  $P_{\eta}$ . The "coding"

Ę

$$\rightarrow \eta$$

and "decoding" operations

$$\eta' \rightarrow \xi'$$

are characterized by the conditional distributions

$$\begin{split} P_{\eta/\xi}(B/x) &= P(\eta \ \epsilon \ B/\xi \ = \ x) \\ P_{\xi'/\eta'}(A'/y') &= P(\xi' \ \epsilon \ A'/\eta' \ = \ y'). \end{split}$$

<sup>4</sup> A. M. Obukhov, *Izv. AN USSR*, *Phys.-Math. Series*, pp. 339-370; 1938.

The fundamental Shannon problem is the following. Given the spaces X, X', Y, Y' of possible values of the "input message"  $\xi$ , the "output message"  $\xi'$ , the input signal  $\eta$ , and the output signal  $\eta'$ ; given the characteristics of the transmitter, *i.e.*, the conditional distribution  $P_{\eta'/\eta}$  and the class V of admissible input signal distributions  $P_{\eta}$ ; finally, given the distribution

$$P_{\xi}(A) = P(\xi \varepsilon A)$$

of the input message and the "fidelity criterion"

where W is a certain class of joint distributions

$$P_{\xi\xi'}(C) = P[(\xi, \xi') \varepsilon C]$$

of the input and output communications. To find: Is it possible, and if it is, by what means, to give a coding and decoding rule (*i.e.*, the conditional distributions  $P_{\eta/\xi}$  and  $P_{\xi'/\eta'}$ ) in such a manner that by calculating the distribution  $P_{\xi\xi'}$  in terms of the distributions  $P_{\xi}$ ,  $P_{\eta'/\eta}$ ,  $P_{\xi'/\xi}$  under the assumption that the sequence

$$\xi, \, \eta, \, \xi', \, \eta'$$

is Markovian, we will obtain

 $P_{\mathbf{F}\mathbf{F}'} \mathbf{\epsilon} W?$ 

As does Shannon, so let us define the "capacity" of the transmitter thus

$$C = \sup_{\mathbf{R} \to V} I(\eta, \eta')$$

and let us introduce the quantity

$$H_{W}(\xi) = \inf_{P_{\xi\xi' \in W}} I(\xi, \xi')$$

which Shannon calls the "rate of creating information relative to a fidelity criterion" when computed per unit time. Then, the *necessary condition of the possibility of transmission* 

$$H_{W}(\xi) \le C \tag{7}$$

results at once from property 5 of Section II.

The incomparably deep idea of Shannon is that (7), when applied to the continuous operation of a "communication channel," is "almost sufficient" in a certain sense and under certain very broad conditions. From the mathematical point of view, it is a matter here of proving a limit theorem of the following type. It is assumed that the space X, X', Y, Y' of the distributions  $P_{\xi}$  and  $P_{\pi'/\pi}$  of the classes V and W, and therefore, of the quantities C and  $H_W(\xi)$ , depend on the parameter T (which plays the role, in applications, of the duration of transmitter operation). It is required to establish that the condition

$$\liminf_{T \to \infty} \frac{C^T}{H_W^T(\xi)} > 1 \tag{8}$$

is sufficient, under a certain sufficiently general character of the assumptions, for the possibility of transmission, satisfying the conditions formulated above, for sufficiently large T. Naturally, in such a formulation the problem is somewhat indistinct (for example, similar to the general problem of studying possible limit distributions for a sum of large numbers of "small" components). However, I intended to avoid any return to the terminology of the theory of stationary, random processes here, since it was shown in the note of the young Romanian mathematician Rozenblat-Rot Milu<sup>5</sup> that interesting results can be obtained in the designated direction without the assumption of stationariness.

