INFORMATION THEORY FOR MATHEMATICIANS[1]

By J. Wolfowitz

Cornell University

A more descriptive term for information theory and one preferred by the present writer is "the theory of coding of messages." In this expository note we will describe briefly some basic concepts of this theory when transmission is through a "noisy channel" (noise = chance errors). We shall assume that both the transmitting alphabet and the receiving alphabet consist of two symbols, 0 and 1, say. This represents no loss in generality because the extension to any other alphabet, say one of twenty-six symbols, is immediate and presents no difficulty at all.

The fundamental paper of the theory is [1]; other important papers are [2], [3], [4], and [5]. The papers most easily intelligible to the mathematician are probably [3], [4], [7], and [8]. The latter three deal with the subject matter of the present paper; [4] and [7] may each be read without any prior reading, and [8] is a sequel of [7]. In the present paper we describe four theorems proved in [7] and [8] and their relation to prior results.

Suppose that a person has a vocabulary of S words, any of which he may want to transmit, in any frequency and in any order, over some channel. We emphasize that we do not assume anything about the frequency with which particular words are transmitted, nor that the words to be transmitted are selected by any random process; in this respect our treatment differs from most of those in the literature.

Let the words be numbered in some fixed but arbitrary manner. Then transmitting a word is equivalent to transmitting one of the integers $1, 2, \cdots, S$. Let $s = \log S$ (all logarithms in this paper are to the base 2). Then there are S sequences of s elements each[2], each element either 0 or 1. If there is no noise, i.e., error of transmission, then, to transmit any word one has only to transmit the appropriate sequence of s zeros or ones.

If there is noise then this is clearly not enough, for the transmitted sequence will usually be incorrectly received. What is needed is that the received sequence, which will usually be a moderately garbled version of the transmitted sequence, should still be different from the moderately garbled version of any other transmitted sequence, so that one can infer what sequence it is that has been transmitted. But this requires that the sequences to be sent be not too similar in some reasonable sense, lest they be confused in transmission. Hence one must employ sequences of length greater than s, and *not all* such sequences (so that "neighboring" sequences be not sent). All these remarks will now be made precise.

Let the integer m ($\geqq 0$) be the "memory". A sequence of n (respectively $(n - m)$,

Received October 31, 1957.

[1] Rietz lecture delivered (under a different title) at the Atlantic City meeting of the Institute of Mathematical Statistics on September 10, 1957, by invitation of the Council of the Institute. Work under contract with the Office of Naval Research.

[2] Obviously, if s is not an integer one should replace it by the smallest integer $\geqq s$.

$(m + 1)$) elements, each zero or one, will be called an x-sequence (resp., a y-sequence, an α-sequence).[3] A transmitted sequence (received sequence) is always an x-sequence (y-sequence). There is given a "channel probability function" p, defined in the domain of all α-sequences, such that, for any α-sequence α, $0 \leqq p(\alpha) \leqq 1$. The "noisy channel" transmits an x-sequence x as follows: Let α_1 be the α-sequence of the first $(m + 1)$ elements of x. The channel "performs" a chance experiment with one of two possible outcomes, 1 and 0, with respective probabilities $p(\alpha_1)$ and $1 - p(\alpha_1)$. The outcome of the experiment is the first element of the received sequence $Y(x)$. Let α_2 be the α-sequence of the 2nd, 3rd, $\cdots$, $(m + 2)$th elements of x. The channel now performs a chance experiment, independent of the first, with possible outcomes 1 and 0 and respective probabilities $p(\alpha_2)$ and $1 - p(\alpha_2)$. This is repeated until $(n - m)$ independent experiments have been performed. The probabilities of the outcomes one and zero in the ith experiment are $p(\alpha_i)$ and $1 - p(\alpha_i)$, respectively, where α_i is the α-sequence of the ith, $(i + 1)$st, $\cdots$, $(i + m)$th elements of x. The received sequence $Y(x)$ is a *chance* y-sequence made up of the outcomes of the experiments in consecutive order. Let y_1 be any y-sequence. If $P\{Y(x) = y_1\} > 0$ (the symbol $P\{\ \}$ denotes the probability of the relation in braces) then y_1 is called a possible received sequence when x is transmitted.

