APPROXIMATION WITH A FIDELITY CRITERION

J. WOLFOWITZ
CORNELL UNIVERSITY

1. Introduction

The present paper, as so many others in information theory, was stimulated by a paper of Shannon's [1]. The interesting theorem 1 below is due to him; the new result is theorem 2. We give a different proof of theorem 1. Actually this proof is not very new and is essentially the one used to prove theorem 1 of [3] (reproduced in [2] as theorem 3.2.1). The relation between the notion of "distortion" and that of "being generated" will be clear from this proof.

In the present paper we keep separate the ideas of approximating and coding. Then theorem 1 says essentially that, by embedding a certain number of sequences one can achieve a prescribed bound on the distortion, and theorem 2 says essentially that this cannot be done with fewer sequences. Shannon's results on coding are described in section 4. Some of his generalizations and additional suggestions for further generalizations are described in section 5.

It may perhaps be of interest to mention that theorem 4.9 of [4] is a special case of (4.3) below (the latter is theorem 1 of [1]). In fact, the probability of error defined in (4-65) of [4] is a special case of Shannon's distortion function ((2.1) below).

In [1], and in the present paper, the "source" digits (components of u below) are chance variables with a given (fixed) distribution. This is also true in the situation treated in theorem 4.9 of [4]. In the strong converse proved in [3], and in the others proved in [2], the messages are *not* stochastic and are chosen arbitrarily by the sender. If they should be chosen by a chance process their distribution can be arbitrary. The claims made in ([4], p. 219) on behalf of theorem 4.9 of [4] are therefore without the least basis in fact.

2. The approximating theorem

Consider the alphabets $M = \{m_1, \cdots, m_a\}$ and $Z = \{z_1, \cdots, z_b\}$. Let M^* (resp. Z^*) be the space of n-sequences (sequences of length n) in the M-alphabet (resp. the Z-alphabet). Let $\pi = (\pi_1, \cdots, \pi_a)$ be a probability a-vector which will be fixed in all that follows. When we speak of the probability distribution on M^*, we shall always mean the distribution implied by n independent chance variables with the common distribution π.

Research under contract with the Office of Naval Research.

Let d be a nonnegative function, called the "distortion" function, defined on $(M \times Z)$. Let $u_o = (x_1, \cdots, x_n)$ be any sequence in M^* and $v_o = (y_1, \cdots, y_n)$ be any sequence in Z^*. We define the distortion $d(u_o, v_o)$ between u_o and v_o by

$$d(u_o, v_o) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i). \tag{2.1}$$

Let D_o and D_{oo} be, respectively, the minimum and maximum values of d. Let D be a variable which temporarily takes values in the open interval (D_o, D_{oo}). For any value of D let $w(j|i|D)$, $i = 1, \cdots, a$; $j = 1, \cdots, b$, be nonnegative numbers (if they exist) such that $w(\cdot|i|D)$ is a probability b-vector with the following properties:

$$\sum_{i,j} \pi_i w(j|i|D)\, d(i, j) = L(w(\cdot|\cdot|D)) \text{ (say)} \leq D \tag{2.2}$$

and

$$\begin{aligned} R(w(\cdot|\cdot|D)) \text{ (say)} &= \sum_{i,j} \pi_i w(j|i|D) \log \left(\frac{w(j|i|D)}{\sum_i \pi_i w(j|i|D)} \right) \\ &\leq \sum_{i,j} \pi_i w(j|i) \log \left(\frac{w(j|i)}{\sum_i \pi_i w(j|i)} \right), \end{aligned} \tag{2.3}$$

where $w(\cdot|\cdot)$ is any channel probability function ((c.p.f.), that is, $w(j|i) \geq 0$, $i = 1, \cdots, a$; $j = 1, \cdots, b$, and $w(\cdot|i)$ is a probability b-vector) such that $L(w(\cdot|\cdot)) \leq D$. Henceforth we write

$$\pi'(D) = \left(\sum_i \pi_i w(1|i|D), \cdots, \sum_i \pi_i w(b|i|D)\right). \tag{2.4}$$

