

A STATISTICAL PROBLEM ARISING IN THE THEORY OF  
DETECTION OF SIGNALS IN THE PRESENCE OF NOISE IN A  
MULTI-CHANNEL SYSTEM AND LEADING TO STABLE  
DISTRIBUTION LAWS

R. L. DOBRUSHIN

(Translated by R. A. Silverman)

1. The following problem is important (see e.g. [1]) in radiolocation. There are  $n$  channels, at the output of each of which there is present a random voltage  $\xi_i$ . We shall assume that the random variables  $\xi_i$  are independent and that each of them has a Rayleigh distribution (i.e. the distribution of the absolute value of a two-dimensional symmetric normal random variable), so that the probability density of the quantity  $\xi_i$  is

$$(1) \quad p_{\xi_i}(x) = \frac{2x}{\lambda_i} e^{-x^2/\lambda_i}, \quad x \geq 0,$$

where

$$(2) \quad \lambda_i = M\xi_i^2.$$

There are two competing hypotheses concerning the parameters  $\lambda_i$ :

**Hypothesis A.** (Absence of signal). All the  $\lambda_i$  are the same and are equal to the number  $\bar{d} > 0$ . This hypothesis means that pure noise is the source of the voltages appearing at the channel outputs. The number  $\bar{d}$  is the average noise power in the channel.

**Hypothesis B.** (Presence of signals). All the  $\lambda_i$  except one  $\lambda_j$  are equal to  $\bar{d}$ , and  $\lambda_j = \bar{d} + \bar{s}$ , where  $\bar{s} > 0$ . The index  $j$  is also random and takes each of the  $n$  possible values with probability  $1/n$ . This hypothesis means that noise is the source of the voltages in all the channels except the  $j$ -th, and that the voltage in the  $j$ -th channel is a useful signal to which noise is also added. The number  $\bar{s}$  is the average signal power<sup>1</sup>. In setting the problem up in this way<sup>2</sup>, we neglect, of course, the possibility of the useful signal appearing simultaneously in several circuits.

The scheme under consideration is of practical interest, since the assumption that the voltages  $\xi_i$  have Rayleigh distributions is a consequence of assuming that the signal and noise are both gaussian. This latter assumption means that we are considering the problem of observing an object which gives a re-

<sup>1</sup> Our considerations would change little if we would assume that the *a priori* probabilities of an occurrence of a signal in distinct channels do not coincide with each other for different channels.

<sup>2</sup> Let us mention the recent paper [5] where one of the possible criteria of distinguishing hypotheses A and B is studied.

flected signal which is the sum of a large number of reflections from randomly distributed reflecting centers.

The joint distributions of the quantities  $\xi_1, \dots, \xi_n$  have the form

$$\begin{aligned} p_A(x_1, \dots, x_n) &= \prod_{i=1}^n \frac{2\alpha_i}{d} e^{-\alpha_i^2/d}, \\ p_B(x_1, \dots, x_n) &= \frac{1}{n} \sum_{j=1}^n \frac{2\alpha_j}{d + \bar{d}} e^{-\alpha_j^2/(d + \bar{d})} \prod_{i \neq j} \frac{2\alpha_i}{d} e^{-\alpha_i^2/d} \end{aligned} \quad (3)$$

under hypotheses A and B, respectively. Examining the likelihood ratio  $p_B/p_A$  we see that the optimal criterion for distinguishing the hypotheses A and B is based on the statistic

$$\varphi(\xi_1, \dots, \xi_n) = \sum_{j=1}^n \exp \left\{ \frac{\bar{d}}{d(d + \bar{d})} \xi_j^2 \right\}. \quad (4)$$

We introduce the ratio

$$\beta = \frac{\bar{d}}{d},$$

where the number  $\beta$  is a dimensionless quantity which characterizes the ratio of the average signal power to the average noise power. A simple calculation shows that if  $\xi_i$  has a Rayleigh distribution with parameter  $\lambda_i$ , then the quantity

$$\eta_i = \exp \left\{ \frac{\bar{d}}{d(d + \bar{d})} \xi_i^2 \right\} \quad (5)$$

is such that

$$\begin{aligned} \mathbf{P}\{\eta_i > x\} &= x^{-\gamma}, & 1 \leq x < \infty, \\ \gamma &= \frac{\bar{d}}{d(d + \bar{d})}. \end{aligned}$$

Hence, if hypothesis A is true, the statistic  $\varphi(\xi_1, \dots, \xi_n)$  has the distribution of the sum

$$S = \eta_1 + \dots + \eta_n$$

of  $n$  independent summands, for each of which

$$\mathbf{P}\{\eta_i > x\} = x^{-(1+\beta)\beta^{-1}}, \quad 1 \leq x < \infty \quad (6)$$

whereas, if hypothesis B is true, the statistic  $\varphi(\xi_1, \dots, \xi_n)$  has the distribution of the sum

$$\bar{S} = \zeta + \eta_2 + \dots + \eta_n,$$

where  $\zeta, \eta_2, \dots, \eta_n$  are independent, and

$$\begin{aligned} \mathbf{P}\{\eta_i > x\} &= x^{-(1+\beta)\beta^{-1}}, & 1 \leq x < \infty, \\ \mathbf{P}\{\zeta > x\} &= x^{-\beta}, & 1 \leq x < \infty. \end{aligned} \quad (7)$$

As was to be expected from physical considerations, the distributions of the statistic  $\varphi(\xi_1, \dots, \xi_n)$  depend only on the parameter  $\beta$ .

As is well known, any criterion for distinguishing hypotheses is based on specifying two complementary sets  $X_A$  and  $X_B$  in  $n$ -dimensional space such that

hypothesis A is considered to be true if  $(\xi_1, \dots, \xi_n) \in X_A$ , whereas hypothesis B is considered to be true if  $(\xi_1, \dots, \xi_n) \in X_B$ . Consider the probability

$$\int \dots \int_{X_B} p_A(x_1, \dots, x_n) dx_1 \dots dx_n \quad (8)$$

that hypothesis B will be accepted when actually hypothesis A is true. We shall call this probability the *probability of false alarm*. Moreover, we introduce the probability

$$\int \dots \int_{X_B} p_B(x_1, \dots, x_n) dx_1 \dots dx_n \quad (9)$$

of accepting hypothesis B when it is actually true. We shall call this probability the *probability of correctly observing the signal*. Suppose two numbers  $F$  and  $D$  are given, where  $0 < F \leq D < 1$ . We shall say that it is *possible to distinguish the hypotheses with probabilities  $F$  and  $D$*  if there exists a criterion for distinguishing these hypotheses for which the probability of false alarm is no greater than  $F$ , and for which the probability of correctly observing the signal is not less than  $D$ . It follows from the general theorems of mathematical statistics that here we can confine ourselves to a consideration of criteria based on the statistic  $\varphi$ . Accordingly, the possibility of distinguishing the hypotheses with the given probabilities depends only on the parameters  $n$  and  $\beta$ . Let us fix  $F$  and  $D$ . Then from continuity considerations it is clear that there exists a boundary value  $\beta_n(F, D)$  such that for  $\beta \geq \beta_n(F, D)$  it is possible to distinguish the hypotheses with probabilities  $F$  and  $D$  and such that for  $\beta_n(F, D) > \beta$  it is no longer possible. If for  $\beta = \beta_n(F, D)$  we use the optimal criterion based on the likelihood ratio for distinguishing the hypotheses A and B, then for some  $y$  the criterion with region

$$X_A = \{\varphi(x_1, \dots, x_n) < y\}$$

will give errors exactly equal to  $F$  and  $D$ .

The quantity  $\beta_n(F, D)$  gives the smallest value of the difference between the signal power and noise power for which it is still possible to distinguish the signal from the background noise with the given error probabilities. The subject of this paper is the study of this quantity. From general information-theoretic considerations we clearly have the qualitative conclusion that  $\beta_n(F, D)$  must depend logarithmically on  $n$ . In this paper we give much more precise quantitative estimates of  $\beta_n(F, D)$ .

2. First of all, we show that for any fixed  $F$  and  $D$

$$\begin{aligned} \beta_n(F, D) &= \frac{\log n}{\log \left( \frac{1-F}{D-F} \right)} - 1 \\ &+ \frac{1}{(1-F) \log \left( \frac{1-F}{D-F} \right)} \int_{-\infty}^K \log(K-y) p(y) dy + o(1) \end{aligned} \quad (10)$$

as  $n \rightarrow \infty$ , where  $p(y)$  is the density of the stable distribution law with para-

meter  $\alpha = 1$  (see [4], Sect. 34), the characteristic function of which is

$$(11) \quad \psi(\lambda) = \exp \left\{ \int_0^\infty \left( e^{i\lambda u} - 1 - \frac{i\lambda u}{1 + u^2} \right) \frac{du}{u^2} \right\},$$

and moreover

$$(12) \quad F = \int_K^\infty \hat{p}(x) dx.$$

Equation (10) is not completely suitable for practical solution of the problem which interests us, since on the one hand the second term in equation (10) yields with difficulty to numerical calculation, and since on the other hand the term  $o(1)$  is of order  $O(1/\log n)$ , as might be expected, so that the approximation is sufficiently accurate only for very large  $n$ . For the same reasons, calculation (as is possible in principle) the remaining terms of the asymptotic expansion in powers of  $\log n$  would not be particularly effective.

3. Fortunately, it is the case of a small false alarm probability which is basically of practical interest (in radiolocation problems it is natural to take  $F$  to be of order  $10^{-5}$ ). Therefore the following asymptotic formula proves useful:

$$(13) \quad \beta_n(F, D) = \frac{\log n + \log 1/F}{\log 1/D} - 1 + o(1),$$

as  $F \rightarrow 0$ , for fixed  $D$  and  $n$ . There are grounds for expecting the even more exact formula

$$(14) \quad \beta_n(F, D) = \frac{\log n + \log \frac{1}{F}}{\log \left( \frac{1-F}{D-F} \right)} - 1 + O(F)$$

to be valid, where  $O(F)$  is uniform with respect to  $n$ . However, we know of no strict proof of this assertion.

4. A simple and feasible criterion for distinguishing the hypotheses A and B is based on the statistic  $\max_i \xi_i$ . Namely, with this criterion A is accepted if  $\max_i \xi_i \leq \gamma$ , and otherwise B is accepted. In analogy to what was done above,  $\hat{\beta}_n(F, D)$  is introduced in such a way that for  $\beta \geq \hat{\beta}_n(F, D)$  the hypotheses can be distinguished with probabilities  $F$  and  $D$  by the use of a criterion of the type just described, while for  $\beta < \hat{\beta}_n(F, D)$  this is no longer possible. By definition

$$\hat{\beta}_n(F, D) \geq \beta_n(F, D).$$

The quantity  $\hat{\beta}_n(F, D)$  can be calculated starting with elementary formulas for the maximum of the sum of independent random variables. The corresponding calculation is well known [9].

It is shown that the same asymptotic formula (14) is valid for  $\hat{\beta}_n(F, D)$  as for  $\beta_n(F, D)$ , where the remainder term  $O(F)$  is uniformly small with respect to  $n$ . Thus, for small probabilities  $F$  the criterion based on the maximum gives practically the same result as the optimal result. It is of interest to find the exact asymptote of the expression

$$\frac{\beta_n(F, D) - \beta_n(F, D)}{\beta_n(F, D)}$$

5. The problem considered here was suggested to the author by Yu. B. Kobzarev, for which the author acknowledges his sincere gratitude.

In Sections 6-10 the result formulated in Section 2 is proved. In Sections 11, 12 the result formulated in Section 3 is proved. These two parts of the paper can be read independently.