Let λ, $0 < \lambda < 1$, be a given number. A "code" of length t is a set $\{(x_i, A_i), i = 1, \cdots, t\}$ where each x_i is an x-sequence, each A_i is a set of y-sequences, the A_i are all disjoint, and for each i, $i = 1, \cdots, t$,

$$P\{Y(x_i) \varepsilon A_i\} \geqq 1 - \lambda.$$

To be able to transmit S words we need a code of length S. The practical application of a code is as follows: When one wishes to transmit the ith word one transmits the x-sequence x_i. Whenever the receiver receives a y-sequence which is in A_j, he always concludes that the jth word has been sent. When the receiver receives a y-sequence not in $A_1 \cup A_2 \cup \cdots \cup A_t$ he may draw any conclusion he wishes about the word that has been sent. The probability that any word transmitted will be incorrectly received is $< \lambda$.

The quantity $(1/n) \log t$ is called the rate of transmission. The practical advantages of a high rate of transmission are obvious. In this paper we shall be concerned with the problem of determining, or at least bounding, the highest possible rate of transmission.

If $p(\alpha_1) = p(\alpha_2)$, then the two α-sequences α_1 and α_2 are indistinguishable in transmission. Barring such cases for simplicity, then, whatever be λ, $0 < \lambda < 1$, it is always possible to find an n and then a code of length S, provided one is willing to transmit at a sufficiently small rate. By sufficient repetition of the word to be transmitted one can insure that the probability of its correct reception exceeds $1 - \lambda$.[4]

[3] These terms are used only in [7] and [8].

[4] For example, "estimation" of the word transmitted may be by the method of maximum likelihood. The words in the vocabulary are the possible "values" of the parameter to be estimated. Since there are only finitely many words in the vocabulary the method of maximum likelihood is uniformly consistent.

What we have called a code in the present paper is usually called "an error correcting" code[5] in the literature of coding theory. The latter often admits as codes systems which do not meet the definition of code given above. Much of the literature of coding theory is concerned with the situation where the words to be transmitted are chosen from the vocabulary by a chance process with known distribution. Without discussing this matter further here we invite the reader to verify that the results cited below about the existence of (error correcting) codes of certain lengths hold a fortiori when the words to be transmitted are chosen by a chance process.

Let M_2 be the class of all stationary, metrically transitive stochastic processes

$$X_1, X_2, X_3, \cdots$$

where the chance variables X_i can take only the values 0 and 1. Let M_1 be the subclass of M_2 in which the X_i constitute a Markov chain. Let M_0 be the subclass of M_1 in which the X_i are independently distributed. We shall shortly define a functional φ on every member of M_2 (more precisely, φ will be a functional of the distribution functions of the stochastic processes). In the meantime, let C_2, C_1, C_0, be, respectively, the supremum of φ over M_2, M_1, M_0, respectively. Then, of course, $C_0 \leqq C_1 \leqq C_2$.

Let ϵ always be an arbitrary positive number. The following Theorem A was first proved by Shannon [1] for the situation when the words to be transmitted are chosen by a known random process and for, in general, not error correcting codes.

THEOREM A. *For sufficiently large n there exists a code of length*

$$2^{n(C_1-\epsilon)}.$$

(In [1] Shannon not only proved this remarkable theorem but brilliantly laid the foundations of the whole subject). Basing himself on the ingenious and important work of Feinstein [2] and McMillan [5], Khintchine in a very important paper [4] rigorously proved THEOREM B. *For sufficiently large n there exists a code of length*

$$2^{n(C_2-\epsilon)}.$$

While Khintchine's paper does not explicitly treat error correcting codes, one can deduce from his proof that Theorem B holds for error correcting codes.

Theorem B obviously implies Theorem A (both for error correcting codes). The question arises whether Theorem B is stronger than Theorem A, i.e., whether $C_1 < C_2$. For $m = 0$ we will see below that the answer is in the negative. For general m the subject is under investigation.

In [7] (Theorem 3) the present author gave an extremely simple and very much briefer proof of Theorem B. Using essentially the same simple methods he proved the following improvement on Theorem A.

---

[5] More about error correcting codes in, for example, [6].

THEOREM 1 of [7]: *For any n there exists a code of length*

$$2^{nC_1 - K_1 n^{1/2}},$$

where K_1 is a positive constant[6] *which does not depend on n.*

We next concern ourselves with the important and interesting question of an upper bound for the length of an (error correcting) code. For the codes considered by Shannon the latter stated[7] ([1]) that there cannot exist a code of length greater than

$$2^{n(C_2+\epsilon)}.$$

Shannon gave a proof to which all others in the literature refer. Khintchine [4] pointed out that neither the argument of [1] nor any of the arguments to be found in the literature constitute a proof or even the outline of a proof; he also pointed out the desirability of proving the result and mentioned some of the difficulties.