Many remarkable works have been devoted to the derivation of a limit theorem of the kind indicated. The work of Khinchin<sup>6</sup> is the contribution of a USSR mathematician in this research direction. It appears to me that much remains to be done here. Namely, results of this kind are intended to give a foundation to the widespread conviction that the expression  $I(\xi, \eta)$  is not just one of the possible methods of measuring the "quantity of information" but it is a measure of the quantity of information having an advantage, in principle, over the others, actually. Since the "information," by its original nature, is not a scalar, then axiomatic investigations, permitting  $I(\xi, \eta)$ [or the entropy  $H(\xi)$ ] to be characterized uniquely by using simple formal properties, in this respect have lesser value, in my opinion. The situation here seems to me to be similar to our being ready to assign, at once, the greatest value to that method, out of all those proposed by Gauss to give a foundation to the normal law of error distribution, which starts from the limit theorem for the sum of a large number of small components. Other methods (for example, based on the principle of the arithmetic mean) demonstrate only why any other error distribution law could not be as acceptable and suitable as the normal law but they do not answer the question of why the normal law is actually encountered often in real problems. Similarly, the beautiful formal properties of the expressions  $H(\xi)$  and  $I(\xi, \eta)$  cannot demonstrate why they are sufficient for the complete (albeit from the asymptotic point of view) solution of many problems in many cases.

## IV. Calculation and Estimation of the $\epsilon$ Entropy in Certain Particular Cases

If the condition

$$P_{\mathfrak{s}\mathfrak{s}'}\mathfrak{e}W$$

is chosen as the certainty of exact coincidence of  $\xi$  and  $\xi' = V$ 

$$P(\xi = \xi') = 1$$

then

$$H_w(\xi) = H(\xi).$$

In conformance with this, it seems to be natural to designate  $H_w(\xi)$  in the general case as the "entropy of the random object  $\xi$  for the accuracy of reproduction W."

Now, let us assume that X is a metric space and that the space X' coincides with X, *i.e.*, methods are investigated of the approximate transmission of information from the point  $\xi \in X$  by using the indication of the point  $\xi'$  of the same space X. It seems natural to require that

$$P\{\rho(\xi,\xi') \leq \epsilon\} = 1 \qquad (W^{0}_{\epsilon})$$

or that

$$M\rho^2(\xi,\xi') \leq \epsilon^2 \qquad (W_{\epsilon}).$$

We will denote these two forms of the " $\epsilon$  entropy" of the distribution  $P_{\epsilon}$  by

$$H_{W,\epsilon^{\circ}}(\xi) = H^{0}_{\epsilon}(\xi)$$
$$H_{W,\epsilon}(\xi) = H_{\epsilon}(\xi).$$

As regards the  $\epsilon$  entropy  $H^0_{\epsilon}$ , I shall only note here a certain estimate for

$$H^0_{\epsilon}(X) = \sup_{P_{\xi}} H^0_{\epsilon}(\xi)$$

where the upper bound is taken over all the probability distributions  $P_{\xi}$  in the space X. As is known, for  $\epsilon = 0$ ,

$$H_0^0(X) = \sup_{P_{\xi}} H(\xi) = \log N_x$$

where  $N_x$  is the number of elements of the manifold X. For  $\epsilon > 0$ ,

$$\log N_x^c(2\epsilon) \le H_{\epsilon}^0(x) \le \log N_x^a(\epsilon)$$

where  $N_x^{\epsilon}(\epsilon)$  and  $N_x^{\epsilon}(\epsilon)$  are characteristics of the space Xwhich are introduced in my note.<sup>2</sup> The asymptotic properties of the function  $N_x(\epsilon)$  as  $\epsilon \to 0$ , studied in my work<sup>2</sup> for a number of specific spaces X, are interesting analogs of the properties, explained later, of the asymptotic behavior of the function  $H_{\epsilon}(\xi)$ .

Let us now turn to the  $\epsilon$  entropy  $H_{\epsilon}(\xi)$ . If X is an *n*-dimensional Euclidean space and if

$$P_{\xi}(A) = \int_{A} P_{\xi}(x) \ dx_1 \ dx_2 \ \cdots \ dx_n$$

then, at least in the case of the sufficiently smooth function  $p_{\xi}(x)$ , the following well-known formula holds:

$$H_{\epsilon}(\xi) = n \log \frac{1}{\epsilon} + [h(\xi) - n \log \sqrt{2\pi e}] + o(1) \qquad (9)$$

where

$$h(\xi) = -\int_X p_{\xi}(x) \log p_{\xi}(x) \, dx_1 \, \cdots \, dx_n$$

is the "differential entropy," already introduced in the first Shannon works. Hence, the asymptotic behavior of  $H_{\epsilon}(\xi)$  in the case of sufficiently smooth continuous distributions in *n*-dimensional space is determined, to a first approximation, by the dimensionality of the space and the differential entropy  $h(\xi)$  only enters as the second term in the expression for  $H_{\epsilon}(\xi)$ .