6. We have already noted that when the hypothesis A is true the statistic  $\varphi(\xi_1, \dots, \xi_n)$  is distributed like the sum of  $n$  independent random variables with a power-law distribution. In this connection, we need some facts about the asymptotic behavior of such sums.

$$S_n^{(a)} = \gamma_1^{(a)} + \dots + \gamma_n^{(a)},$$

where

$$\mathbf{P}\{\gamma_i^{(a)} > x\} = x^{-a}, \quad x \geq 1, \quad 1 < a < 2.$$

Then  $\mathbf{M}\gamma_i^{(a)} = 1/(a-1)$ . It follows from known theorems on the convergence to stable laws (see [4], p. 195) that

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \frac{S_n^{(a)} - \frac{n}{a-1}}{n^{1/a}} < x \right\} = G_a(x),$$

where the distribution  $G_a(x)$  has the characteristic function

$$(15) \quad \varphi_a(t) = \exp \left\{ \int_0^\infty \left( e^{itx} - 1 - \frac{itx}{1+x^2} \right) \frac{dx}{x^{1+a}} \right\}.$$

We need some additional facts about stable distributions and their convergence. All these facts are special cases of much more general results, kindly communicated to the author by B. M. Zolotarev. Here they are given in a form which is convenient for us.

**Fact 1.** There exists a constant  $C$  such that

$$(16) \quad \left| \mathbf{P} \left\{ \frac{S_n^{(a)} - \frac{n}{a-1}}{n^{1/a}} < x \right\} - G_a(x) \right| \leq \frac{C}{n^{1/2}}$$

for all  $x, n, a > 1$ .

**Fact 2.** For any  $a, 1 \leq a < 2$ , the function  $G_a(x)$  has an everywhere positive density  $g_a(x)$ , such that there exists a constant  $C$  for which

$$(17) \quad g_a(x) \leq \frac{C}{|x|+1}$$

simultaneously for all  $x$  and all  $a$  sufficiently close to 1.

**Fact 3.** There exists a constant  $C$  such that

$$(18) \quad G_a(x) \leq Ce^x$$

simultaneously for all  $x$  and all  $a$  sufficiently close to 1.

as  $n \rightarrow \infty$ . As already noted (Fact 2, Section 6), there exists a non-null density  $g_1(x)$ , and consequently the equation

$$1 - G_1(\bar{K}) = F$$

has a unique solution  $\bar{K}$ . Therefore, it follows from (26) that

$$\lim_{n \rightarrow \infty} \bar{K}_n = \bar{K}. \tag{28}$$

9. We now turn to a study of the distribution of the statistic  $\bar{\varphi}(\xi_1, \dots, \xi_n)$  when the hypothesis B is true. Here the distribution of  $\bar{\varphi}(\xi_1, \dots, \xi_n)$  coincides with the distribution of the sum

$$\frac{\zeta + \eta_1 + \dots + \eta_{n-1} - n\bar{\beta}_n^{-1}}{n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}},$$

where  $\eta_i$  and  $\zeta$  are independent and their distributions are given by equation (7). For brevity we set

$$L_n(x) = \mathbf{P} \left\{ \frac{\eta_1 + \dots + \eta_{n-1} - n\bar{\beta}_n^{-1}}{n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} < x \right\}.$$

From the identity

$$\frac{\eta_1 + \dots + \eta_{n-1} - n\bar{\beta}_n^{-1}}{n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} = \frac{\eta_1 + \dots + \eta_{n-1} - (n-1)\bar{\beta}_n^{-1}}{(n-1)\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} - \frac{1}{\bar{\beta}_n n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}}$$

and the relation applied to  $S_{n-1}^{(\bar{\beta}_n)}$ , it follows that for all  $x$

$$\left| G_{\bar{\beta}_n} \left( x \left( 1 - \frac{1}{\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} \right) + \bar{\beta}_n^{-1} n^{-\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} \right) - L_n(x) \right| \leq \frac{C}{\sqrt{n-1}}.$$

From Fact 2 of Section 6 and from the fact that  $\bar{\beta}_n \rightarrow 1$  as  $n \rightarrow \infty$ , it follows that

$$\left| G_{\bar{\beta}_n}(x + \Delta x) - G_{\bar{\beta}_n}(x) \right| \leq \frac{C|\Delta x|}{|x| - |\Delta x| + 1}$$

for sufficiently large  $n$ . Therefore (29) shows that

$$L_n(x) = G_{\bar{\beta}_n}(x) + O\left(\frac{1}{\sqrt{n}}\right)$$

uniformly in  $x$  as  $n \rightarrow \infty$ . For brevity we set

$$H_n(x) = \mathbf{P} \{ \zeta n^{-\bar{\beta}_n(\bar{\beta}_n+1)^{-1}} < x \}.$$

Then (see (7))

$$H_n(x) = 1 - n^{-(1+\bar{\beta}_n)^{-1}} x^{-\bar{\beta}_n^{-1}}, \quad n^{-\bar{\beta}_n^{-1}} \cong x < \infty.$$

It follows from (20), (23) and (25) that

$$\bar{D}_n(F) = P_B \{ \bar{\varphi}(\xi_1, \dots, \xi_n) \geq \bar{K}_n \}.$$

7. We fix the parameters  $n$  and  $\beta$ , and specify the probability  $F$ . The condition (cf. (8))

$$\int \dots \int_{\varphi(x_1, \dots, x_n) \geq K} p_A(x_1, \dots, x_n) dx_1 \dots dx_n = F$$

gives the level  $F$  as a function  $K_n(\beta, F)$  of  $n, \beta, F$ . We set (cf. (9))

$$D_n(\beta, F) = \int \dots \int_{\varphi(x_1, \dots, x_n) \geq K_n(\beta, F)} p_B(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Consider the sequence

$$\bar{\beta}_n = c \log n + d + o(1),$$

where  $c > 0$  and  $d$  are constants. Our next job is to study the asymptotic behavior of the quantity

$$\bar{D}_n(F) = D_n(\bar{\beta}_n, F)$$

as  $n \rightarrow \infty$ . The result (10) which interests us will then follow without any further work.

8. We set

$$\bar{\varphi}(x_1, \dots, x_n) = \frac{\varphi(x_1, \dots, x_n) - n\bar{\beta}_n^{-1}}{n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}}.$$

Writing for brevity

$$\bar{\alpha}_n = \bar{\beta}_n^{-1}(\bar{\beta}_n + 1)$$

and noting that when the hypothesis A is true the statistic  $\bar{\varphi}(\xi_1, \dots, \xi_n)$  has the same distribution as the normalized sum  $S_n^{(\bar{\alpha}_n)}$ , we deduce from (16) that

$$\left| \mathbf{P}_A \{ \bar{\varphi}(\xi_1, \dots, \xi_n) < x \} - G_{\bar{\alpha}_n}(x) \right| \leq \frac{C}{n^{1/2}}.$$

Moreover, we set

$$\bar{K}_n = \frac{K_n(\bar{\beta}_n, F) - n\bar{\beta}_n^{-1}}{n\bar{\beta}_n(\bar{\beta}_n+1)^{-1}}.$$

Then, corresponding to (19) we have

$$\mathbf{P}_A \{ \bar{\varphi}(\xi_1, \dots, \xi_n) > \bar{K}_n \} = F.$$

Now the following fact which is basic for our purposes follows from the inequality (24) by setting  $x = \bar{K}_n$ :

$$\left| G_{\bar{\alpha}_n}(\bar{K}_n) - (1 - F) \right| \leq \frac{C}{\sqrt{n}}.$$

As  $n \rightarrow \infty$ , the quantity  $\bar{\alpha}_n \rightarrow 1$ . It is clear from the formula (15) for the characteristic function that

$$G_{\bar{\alpha}_n}(x) \rightarrow G_1(x)$$

Therefore

$$\bar{D}_n(F) = 1 - \int_{-\infty}^{\bar{K}_n} L_n(\bar{K}_n - y) dH_n(y).$$

It follows from (31) that

$$(33) \quad \bar{D}_n(F) = 1 - \int_{-\infty}^{\bar{K}_n} G_{\bar{a}_n}(\bar{K}_n - y) dH_n(y) + o\left(\frac{1}{\sqrt{\log n}}\right).$$

In view of (32), we discover that

$$\begin{aligned} \bar{D}_n(F) &= G_{\bar{a}_n}(\bar{K}_n) + \int_{-\infty}^{\bar{K}_n} [1 - \bar{H}_n(\bar{K}_n - y)] dG_{\bar{a}_n}(y) + o\left(\frac{1}{\sqrt{\log n}}\right) \\ &= G_{\bar{a}_n}(\bar{K}_n - n^{-\bar{a}_n}) + n^{-(1-\bar{a}_n)} \int_{-\infty}^{\bar{K}_n - n^{-\bar{a}_n}} (\bar{K}_n - y)^{-\bar{a}_n} dG_{\bar{a}_n}(y) + o\left(\frac{1}{\sqrt{\log n}}\right). \end{aligned}$$

Moreover, since  $\bar{a}_n \rightarrow 1$ , taking into account first (30) and then (26), we see that

$$(35) \quad G_{\bar{a}_n}(\bar{K}_n - n^{-\bar{a}_n}) = F + o\left(\frac{1}{\sqrt{\log n}}\right).$$

Recalling the basic condition (21), we note further that

$$(36) \quad n^{-(1-\bar{a}_n)} = \exp\left\{-\frac{\log n}{c \log n + d + o(1)}\right\} = e^{-1/c} \left[1 + \frac{d+1}{c^2 \log n} + o\left(\frac{1}{\log n}\right)\right].$$

Since  $\bar{\beta}_n \sim c \log n$ ,

$$(\bar{K}_n - y)^{-\bar{\beta}_n} \leq 1$$

for  $\infty > \bar{K}_n - y > \sqrt{\log n}$  and for sufficiently large  $n$ . Thus, in view of (18) and (28), we find that

$$(37) \quad \int_{-\infty}^{\bar{K}_n - \sqrt{\log n}} (\bar{K}_n - y)^{-\bar{\beta}_n} dG_{\bar{a}_n}(y) \leq G_{\bar{a}_n}(\bar{K}_n - \sqrt{\log n}) = o\left(\frac{1}{\log n}\right).$$

On the other hand, it follows from (17) that

$$(38) \quad \begin{aligned} \int_{\bar{K}_n - (\log n)^{-2}}^{\bar{K}_n - n^{-\bar{a}_n}} (\bar{K}_n - y)^{-\bar{\beta}_n} dG_{\bar{a}_n}(y) &\leq C \int_{\bar{K}_n - (\log n)^{-2}}^{\bar{K}_n - n^{-\bar{a}_n}} (\bar{K}_n - y)^{-\bar{\beta}_n} dy \\ &\leq C \int_0^{(\log n)^{-2}} z^{-\bar{\beta}_n} dz \leq C \int_0^{(\log n)^{-2}} z^{-1/d} dz = o\left(\frac{1}{\log n}\right) \end{aligned}$$

for sufficiently large  $n$ . Finally

$$(39) \quad \begin{aligned} (\bar{K}_n - y)^{-\bar{\beta}_n} &= \exp\left\{-\frac{\log(\bar{K}_n - y)}{c \log n + d + o(1)}\right\} \\ &= 1 - \frac{\log(\bar{K}_n - y)}{c \log n + d + o(1)} (1 + o(1)) = 1 - \frac{\log(\bar{K}_n - y)}{c \log n} [1 + o(1)] \end{aligned}$$

for  $\sqrt{\log n} \leq \bar{K}_n - y \leq (\log n)^{-2}$ , where the term  $o(1)$  is uniform in the region considered. Thus, collecting together (37), (38) and (39), we see that

$$(40) \quad \begin{aligned} \int_{-\infty}^{\bar{K}_n - n^{-\bar{a}_n}} (\bar{K}_n - y)^{-\bar{\beta}_n} dG_{\bar{a}_n}(y) &= \int_{\bar{K}_n - \sqrt{\log n}}^{\bar{K}_n - (\log n)^{-2}} (\bar{K}_n - y)^{-\bar{\beta}_n} dG_{\bar{a}_n}(y) + o\left(\frac{1}{\log n}\right) \\ &= [G_{\bar{a}_n}(\bar{K}_n - (\log n)^{-2}) - G_{\bar{a}_n}(\bar{K}_n - \sqrt{\log n})] \\ &\quad - \frac{1}{c \log n} \int_{\bar{K}_n - \sqrt{\log n}}^{\bar{K}_n - (\log n)^{-2}} \log(\bar{K}_n - y) dG_{\bar{a}_n}(y) [1 + o(1)] + o\left(\frac{1}{\log n}\right). \end{aligned}$$