When we use π and $\pi'(D)$ to multiply matrices we shall consider them to be column vectors. Let $W(D)$ be the $(b \times a)$-matrix with element $w(j|i|D)$ in the j-th row and i-th column. Then

$$\pi'(D) = W(D)\pi. \tag{2.5}$$

To simplify the notation, we shall write $R(D)$ for $R(w(\cdot|\cdot|D))$. From the definition of $R(D)$ it is obvious that $R(D)$ is a monotonically nonincreasing function of D. Let $D_1 < D_2$ be any two values of D, and consider

$$w_o(\cdot|\cdot) = \tfrac{1}{2} w(\cdot|\cdot|D_1) + \tfrac{1}{2} w(\cdot|\cdot|D_2). \tag{2.6}$$

We have

$$L(w_o(\cdot|\cdot)) = \tfrac{1}{2} L(w(\cdot|\cdot|D_1)) + \tfrac{1}{2} L(w(\cdot|\cdot|D_2)) \tag{2.7}$$

and

$$R(w_o) \leq \tfrac{1}{2}(R(D_1) + R(D_2)). \tag{2.8}$$

Hence, $R(D)$ is a convex function of D, and hence, a (strictly) monotonically decreasing function of D.

The minimum value $D_{\min}$ of D, which we shall need to consider, can be found as follows: fix i; let j_o be such that

$$d(i, j_o) = \min_j d(i, j), \tag{2.9}$$

and let $w(j_o|i) = 1$. Then $D_{\min} = \sum_i \pi_i \min_j d(i, j)$. The maximum value $D_{\max}$ of D which we shall need to consider is the smallest value of D for which $R = 0$. If $R = 0$, then $w(j|i|D_{\max})$ is independent of i, say $w_o(j)$. Then

$$L(w(\cdot|\cdot|D_{\max})) = \min_{w_0} \sum_j w_o(j) \sum_i \pi_i\, d(i, j) = D_{\max} = \min_j \sum_i \pi_i\, d(i, j). \tag{2.10}$$

Henceforth, D will be a variable with values in the open interval $(D_{\min}, D_{\max})$. What happens at the ends of the interval will be discussed separately later, or else will be obvious.

Let S be a set of n-sequences in Z^*. For any element u_o of M^*, let

$$d(u_o, S) = \min_{v_o \in S} d(u_o, v_o). \tag{2.11}$$

Let u be a chance sequence with values in M^* and the distribution already defined on M^*. For any set S, the expected value $E\, d(u, S)$ is thus defined.

THEOREM 1. (The approximating theorem.) *Let $\epsilon^* > 0$ be arbitrary. There exists a function $n_o(\epsilon^*)$ of ϵ^* such that, for $n > n_o(\epsilon^*)$, we have the following: for any $D(D_{\min} < D < D_{\max})$ there exists a set $S(D) \subset Z^*$ containing N elements such that*

$$E\, d(u, S(D)) < D + \epsilon^* \tag{2.12}$$

and

$$N \leq \exp_2 \{nR(D)\}. \tag{2.13}$$

PROOF. We may assume that $D < D_{\max} - \epsilon^*$, or the theorem is trivially true. Let

$$\epsilon = \frac{\epsilon^*}{2(1 + D_{oo})}, \qquad D' = D + \frac{\epsilon^*}{2}, \tag{2.14}$$

and

$$h = \min_y \left[R(y) - R\left(y + \frac{\epsilon^*}{2}\right)\right] \tag{2.15}$$

where the minimum is taken over the range $D_{\min} \leq y \leq D_{\max} - \epsilon^*/2$. Hence, $h > 0$. Throughout the course of the present proof (and only then), write π', for short, in place of $\pi'(D') = W(D')\pi$. Let $w'(\cdot|\cdot|D')$ be defined by

$$w'(i|j|D') = \frac{\pi_i w(j|i|D')}{\pi_j'}, \qquad i = 1, \cdots, a; \quad j = 1, \cdots, b. \tag{2.16}$$

We define the chance variable (u, v) (u has already been defined) with values in $M^* \times Z^*$ and distribution determined by either of the following (which give the same result):

(i) the (marginal) distribution of u is as given above, and the conditional distribution of the k-th component of v, $(k = 1, \cdots, n)$, given $u = u_o$ and all the other components of v, is $w(\cdot|x_k|D')$, or

(ii) the (marginal) distribution of v is that of a sequence of independent chance variables with common distribution π', and the conditional distribution of the k-th component of u, $(k = 1, \cdots, n)$, given $v = v_o$ and all the other components of u, is $w'(\cdot|y_k|D')$.

