THEORY OF PROBABILITY
Volume III AND ITS APPLICATIONS Number 4
1958

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.

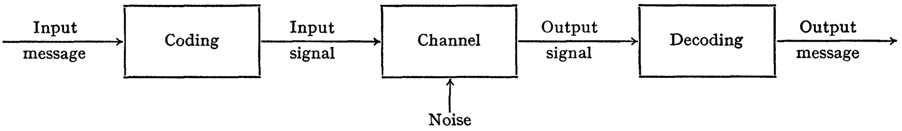

Fig. 1

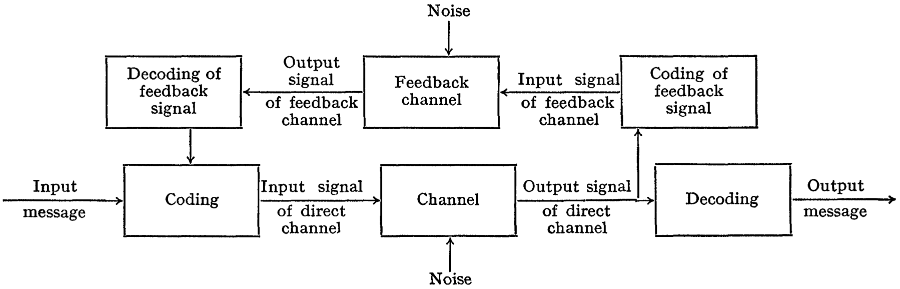
Noise
Decoding of feedback signal
Output signal of feedback channel
Feedback channel
Input signal of feedback channel
Coding of feedback signal
Input message
Coding
Input signal of direct channel
Channel
Output signal of direct channel
Decoding
Output message
Noise

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 α is associated, the channel must reduce to a channel without memory for any fixed value $\alpha = a_e$ 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 $(E_1, \cdots, E_m)$ be the alphabet of symbols at the channel input and let $(\bar{E}_1, \cdots, \bar{E}_{\bar{m}})$ be the alphabet of symbols at the channel output. We fix an integer n, the *length of transmission*. We denote by $\mathscr{E}^{(n)}$ the space of all possible input signals, which consists of m^n sequences of the form $(E_{i_1}, E_{i_2}, \cdots, E_{i_n})$, and we denote by $\bar{\mathscr{E}}^{(n)}$ the space of all possible output signals, which consists of $\bar{m}^n$ sequences of the form $(\bar{E}_{j_1}, \bar{E}_{j_2}, \cdots, \bar{E}_{j_n})$. The communication channel is defined by giving the system of conditional probabilities

$$p(j_1, \cdots, j_n | i_1, \cdots, i_n), \qquad \begin{aligned} i_k &= 1, \cdots, m, \\ j_k &= 1, \cdots, \bar{m}, \end{aligned}$$

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

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

for any $i_1, \cdots, i_n$.

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

$$(2) \qquad \sum_{j_{k+1}, \cdots, j_n} p(j_1, \cdots, j_n | i_1, \cdots, i_n) = q(j_1, \cdots, j_k | i_1, \cdots, i_k),$$

for any fixed integral k, $0 \leqq k \leqq n$, and indices $j_1, \cdots, j_k, i_1, \cdots, i_k$, i.e., if (2) does not depend on the choice of the indices $i_{k+1}, \cdots, i_n$. The meaning of the condition (2) is that the conditional probability of the first k output symbols taking fixed values $\bar{E}_{j_1}, \cdots, \bar{E}_{j_k}$, given that the input signal is $E_{i_1}, \cdots, E_{i_n}$, 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 $p_{ij}^{(k)}$, $i = 1. \cdots, m$, $j = 1, \cdots, \bar{m}$, such that

$$(3) \qquad p(j_1, \cdots, j_n | i_1, \cdots, i_n) = p_{i_1 j_1}^{(1)} p_{i_2 j_2}^{(2)} p_{i_3 j_3}^{(3)} \cdots p_{i_n j_n}^{(n)}$$