It follows from (17) that

$$(41) \quad G_{\bar{a}_n}(\bar{K}_n) - G_{\bar{a}_n}(\bar{K}_n - (\log n)^{-2}) \leq \frac{C}{(\log n)^2} = o\left(\frac{1}{\log n}\right).$$

Moreover, it follows from (18) that

$$(42) \quad G_{\bar{a}_n}(\bar{K}_n - \sqrt{\log n}) = o\left(\frac{1}{\log n}\right).$$

Finally, taking into account (26), we discover that

$$(43) \quad G_{\bar{a}_n}(\bar{K}_n - (\log n)^{-2}) - G_{\bar{a}_n}(\bar{K}_n - \sqrt{\log n}) = 1 - F + o\left(\frac{1}{\log n}\right).$$

Furthermore, from (27), (28) and the convergence of the integral

$$\int_{-\infty}^{\bar{K}} \log(\bar{K} - y) dG_1(y),$$

which is a consequence of (17) and (18), it follows that

$$(44) \quad \int_{\bar{K}_n - \sqrt{\log n}}^{\bar{K}_n - (\log n)^{-2}} \log(\bar{K}_n - y) dG_{\bar{a}_n}(y) \rightarrow \int_{-\infty}^{\bar{K}} \log(\bar{K} - y) dG_1(y)$$

as  $n \rightarrow \infty$ . In view of (40), (43) and (44), we see that

$$(45) \quad \int_{-\infty}^{\bar{K}_n - n^{-\bar{a}_n}} (\bar{K}_n - y)^{\bar{\beta}_n} dG_{\bar{a}_n}(y) = 1 - F - \frac{1}{c \log n} \int_{-\infty}^{\bar{K}} \log(\bar{K} - y) dG_1(y) + o\left(\frac{1}{\log n}\right)$$

Finally, (34), (35), (36) and (45) together show that

$$(46) \quad \begin{aligned} \bar{D}_n(F) &= F + e^{-1/c} \left\{ 1 - F + \frac{1}{\log n} \left[ \frac{d+1}{c^2} (1-F) - \frac{1}{c} \int_{-\infty}^{\bar{K}} \log(\bar{K} - y) dG_1(y) \right] \right. \\ &\quad \left. + o\left(\frac{1}{\log n}\right) \right\}. \end{aligned}$$

10. Now we recall that by definition

$$(47) \quad D_n(\beta_n(F, D), F) \equiv D,$$

and that  $D_n$  grows monotonically as the first argument grows. We now define the constants  $\bar{c}$  and  $\bar{d}$  in such a way that by putting these numbers in (46) instead of  $c$  and  $d$ , the expression (46) gives  $D + o(1/\log n)$ , i.e. we find  $\bar{c}$  and  $\bar{d}$  from the equations

$$(48) \quad \begin{aligned} F + e^{1/c} (1-F) &= D, \\ e^{-1/\bar{c}} \left[ \frac{\bar{d}+1}{\bar{c}^2} (1-F) - \frac{1}{\bar{c}} \int_{-\infty}^{\bar{K}} \log(\bar{K} - y) dG_1(y) \right] &= 0. \end{aligned}$$

If

$$\overline{\lim}_{n \rightarrow \infty} \frac{\beta_n(F, D)}{\log n} > \bar{\varepsilon},$$

then for some subsequence  $n_i$ ,

$$\beta_{n_i}(F, D) \geq \bar{\varepsilon} \log n_i,$$

where  $\bar{\varepsilon} > c$ . But for  $\bar{\beta}_{n_i} = \bar{\varepsilon} \log n_i$ , as can be seen from (46), we have

$$\lim_{i \rightarrow \infty} D_{n_i}(\bar{\beta}_{n_i}, F) > D,$$

and we arrive at a contradiction with (47). Similarly, it is shown that

$$\lim_{n \rightarrow \infty} \frac{\beta_n(F, D)}{\log n} \leq \bar{\varepsilon}$$

and therefore  $\beta_n(F, D) \sim \bar{\varepsilon} \log n$ . Moreover, if

$$\overline{\lim}_{n \rightarrow \infty} [\beta_n(F, D) - \bar{\varepsilon} \log n] > \bar{d},$$

then for the subsequence we have  $\bar{\beta}_{n_i} \geq c \log n_i + \bar{d}$ , where  $\bar{d} > \bar{d}$ , and, as can again be seen from (47),

$$D_{n_i}(c \log n_i + \bar{d}, F) = D + \frac{K}{\log n_i} (1 + o(1)),$$

where  $K > 0$ . We have again arrived at a contradiction. Carrying out the same argument for the lower limit, we discover that

$$\bar{\beta}_n(F, D) = \bar{\varepsilon} \log n + \bar{d} + o(1),$$

and from (48) it is clear that this assertion is equivalent to the assertion (10), formulated in Section 2.

11. We now undertake the proof of the results formulated in Section 3. First of all, we shall prove a lemma about the sums of independent random variables which have power-law distributions.

**Lemma.** Let  $\xi_1, \dots, \xi_n$  be independent random variables such that

$$(49) \quad \mathbf{P}\{\xi_i \geq x\} = x^{-\rho}, \quad x \geq 1, \rho \geq 1.$$

Then for fixed  $n$  and  $x \rightarrow \infty$

$$\mathbf{P}\{\xi_1 + \dots + \xi_n \geq x\} = nx^{-\rho}[1 + o(1)],$$

where the term  $o(1)$  goes uniformly to zero for  $1 \leq \rho \leq 1 + \varepsilon$ ,  $\varepsilon > 0$ .

**PROOF.** We note that the following equality between events is valid for  $\log x \geq n$

$$(50) \quad \{\xi_1 + \dots + \xi_n \geq x\} = \bigcup_{i=1}^n \{\xi_1 + \dots + \xi_n \geq x, \xi_i \geq \frac{x}{\log x}, \xi_j \leq \frac{x}{\log x}, j \leq i\} \cup A_n,$$

where the event  $A_n$  means that

$$\xi_1 + \dots + \xi_n \geq x,$$

and at least two of the summands  $\xi_i$  exceed  $x/\log x$ . It is obvious that

$$(51) \quad \mathbf{P}\{A_n\} \leq n(n-1) \left(\frac{x}{\log x}\right)^{-2\rho} = x^{-\rho} o(1),$$

where  $o(1)$  approaches zero uniformly for  $1 \leq \rho \leq 1 + \varepsilon$ ,  $x \rightarrow \infty$ . Moreover

$$\begin{aligned} & \left\{ \xi_1 + \dots + \xi_n \geq x, \xi_i \geq \frac{x}{\log x}, \xi_j < \frac{x}{\log x}, j \leq i \right\} \\ &= \left\{ \xi_1 + \dots + \xi_n \geq x, \xi_i \geq x - \frac{2x/n}{\log x}, \xi_j < \frac{x}{\log x}, j \neq i \right\} \\ &= \left\{ \xi_i \geq x, \xi_j < \frac{x}{\log x}, j \neq i \right\} + \left\{ \xi_1 + \dots + \xi_n \geq x, x > \xi_i \geq x - \frac{2x/n}{\log x}, \xi_j < \frac{x}{\log x}, j \neq i \right\}. \end{aligned}$$

The probability of the second term of this sum does not exceed

$$(53) \quad \begin{aligned} \mathbf{P}\left\{x \geq \xi_i > x - \frac{2x/n}{\log n}\right\} &= x^{-\rho} - \left(x - \frac{2x/n}{\log n}\right)^{-\rho} \\ &= x^{-\rho} \left[1 - \left(1 - \frac{2/n}{\log n}\right)^{-\rho}\right] = x^{-\rho} o(1), \end{aligned}$$

where the term  $o(1)$  has the same property as above. Finally

$$(54) \quad \begin{aligned} \mathbf{P}\left\{\xi_i \geq x, \xi_j < \frac{x}{\log x}, j \neq i\right\} &= \mathbf{P}\{\xi_i \geq x\} \prod_{j \neq i} \mathbf{P}\left\{\xi_j < \frac{x}{\log x}\right\} \\ &= x^{-\rho} [1 + o(1)]. \end{aligned}$$

Collecting together the assertions (50)–(54), we deduce the stated lemma.

12. We now denote by  $x(F, \beta)$  the number  $x$  such that

$$\mathbf{P}\{S_\beta = \eta_1 + \dots + \eta_n > x\} = F,$$

where the quantities  $\eta_i$  are independent and have the distribution (49) with  $\rho = 1 + 1/\beta$  (cf. (6)). Below we shall keep fixed the number of summands  $n$  (the number of channels). In view of this, we shall drop the corresponding index. We note that, as  $F \rightarrow 0$ , the quantity  $x(F, \beta) \rightarrow \infty$  uniformly in  $\beta$ , whence it follows from the lemma of Section 11 that as  $F \rightarrow 0$

$$(55) \quad x(F, \beta) = (F/n)^{-\beta/\beta+1} (1 + o(1))$$

uniformly in  $\beta > \varepsilon > 0$ .

We now consider the sum (cf. (7))

$$\bar{S}_\beta = \zeta + \eta_2 + \dots + \eta_n,$$

where  $\eta_i, \zeta$  are independent and have power-law distributions (49) with indices  $\rho = 1 + 1/\beta$  and  $\rho = 1/\beta$ , respectively. We study

$$(56) \quad \mathbf{P}\{\bar{S}_\beta \geq x(F, \beta)\}$$

as  $F \rightarrow 0$ . We note that the following equality between events is valid

$$(57) \quad \{\bar{S}_\beta \geq x(F, \beta)\} = \{\zeta \geq x(F, \beta)\} \cup \{x(F, \beta) > \zeta \geq x(F, \beta) - \log x(F, \beta)\}, \\ \bar{S}_\beta \geq x(F, \beta) \cup \{\zeta < x(F, \beta) - \log x(F, \beta)\}, \bar{S}_\beta \geq x(F, \beta).$$

The probability of the first term in (57) is equal to  $[x(F, \beta)]^{-1/\beta}$ . The probability of the second term does not exceed

$$\mathbf{P}\{x(F, \beta) \geq \zeta > x(F, \beta) - \log x(F, \beta)\} = [x(F, \beta)]^{-1/\beta} \left[ \left( 1 - \frac{\log x(F, \beta)}{x(F, \beta)} \right)^{-1/\beta} - 1 \right].$$

Here the first factor does not exceed unity, and the second approaches zero uniformly in  $\beta$  for  $\beta > \varepsilon > 0$  as  $x(F, \beta) \rightarrow \infty$ . Consequently, the probability of the second term in (57) approaches zero uniformly as  $F \rightarrow 0$  for  $\beta > \varepsilon > 0$ . Finally, the event

$$(58) \quad \{\zeta < x(F, \beta) - \log x(F, \beta), \bar{S}_\beta \geq x(F, \beta)\} \subset \{\eta_2 + \dots + \eta_n > \log x(F, \beta)\},$$

and the probability of the right-hand side of the relation (58) approaches zero uniformly in  $\beta$  for  $\beta > \varepsilon > 0$ . To sum up, we discover, taking into account (55) also, that

$$(59) \quad \mathbf{P}\{\bar{S}_\beta > x(F, \beta)\} = [x(F, \beta)]^{-1/\beta} + o(1) = (\eta_1/F)^{-1/\beta} + o(1),$$

where the term  $o(1)$  approaches zero uniformly in  $\beta > \varepsilon > 0$  as  $F \rightarrow 0$ .