Let $N(i|u_o)$ be the number of elements i in u_o, and similarly for v_o. Let $N(i, j|u_o, v_o)$ be the number of k, $k = 1, \cdots, n$, such that $x_k = i$ and $y_k = j$. We shall say that u_o is generated by v_o if

$$(2.17)\qquad |N(i, j|u_o, v_o) - N(j|v_o)w'(i|j|D')| \leq \delta[N(j|v_o)w'(i|j|D'))(1 - w'(i|j|D'))]^{1/2}$$

for $i = 1, \cdots, a; j = 1, \cdots, b$. Here $\delta > 0$ is such that

$$(2.18)\qquad P\{u \text{ is generated by } v_o|v = v_o\} > 1 - \frac{\epsilon}{4}.$$

(The symbol $P\{\ \}$ denotes the probability of the relation in braces. The symbol $P\{A|B\}$ denotes the probability of A, conditional upon B.) We shall say that a sequence v_o in Z^* is a π'-sequence if

$$(2.19)\qquad |N(j|v_o) - n\pi_j'| \leq z\sqrt{n\pi_j'(1 - \pi_j')}, \qquad j = 1, \cdots, b$$

where $z > 0$ is such that

$$(2.20)\qquad P\{v \text{ is a } \pi'\text{-sequence}\} > 1 - \frac{\epsilon}{4}.$$

It follows from (2.2), (2.16), (2.17), and (2.19) that, for all n sufficiently large and any pair (u_o, v_o) such that v_o is a π'-sequence and u_o is generated by v_o, we have

$$(2.21)\qquad d(u_o, v_o) < D' + \epsilon.$$

Let

$$(2.22)\qquad \{(v_1, A_1), \cdots, (v_N, A_N)\}$$

be a code $(n, N, 1 - \epsilon/4)$ as follows:

$$(2.23)\qquad v_1, \cdots, v_N \text{ are } \pi'\text{-sequences;}$$

$$(2.24)\qquad A_i,\ i = 1, \cdots, N, \text{ consists of all } n\text{-sequences in } M^* \text{ generated by } v_i \text{ and not in } A_1 \cup \cdots \cup A_{i-1};$$

$$(2.25)\qquad P\{u \in A_i|v = v_i\} \geq \frac{\epsilon}{4}, \qquad i = 1, \cdots, N;$$

(2.26) it is impossible to increase N and maintain (2.23)–(2.25).

As in ([2], (3.2.5)), we conclude that, when v_o is any π'-sequence not in the set $\{v_1, \cdots, v_N\}$, we have

$$(2.27)\qquad P\{u \text{ is generated by } v_o \text{ and is in } A_1 \cup \cdots \cup A_N|v = v_o\} > 1 - \frac{\epsilon}{2}.$$

For, if (2.27) did not hold, we could increase N by adding to (2.22) the pair (v_o, A_o), where A_o is the set of sequences generated by v_o and not in $A_1 \cup \cdots \cup A_N$.

Now, to each A_i, add enough sequences generated by v_i so that, calling the enlarged set B_i,

$$P\{u \in B_i | v = v_i\} > 1 - \frac{\epsilon}{2}, \qquad i = 1, \cdots, N. \tag{2.28}$$

We conclude from (2.27) and (2.28) that, for *any* π'-sequence v_o we have

$$P\{u \in (B_1 \cup \cdots \cup B_N) | v = v_o\} > 1 - \frac{\epsilon}{2}. \tag{2.29}$$

From (2.20) and (2.29) we conclude that

$$P\{u \in (B_1 \cup \cdots \cup B_N)\} > 1 - \epsilon. \tag{2.30}$$

Let

$$S(D) = \{v_1, \cdots, v_N\}. \tag{2.31}$$

From (2.30), (2.21), and the fact that every sequence in B_i is generated by the π'-sequence v_i, we obtain that

$$E\, d(u, S(D)) < D' + \epsilon + \epsilon D_{oo} = D + \epsilon^*. \tag{2.32}$$

From ([2], lemma 3.3.1) we obtain that

$$N < \exp_2 \{n[R(D') + h]\} \le \exp_2 \{nR(D)\} \tag{2.33}$$

for n sufficiently large. In the above argument, whenever n had to be sufficiently large, its lower bound could be made to depend only on ϵ^* and not on D or D'. (The lemma of [2] which we invoked is valid with constants which do not depend on the channel probability function.) The theorem is therefore proved.