for all $i_1, \cdots, i_n$ and $j_1, \cdots, j_n$. 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 $\mathscr{F} = (F_1, \cdots, F_e)$ and $\bar{\mathscr{F}} = (\bar{F}_1, \cdots, \bar{F}_{\bar{e}})$ be the spaces of input messages and output messages, respectively. Ordinarily, in the applications, the spaces $\mathscr{F}$ and $\bar{\mathscr{F}}$ have just as complicated a structure as the spaces $\mathscr{E}^{(n)}$, $\bar{\mathscr{E}}^{(n)}$, but this fact is not important for us. We specify a random quantity η which takes values in the space $\mathscr{F}$ of input messages; we shall call η the *input message*. We also consider random quantities ξ, $\bar{\xi}$ and $\bar{\eta}$ which take values in the spaces $\mathscr{E}^{(n)}$, $\bar{\mathscr{E}}^{(n)}$ and $\bar{\mathscr{F}}$, respectively. We call these quantities the input signal, the output signal and the output message, respectively. We set (see (2))

$$r_k(j_k|i_1, \cdots, i_k; j_1, \cdots, j_{k-1}) = \frac{q(j_1, \cdots, j_k|i_1, \cdots, i_k)}{\sum_{j_k=1}^{\bar{m}} q(j_1, \cdots, j_k|i_1, \cdots, i_k)}. \tag{4}$$

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

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

We shall say that *the input message η 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 $\xi = (\xi_1, \cdots, \xi_n)$ and an output signal $\bar{\xi} = (\bar{\xi}_1, \cdots, \bar{\xi}_n)$ such that the following two conditions are satisfied. First of all, the conditional probability

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

for any $j = 1, \cdots, e$, $i_r = 1, \cdots, m$, $j_r = 1, \cdots, \bar{m}$. The intuitive meaning of this condition is that the conditional probability of the k-th output symbol taking the value $\bar{E}_{j_k}$ 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

$$\begin{aligned} \mathbf{P}\{\bar{\eta} = \bar{F}_\alpha|\bar{\xi} = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_n}), \xi = (E_{i_1}, \cdots, E_{i_n}), \eta = F_j\} \\ = \mathbf{P}\{\bar{\eta} = \bar{F}_\alpha|\bar{\xi} = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_n})\} \end{aligned} \tag{7}$$

for any $\alpha = 1, \cdots, \bar{e}$, $j = 1, \cdots, e$, $i_r = 1, \cdots, m$, $j_r = 1, \cdots, \bar{m}$. 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 $(\bar{E}_{j_1}, \cdots, \bar{E}_{j_n})$ 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 η 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 ξ 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:

$$\tag{8} \begin{gathered} \mathbf{P}\{\xi_k = E_{i_k} | \xi_1 = E_{i_1}, \cdots, \xi_{k-1} = E_{i_{k-1}}, \bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}, \eta = F_j\} \\ = \mathbf{P}\{\xi_k = E_{i_k} | \xi_1 = E_{i_1}, \cdots, \xi_{k-1} = E_{i_{k-1}}, \eta = F_j\} \end{gathered}$$

for any $k = 1, \cdots, n$, $i_r = 1, \cdots, m$, $j_r = 1, \cdots, \bar{m}$, $j = 1, \cdots, e$. If condition (8) is not satisfied then we shall say that *the message η 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 $\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}$ 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 $\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}$ 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 ζ and $\bar{\zeta}$ take the values $(G_1, \cdots, G_k)$ and $(\bar{G}_1, \cdots, \bar{G}_k)$, respectively, then by *the information of the pair* ζ and $\bar{\zeta}$ is meant the number

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

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

$$\tag{10} \mathbf{P}\{\bar{\xi} = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_n}) | \xi = (E_{i_1}, \cdots, E_{i_n})\} = p(j_1, \cdots, j_n | i_1, \cdots, i_n)$$