We now note that according to the definition of the desired quantity  $\beta_n(F, D)$  the equality

$$(60) \quad \mathbf{P}\{\bar{S}_{\beta_n(F, D)} \geq x(F, D)\} = D$$

is valid for

$$\bar{x}(F, D) = x(F, \beta_n(F, D)).$$

Since as  $F \rightarrow 0$  the quantity  $\beta_n(F, D)$  grows monotonically, we can use equation (60), which shows that

$$(61) \quad (\eta_1/F)^{\frac{1}{\beta_n(F, D)+1}} = D.$$

By elementary transformations we deduce the desired formula (13) from (61).

Received by the editors  
November 2, 1957.

#### REFERENCES

1. A. H. BENNER and R. F. DRENNICK, *On the problem of optimum detection of pulsed signals in noise*, RCA Review, 16, N. 3, 1955, pp. 461-479; Russian translation: *Problems of Radio-location*, No. 5, 1955, pp. 3-167.
2. *Threshold Signals*, Soviet Radio, 9, 1952, Moscow. (In Russian.)
3. Report of the Institute of Radio Engineering and Electronics, Akad. Nauk SSSR, No. 121-56-1. (In Russian.)
4. B. V. GNEDENKO and A. N. KOLMOGOROV, *Limit Distributions for Sums of Independent Random Variables*, Gostekhizdat, Moscow-Leningrad, 1949. (In Russian.)
5. R. DOORNBOS and H. J. PEUNIS, *Stippage test for a set of gamma-variables*, Proc. Koninkl. Nederl. Akad. Wetensch., A59, No. 3, 1956, pp. 329-337.

#### A STATISTICAL PROBLEM ARISING IN THE THEORY OF DETECTION OF SIGNALS IN THE PRESENCE OF NOISE IN A MULTI-CHANNEL SYSTEM AND LEADING TO STABLE DISTRIBUTION LAWS

R. L. DOBRUSHIN (MOSCOW)

(Summary)

The following statistical problem arising in the theory of radar is solved. There are  $t$  hypotheses concerning joint densities of random variables  $\xi_1, \dots, \xi_n$ .

$$A: P_A(x_1, \dots, x_n) = \prod_{i=1}^n \frac{2x_i}{d} e^{-x_i^2/d},$$

$$B: P_B(x_1, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n \frac{2x_j}{d+d} e^{-x_j^2/(d+d)} \prod_{i \neq j} \frac{2x_i}{d} e^{-x_i^2/d},$$

where  $d > 0, d > 0$ . Let  $\beta = d/d$ . Denote by  $\beta_n(F, D)$  such a number that there are no tests distinguishing between hypotheses A and B with error probabilities which are smaller than  $F$  at  $D$  for  $\beta \leq \beta_n(F, D)$  and that tests exist for  $\beta > \beta_n(F, D)$ . It is proved that for  $n = \text{const}$  and  $F \rightarrow 0$

$$\beta_n(F, D) = \frac{\log n + \log 1/F}{\log 1/D} - 1 + o(1).$$

An asymptotic formula containing stable distributions for  $\beta_n(F, D)$ , in which  $F = \text{const}$  and  $n \rightarrow \infty$ , is derived.

INFORMATION TRANSMISSION IN A CHANNEL  
WITH FEEDBACK

R. L. DOBRUSHIN

*(Translated by R. A. Silverman)*

## 1. Introduction

In information theory one usually considers the following situation, shown in Figure 1 (see e.g. [5] and [3]). The statistical properties of the channel are assumed to be given, and one studies the possibility of converting a given input message into a given output message using optimum methods of coding and detection.

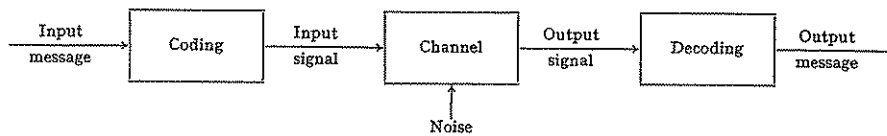


Fig. 1

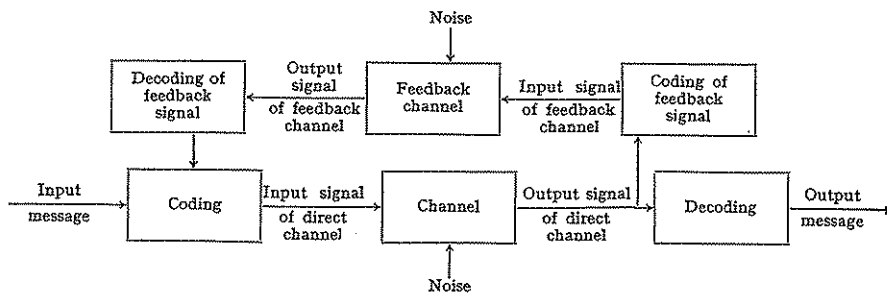


Fig. 2

In many actual communication systems there is also the possibility of transmitting signals in the reverse direction, from the output to the input. In such cases one speaks of systems with feedback (see Figure 2). The feedback is often used to improve the quality of transmission in the forward direction. Thus, for example, if two parties A and B are talking by telephone, a feedback channel is ordinarily used even in the case where the only purpose of their conversation is to convey information emanating at A and destined for B, and B has no information for A and A does not intend to receive any such information. Thus, B interrogates A and asks A to repeat words that he has not heard

distinctly. If the information is easily distorted or especially important, then  $B$  repeats all that he has heard, so that  $A$  can verify that  $B$  has received the message correctly. In recent years, telegraph communication systems handling discrete information have been developed which are based on just this principle. The papers [1], [2], [6] and [8] are devoted to a discussion of concrete systems of this kind. They also pose the problem (which arises naturally) of comparing the channel capacities of systems with feedback with systems without feedback for the same noise level, but this problem is not completely solved in these papers.

In the basic part of the present paper (just as in the earlier papers [1], [2], [6], [8]) we consider a communication channel with discrete time, in which the symbols transmitted at different instants of time are perturbed independently of one another. We shall call such a channel a channel without memory. We show that if the information contained in the messages exceeds the channel capacity in the forward direction, then this message cannot be transmitted over a channel with feedback, regardless of the capacity of the feedback channel. Thus, use of feedback does not increase the capacity of a channel without memory. Of course, it should not be forgotten that this result, like all results of information theory, is of an asymptotic character and pertains to optimum methods of coding and detection. The use of feedback does not permit the transmission of a message which cannot be transmitted without using feedback; however, it can simplify the method of coding signals, which is used for transmission.

We show that the situation is fundamentally different if we consider a channel with memory. In this case, the feedback channel can convey information about how the signals have gone through the channel, and therefore also about the character of past noise; this information, which makes more precise our information about the character of the noise in the present and future, can help in selecting an optimum method for coding the messages. However, in this case, it is essential that the communication channel be sufficiently fast-acting, since the ergodic character of real noise leads to the fact that delayed information about the character of the noise in the remote past does not increase our information about it in the present, and the situation may become the same as for the channel without memory.

The problem of giving a complete quantitative treatment of the questions raised here remains unsolved. It is solved only for a special class of channels with memory, where we assume that the capacity of the feedback channel is so large that the feedback loop instantaneously conveys to the channel input complete and errorless information about the input signal and where we assume that the channel is slow-acting; in particular, for a channel with which some random parameter  $\alpha$  is associated, the channel must reduce to a channel without memory for any fixed value  $\alpha = \alpha_0$  of the parameter. In the case of such a channel, we also find the formula for its asymptotic capacity without using feedback. This formula is evidently new and may also be of independent interest. Calculations of the gain in capacity achieved by using feedback are given for some examples.

Summarizing the above considerations, we can say that using feedback for

"interrogation" does not increase the channel capacity, but can simplify the method of transmission. The channel capacity can be increased by using feedback for "tracking the state of the direct channel".

## 2. Mathematical Formulation of the Problem

For the sake of notational simplicity, we shall assume that every one of the random variables under consideration takes a finite number of values. However, all the constructions which follow can be carried over without change to the case of channels with an arbitrary range of values.

Let  $(E_1, \dots, E_m)$  be the alphabet of symbols at the channel input and let  $(\bar{E}_1, \dots, \bar{E}_n)$  be the alphabet of symbols at the channel output. We fix an integer  $n$ , the length of transmission. We denote by  $\mathcal{E}^{(n)}$  the space of all possible input signals, which consists of  $m^n$  sequences of the form  $(E_{i_1}, E_{i_2}, \dots, E_{i_n})$ , and we denote by  $\bar{\mathcal{E}}^{(n)}$  the space of all possible output signals, which consists of  $\bar{m}^n$  sequences of the form  $(\bar{E}_{j_1}, \bar{E}_{j_2}, \dots, \bar{E}_{j_n})$ . The communication channel is defined by giving the system of conditional probabilities

$$p(j_1, \dots, j_n | i_1, \dots, i_n),$$

$$i_k = 1, \dots, m,$$

$$j_k = 1, \dots, \bar{m},$$

for the probability that the input signal  $(E_{i_1}, \dots, E_{i_n})$  is transformed into the output signal  $(\bar{E}_{j_1}, \dots, \bar{E}_{j_n})$ . It is obvious that

$$(1) \quad \sum_{j_1, \dots, j_n} p(j_1, \dots, j_n | i_1, \dots, i_n) = 1$$

for any  $i_1, \dots, i_n$ .

Following Khinchin [7], we shall call a communication channel a *channel without anticipation* if

$$(2) \quad \sum_{i_{k+1}, \dots, i_n} p(j_1, \dots, j_n | i_1, \dots, i_n) = q(j_1, \dots, j_n | i_1, \dots, i_k),$$

for any fixed integral  $k$ ,  $0 \leq k \leq n$ , and indices  $j_1, \dots, j_n, i_1, \dots, i_n$ , i.e., if (2) does not depend on the choice of the indices  $i_{k+1}, \dots, i_n$ . The meaning of the condition (2) is that the conditional probability of the first  $k$  output symbols taking fixed values  $\bar{E}_{j_1}, \dots, \bar{E}_{j_k}$ , given that the input signal is  $E_{i_1}, \dots, E_{i_n}$ , should depend only on the first  $k$  symbols of this signal. In cases where the parameter has the meaning of physical time, the assumption that the communication channel is a channel without anticipation is natural, since it means that at a given instant of time the output signal does not depend on which values are taken by the symbols of the input signal at future instants of time. In what follows, we shall consider only channels without anticipation.

We shall say that the channel under consideration is a *channel without memory* if we can assign a set of probabilities  $p_{ij}^{(k)}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, \bar{m}$ , such that

$$(3) \quad p(j_1, \dots, j_n | i_1, \dots, i_n) = p_{i_1 j_1}^{(1)} p_{i_2 j_2}^{(2)} \dots p_{i_n j_n}^{(n)}$$

for all  $i_1, \dots, i_n$  and  $j_1, \dots, j_n$ . It is clear that a channel without memory is a

channel without anticipation. We shall call a memoryless channel homogeneous in time if  $p_{ij}^{(k)}$  does not depend on  $k$ .