It is natural to expect that the growth of  $H_{\epsilon}(\xi)$  as  $\epsilon \to 0$ will be substantially more rapid for typical distributions

<sup>&</sup>lt;sup>5</sup> Rozenblat-Rot Milu, Trudy, Third All-Union Math. Conf., vol. 2, pp. 132–133; 1956. <sup>6</sup> A. Ia. Khinchin, Usp. Mate. Nauk, vol. 11, no. 1 (67), pp.

<sup>&</sup>lt;sup>o</sup> A. Ia. Khinchin, Usp. Mate. Nauk, vol. 11, no. 1 (67), pp. 17–75; 1956.

(10)

in infinite-dimensional spaces. As the simplest example, let us consider the Wiener random function  $\xi(t)$ , defined for  $0 \le t \le 1$ , with the normally-distributed independent increments:

$$\Delta\xi = \xi(t + \Delta t) - \xi(t)$$

for which

$$\xi(0) = 0;$$
  $M\Delta\xi = 0;$   $M(\Delta\xi)^2 = \Delta t.$ 

Iaglom found that in this case, in the  $L^2$  metric

$$H_{\epsilon}(\xi) = rac{4}{\pi}rac{1}{\epsilon^2} + o\!\!\left(\!rac{1}{\epsilon^2}\!
ight)$$

Under certain natural assumptions, the formula

$$H_{\epsilon}(\xi) = \frac{4}{\pi} \chi \left( \frac{1}{\epsilon^2} \right) + o\left( \frac{1}{\epsilon^2} \right)$$
(1)

where

$$\chi(\xi) = \int_{t_0}^{t_1} Mb \mid t, \, \xi(t) \mid dt$$

can be obtained in a more general way for the diffuse kind of Markov process on the  $t_0 \leq t \leq t_1$  time segment with

$$M_{\Delta\xi} = A[t, \xi(t)] \Delta t + o(\Delta t);$$
  
$$M(\Delta\xi)^2 = B[t, \xi(t)] \Delta t + o(\Delta t).$$

The  $\epsilon$ -entropy  $H_{\epsilon}$  can be calculated exactly for the case of the normal distribution in an n-dimensional space or in Hilbert space. After a suitable orthogonal coordinate transformation, the *n*-dimensional vector  $\boldsymbol{\xi}$  assumes the form

$$\xi = (\xi_1, \xi_2, \cdots, \xi_n)$$

where the coordinates  $\xi_k$  are mutually independent and distributed normally. The parameter  $\theta$ , for given  $\epsilon$ , is determined from the equation

$$\epsilon^2 = \sum \min(\theta^2, D^2 \xi_k)$$

and in the case of  $\xi$  distributed normally

$$H_{\epsilon}(\xi) = rac{1}{2} \sum_{D^{2} \xi_{k} o heta^{2}} \log rac{D^{2} \xi_{k}}{ heta^{2}} \cdot$$

The approximating vector

$$\xi' = (\xi'_1, \xi'_2, \cdots, \xi'_n)$$

should be chosen such that

$$\xi'_k = 0$$

for  $D^2 \xi_k \leq \theta^2$  and

$$\xi_k = \xi'_k + \Delta_k;$$
  $D^2 \Delta_k = \theta^2;$   $D^2 \xi'_k = D^2 \xi_k - \theta^2$ 

for  $D^2\xi_k > \theta$  and the vectors  $\xi_k$  and  $\Delta_k$  are mutually independent. The infinite-dimensional case is in no way different from the finite dimensional.