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

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

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

$$\tag{11} C^{(n)} = D_1 + \cdots + D_n,$$

where

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

and the upper bound is taken over all possible probability distributions $(p_1, \cdots, p_m)$. 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 η is transformed into the output message $\tilde{\eta}$ as a result of transmission through a channel of length n without using feedback, then*

$$I(\eta, \tilde{\eta}) \leqq C^{(n)}, \tag{13}$$

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 η to be transformed into the message $\tilde{\eta}$ as a result of transmission through a channel of length n, in a certain asymptotic sense as $n \to \infty$, 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 η is transformed into the output message $\tilde{\eta}$ as a result of transmission through a memoryless channel of length n using feedback, then*

$$I(\eta, \tilde{\eta}) \leqq C^{(n)}, \tag{14}$$

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 $p_{ij}(\alpha)$, which specify for any fixed $\alpha = a$ a memoryless channel which is homogeneous in time, and that the parameter α 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 α takes only a finite number of values $a_1, \cdots, a_s$ with probabilities $q(a_1), \cdots, q(a_s)$, respectively. Then we can use the formula

$$(15)\qquad p(j_1,\cdots,j_n|i_1,\cdots,i_n)=\sum_{k=1}^{s} q(a_k)p_{i_1j_1}(a_k)p_{i_2j_2}(a_k)\cdots p_{i_nj_n}(a_k)$$

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

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

as $n\to\infty$, where

$$(16')\qquad \bar{C}=\sup\sum_k q(a_k)\sum_{i,j}p_ip_{ij}(a_k)\log\frac{p_{ij}(a_k)}{\bar{p}_j(a_k)},\quad \bar{p}_j(a_k)=\sum_i p_ip_{ij}(a_k),$$

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

Theorem 3. *If the input message η 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*

$$(17)\qquad I(\eta,\bar{\eta})\leqq\bar{K}n,$$

where

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

and

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

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 α 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.

Theorem 3 says nothing about the possibility of transmitting a message with information $I(\eta, \tilde{\eta}) < \bar{K}n$, 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 u a random change of operating regime which is independent of the previous operation of the channel. If u 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

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

then the message η can be transformed into the message $\tilde{\eta}$ as a result of transmission using feedback, if u 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 u is small, then for a long interval of time (of order $1/u$) 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 $\alpha = a_k$ which was found, and we can transmit the message at a rate of $D(a_k)$ bits per second. Since for approximately one $q(a_k)$th of the time the channel operates in the regime a_k, the average message transmission rate will be close to

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

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 β, taking the values $B_1, \cdots, B_r$. By the conditional information of ζ 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 β we mean the number

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

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

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

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

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

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

$$I((\zeta, \beta), \bar{\zeta}) \geqq I(\beta, \bar{\zeta}) \tag{21}$$

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

$$I(\gamma, \delta) \leqq I(\gamma, \gamma) \tag{22}$$

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

$$I(\gamma, \delta) \leqq I(\gamma, (\gamma, \delta)) = I((\gamma, \delta), \gamma) = I(\gamma, \gamma) + \mathbf{M}I(\gamma, \delta|\gamma),$$

and that $\mathbf{M}I(\delta, \gamma|\gamma) = 0$, 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 γ takes a finite number s of values, then its entropy

$$I(\gamma, \gamma) \leqq \log s. \tag{23}$$

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

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

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

$$\mathbf{P}\{\bar\zeta = \bar G_j, \zeta = G_i|\beta = B_u\} = \mathbf{P}\{\bar\zeta = \bar G_j|\beta = B_u\}\mathbf{P}\{\zeta = G_i|\beta = B_u\} \tag{25}$$

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 $\bar\zeta$, β, ζ form a Markov chain, then the conditional information $I(\zeta, \bar\zeta|\beta = B_u) = 0$, and consequently, the conditional information formula (20) shows that

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

In what follows this important fact will be used repeatedly.