Now in addition let  $\mathcal{F} = (F_1, \dots, F_e)$  and  $\bar{\mathcal{F}} = (\bar{F}_1, \dots, \bar{F}_e)$  be the spaces of input messages and output messages, respectively. Ordinarily, in the applications, the spaces  $\mathcal{F}$  and  $\bar{\mathcal{F}}$  have just as complicated a structure as the spaces  $\mathcal{E}^{(n)}$ ,  $\bar{\mathcal{E}}^{(n)}$ , but this fact is not important for us. We specify a random quantity  $\eta$  which takes values in the space  $\mathcal{F}$  of input messages; we shall call  $\eta$  the *input message*. We also consider random quantities  $\xi, \bar{\xi}$  and  $\bar{\eta}$  which take values in the spaces  $\mathcal{E}^{(n)}$ ,  $\bar{\mathcal{E}}^{(n)}$  and  $\bar{\mathcal{F}}$ , respectively. We call these quantities the input signal, the output signal and the output message, respectively. We set (see (2))

$$(4) \quad r_k(j_k | i_1, \dots, i_k; j_1, \dots, j_{k-1}) = \frac{q(j_1, \dots, j_k | i_1, \dots, i_k)}{\sum_{j_{k-1}} q(j_1, \dots, j_k | i_1, \dots, i_k)}$$

The quantity  $r_k(j_k | i_1, \dots, i_k; j_1, \dots, j_{k-1})$  gives the conditional probability that the  $k$ -th output symbol takes the value  $E_{j_k}$  given the  $k$  symbols  $E_{i_1}, \dots, E_{i_k}$  of the input signal and the  $k-1$  symbols  $E_{j_1}, \dots, E_{j_{k-1}}$  of the output signal. In the case of a memoryless channel

$$(5) \quad r_k(j_k | i_1, \dots, i_k; j_1, \dots, j_{k-1}) = p_{j_k i_k}^{(k)}$$

We shall say that the *input message*  $\eta$  is transformed into the *output message*  $\bar{\eta}$  as a result of transmission through a channel of length  $n$ , if it is possible to choose an input signal  $\xi = (\xi_1, \dots, \xi_n)$  and an output signal  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  such that the following two conditions are satisfied. First of all, the conditional probability

$$(6) \quad \mathbf{P}\{\xi_k = \bar{E}_{j_k} | \xi_1 = E_{i_1}, \dots, \xi_k = E_{i_k}, \bar{\xi}_1 = \bar{E}_{j_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}, \eta = F_j\} = r(j_k | i_1, \dots, i_k; j_1, \dots, j_{k-1})$$

for any  $j = 1, \dots, e, i_r = 1, \dots, m, j_r = 1, \dots, \bar{m}$ . The intuitive meaning of this condition is that the conditional probability of the  $k$ -th output symbol taking the value  $E_{j_k}$  for the  $k$  input symbols and the  $k-1$  previous output symbols is just what is predetermined by the properties of the given channel, regardless of which particular message is coded into the signal. Secondly, the conditional probability

$$(7) \quad \mathbf{P}\{\bar{\eta} = \bar{F}_\alpha | \bar{\xi} = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n}), \xi = (E_{i_1}, \dots, E_{i_n}), \eta = F_j\} = \mathbf{P}\{\bar{\eta} = \bar{F}_\alpha | \bar{\xi} = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n})\}$$

for any  $\alpha = 1, \dots, \bar{e}, j = 1, \dots, e, i_r = 1, \dots, m, j_r = 1, \dots, \bar{m}$ . The intuitive meaning of this second condition is that the particular output message  $\bar{F}_\alpha$  which is obtained as a result of decoding the output signal  $(\bar{E}_{i_1}, \dots, \bar{E}_{i_n})$  does not depend on what the input message and input signal were. This condition is natural, since in carrying out the decoding we cannot make use of any additional information about the message and signal at the channel input other than the information contained in the output signal.

We shall say that the *input message*  $\eta$  is transformed into the *output message*

$\bar{\eta}$  as a result of transmission through a channel of length  $n$  without using feedback if the input signal  $\xi$  and the output signal  $\bar{\xi}$  can be chosen in such a way that besides conditions (6) and (7) the following supplementary condition is also satisfied:

$$(8) \quad \mathbf{P}\{\xi_k = E_{i_k} | \xi_1 = E_{i_1}, \dots, \xi_{k-1} = E_{i_{k-1}}, \bar{\xi}_1 = \bar{E}_{j_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}, \eta = F_j\} = \mathbf{P}\{\xi_k = E_{i_k} | \xi_1 = E_{i_1}, \dots, \xi_{k-1} = E_{i_{k-1}}, \eta = F_j\}$$

for any  $k = 1, \dots, n, i_r = 1, \dots, m, j_r = 1, \dots, \bar{m}, j = 1, \dots, e$ . If condition (8) is not satisfied then we shall say that the message  $\eta$  is transformed into the message  $\bar{\eta}$  as a result of transmission through a channel of length  $n$  using feedback. The intuitive meaning of the assumption (8) is that (in the case of transmission without feedback) when it is necessary to specify the  $k$ -th symbol of the input signal during the process of coding, we can only use the message  $F_j$  and the  $k-1$  previous input symbols, whereas the symbols  $\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}$  of the output signal are not available to us. On the other hand, the advantages conferred by using feedback consists in the fact (and only in the fact) that the feedback loop furnishes us information at the input about the symbols  $E_{i_1}, \dots, E_{i_{k-1}}$  of the output signal at earlier instants of time, information which can be used in coding the  $k$ -th symbol. It is just this scheme which comprises the examples of using feedback adduced in Section 1.

### 3. Formulation of the Basic Theorem

We now recall that if two random quantities  $\zeta$  and  $\bar{\zeta}$  take the values  $(G_1, \dots, G_k)$  and  $(\bar{G}_1, \dots, \bar{G}_k)$ , respectively, then by the information of the pair  $\zeta$  and  $\bar{\zeta}$  is meant the number

$$(9) \quad I(\zeta, \bar{\zeta}) = \sum \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j\} \log \frac{\mathbf{P}\{\zeta = G_i, \bar{\zeta} = G_j\}}{\mathbf{P}\{\zeta = G_j\} \mathbf{P}\{\bar{\zeta} = G_j\}}$$

We shall say that two random quantities  $\xi = (\xi_1, \dots, \xi_n)$  and  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  are related by a channel of length  $n$  if these quantities take values in the spaces  $\mathcal{E}^{(n)}$  and  $\bar{\mathcal{E}}^{(n)}$ , respectively, and if

$$(10) \quad \mathbf{P}\{\bar{\xi} = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n}) | \xi = (E_{i_1}, \dots, E_{i_n})\} = p(i_1, \dots, i_n | i_1, \dots, i_n)$$

for any  $i_1, \dots, i_n, j_1, \dots, j_n$ . As usual, by the capacity of the channel of length  $n$  without feedback we mean the number

$$(10') \quad C^{(n)} = \sup I(\xi, \bar{\xi}),$$

where the upper bound is taken over all pairs of random quantities  $\xi, \bar{\xi}$  related by the channel of length  $n$ . If the channel in question is a memoryless channel, then

$$(11) \quad C^{(n)} = D_1 + \dots + D_n,$$

where

$$(12) \quad D_k = \sup_{\{e_i\}} \sum_{i,j} p_i p_j^{(k)} \log \frac{p_{ij}^{(k)}}{\bar{p}_j}, \bar{p}_j = \sum_{i=1}^m p_i p_j^{(k)}$$

the formula

$$(15) \quad p(j_1, \dots, j_n | i_1, \dots, i_n) = \sum_{k=1}^n q(a_k) p_{i_1 j_1}(a_k) p_{i_2 j_2}(a_k) \dots p_{i_n j_n}(a_k)$$

for the conditional probabilities specifying the channel as a formal mathematical definition of the random memoryless channel. It is obvious that we get a channel without anticipation.

In Section 7 we shall establish the following asymptotic expression for the capacity of the random memoryless channel without feedback

$$(16) \quad C^{(n)} = n\bar{C} + O(1)$$

as  $n \rightarrow \infty$ , where

$$(16') \quad \bar{C} = \sup_k \sum_{i,j} q(a_k) \sum_{i,j} p_i p_{ij}(a_k) \log \frac{p_{ij}(a_k)}{\bar{p}_j(a_k)},$$

$$\bar{p}_j(a_k) = \sum_i p_i p_{ij}(a_k),$$

and the upper bound is taken over all possible probability distributions  $p_1, \dots, p_m$ . We emphasize that the  $p_i$  unlike the  $p_j(a)$  do not depend on the random parameter  $\alpha$ . In addition, we shall prove the following theorem about the random memoryless channel with feedback.

**Theorem 3.** *If the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a random memoryless channel of length  $n$  using feedback, then the information*

$$(17) \quad I(\eta, \bar{\eta}) \leq \bar{K}n,$$

where

$$(18) \quad \bar{K} = \sum_k q(a_k) D(a_k)$$

and

$$(18') \quad D(a_k) = \sup_{\{p_i\}} \sum_{i,j} p_i p_{ij}(a_k) \log \frac{p_{ij}(a_k)}{\bar{p}_j}, \quad \bar{p}_j = \sum_i p_i p_{ij}(a_k)$$

is equal to the capacity per unit time without feedback of the memoryless channel specified by the transition probabilities  $p_{ij}(a_k)$ . (Cf. (11) and (12).)

For a non-random channel, the quantity  $\alpha$  takes only one value and  $\bar{K}n = C^{(n)}$ . Therefore Theorem 3 is a generalization of Theorem 2. It is clear that  $\bar{C}$  is always less than or equal to  $\bar{K}$ ; some examples given in Section 8 show that the difference  $\bar{K} - \bar{C}$ , which characterizes the gain in capacity by using feedback, is generally speaking positive, but usually small. Intuitively, we can explain the difference between  $\bar{C}$  and  $\bar{K}$  by saying that in the case where feedback is absent the optimum probabilities  $p_i$  for the channel input signals should be chosen in common for all the operating regimes of the channel, whereas feedback furnishes information at the input about the operating regime of the channel, so that the input signal probabilities  $p_i$  can therefore be chosen to be optimum for the regime in question.

and the upper bound is taken over all possible probability distributions  $(p_1, \dots, p_m)$ . Equation (11), which is natural from an intuitive point of view, will be proved in Section 6.

We now formulate the first statement of Shannon's well-known theorem about a channel without feedback.

**Theorem 1.** *If the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a channel of length  $n$  without using feedback, then*

$$(13) \quad I(\eta, \bar{\eta}) \leq C^{(n)},$$

where  $C^{(n)}$  is the capacity of the channel of length  $n$  without feedback.

The familiar proof of this statement will be given in Section 5, since it has not been given heretofore in just the form we require. The second statement of Shannon's theorem says that the condition (13) is also sufficient for the message  $\eta$  to be transformed into the message  $\bar{\eta}$  as a result of transmission through a channel of length  $n$ , in a certain asymptotic sense as  $n \rightarrow \infty$ , if certain additional regularity conditions imposed on the ergodic channel are met. We shall give neither a proof nor even a precise formulation of this assertion, since it is not used in what follows and we need it only for general orientation. For details see [3], [5], [7], and [8].

A basic result of this paper is the following theorem about a memoryless channel with feedback.

**Theorem 2.** *If the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a memoryless channel of length  $n$  using feedback, then*

$$(14) \quad I(\eta, \bar{\eta}) \leq C^{(n)},$$

where  $C^{(n)}$  is the capacity of the channel of length  $n$  without feedback.

Thus, if information can be transmitted through a memoryless channel using feedback, this information cannot exceed the capacity of the channel without feedback. Since, according to the second statement of Shannon's theorem it is true that in a certain asymptotic sense information less than the channel capacity can be transmitted over the channel without using feedback, in an analogous asymptotic sense, we can say that any information which can be transmitted through a memoryless channel using feedback can be transmitted through the same channel without using feedback.