Finally, it is very essential that the maximum value of  $H_{\epsilon}(\xi)$  for the vector  $\xi$  (*n* dimensional or infinite dimensional)

be attained in the normal distribution case for given second central moments. This result can be obtained directly or from the following proposition of Pinsker.<sup>7</sup>

Theorem: Let the positive-definite symmetric matrix of the  $s_{ij}$ ,  $0 \le i, j \le m + n$  quantities and the distribution  $P_{\xi}$  of the vector be given

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_m)$$

for which the central second moments equal  $s_{i,i}$  (for  $0 \leq i, j \leq m$ ). Let the condition W on the joint distribution  $P_{\xi\xi'}$  of the vector  $\xi$  and the vector

$$\xi' = (\xi_{m+1}, \xi_{m+2}, \cdots, \xi_{m+n})$$

be that the central second moments of the quantities

1) 
$$\xi_1, \xi_2, \cdots, \xi_{m+n}$$

equal  $s_{ii}$  (for  $0 \le i, j \le m + n$ ). Then

$$H_{w}(\xi) \leq \frac{1}{2} \log \frac{AB}{C}.$$
 (13)

The notation in (13) corresponds to the explanation of Section II. It is seen from a comparison with the results of Section II, that inequality (13) becomes the equality in the case of the normal distribution  $P_{\xi}$ .

The principles of solving the variational problems arising in the calculation of the "rate of creating information" were indicated sufficiently long ago by Shannon. Shannon and Weaver<sup>8</sup> write: "Unfortunately these formal solutions are difficult to evaluate in particular cases and seem to be of little value."<sup>9</sup> In substance, however, many problems of this kind are simple enough, as is seen from the above. It is possible that the slow development of investigations in this direction is related to insufficient understanding of the fact that the solution of the variation problem often appears to be degenerate in typical cases: For example, the evaluation of  $H_{\epsilon}(\xi)$  in the problem selected above, for the normally distributed vector  $\xi$  in the *n*-dimensional case, the vector  $\xi'$  often appears to be not n dimensional but only k dimensional with k < n; in the infinite-dimensional case, the vector  $\xi'$  always appears to be finite dimensional.

#### V. QUANTITY OF INFORMATION AND RATE OF CREATING (12)INFORMATION IN THE STATIONARY PROCESS CASE

Let us consider two stationary and stationarily-related processes,

$$\xi(t), \ \eta(t) \qquad -\infty \ < \ t \ < \ +\infty.$$

Let us denote by  $\xi_T$  and  $\eta_T$  the segments of the  $\xi$  and  $\eta$ processes in the time  $0 < t \leq T$  and by  $\xi_{-}$  and  $\eta_{-}$  the flow of the  $\xi$  and  $\eta$  processes on the negative semiaxis  $-\infty <$  $t \leq 0$ . To give the pair  $(\xi, \eta)$  of the stationarily-related  $\xi$ and  $\eta$  processes means to give the probability distribution

- 125; 1956.
  <sup>8</sup> C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press; 1949.
  <sup>9</sup> Ibid., sec. 28, p. 79 of the Russian translation.

<sup>&</sup>lt;sup>7</sup> M. S. Pinsker, Trudy, Third All-Union Math. Conf., vol. 1, p.

 $P_{in}$ , invariant to shift along the t axis, in the space of the under certain assumptions on the regularity of the process function pair  $\{x(t), y(t)\}$ . If  $\xi_{-}$  is fixed, then the following  $\xi$  and for certain natural types of conditions W. conditional probability

$$P_{\xi_T \eta/\xi_-}(C/x_-) = P\{(\xi_T, \eta) \in C/\xi_- = x_-\}$$

arises from the distribution  $P_{\xi_{7}}$ . Using this distribution, the conditional quantity of information

$$I(\xi_T, \eta/x)$$

is calculated in conformance with Section II. If the mathematical expectation

$$MI(\xi_T, \eta/\xi_-)$$

is finite for any T > 0, then it is finite for all other T > 0and

$$MI(\xi_T, \eta/\xi_-) = TI(\xi, \eta)$$

It is natural to call the quantity  $\vec{I}(\xi, \eta)$  the "rate of creating information of the process  $\eta$  for compliance with the process  $\xi$ ." If the process  $\xi$  can be extrapolated with complete accuracy to future occurrences, then

$$\dot{I}(\xi,\,\eta)\,=\,0.$$

In particular, this will be so if the process  $\xi$  has a bounded spectrum. Generally speaking, the following equality

$$\vec{I}(\xi, \eta) = \vec{I}(\eta, \xi) \tag{14}$$

does not hold. However, under sufficiently broad conditions on the "regularity" of the process  $\xi_{i}^{10}$  the equality