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

$$I((\beta, \gamma), \delta) + I(\beta, \gamma) = I(\beta, (\gamma, \delta)) + I(\gamma, \delta) \tag{27}$$

for any three random quantities β, γ, δ. 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 η 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 $\xi = (\xi_1, \cdots, \xi_n)$ and output signal $\bar\xi = (\bar\xi_1, \cdots, \bar\xi_n)$. Applying the formula for total probability to the system of events $\xi = (E_{i_1}, \cdots, E_{i_n})$, we deduce from equation (7) that

$$\mathbf{P}\{\bar\eta = \bar F_\alpha|\bar\xi = (\bar E_{j_1}, \cdots, \bar E_{j_n}), \eta = F_j\} = \mathbf{P}\{\bar\eta = \bar F_\alpha|\bar\xi = (\bar E_{j_1}, \cdots, \bar E_{j_n})\}.$$

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

$$I(\eta, \bar\eta) \leqq I((\bar\eta, \bar\xi), \eta) = I(\bar\xi, \eta). \tag{28}$$

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

$$\begin{aligned}\mathbf{P}\{(\bar\xi_1, \cdots, \bar\xi_k) = (\bar E_{j_1}, \cdots, \bar E_{j_k})|(\xi_1, \cdots, \xi_k) = (E_{i_1}, \cdots, E_{i_k}), \eta = F_j\}\\ = \prod_{\alpha=1}^{k} r(j_\alpha|i_k, \cdots, i_\alpha; j_1, \cdots, j_{\alpha-1})\end{aligned} \tag{29}$$

for all $j_1, \cdots, j_k, i_1, \cdots, i_k$ and any $1 \leqq k \leqq n$. 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 $\{(\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}})\}$ are independent under the condition that $\{(\xi_1, \cdots, \xi_{k-1}) = (E_{i_1}, \cdots, E_{i_{k-1}})\}$ and $\{\eta = F_j\}$. Hence it follows that

$$\begin{aligned} &\mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}) | (\xi_1, \cdots, \xi_k) = (E_{i_1}, \cdots, E_{i_k}), \eta = F_j\} \\ &\quad = \mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}) | (\xi_1, \cdots, \xi_{k-1}) = (E_{i_1}, \cdots, E_{i_{k-1}}), \eta = F_j\}. \end{aligned} \tag{30}$$

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

$$\begin{aligned} &\mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_k) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_k}) | (\xi_1, \cdots, \xi_k) = (E_{i_1}, \cdots, E_{i_k}), \eta = F_j\} \\ &= \mathbf{P}\{\bar{\xi}_k = \bar{E}_{j_k} | (\xi_1, \cdots, \xi_k) = (E_{i_1}, \cdots, E_{i_k}), (\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}), \eta = F_j\} \\ &\quad \cdot \mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}) | (\xi_1, \cdots, \xi_k) = (E_{i_1}, \cdots, E_{i_k}), \eta = F_j\} \\ &= r(j_k | i_1, \cdots, i_k; j_1, \cdots, j_{k-1}) \\ &\quad \cdot \mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_{k-1}) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_{k-1}}) | (\xi_1, \cdots, \xi_{k-1}) = (E_{i_1}, \cdots, E_{i_{k-1}}), \eta = F_j\} \\ &= \prod_{\alpha=1}^{n} r(j_\alpha | i_1, \cdots, i_\alpha; j_1, \cdots, j_{\alpha-1}), \end{aligned} \tag{31}$$

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

$$I(\eta, \bar{\xi}) \leqq I(\xi, \bar{\xi}) \tag{32}$$

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

$$\begin{aligned} &\mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_n) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_n}) | (\xi_1, \cdots, \xi_n) = (E_{i_1}, \cdots, E_{i_n})\} \\ &\quad = \prod_{k=1}^{n} r(j_k | i_1, \cdots, i_k; j_1, \cdots, j_{k-1}). \end{aligned} \tag{33}$$