In a subsequent section (Section 8) we consider a special class of channels without memory. We shall call channels of this class random channels without memory. We assume that we are given a set of transition probabilities  $p_{ij}(a)$ , which specify for any fixed  $\alpha = a$  a memoryless channel which is homogeneous in time, and that the parameter  $\alpha$  is a random variable. Intuitively, we can imagine that the random memoryless channel is an ordinary memoryless channel, but with a statistical operating regime which is unknown *a priori* and with a given *a priori* probability distribution for this regime. For simplicity, we assume that the random parameter  $\alpha$  takes only a finite number of values  $a_1, \dots, a_s$  with probabilities  $q(a_1), \dots, q(a_s)$ , respectively. Then we can use

Theorem 3 says nothing about the possibility of transmitting a message with information  $I(\eta, \bar{\eta}) < \bar{K}\eta$ , and therefore the question naturally arises as to whether the estimate (17) is optimal. The random communication channel is ergodic only in the case where it reduces to an ordinary channel. Therefore we can get no results for it of the type of the converse of Shannon's theorem. However, by slightly varying the channel considered above, we can make it ergodic. To do so we must assume that at every instant of time there occurs with probability  $\eta$  a random change of operating regime which is independent of the previous operation of the channel. If  $\eta$  is close to 0, then our new channel will "almost coincide" with the random memoryless channel considered above, and the characteristics which we find for the random memoryless channel will be limits of the corresponding characteristics for the new channel. It can be shown that if

$$\frac{I(\eta, \bar{\eta})}{\bar{K}\eta} < 1,$$

then the message  $\eta$  can be transformed into the message  $\bar{\eta}$  as a result of transmission using feedback, if  $\eta$  is small enough and the channel length  $n$  is large enough. We shall give here neither a complete proof nor a precise statement of this fact, but we shall indicate the intuitive idea of the proof. If  $\eta$  is small, then for a long interval of time (of order  $1/\eta$ ) the channel has a constant operating regime. Transmitting various signals for tracking purposes and using the feedback channel, in a small part of this interval we can find in just which regime the channel is operating. Then, we can use the coding method which is optimum for the regime  $\alpha = \alpha_k$  which was found, and we can transmit the message at a rate of  $D(\alpha_k)$  bits per second. Since for approximately one  $q(\alpha_k)$ th of the time the channel operates in the regime  $\alpha_k$ , the average message transmission rate will be close to

$$\sum_k q(\alpha_k) D(\alpha_k) = \bar{K},$$

as was to be shown.

#### 4. Some Properties of Information

We recall some familiar properties of the information  $I(\zeta, \bar{\zeta})$  defined by equation (9). Consider a third random quantity  $\beta$ , taking the values  $B_1, \dots, B_u$ . By the conditional information of  $\zeta$  and  $\bar{\zeta}$ , under the condition  $\beta = B_u$ , we mean the number

$$I(\zeta, \bar{\zeta}|\beta = B_u) = \sum_{i,j} \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j|\beta = B_u\} \log \frac{\mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j|\beta = B_u\}}{\mathbf{P}\{\zeta = G_i|\beta = B_u\} \mathbf{P}\{\bar{\zeta} = \bar{G}_j|\beta = B_u\}}.$$

By the average value of the conditional information of the pair  $\zeta, \bar{\zeta}$  for a given  $\beta$  we mean the number

$$(19) \quad \mathbf{MI}(\zeta, \bar{\zeta}|\beta) = \sum_{u=1}^u I(\zeta, \bar{\zeta}|\beta = B_u) \mathbf{P}\{\beta = B_u\}.$$

The pair of random quantities  $\zeta, \beta$  can be regarded as a single random quantity  $(\zeta, \beta)$ , the values of which are the  $kv$  pairs  $(G_i, B_u)$ . With this notation, the following important formula for conditional information is valid

$$(20) \quad I((\zeta, \beta), \bar{\zeta}) = I(\beta, \bar{\zeta}) + \mathbf{MI}(\zeta, \bar{\zeta}|\beta).$$

Evidently this formula was first stated explicitly in the work of Kolmogorov [3]. In the case which interests us (i.e., discrete quantities) equation (20) is a consequence of the following simple calculation

$$\begin{aligned} I((\zeta, \beta), \bar{\zeta}) &= \sum_{i,j,u} \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j, \beta = B_u\} \times \log \frac{\mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j, \beta = B_u\}}{\mathbf{P}\{\zeta = G_i, \beta = B_u\} \mathbf{P}\{\bar{\zeta} = \bar{G}_j\}} \\ &= \sum_{i,j,u} \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j, \beta = B_u\} \\ &\quad \times \log \frac{\mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j|\beta = B_u\} \mathbf{P}\{\beta = B_u\} \mathbf{P}\{\bar{\zeta} = \bar{G}_j, \beta = B_u\}}{\mathbf{P}\{\zeta = G_i|\beta = B_u\} \mathbf{P}\{\bar{\zeta} = \bar{G}_j, \beta = B_u\}} \\ &= \sum_{i,j,u} \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j, \beta = B_u\} \log \frac{\mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j|\beta = B_u\}}{\mathbf{P}\{\zeta = G_i|\beta = B_u\} \mathbf{P}\{\bar{\zeta} = \bar{G}_j|\beta = B_u\}} \\ &\quad + \sum_{i,j,u} \mathbf{P}\{\zeta = G_i, \bar{\zeta} = \bar{G}_j, \beta = B_u\} \log \frac{\mathbf{P}\{\bar{\zeta} = \bar{G}_j, \beta = B_u\}}{\mathbf{P}\{\bar{\zeta} = \bar{G}_j\} \mathbf{P}\{\beta = B_u\}} \\ &= I(\beta, \bar{\zeta}) + \mathbf{MI}(\zeta, \bar{\zeta}|\beta). \end{aligned}$$

We now note some important consequences of the conditional information formula (20). First of all, we note that since the conditional information  $I(\zeta, \bar{\zeta}|\beta) \geq 0$ , then

$$(21) \quad I((\zeta, \beta), \bar{\zeta}) \geq I(\beta, \bar{\zeta})$$

for any three random quantities  $\zeta, \beta, \bar{\zeta}$ . We now observe that

$$(22) \quad I(\gamma, \delta) \leq I(\gamma, \gamma)$$

for any two random quantities  $\gamma$  and  $\delta$ . To derive the inequality (22) it is sufficient to note that using the conditional information formula and the inequality (21) we have

$$I(\gamma, \delta) \leq I(\gamma, \gamma, \delta) = I((\gamma, \delta), \gamma) = I(\gamma, \gamma) + \mathbf{MI}(\gamma, \delta|\gamma),$$

and that  $\mathbf{MI}(\delta, \gamma|\gamma) = 0$ , since the information of any random quantity concerning a constant is zero. We also find useful the widely known fact (see [5], [7]) that if a quantity  $\gamma$  takes a finite number  $s$  of values, then its entropy

$$(23) \quad I(\gamma, \gamma) \leq \log s.$$

We now recall that we say that the three random quantities  $\zeta, \beta, \bar{\zeta}$  form a Markov chain if

$$(24) \quad \mathbf{P}\{\bar{\zeta} = \bar{G}_j|\beta = B_u, \zeta = G_i\} = \mathbf{P}\{\bar{\zeta} = \bar{G}_j|\beta = B_u\}$$

for any  $i, j, u$ . The condition (24) is equivalent to the following condition:

$$(25) \quad \mathbf{P}\{\bar{\zeta} = \bar{G}_i, \zeta = G_i | \beta = B_i\} = \mathbf{P}\{\bar{\zeta} = \bar{G}_j | \beta = B_j\} \mathbf{P}\{\zeta = G_j | \beta = B_j\}$$

for all  $i, j, u$ . The familiar fact that the conditions (25) and (24) are equivalent is not hard to verify by direct calculation. It clearly follows from (25) that if the quantities  $\bar{\zeta}, \beta, \zeta$  form a Markov chain, then the conditional information  $I(\bar{\zeta}, \bar{\zeta} | \beta = B_u) = 0$ , and consequently, the conditional information formula (20) shows that

$$(26) \quad I(\bar{\zeta}, \zeta, \beta) = I(\bar{\zeta}, \beta).$$

In what follows this important fact will be used repeatedly.

Finally, we need the following formula for triple information, i.e.,

$$(27) \quad I((\beta, \gamma), \delta) + I(\beta, \gamma) = I(\beta, (\gamma, \delta)) + I(\gamma, \delta)$$

for any three random quantities  $\beta, \gamma, \delta$ . To derive this formula it is sufficient to observe that by direct calculation both the left-hand and right-hand sides of this formula equal

$$\sum_{i,j,k} \mathbf{P}\{\beta = B_i, \gamma = C_j, \delta = D_k\} \log \frac{\mathbf{P}\{\beta = B_i, \gamma = C_j, \delta = D_k\}}{\mathbf{P}\{\beta = B_i\} \mathbf{P}\{\gamma = C_j\} \mathbf{P}\{\delta = D_k\}}.$$

### 5. Proof of Shannon's Theorem for a Channel Without Feedback

We suppose that the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a channel of length  $n$  and we consider the corresponding input signal  $\xi = (\xi_1, \dots, \xi_n)$  and output signal  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$ . Applying the formula for total probability to the system of events  $\xi = (E_{i_1}, \dots, E_{i_n})$ , we deduce from equation (7) that

$$\mathbf{P}\{\bar{\eta} = \bar{F}_a | \bar{\xi} = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n}), \eta = F_j\} = \mathbf{P}\{\bar{\eta} = \bar{F}_a | \xi = (E_{i_1}, \dots, E_{i_n})\}.$$

Comparing this equality and the definition (24) of a Markov chain, we see that the quantities  $\eta, \bar{\xi}, \bar{\eta}$  form a Markov chain. Therefore, it follows from equations (21) and (26) that the information

$$(28) \quad I(\eta, \bar{\eta}) \leq I((\bar{\eta}, \bar{\xi}), \eta) = I(\bar{\xi}, \eta).$$

We emphasize that we derived the inequality (28) starting with the condition (7), which is true both when feedback is used and when feedback is not used, and therefore (28) is also true for channels with feedback.

We now consider transmission without using feedback. Then equation (8) is true. We wish to show, starting from the general condition (6) and equation (8), that

$$(29) \quad \mathbf{P}\{(\bar{\xi}_1, \dots, \bar{\xi}_k) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_k}) | (\xi_1, \dots, \xi_k) = (E_{i_1}, \dots, E_{i_k}), \eta = F_j\} \\ = \prod_{\alpha=1}^k \mathbf{P}\{j_\alpha | i_\alpha, \dots, i_\alpha; j_1, \dots, j_{\alpha-1}\}$$

for all  $j_1, \dots, j_k, i_1, \dots, i_k$  and any  $1 \leq k \leq n$ . We shall prove equation (29) by induction on  $k$ . For  $k = 1$  it reduces at once to equation (6) for  $k = 1$ .

Suppose it is valid for  $k-1$  factors. Now we note that it follows from equation (8) that the events  $\{\xi_k = E_{i_k}\}$  and  $\{(\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}})\}$  are independent under the condition that  $\{(\xi_1, \dots, \xi_{k-1}) = (E_{i_1}, \dots, E_{i_{k-1}})\}$  and  $\{\eta = F_j\}$ . Hence it follows that

$$(30) \quad \mathbf{P}\{(\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}) | (\xi_1, \dots, \xi_k) = (E_{i_1}, \dots, E_{i_k}), \eta = F_j\} \\ = \mathbf{P}\{(\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}) | (\xi_1, \dots, \xi_{k-1}) = (E_{i_1}, \dots, E_{i_{k-1}}), \eta = F_j\}.$$

Now applying the conditional probability formula, equation (6), equation (30) and the induction hypothesis, we find that

$$(31) \quad \mathbf{P}\{(\xi_1, \dots, \xi_k) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_k}) | (\xi_1, \dots, \xi_k) = (E_{i_1}, \dots, E_{i_k}), \eta = F_j\} \\ = \mathbf{P}\{\xi_k = \bar{E}_{i_k} | (\xi_1, \dots, \xi_k) = (E_{i_1}, \dots, E_{i_k}), (\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}), \eta = F_j\} \\ \cdot \mathbf{P}\{(\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}) | (\xi_1, \dots, \xi_k) = (E_{i_1}, \dots, E_{i_k}), \eta = F_j\} \\ = \mathbf{P}\{j_k | i_1, \dots, i_k, j_1, \dots, j_{k-1}\} \\ \cdot \mathbf{P}\{(\xi_1, \dots, \xi_{k-1}) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_{k-1}}) | (\xi_1, \dots, \xi_{k-1}) = (E_{i_1}, \dots, E_{i_{k-1}}), \eta = F_j\} \\ = \prod_{\alpha=1}^k \mathbf{P}\{j_\alpha | i_1, \dots, i_\alpha, j_1, \dots, j_{\alpha-1}\},$$

as was to be proved.