$$\overline{I}(\xi, \eta) = \overline{I}(\xi, \eta)$$

holds, where

$$\overline{I}(\xi, \eta) = \lim_{T\to\infty} \frac{1}{T} I(\xi_T, \eta_T).$$

Since  $I(\xi_T, \eta_T) = I(\eta_T, \xi_T)$ , then always

$$I(\xi, \eta) = I(\eta, \xi)$$

and, therefore, when both equalities  $\vec{I}(\xi, \eta) = \vec{I}(\xi, \eta)$  and  $\vec{I}(\eta, \xi) = \vec{I}(\eta, \xi)$  are correct, equality (14) holds. Now, let W be a certain class of joint distributions  $P_{\xi\xi'}$  of two stationary and stationarily-related processes  $\xi$  and  $\xi'$ . It is natural to call the equantity

$$\vec{H}_{W}(\xi) = \inf_{P_{\xi\xi',\xiW}} \vec{I}(\xi',\xi)$$

the "rate of creating information in the process  $\xi$  under the accuracy of reproduction W." It can be shown that

$$\vec{H}_{W}(\xi) = \vec{H}_{W}(\xi)$$

where

$$\overline{H}_{W} = \inf_{P_{\xi\xi'} \in W} \overline{I}(\xi, \xi')$$

<sup>10</sup> Here and later, the regularity of the process means, roughly speaking, that the segments of the process nearby roughly segments of the taxis sufficiently removed from each other, are almost independent. In the case of Gaussian processes, the well-known definition of regularity introduced in my work<sup>14</sup> is applicable here.

VI. CALCULATION AND ESTIMATION OF THE AMOUNT OF INFORMATION AND THE RATE OF CREATING

INFORMATION IN TERMS OF THE SPECTRUM

Pinsker<sup>11</sup> established the formula:

$$\overline{I}(\xi, \eta) = -\frac{1}{4\pi} \int_{-\infty}^{\infty} \log \left[1 - r^2(\lambda)\right] d\lambda \qquad (15)$$

$$r^{2}(\lambda) = \frac{|f_{\xi\eta}(\lambda)|^{2}}{f_{\xi\xi}(\lambda)f_{\eta\eta}(\lambda)}$$

and  $f_{\xi\xi}$ ,  $f_{\xi\eta}$ ,  $f_{\eta\eta}$  are spectral densities; for the case when the distribution  $P_{\xi\eta}$  is normal and at least one of the processes  $\xi$  or  $\eta$  is regular. In connection with the review by Doob,<sup>12</sup> we would like to note that the novelty, in principle, of the Pinsker result is somewhat greater than can be expected on the basis of this review. The expression

$$\bar{h}(\xi) = \log \left(2\pi \sqrt{e}\right) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \log f_{\xi\xi}(\lambda) \, d\lambda \qquad (16)$$

is known in the case of processes with discrete time t for the differential entropy of a normal process per unit time:

$$\bar{h}(\xi) = \lim_{T\to\infty} \frac{1}{T} h(\xi_1, \xi_2, \cdots, \xi_T).$$

However, no analog of the expression  $\bar{h}(\xi)$  exists in the continuous time and unbounded spectrum case and the Pinsker formula requires independent derivation.

It is natural to characterize the accuracy of reproducing the stationary process  $\xi$ , using the stationary process  $\xi$ stationarily related to  $\xi$ , by the quantity

$$\sigma^2 = M[\xi(\alpha) - \xi'(\alpha)]^2$$

and in the case of a W condition of the form

$$\sigma^2 \leq \epsilon^2$$

it is natural to call the quantity

$$\overline{H}_{\epsilon}(\xi) = \overline{H}_{W}(\xi)$$

the  $\epsilon$  entropy per unit time of the process  $\xi$  and under the assumption that

$$\vec{H}_{W}(\xi) = \vec{H}_{W}(\xi)$$

the rate of creating information in the process  $\xi$  for average accuracy of transmission  $\epsilon$ . It can be concluded from the appropriate statement for finite-dimensional distributions (see Section IV) that the quantity  $H_{\epsilon}(\xi)$  attains a maximum in the case of the normal process  $\xi$  for a given spectral density  $f_{\xi\xi}(\lambda)$ . In the normal case,  $\overline{H}_{\xi}(\xi)$  can be calculated easily in terms of the spectral density exactly as was explained in Section IV applied to  $H_{\epsilon}(\xi)$  for the n-di-

<sup>&</sup>lt;sup>11</sup> M. S. Pinsker, Doklady, AN USSR, vol. 98, 213-216; 1954.