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

$$r(j_k | i_1, \cdots, i_k; j_1, \cdots, j_{k-1}) = \frac{\sum_{j_{k+1}, \cdots, j_n} p(j_1, \cdots, j_n | i_1, \cdots, i_n)}{\sum_{j_k, \cdots, j_n} p(j_1, \cdots, j_n | i_1, \cdots, i_n)}. \tag{34}$$

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

$$\begin{aligned} &\mathbf{P}\{(\bar{\xi}_1, \cdots, \bar{\xi}_n) = (\bar{E}_{j_1}, \cdots, \bar{E}_{j_n}) | (\xi_1, \cdots, \xi_n) = (E_{i_1}, \cdots, E_{i_n})\} \\ &\quad = p(j_1, \cdots, j_n | i_1, \cdots, i_n). \end{aligned} \tag{35}$$

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

$$I(\xi, \bar{\xi}) \leqq C^{(n)}. \tag{36}$$

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

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

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 $\xi = (\xi_1, \cdots, \xi_n)$ and $\bar{\xi} = (\bar{\xi}_1, \cdots, \bar{\xi}_n)$ are two random quantities related by a memoryless channel, then

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

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

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

Equation (3) for a memoryless channel shows that

$$\mathbf{P}\{\bar{\xi}_1 = \bar{E}_{j_1} | \xi_1 = E_{i_1}, \xi_2 = E_{i_2}, \bar{\xi}_2 = \bar{E}_{j_2}\} = p^{(1)}_{i_1 j_1},$$
$$\mathbf{P}\{\bar{\xi}_2 = \bar{E}_{j_2} | \xi_1 = E_{i_1}, \xi_2 = E_{i_2}\} = p^{(2)}_{i_2 j_2}.$$

Consequently, the quantities $\bar{\xi}_1$, ξ_1, $(\xi_2, \bar{\xi}_2)$ and the quantities $\bar{\xi}_2$, ξ_2, ξ_1 form Markov chains. Then according to equation (26)

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

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 $\xi'_1 = (\xi_1, \cdots, \xi_{n-1})$, $\bar{\xi}'_1 = (\bar{\xi}_1, \cdots, \bar{\xi}_{n-1})$, $\xi'_2 = \xi_2$ and $\bar{\xi}'_2 = \bar{\xi}_2$ form a channel without memory, so that according to what was shown above we have

$$I((\xi_1, \cdots, \xi_n), (\bar{\xi}_1, \cdots, \bar{\xi}_n)) \leqq I((\xi_1, \cdots, \xi_{n-1}), (\bar{\xi}_1, \cdots, \bar{\xi}_{n-1})) + I(\xi_n, \bar{\xi}_n). \tag{40}$$

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 $(\xi_1, \cdots, \xi_n)$ and $(\bar{\xi}_1, \cdots, \bar{\xi}_n)$ are related by the channel, we have

$$\mathbf{P}\{\bar{\xi}_k = E_{j_k} | \xi_k = E_{i_k}\} = p^{(k)}_{ij},$$

then $I(\xi_k, \bar{\xi}_k) \leqq D_k$ (see (12)), whence it follows according to (37) that

$$I(\xi, \bar{\xi}) \leqq D_1 + \cdots + D_n = C^{(n)}. \tag{41}$$

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

$$\mathbf{P}\{\xi = (E_{i_1}, \cdots, E_{i_n}), \bar{\xi} = (E_{j_1}, \cdots, E_{j_n})\} = \prod_{k=1}^{n} \tilde{p}^{(k)}_i p^{(k)}_{i_k j_k},$$

then we obtain quantities related by the channel for which

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

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

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

whence it follows that

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

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

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

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

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

(see (16′)), where the upper bound is taken over all pairs of random quantities ξ 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 ξ and $\bar{\xi}$ are related by a memoryless channel with transition probabilities $p_{ij}^{(l)} = p_{ij}(a_k)$. Applying the inequality (37), we discover that

$$I(\xi, \bar{\xi}|\alpha = a_k) \leqq I(\xi_1, \bar{\xi}_1|\alpha = a_k) + I(\xi_2, \bar{\xi}_2|\alpha = a_k) + \cdots + I(\xi_n, \bar{\xi}_n|\alpha = a_k). \tag{45}$$