Applying equation (29) for  $k = n$ , we see that the sequence of quantities  $\eta, \xi, \bar{\xi}$  forms a Markov chain. Arguing just as in the derivation of equation (28), we discover that

$$(32) \quad I(\eta, \bar{\xi}) \leq I(\xi, \bar{\xi})$$

in the case of transmission without using feedback. Again applying equation (29) for  $k = n$ , we find that

$$(33) \quad \mathbf{P}\{(\xi_1, \dots, \xi_n) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n}) | (\xi_1, \dots, \xi_n) = (E_{i_1}, \dots, E_{i_n})\} \\ = \prod_{k=1}^n \mathbf{P}\{j_k | i_1, \dots, i_k, j_1, \dots, j_{k-1}\}.$$

It follows from equations (2) and (4) that

$$(34) \quad \mathbf{P}\{j_k | i_1, \dots, i_k, j_1, \dots, j_{k-1}\} = \frac{\sum_{i_{k+1}, \dots, i_n} \mathbf{P}\{j_1, \dots, j_n | i_1, \dots, i_n\}}{\sum_{i_k, \dots, i_n} \mathbf{P}\{j_1, \dots, j_n | i_1, \dots, i_n\}}.$$

Setting (34) in (33), we find that

$$(35) \quad \mathbf{P}\{(\xi_1, \dots, \xi_n) = (\bar{E}_{i_1}, \dots, \bar{E}_{i_n}) | (\xi_1, \dots, \xi_n) = (E_{i_1}, \dots, E_{i_n})\} \\ = \mathbf{P}\{j_1, \dots, j_n | i_1, \dots, i_n\}.$$

Equation (35) shows that in the case of transmission without using feedback the quantities  $\xi$  and  $\bar{\xi}$  are related by a channel of length  $n$ , and therefore it follows from the definition of capacity (10') that

$$(36) \quad I(\xi, \bar{\xi}) \leq C^{(n)}.$$

From the inequalities (36), (32) and (28) we infer that

$$I(\eta, \bar{\eta}) \leq C^{(n)},$$

whereby we have proved Shannon's theorem, formulated in Section 3.

### 6. Derivation of the Formulas for the Capacity of the Memoryless Channel and the Random Memoryless Channel in the Absence of Feedback

We begin by proving the following important fact. If  $\xi = (\xi_1, \dots, \xi_n)$  and  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  are two random quantities related by a memoryless channel, then

$$I(\xi, \bar{\xi}) \leq I(\xi_1, \bar{\xi}_1) + I(\xi_2, \bar{\xi}_2) + \dots + I(\xi_n, \bar{\xi}_n). \tag{37}$$

Consider first the case  $n = 2$ . In this case the triple information formula (27) shows that

$$(38) \quad I((\bar{\xi}_1, \bar{\xi}_2), (\xi_1, \xi_2)) = I(\bar{\xi}_1, (\xi_1, \xi_2, \bar{\xi}_2)) + I((\xi_1, \xi_2), \bar{\xi}_2) - I(\bar{\xi}_1, \bar{\xi}_2).$$

Equation (3) for a memoryless channel shows that

$$\begin{aligned} \mathbf{P}\{\bar{\xi}_1 = \bar{E}_1 | \xi_1 = E_1, \xi_2 = E_2\} &= \mathbf{P}\{\bar{E}_1 | E_1, \bar{\xi}_2 = \bar{E}_2\} = p_{\bar{E}_1|E_1}^{(1)}, \\ \mathbf{P}\{\bar{\xi}_2 = \bar{E}_2 | \xi_1 = E_1, \xi_2 = E_2\} &= p_{\bar{E}_2|E_2}^{(2)}. \end{aligned}$$

Consequently, the quantities  $\bar{\xi}_1, \xi_1, (\xi_2, \bar{\xi}_2)$  and the quantities  $\bar{\xi}_2, \xi_2, \xi_1$  form Markov chains. Then according to equation (26)

$$(39) \quad \begin{aligned} I(\bar{\xi}_1, (\xi_1, \xi_2, \bar{\xi}_2)) &= I(\xi_1, \bar{\xi}_1), \\ I((\xi_1, \xi_2), \bar{\xi}_2) &= I(\xi_2, \bar{\xi}_2), \end{aligned}$$

and (37) for  $n = 2$  follows from (38) and (39). To generalize this formula to arbitrary  $n$  it is sufficient to observe that the four quantities  $\xi'_1 = (\xi_1, \dots, \xi_{n-1})$ ,  $\bar{\xi}'_1 = (\bar{\xi}_1, \dots, \bar{\xi}_{n-1})$ ,  $\xi'_2 = \xi_2$  and  $\bar{\xi}'_2 = \bar{\xi}_2$  form a channel without memory, so that according to what was shown above we have

$$(40) \quad I((\xi_1, \dots, \xi_n), (\bar{\xi}_1, \dots, \bar{\xi}_n)) \leq I((\xi_1, \dots, \xi_{n-1}), (\bar{\xi}_1, \dots, \bar{\xi}_{n-1})) + I(\xi_n, \bar{\xi}_n).$$

Applying induction, we deduce from (40) the desired inequality (37).

We now use (37) to derive equation (11) for the capacity of a channel without memory. Since if  $(\xi_1, \dots, \xi_n)$  and  $(\bar{\xi}_1, \dots, \bar{\xi}_n)$  are related by the channel, we have

$$\begin{aligned} \mathbf{P}\{\bar{\xi}_k = E_{j_k} | \xi_k = E_k\} &= p_{j_k}^{(k)}, \\ \text{then } I(\bar{\xi}_k, \xi_k) &\leq D_k \text{ (see (12)), whence it follows according to (37) that} \end{aligned}$$

$$(41) \quad I(\xi, \bar{\xi}) \leq D_1 + \dots + D_n = C^{(n)}.$$

To show that the equality sign is achieved in (41), it is sufficient to note that if we denote by  $p_i^{(k)}$  the set of probabilities  $p_i$  for which the upper bound  $D_k$  in (12) is achieved and if we take for  $\xi, \bar{\xi}$  quantities with the joint distribution

$$\mathbf{P}\{\xi = (E_{i_1}, \dots, E_{i_n}), \bar{\xi} = (E_{j_1}, \dots, E_{j_n})\} = \prod_{k=1}^n p_{i_k}^{(k)} p_{j_k}^{(k)},$$

then we obtain quantities related by the channel for which

$$I(\xi, \bar{\xi}) = D_1 + \dots + D_n = C^{(n)}.$$

Now let  $\xi = (\xi_1, \dots, \xi_n)$  and  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  be random quantities related by a random memoryless channel of length  $n$ , and let  $\alpha$  be the random variable used in the definition of the random memoryless channel. Applying the conditional information formula (20), we note that

$$I((\bar{\xi}, \alpha), \xi) = I(\bar{\xi}, \bar{\xi}) + \mathbf{M}I(\alpha, \xi | \bar{\xi}) = I(\bar{\xi}, \alpha) + \mathbf{M}I(\xi, \bar{\xi} | \alpha),$$

whence it follows that

$$(42) \quad I(\xi, \bar{\xi}) = \mathbf{M}I(\bar{\xi}, \xi | \bar{\xi}) + I(\bar{\xi}, \alpha) - \mathbf{M}I(\alpha, \xi | \bar{\xi}).$$

Moreover, we note that because of the general inequality (22) and the inequality (23), we have

$$(43) \quad I(\xi, \bar{\xi}) = \mathbf{M}I(\xi, \bar{\xi} | \alpha) + O(1),$$

where the term  $O(1)$  is uniformly bounded for all  $n$  and all pairs  $\xi$  and  $\bar{\xi}$ . We shall show below that

$$(44) \quad \sup_{(\xi, \bar{\xi})} \mathbf{M}I(\xi, \bar{\xi} | \alpha) = n\bar{C}$$

(see (16')), where the upper bound is taken over all pairs of random quantities  $\xi$  and  $\bar{\xi}$  related by a random memoryless channel of length  $n$ . Equation (16) will follow in an obvious way from (43) and (44). Now we use the fact that for any fixed  $\alpha = a_k$  the quantities  $\xi$  and  $\bar{\xi}$  are related by a memoryless channel with transition probabilities  $p_{ij}^{(k)} = p_{ij}(a_k)$ . Applying the inequality (37), we discover that

$$(45) \quad I(\xi, \bar{\xi} | \alpha = a_k) \leq I(\xi_1, \bar{\xi}_1 | \alpha = a_k) + I(\xi_2, \bar{\xi}_2 | \alpha = a_k) + \dots + I(\xi_n, \bar{\xi}_n | \alpha = a_k).$$

From this it follows that

$$(46) \quad \mathbf{M}I(\xi, \bar{\xi} | \alpha) \leq \sum_{\alpha=1}^n I(\xi_\alpha, \bar{\xi}_\alpha | \alpha = a_k) q(a_k),$$

where the equality sign is achieved in (45) and (46) if the quantities  $\xi_1, \dots, \xi_n$  are independent for any condition of the form  $\alpha = a_k$ . To derive the desired equality (43) it remains to observe that

$$\sup_{\xi, \bar{\xi}} \sum_{\alpha=1}^n q(a_k) I(\xi_\alpha, \bar{\xi}_\alpha | \alpha = a_k) = \bar{C}$$

by definition of the quantity  $\bar{C}$  (see (16')) and that if we take for  $\xi$  the set  $(\xi_1, \dots, \xi_n)$  of independent quantities  $\xi_\alpha$  which have the distribution  $\{p_j\}$  for which the upper bound in (16') is achieved and if we set

$$\mathbf{P}\{\xi_1 = E_{i_1}, \dots, \xi_n = E_{i_n} | \xi_1 = E_{i_1}, \dots, \xi_n = E_{i_n}\} = p_{j_1}^{(1)} \dots p_{j_n}^{(n)},$$

(see (15)) then we obtain a pair of quantities  $(\xi, \bar{\xi})$  which are related by a random memoryless channel of length  $n$  and which are such that

$$\mathbf{M}I(\xi, \bar{\xi} | \alpha) = n\bar{C}.$$

### 7. Information Transmission Using Feedback

First of all we note that the method of proving Shannon's theorem which was used in Section 5 for a channel without feedback is fundamentally inapplicable to the case of a channel with feedback. The point is that in the case of transmission using feedback the input signal  $\xi$  and the output signal  $\bar{\xi}$  may not be related by the channel even for a memoryless channel. As an example demonstrating this statement, consider the memoryless channel specified by the transition probabilities

$$(47) \quad p_{ij}^{(k)} = \frac{1}{2}$$

for all  $i, j$  and  $k$ . This means that the output symbol does not depend on the corresponding input symbol. It is clear that the capacity of such a channel is  $C^{(n)} = 0$ . At the same time, it is easy to verify that if we choose the quantities  $\xi = (\xi_1, \dots, \xi_n)$  and  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  such that  $\xi$  and  $\bar{\xi}$  do not depend on the input message  $\eta$  and the output message  $\bar{\eta}$ , such that

$$\mathbf{P}\{\xi_k = E_j\} = \mathbf{P}\{\xi_k = E_a\} = \mathbf{P}\{\xi_k = E_b\} = \frac{1}{2},$$

for all  $k$ , and such that  $\xi_k = \bar{\xi}_{k-1}$  and does not depend on the remaining  $\xi_i, i \neq k$ , and  $\bar{\xi}_j, j \neq k-1$ , then the four quantities  $\eta, \xi, \bar{\xi}$  and  $\bar{\eta}$  will satisfy the conditions (6) and (7). Intuitively, our construction means that at each instant of time we transmit the signal received at the output at the preceding instant of time, independently of the value of the input message. It is clear that in this case

$$I(\xi, \bar{\xi}) = \sum_{k=2}^n I(\xi_k, \bar{\xi}_{k-1}) = n-1 > C^{(n)} = 0.$$