It is obvious that we can replace ϵ^* by zero in (2.12) if we replace $R(D)$ by $(R(D) + \epsilon^*)$ in (2.13). From this, one can easily conclude what the theorem is when $D = D_{\min}$ or $D_{\max}$.

THEOREM 1′. *The set $S(D)$ whose existence is proved in theorem* 1 *may consist only of π'-sequences.*

This is a consequence of (2.23).

3. Converse of the approximating theorem

THEOREM 2. *Let $\epsilon^* > 0$ be arbitrary. There exists a function $n_{oo}(\epsilon^*)$ of ϵ^* such that, for $n > n_{oo}(\epsilon^*)$, we have the following: for any D ($D_{\min} < D < D_{\max}$), any set $S(D) \subset Z^*$ which contains N elements and satisfies*

$$E\, d(u, S(D)) \le D, \tag{3.1}$$

must also satisfy

$$N > \exp_2 \{n[R(D) - \epsilon^*]\}. \tag{3.2}$$

PROOF. Let $\epsilon > 0$ be a number to be chosen later. Write $D + \epsilon = D^*$. We have

$$P\{d(u, S(D)) < D^*\} \ge \frac{\epsilon}{D^*} = 2\alpha \text{ (say)} \tag{3.3}$$

by (3.1). Define the set G' by

$$G' = \{u_o \in M^* | d(u_o, S(D)) < D^*\}. \tag{3.4}$$

Let u_o be any sequence in M^*. We shall say that u_o is a π-sequence if

$$|N(i|u_o) - n\pi_i| \leq z'\sqrt{n\pi_i(1 - \pi_i)}, \qquad i = 1, \cdots, a \tag{3.5}$$

where $z' > 0$ is such that

$$P\{u \text{ is a } \pi\text{-sequence}\} > 1 - \alpha. \tag{3.6}$$

Hence,

$$P\{u \in G\} > \alpha, \tag{3.7}$$

where G is the set of π-sequences which are members of G'. For n sufficiently large the number of sequences in G exceeds

$$\alpha \cdot \exp_2 \{n[H(\pi) - \epsilon]\}, \tag{3.8}$$

where $H(\pi) = -\sum_i \pi_i \log \pi_i$; this is proved exactly as in lemma 2.1.7 of [2]. The lower bound on n does not depend on D.

Let u_o be any sequence in G and v_o any sequence in $S(D)$ such that

$$d(u_o, v_o) < D^*. \tag{3.9}$$

Now

$$\begin{aligned} nd(u_o, v_o) &= \sum_{i,j} N(i, j|u_o, v_o)\, d(i, j) \\ &= \sum_{i,j} N(i|u_o)w(j|i|u_o, v_o)\, d(i, j) \end{aligned} \tag{3.10}$$

where

$$w(j|i|u_o, v_o) = \frac{N(i, j|u_o, v_o)}{N(i|u_o)}. \tag{3.11}$$

(We stop for a moment to dispose of the case $N(i|u_o) = 0$. If no component of π is zero, then for n sufficiently large this can never occur. If $\pi_i = 0$ let the probability vector $w(\cdot|i|u_o, v_o)$ be defined arbitrarily.) Let $W(u_o, v_o)$ be the $(b \times a)$-matrix whose (j, i)-th element is $w(j|i|u_o, v_o)$. Since u_o is a π-sequence, it follows from (3.9), (3.10), and the definition of R in (2.3), that

$$R(w(\cdot|\cdot|u_o, v_o)) > R(D^*) - \psi_1(\epsilon), \tag{3.12}$$

where $\psi_1(\epsilon) \to 0$ as $\epsilon \to 0$ and $n \to \infty$.