From this it follows that

$$\mathbf{M}I(\xi, \bar{\xi}|\alpha) \leqq \sum_{e=1}^{n} \sum_{k} I(\xi_e, \bar{\xi}_e|\alpha = a_k) q(a_k), \tag{46}$$

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

$$\sup_{\xi_e} \sum_{k} q(a_k) I(\xi_e, \bar{\xi}_e|\alpha = a_k) = \bar{C}$$

by definition of the quantity $\bar{C}$ (see (16′)) and that if we take for ξ the set $(\xi_1, \cdots, \xi_n)$ of independent quantities ξ_e which have the distribution $\{p_i\}$ for which the upper bound in (16′) is achieved and if we set

$$\mathbf{P}\{\bar{\xi}_1 = E_{j_1}, \cdots, \bar{\xi}_n = E_{j_n}|\xi_1 = E_{i_1}, \cdots, \xi_n = E_{i_n}\} = p(j_1, \cdots, j_n|i_1, \cdots, i_n),$$

(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 ξ 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

$$p_{ij}^{(k)} = \tfrac{1}{2} \tag{47}$$

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 $\xi = (\xi_1, \cdots, \xi_n)$ and $\bar{\xi} = (\bar{\xi}_1, \cdots, \xi_n)$ such that ξ and $\bar{\xi}$ do not depend on the input message η and the output message $\bar{\eta}$, such that

$$\mathbf{P}\{\xi_k = E_1\} = \mathbf{P}\{\xi_k = E_2\} = \mathbf{P}\{\bar{\xi}_k = E_1\} = \mathbf{P}\{\bar{\xi}_k = E_2\} = \tfrac{1}{2},$$

for all k, and such that $\xi_k = \bar{\xi}_{k-1}$ and does not depend on the remaining ξ_i, $i \neq k$, and $\bar{\xi}_j$, $j \neq k-1$, then the four quantities η, ξ, $\bar{\xi}$ 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 η is transformed into the output message $\bar{\eta}$ as a result of transmission through a memoryless channel of length n using feedback, and that ξ and $\bar{\xi}$ are the corresponding input and output signals. As we already noted, the inequality (28) which says that

$$I(\eta, \bar{\eta}) \leqq I(\eta, \xi). \tag{28'}$$

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

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

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

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

Equation (6) shows that under the condition $(\bar{\xi}_1, = \bar{E}_{j_1}, \cdots, \xi_{k-1} = \bar{E}_{j_{k-1}})$ the three random quantities η, $(\xi_1, \cdots, \xi_k)$, $\bar{\xi}_k$ form a Markov chain. Therefore, it follows from the general formulas (21) and (25) that

$$\begin{aligned} &I(\bar{\xi}_k, \eta|\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \\ &\qquad \leqq I(\bar{\xi}_k, (\eta, \xi_1, \cdots, \xi_k)|\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \\ &\qquad = I(\bar{\xi}_k, (\xi_1, \cdots, \xi_k)|\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}). \end{aligned} \tag{49}$$

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

$$\mathbf{P}\{\bar{\xi}_k = \bar{E}_{j_k}|\xi_1 = E_{i_1}, \cdots, \xi_k = E_{i_k}, \bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}\} = p^{(k)}_{i_k j_k}. \tag{50}$$