We now turn to the proof of the basic theorem about the memoryless channel with feedback. We assume that the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a memoryless channel of length  $n$  using feedback, and that  $\xi$  and  $\bar{\xi}$  are the corresponding input and output signals. As we already noted, the inequality (28) which says that

$$(28') \quad I(\eta, \bar{\eta}) \leq I(\eta, \xi).$$

is also true for the case of transmission using feedback. Because of this we shall evaluate the information

$$I(\eta, \bar{\xi}) = I(\eta, (\bar{\xi}_1, \dots, \bar{\xi}_n)).$$

Applying  $n$  times the conditional information formula (20), we deduce that

$$(48) \quad \begin{aligned} I(\eta, \bar{\xi}) &= I((\bar{\xi}_1, \dots, \bar{\xi}_n), \eta) = I((\bar{\xi}_1, \dots, \bar{\xi}_{n-1}), \eta) + \mathbf{M}I(\bar{\xi}_n, \eta | (\bar{\xi}_1, \dots, \bar{\xi}_{n-1})) \\ &= I((\bar{\xi}_1, \dots, \bar{\xi}_{n-2}), \eta) + \mathbf{M}I(\bar{\xi}_{n-1}, \eta | (\bar{\xi}_1, \dots, \bar{\xi}_{n-2})) + \mathbf{M}I(\bar{\xi}_n, \eta | (\bar{\xi}_1, \dots, \bar{\xi}_{n-1})) \\ &= I(\bar{\xi}_1, \eta) + \mathbf{M}I(\bar{\xi}_2, \eta | \bar{\xi}_1) + \dots + \mathbf{M}I(\bar{\xi}_n, \eta | (\bar{\xi}_1, \dots, \bar{\xi}_{n-1})). \end{aligned}$$

Equation (6) shows that under the condition  $(\bar{\xi}_1, \dots, \bar{\xi}_{k-1} = \bar{E}_{i_{k-1}})$  the three random quantities  $\eta, (\xi_1, \dots, \xi_k), \bar{\xi}_k$  form a Markov chain. Therefore, it

follows from the general formulas (21) and (25) that

$$(49) \quad \begin{aligned} I(\xi_k, \eta | \bar{\xi}_1) &= \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}} \\ &\leq I(\xi_k, (\eta, \xi_1, \dots, \xi_k) | \bar{\xi}_1) = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}} \\ &= I(\xi_k, (\xi_1, \dots, \xi_k) | \bar{\xi}_1) = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}. \end{aligned}$$

Using equations (6) and (5), we discover that in the case of a memoryless channel

$$(50) \quad \mathbf{P}\{\xi_k = \bar{E}_{i_k} | \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \xi_k = \bar{E}_{i_k}, \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}\} = p_{i_k j_{k-1}}^{(k)}.$$

This equality shows that under the condition  $(\bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}})$  the three quantities  $(\xi_1, \dots, \xi_{k-1}), \xi_k, \bar{\xi}_k$  form a Markov chain, and therefore,

$$(51) \quad I(\xi_k | (\xi_1, \dots, \xi_k) | \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) = I(\xi_k, \xi_k | \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}).$$

It follows from equation (50) that

$$(52) \quad \mathbf{P}\{\xi_k = \bar{E}_{i_k} | \bar{\xi}_k = \bar{E}_{i_k}, \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}\} = p_{i_k j_{k-1}}^{(k)}.$$

Therefore, it is a consequence of the definition (12) that the conditional information

$$(53) \quad I(\xi_k, \bar{\xi}_k | \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \leq D_k.$$

It follows from (53), (51) and (49) that

$$(54) \quad I(\xi_k, \eta | \bar{\xi}_1 = \bar{E}_{i_1}, \dots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \leq D_k$$

for all  $k = 1, \dots, n$ . Applying (54) and (48), we deduce that

$$(55) \quad I(\eta, \bar{\xi}) \leq D_1 + \dots + D_n = C^{(n)}.$$

The inequality (28') shows that the assertion of the basic theorem for a memoryless channel with feedback follows from the inequality (55).

We now turn to a consideration of the random memoryless channel. We wish to prove Theorem 3, formulated in Section 3. Suppose the input message  $\eta$  is transformed into the output message  $\bar{\eta}$  as a result of transmission through a random memoryless channel of length  $n$  using feedback. Let  $\xi = (\xi_1, \dots, \xi_n)$  and  $\bar{\xi} = (\bar{\xi}_1, \dots, \bar{\xi}_n)$  be the corresponding input and output signals. Here equation (28) is again applicable and shows that

$$(56) \quad I(\eta, \bar{\eta}) \leq I(\eta, \bar{\xi}).$$

Arguing just as in the derivation of equation (42) (we need only replace  $\xi$  everywhere by  $\eta$ ) we find that

$$(57) \quad I(\eta, \bar{\xi}) = \mathbf{M}I(\eta, \bar{\xi} | \alpha) + I(\eta, \alpha) - \mathbf{M}I(\alpha, \eta | \bar{\xi}).$$

From the intuitive probabilistic interpretation of the quantity  $\alpha$  it follows<sup>1</sup>

<sup>1</sup> Equations (6) and (16) serve as a formal definition of transmission over a random memoryless channel with feedback. Therefore, strictly speaking, we must show that for any three quantities  $\eta, \xi, \bar{\xi}$  satisfying the conditions (6) and (16), we can construct a quantity  $\alpha$  with distribution  $\{g(\alpha, i)\}$  such that  $\eta$  and  $\alpha$  are independent and such that equation (58) is true. Because of its simplicity, we do not give this straightforward argument.

in an obvious way that the input message and the parameter  $\alpha$  are independent quantities and that the condition (6) can now be replaced by the condition

$$(58) \quad \mathbf{P}\{\xi_k = \bar{E}_k | \xi_1 = E_{i_1}, \dots, \xi_k = E_{i_k}, \bar{\xi}_1 = E_{j_1}, \dots, \bar{\xi}_{k-1} = E_{j_{k-1}}, \eta = F_j, \alpha = a_\alpha\} \\ = \bar{p}_{i_k, j_k}(a_\alpha).$$

It follows from the independence of  $\eta$  and  $\alpha$  that  $I(\eta, \alpha) = 0$  and therefore according to (57)

$$(59) \quad I(\eta, \xi) \leq MI(\eta, \xi | \alpha).$$

For any fixed condition  $\alpha = a_\alpha$ , equation (58) shows that the quantities  $\eta, \xi, \bar{\xi}$  accomplish transmission of the message  $\eta$  through the memoryless channel specified by the transition probabilities  $\bar{p}_{ij}(a_\alpha)$  using feedback. Therefore, repeating the argument given at the beginning of this section, we come to the conclusion that

$$(60) \quad I(\eta, \xi | \alpha = a_\alpha) \leq nD(a_\alpha),$$

where  $D(a_\alpha)$  is the capacity per unit time of the memoryless channel with transition probabilities  $\bar{p}_{ij}(a_\alpha)$  (see (18)). It follows from (60) that

$$MI(\eta, \xi | \alpha) \leq n \sum_{\alpha} q(\alpha) D(a_\alpha),$$

and this and the inequality (59) yield the assertion of Theorem 3.

### 8. Some Examples

For some concrete random memoryless channels we compare the channel capacity per unit time  $\bar{C}$  without using feedback (see (16')) and the channel capacity per unit time  $\bar{K}$  using feedback (see (18)).

Suppose that the number of states at the input and output equals 2 and that for any value of the parameter  $a_k$  the matrix  $\bar{p}_{ij}(a_k)$  is symmetric, i.e.,  $\bar{p}_{11}(a_k) = \bar{p}_{22}(a_k)$ ,  $\bar{p}_{12}(a_k) = \bar{p}_{21}(a_k)$ . Then for all  $k$  the upper bound in the expression (18') for  $D(a_k)$  is attained for  $\bar{p}_1 = \bar{p}_2 = \frac{1}{2}$ . Therefore for  $\bar{p}_1 = \bar{p}_2 = \frac{1}{2}$ , the expression

$$\sum_{i,j} q(a_k) \sum_{i,j} \bar{p}_{ij}(a_k) \log \frac{\bar{p}_{ij}(a_k)}{\bar{p}_j(a_k)},$$

standing under the upper bound sign in (16'), is equal to  $\bar{K}$ . Since it is obvious that  $\bar{C}$  is always less than  $\bar{K}$ , we come to the conclusion that for the example in question  $\bar{C} = \bar{K}$ , i.e., using feedback does not give a gain in channel capacity. An analogous argument holds, of course, in any case where the optimum signal distribution does not depend on the parameter  $\alpha$ . If the transition probability matrix is not assumed to be symmetric, then the optimum distribution will depend on  $\alpha$ , and the use of feedback will in this case give a gain in capacity.

We now consider an example of a channel in which use of feedback can substantially increase capacity. Let the input states be  $E_1, \dots, E_n$  and the output states  $\bar{E}_1, \dots, \bar{E}_n$ . Let the parameter  $\alpha$  take the  $n$  values  $a_1, \dots, a_n$  each with probability  $1/n$ . We assume that the transition probabilities  $\bar{p}_{ij}(a_\alpha)$  are given

by the formula

$$\bar{p}_{i1}(a_k) = \begin{cases} 0, & i \neq k, \\ 1, & i = k, \end{cases} \\ \bar{p}_{i2}(a_k) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases}$$

Then the capacity  $D(a_k)$  will be  $1/2$  for any fixed  $k$ . Consequently, the capacity per unit time with feedback will be

$$\bar{K} = \log 2.$$

On the other hand (16') takes its maximum value for  $\bar{p}_1 = \dots = \bar{p}_n = 1/n$ , and therefore

$$\bar{C} = \frac{1}{n} \log n - \left(1 - \frac{1}{n}\right) \log \left(1 - \frac{1}{n}\right).$$

The gain in capacity  $(\bar{K} - \bar{C})/\bar{K}$  is small for small  $n$ . As  $n \rightarrow \infty$ , the quantity

$$\bar{C} \sim \frac{\log n + 1}{n}$$

and the gain approaches 1.

Received by the editors  
March 6, 1958

### REFERENCES

- [1] W. B. BISHOP and B. L. BUCHANAN, *Message redundancy vs. feedback for reducing message uncertainty*, IRE National Convention Record, 1957, Part 2, pp. 33-39.
- [2] E. S. GORBUNOV, *A comparison of some noise-combating codes*, *Electrosvyaz*, No. 12, 1956, pp. 42-47. (In Russian.)
- [3] A. N. KOLMOGOROV, *Theory of information transmission*, Collection: "Session of the Academy of Sciences of the USSR on scientific problems of automatic production", Plenary sessions, Oct. 15-20, 1956; *Izv. Akad. Nauk SSSR*, 1957, pp. 66-99. (In Russian.)
- [4] A. N. KOLMOGOROV, *A new metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces*, *Dokl. Akad. Nauk SSSR*, 119, 5, 1958, pp. 861-864. (In Russian.)
- [5] C. E. SHANNON, *The mathematical theory of communication*, *Bell System Tech. J.*, 27, 1948, pp. 379-423, pp. 623-656.
- [6] B. HARRIS, A. HAUFSCHEIN, L. S. SCHWARTZ, *Optimum decision feedback systems*, IRE National Convention Record, 1957, Part 2, pp. 3-10.
- [7] A. YA. KHINCHIN, *On the basic theorems of information theory*, *Uspekhi Mat. Nauk*, 11, 1, 1956, pp. 17-75.
- [8] I. P. TSAREGRADSKI, *A note on the capacity of a stationary channel with finite memory*, *Theory Prob. Applications*, 3, 1958, pp. 79-91. (English Translation.)
- [9] S. L. CHANG, *Theory of information feedback systems*, I. R. E. Transactions on Information Theory, Vol. IT-2, No. 3, 1956, pp. 29-40.

### TRANSMISSION OF INFORMATION IN CHANNELS WITH FEEDBACK

R. L. DOBRUSHIN (MOSCOW)

(Summary)

In this paper we prove that the use of feedback does not increase the capacity of channels without memory. We also consider some simple channels with memory and compare their capacities when feedback is and is not used.