To each u_o in G we may assign some v_o which satisfies (3.9). Since the right member of (3.11) is the ratio of two integers, it follows that the number of possible matrices $W(u_o, v_o)$ is at most $n^{a(b+1)}$. Let $W = \{w(j|i)\}$ be any matrix obtained as in (3.11), and let B be the set of pairs (u_o, v_o) such that $w(j|i|u_o, v_o) = w(j|i)$, $i = 1, \cdots, a; j = 1, \cdots, b$. Let K be the set of different v_o which occur among the elements of B. Suppose (3.2) does not hold. Then the number N_o of elements in K satisfies

$$N_o \leq \exp_2 \{n[R(D) - \epsilon^*]\}. \tag{3.13}$$

Let v_o be any sequence in K. It follows from the definition of G that

$$(3.14) \qquad |N(j|v_o) - n\varphi_j| < n\psi_2(\epsilon), \qquad j = 1, \cdots, b$$

where

$$(3.15) \qquad \varphi = W\pi,$$

$\psi_2(\epsilon) \to 0$ as $\epsilon \to 0$ and $n \to \infty$, and $W = W(u_o, v_o)$ is a matrix which corresponds to any pair (u_o, v_o) whose second element is the present v_o. Let $w(j|i)$ be the element in the j-th row and i-th column of W. Define

$$(3.16) \qquad w'(i|j) = \frac{\pi_i w(j|i)}{\varphi_j}, \qquad i = 1, \cdots, a, \quad j = 1, \cdots, b.$$

From (3.11), (3.16), and the fact that u_o is a π-sequence, we obtain

$$(3.17) \qquad |N(i, j|u_o, v_o) - n\varphi_j w'(i|j)| < n\psi_3(\epsilon), \qquad i = 1, \cdots, a, \quad j = 1, \cdots, b,$$

where $\psi_3(\epsilon) \to 0$ as $\epsilon \to 0$ and $n \to \infty$. It follows from (3.14), (3.17), and ([2], lemma 2.1.6), that the number of pairs in B, *with the same* v_o, is less than

$$(3.18) \qquad \exp_2 \{n[\sum_j \varphi_j H(w'(\cdot|j)) + \psi_4(\epsilon)]\}$$

where $\psi_4(\epsilon) \to 0$ as $\epsilon \to 0$ and $n \to \infty$. From (3.12) we obtain

$$(3.19) \qquad \begin{aligned} \sum \varphi_j H(w'(\cdot|j)) &< \sum_j \pi_j'(D^*) H(w'(\cdot|j|D^*)) + \psi_1(\epsilon) \\ &= \sum_j \pi_j'(D) H(w'(\cdot|j|D)) + \psi_5(\epsilon), \end{aligned}$$

where $\psi_5(\epsilon) \to 0$ as $\epsilon \to 0$ and $n \to \infty$. The right member of (3.13) is equal to

$$(3.20) \qquad \exp_2 \{n[H(\pi) - \sum_j \pi_j'(D) H(w'(\cdot|j|D)) - \epsilon^*]\}.$$

From (3.13), (3.20), and (3.18), we conclude that the number of different sequences in G is less than

$$(3.21) \qquad \begin{aligned} \exp_2 \{n[H(\pi) &- \sum_j \pi_j'(D) H(w'(\cdot|j|D)) - \epsilon^* \\ &+ \max_W \sum_j \varphi_j H(w'(\cdot|j)) + \psi_4(\epsilon)] + a(b + 1) \log_2 n\} \end{aligned}$$

which, by (3.19), is less than

$$(3.22) \qquad \exp_2 \{n[H(\pi) + \psi_5(\epsilon) + \psi_4(\epsilon) - \epsilon^*] + a(b + 1) \cdot \log_2 n\}.$$

From (3.8) and (3.22), we obtain

$$(3.23) \qquad -n\epsilon + \log_2 \alpha < n(\psi_5(\epsilon) + \psi_4(\epsilon) - \epsilon^*) + a(b + 1) \cdot \log n.$$

Now $\epsilon^* > 0$ is fixed. Let ϵ be sufficiently small and n sufficiently large. We obtain that (3.23) cannot hold. This contradiction and the fact that, whenever n had to be sufficiently large in the above proof, the lower bound on n did not depend on D, complete the proof of theorem 2.

4. Coding and approximating

Suppose that a discrete memoryless channel Γ of capacity C per letter is given. Each sequence in M^* is coded into an n'-sequence in the input alphabet of Γ, the latter is sent over Γ, and the then received n'-sequence in the output alphabet of Γ is decoded by the receiver into a sequence of Z^*. What is the expected value of the distortion d^* between the sequence in M^* and the one in Z^*? Essentially, the answer to this question has been given in [1].