In [7] (Theorem 2) the author proved the following theorem: *When $m = 0$ there is a positive constant*[8] *K_2 such that there cannot exist an (error correcting) code of length greater than*

$$2^{nC_0 + K_2 n^{1/2}}.$$

An immediate consequence of this theorem is that, when $m = 0$, $C_0 = C_1 = C_2$. Hence, when $m = 0$, Theorem B adds nothing to Theorem A, and both are weaker than Theorem 1.

Before passing to the case $m > 0$ we complete the above discussion by defining the functional φ. Let

$$X = (X_1, \cdots, X_n)$$

and define

$$Y(X) = (Y_1, \cdots, Y_{n-m}) = Y \text{ (say)}.$$

(More detailed definitions in [7]; Y is essentially the chance sequence received when the chance sequence X is sent.) Define the symbol

$$P\{Y = y \mid X = x\} \ (P\{X = x \mid Y = y\})$$

as the conditional probability that $Y = y$, given $X = x$ (that $X = x$, given $Y = y$). We define the following functions of the chance variables X and Y: When $X = x$ and $Y = y$, then

function	*equals*
$P\{X\}$	$P\{X = x\}$
$P\{Y\}$	$P\{Y = y\}$
$P\{X \mid Y\}$	$P\{X = x \mid Y = y\}$
$P\{Y \mid X\}$	$P\{Y = y \mid X = x\}$

[6] K_1 depends on the channel probability function.

[7] There is some ambiguity about the theorem actually stated.

[8] K_2 depends on the channel probability function.

Let E denote the expected value operator. It is proved that the following limits all exist:

$$\lim -\frac{1}{n} E[\log P\{X\}] = D_1$$

$$\lim -\frac{1}{n} E[\log P\{Y\}] = D_2$$

$$\lim -\frac{1}{n} E[\log P\{X \mid Y\}] = D_3$$

$$\lim -\frac{1}{n} E[\log P\{Y \mid X\}] = D_4$$

Also it is true and easy to prove that

$$D_1 + D_4 = D_2 + D_3 .$$

Then

$$\varphi = D_1 - D_3 = D_2 - D_4 .$$

We now turn to the general case $m \geqq 0$. For this case Theorem 4 of [8] gives a general upper bound for the length of an (error correcting) code. When $m = 0$ Theorem 4 specializes to Theorem 2. Whether Theorem 4 gives the "best" upper bound (as Theorem 2 does for $m = 0$) is still under investigation. Unfortunately, to state Theorem 4 one needs a page of preliminary definitions and then the theorem is stated in terms which require the reader to be familiar with the theory of Markov chains. (However, the application of the theorem as described in the discussion of [8] which follows its proof is little more difficult than that of Theorem 2). For these reasons it seems best to refer the interested reader to [8].

Postscript added in December, 1957.

Since this paper was submitted for publication the author has obtained the following result: A number J is defined by means of certain algebraic and analytic operations on the channel probability function which we shall not describe here. For any positive ϵ and n sufficiently large, there exists a code of length $2^{n(J-\epsilon)}$, and there cannot exist a code of length greater than $2^{n(J+\epsilon)}$. This result can be approximately described by saying that 2^{nJ} is the maximum achievable code length.

REFERENCES

[1] C. E. SHANNON, "A mathematical theory of communication," *Bell System Tech. Jour.* Vol. 27 (1948), pp. 379–423, 623–656.

[2] A. FEINSTEIN, "A new basic theorem of information theory," *Trans. Inst. Radio Eng.*, Professional Group on Information Theory (1954), pp. 2–22.

[3] A. KHINTCHINE, "The concept of entropy in the theory of probability," *Uspyekhi Mat. Nauk* VIII, 3(55), (1953), 3–20.

[4] A. KHINTCHINE, "On the fundamental theorems of the theory of information," Uspyekhi Mat. Nauk XI, 1(67), (1956), 17-75.
[5] B. MCMILLAN, "The basic theorems of information theory," **Ann. Math. Stat. 24, No.** 2 (1953), 196–219.
[6] E. N. GILBERT, "A comparison of signaling alphabets," Bell System Tech. Journal, 31 (1952), pp. 504–522.
[7] J. WOLFOWITZ, "The coding of messages subject to chance errors," to appear in Illinois Journal of Mathematics.
[8] J. WOLFOWITZ, "An upper bound on the rate of transmission of messages," to appear in Illinois Journal of Mathematics.