<sup>&</sup>lt;sup>12</sup> J. L. Doob, Math. Revs., vol. 16, p. 495; 1955.

mensional distribution. The parameter  $\theta$  is determined Here, however, the novelty, in principle, is that now we see why and within what limits (for not too small  $\epsilon$ ) this

$$\epsilon^{2} = \int_{-\infty}^{\infty} \min \left[\theta^{2}, f_{\xi\xi}(\lambda)\right] d\lambda.$$
 (17)

Using this parameter, the quantity  $H_{\epsilon}(\xi)$  is found from the formula

$$H_{\epsilon}(\xi) = \frac{1}{2} \int_{f_{\xi\xi}(\lambda)\theta^2} \log \frac{f(\lambda)}{\theta^2} d\lambda.$$
 (18)

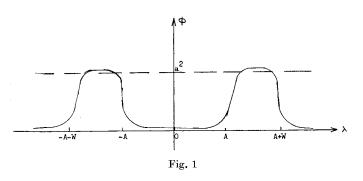
Spectral densities of the kind shown in Fig. 1 which are approximated well by the function:

$$\varphi(\lambda) = \begin{cases} a^2 & \text{for } A \leq |\lambda| \leq A + W \\ 0 & \text{in the remaining cases} \end{cases}$$

are of practical interest. It is easy to calculate that

$$\frac{\theta^2}{H_{\epsilon}(\xi)} \sim \frac{\epsilon^2}{2W}$$
(19)  
$$\overline{H}_{\epsilon}(\xi) \sim W \log \frac{2Wa^2}{\epsilon^2}$$

approximately, in this case for *not too small*  $\epsilon$  for a normal process.



Certainly, (19) is none other than the well-known Shannon formula

$$R = W \log \frac{Q}{N}.$$
 (20)

Here, however, the novelty, in principle, is that now we see why and within what limits (for not too small  $\epsilon$ ) this formula can be used for a process with an unbounded spectrum and such are all the processes in the theory of information transmission which really interest us.

Writing (19) thus

. .

$$\overline{H}_{\epsilon}(\xi) \sim 2W \log \left(a \sqrt{2W}\right) \log \frac{1}{\epsilon}$$
(21)

and comparing with (9), we see that the double width 2W of the useful frequency band plays the role of the number of measurements. This idea of the equivalence of twice the frequency bandwidth to the number of measurements, occurring in a certain sense of the word, per unit time, was apparently first expressed by Kotel'nikov.<sup>13</sup> On the basis of this idea, Kotel'nikov indicated the fact that a function, for which the spectrum is limited to bandwidth 2W, is determined uniquely by the values of the function at the points

$$\cdot \, , \, -\frac{2}{2W} \, , \, -\frac{1}{2W} \, , \, 0, \, \frac{1}{2W} \, , \, \frac{2}{2W} \, , \, \cdots \, , \, \frac{k}{2W} \, , \, \cdots \, .$$

Shannon retained this argumentation, using the representation obtained in this manner to derive (20). Since a function with a bounded spectrum is always singular in the sense of my work<sup>14</sup> and the observation of such a function is not related, generally, to the stationary flow of new information, then the sense of this kind of argumentation does not remain completely clear so that the new derivation of the *approximate* formula (21), cited here, seems to me to be not devoid of interest.

The growth of  $H_{\epsilon}(\xi)$  as  $\epsilon$  decreases occurs, for small  $\epsilon$  for any normally distributed regular random function, substantially more rapidly than would be obtained according to (21). In particular, if  $f_{\xi\xi}(\lambda)$  has order  $1/\lambda\beta$  as  $\lambda \to \infty$ , then  $\bar{h}_{\epsilon}(\xi)$  has order  $1/(2/\epsilon(\beta - 1))$ .

<sup>13</sup> V. A. Kotel'nikov, Material for the First All-Union Conf. on questions of communications; 1933.

<sup>14</sup> A. N. Kolmogorov, Bulletin, Moscow Univ. I, no. 6; 1941. English translation available.