This equality shows that under the condition $(\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}})$ the three quantities $(\xi_1, \cdots, \xi_{k-1})$, ξ_k, $\bar{\xi}_k$ form a Markov chain, and therefore,

$$I(\bar{\xi}_k(\xi_1,\cdots,\xi_k)|\bar{\xi}_1 = \bar{E}_{j_1},\cdots,\bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) = I(\bar{\xi}_k, \xi_k|\bar{\xi}_1 = \bar{E}_{j_1},\cdots,\bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}). \tag{51}$$

It follows from equation (50) that

$$\mathbf{P}\{\bar{\xi}_k = \bar{E}_{j_k}|\xi_k = E_{i_k}, \bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}\} = p^{(k)}_{i_k j_k}. \tag{52}$$

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

$$I(\xi_k, \bar{\xi}_k|\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \leqq D_k. \tag{53}$$

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

$$I(\bar{\xi}_k, \eta|\bar{\xi}_1 = \bar{E}_{j_1}, \cdots, \bar{\xi}_{k-1} = \bar{E}_{j_{k-1}}) \leqq D_k \tag{54}$$

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

$$I(\eta, \bar{\xi}) \leqq D_1 + \cdots + D_n = C^{(n)}. \tag{55}$$

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 η is transformed into the output message $\bar{\eta}$ as a result of transmission through a random memoryless channel of length n using feedback. Let $\xi = (\xi_1, \cdots, \xi_n)$ and $\bar{\xi} = (\bar{\xi}_1, \cdots, \bar{\xi}_n)$ be the corresponding input and output signals. Here equation (28) is again applicable and shows that

$$I(\eta, \bar{\eta}) \leqq I(\eta, \bar{\xi}). \tag{56}$$

Arguing just as in the derivation of equation (42) (we need only replace ξ everwhere by η) we find that

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

From the intuitive probabilistic interpretation of the quantity α it follows[1]

[1] Equations (6) and (15) 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 η, ξ, $\bar{\xi}$ satisfying the conditions (6) and (15), we can construct a quantity α with distribution $\{q(a_e)\}$ such that η and α 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 α are independent quantities and that the condition (6) can now be replaced by the condition

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

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

$$I(\eta,\bar{\xi}) \leqq \mathbf{M}I(\eta,\bar{\xi}|\alpha). \tag{59}$$

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

$$I(\eta,\bar{\xi}|\alpha=a_e) \leqq nD(a_e), \tag{60}$$

where $D(a_e)$ is the capacity per unit time of the memoryless channel with transition probabilities $p_{ij}(a_e)$ (see (18)). It follows from (60) that

$$\mathbf{M}I(\eta,\bar{\xi}|\alpha) \leqq n\sum_e q(a_e)D(a_e),$$

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 $p_{ij}(a_k)$ is symmetric, i.e., $p_{11}(a_k)\equiv p_{22}(a_k)$, $p_{12}(a_k)\equiv p_{21}(a_k)$. Then for all k the upper bound in the expression (18′) for $D(a_k)$ is attained for $p_1=p_2=\frac{1}{2}$. Therefore for $p_1=p_2=\frac{1}{2}$, the expression

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

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 α. If the transition probability matrix is not assumed to be symmetric, then the optimum distribution will depend on α, 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 $E_1,\cdots,E_n$ and the output states $\bar{E}_1,\cdots,\bar{E}_n$. Let the parameter α take the n values $a_1,\cdots,a_n$ each with probability $1/n$. We assume that the transition probabilities $p_{ij}(a_k)$ are given by the formula

$$p_{i1}(a_k) = \begin{cases} 0, & i \neq k, \\ 1, & i = k, \end{cases}$$

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

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 $p_1 = \cdots = p_n = 1/n$, 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 $n \to \infty$, 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. Hauptschein, 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. Tsaregradskii, *A note on the capacity of a stationary channel with finite memory*, Theory Prob. Applications, 3, 1958, pp. 79–91. (English Translation.)

[9] S. 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.