Suppose that $D(D_{\min} < D < D_{\max})$ and an arbitrary $\epsilon > 0$ are given. If n and n' satisfy

$$\frac{n'C}{nR(D)} > 1 + \epsilon, \tag{4.1}$$

then, for all such n and n' greater than lower bounds which depend on D and ϵ, the expected value of d^* is less than $D + \epsilon$. To see this, one takes the set $S(D)$ embedded in Z^* according to theorem 1 ($\epsilon^* = \epsilon/2$). Let K_o be a code $(n', 2^{nR(D)}, \lambda)$ for channel Γ (that is, word length n', code length $2^{nR(D)}$, probability of error $\leq \lambda$). Because of (4.1), we can, by making n and n' sufficiently large, make λ as small as we wish. To each sequence in $S(D)$ we make correspond a transmitted (message) sequence of K_o in any manner, provided only that no transmitted sequence corresponds to more than one sequence in $S(D)$. Let $u_o \in M^*$ be any sequence. We code u_o into that transmitted sequence of K_o which corresponds to that (or any) sequence v_o in $S(D)$ such that

$$d(u_o, v_o) = d(u_o, S(D)). \tag{4.2}$$

After receiving the received sequence, the receiver decides which sequence was transmitted and then decodes the latter into its inverse in $S(D)$, if such an inverse exists. If it does not exist then he decodes into an arbitrary element of $S(D)$. Since λ can be made arbitrarily small the desired result is obvious.

(After this paper was completed, the author concluded that the problem described in this paragraph was considered in greater generality by Dobrushin [5]. It is extremely likely that the result attributed to Shannon in our paragraph above was also obtained by Dobrushin. (The verification of the latter's conditions is a formidable task.) Theorem 1 of our paper does not seem to be in [5] and seems, therefore, to be due to Shannon alone, as ascribed above. Our theorem 2 is not in [5] and, as stated in the introduction, is the new result of the present paper.)

Shannon ([1]) has proved the following nonasymptotic result:

$$E\, d^* \geq R^{-1}\left(\frac{n'}{n} C\right). \tag{4.3}$$

An intuitive explanation of this result is easy to give. According to theorem 2, we must embed approximately $\exp_2 \{nR(D)\}$ sequences in Z^* in order to attain $Ed(u, S(D)) \leq D$. Only $\exp_2 \{n'C\}$ sequences (approximately) can be sent over Γ and be distinguished from each other. Hence, operationally speaking, $S(D)$ acts as if it contained $\exp_2 \{n'C\}$ sequences. By theorem 2 the minimum distortion S which can be achieved with this many sequences in $S(D)$, satisfies $nR(s) = n'C$.

5. Generalizations

Theorems 1 and 2 can be generalized. For example, we used the fact that the distribution on M^* is that implied by independent, identically distributed chance variables in order to obtain (2.20) and (3.6). These will hold if the process on M^* is merely stationary and ergodic. Suppose now that the distortion function d is defined over the Cartesian product of g copies of M and g copies of Z, and, in place of (2.1), one defines

$$d(u_o, v_o) = \frac{1}{n-g+1} \sum_{k=1}^{n-g+1} d(x_k, x_{k+1}, \cdots, x_{k+g-1}, y_k, \cdots, y_{k+g-1}). \tag{5.1}$$

Shannon [1] has treated both these generalizations. Finally, it is not necessary that the channel Γ be discrete memoryless. This case, too, has been treated in [1].

The reader will recognize that the distortion function of (5.1) corresponds to the discrete finite-memory channel of ([2], chapter 5), just as the distortion function of (2.1) corresponds to the discrete memoryless channel. This suggests that one could employ even "nonlocal" distortion measures which correspond to other channels, for example, that of ([2], section 6.6).

REFERENCES

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *Information and Decision Processes*, edited by Robert E. Machol, New York, McGraw-Hill, 1960.

[2] J. Wolfowitz, *Coding Theorems of Information Theory*, Berlin, Springer-Verlag, 1961 (1st ed.), or 1964 (2d ed.).

[3] ———, "The coding of messages subject to chance errors," *Illinois J. Math.*, Vol. 1 (1957), pp. 591–606.

[4] R. G. Gallager, "Information theory," *The Mathematics of Physics and Chemistry*, edited by H. Margenau and G. M. Murphy, Chapter IV, Vol. 2, Princeton, Van Nostrand, 1964.

[5] R. L. Dobrushin, "General formulation of the fundamental theorem of Shannon in information theory," *Uspehi Mat. Nauk.*, Vol. 14 (1959), pp. 3–104.