#2

A NEW BASIC THEOREM OF INFORMATION THEORY

AMIEL FEINSTEIN

TECHNICAL REPORT NO. 282

JUNE 1, 1954

RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the Army Signal Corps, the Navy Department (Office of Naval Research), and the Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039 sc-100, Project 8-102B-0; Department of the Army Project 3-99-10-022.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report No. 282 June 1, 1954

A NEW BASIC THEOREM OF INFORMATION THEORY

Amiel Feinstein

This report is identical with a thesis submitted to the Department of Physics, M. I. T., 1954, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

A new theorem for noisy channels, similar to Shannon's in its general statement but giving sharper results, is formulated and proven. It is shown that the equivocation of the channel defined by the present theorem vanishes with increasing code length. A continuous channel is defined in a manner that permits the application of these results. Detailed proof of the equivalence of this definition and Shannon's is given in an appendix.

INTRODUCTION

Information theory, in the restricted sense used in this paper, originated in the classical paper of C. E. Shannon, in which he gave a precise mathematical definition for the intuitive notion of information. In terms of this definition it was possible to define precisely the notion of a communication channel and its capacity. Like all definitions that purport to deal with intuitive concepts, the reasonability and usefulness of these definitions depend for the most part on theorems whose hypotheses are given in terms of the new definitions but whose conclusions are in terms of previously defined concepts. The theorems in question are called the fundamental theorems for noiseless and noisy channels. We shall deal exclusively with noisy channels.

By a communication channel we mean, in simplest terms, an apparatus for signaling from one point to another. The abstracted properties of a channel that will concern us are: (a) a finite set of signals that may be transmitted; (b) a set (not necessarily finite) of signals that may be received; (c) the probability (or probability density) of the reception of any particular signal when the signal transmitted is specified. A simple telegraph system is a concrete example. The transmitted signals are a long pulse, a short pulse, and a pause. If there is no noise in the wire, the possible received signals are identical with the transmitted signals. If there is noise in the wire, the received signals will be mutilations of the transmitted signals, and the conditional probability will depend on the statistical characteristics of the noise present.

We shall now sketch the definitions and theorems mentioned. Let X be a finite abstract set of elements x, and let p( ) be a probability distribution on X. We define the "information content" of X by the expression $-\sum_X p(x) \log_2 p(x)$, where the base 2 simply determines the fundamental unit of information, called "bit". One intuitive way of looking at this definition is to consider a machine that picks, in a purely random way but with the given probabilities, one x per second from X. Then $-\log_2 p(x_o)$ may be considered as the information or surprise associated with the event that x_o actually came up. If each event x consists of several events, that is, if $x = \{a, b, \ldots\}$, we have the following meaningful result: $H(X) \leq H(A) + H(B) + \ldots$ with equality if, and only if, the events a, b, ... are mutually independent.

We are now in a position to discuss the first fundamental theorem. We set ourselves the following situation. We have the set X. Suppose, further, that we have some "alphabet" of D "letters" which we may take as $0, \ldots, D-1$.

We wish to associate to each x a sequence of integers $0, \ldots, D-1$ in such a way that no sequence shall be an extension of some shorter sequence (for otherwise they would be distinguishable by virtue of their length, which amounts to introducing a $D+1^{th}$ "variable"). Now it is easy to show that a set of D elements has a maximum information content when each element has the same probability, namely $1/D$. Suppose now that with each x we associate a sequence of length N_x. The maximum amount of information obtainable by "specifying" that sequence is $N_x \log_2 D$ bits. Suppose $N_x \log_2 D = -\log_2 p(x)$; then $\sum_X p(x) N_x = H(X)/\log_2 D$ is the average length of the sequence. The first fundamental theorem now states that if we content ourselves with representing sequences of x's by sequences of integers $0, \ldots, D-1$, then if we choose our x-sequences sufficiently long, the sequences of integers representing them will have an average length as little greater then $H(X)/\log_2 D$ as desired, but that it is not possible to do any better than this.

To discuss the second fundamental theorem, we now take, as usual, X to be the set of transmitted messages and Y the set of received signals. For simplicity we take Y finite. The conditional probability mentioned above we denote by $p(y/x)$. Let $p(\ )$ be a probability distribution over X, whose meaning is the probability of each x being transmitted. Then the average amount of information being fed into the channel is $H(X) = -\sum_X p(x) \log_2 p(x)$. Since in general the reception of a y does not uniquely specify the x transmitted, we inquire how much information was lost in transmission. To determine this, we note that, inasmuch as the x was completely specified at the time of transmission, the amount of information lost is simply the amount of information necessary (on the average, of course) to specify the x. Having received y, our knowledge of the respective probability of each x having been the one transmitted is given by $p(x/y)$. The average information needed to specify x is now $-\sum_X p(x/y) \log_2 p(x/y)$. We must now average this expression over the set of possible y's. We obtain finally

$$\sum_Y p(y) \left[-\sum_X p(x/y) \log_2 p(x/y) \right] = -\sum_Y \sum_X p(x, y) \log_2 p(x/y) \equiv H(X/Y)$$

often called the equivocation of the channel. The rate at which information is received through the channel is therefore $R = H(X) - H(X/Y)$. A precise statement of the fundamental theorem for noisy channels is given in section II.

I. For the sake of definiteness we begin by stating a few definitions and subsequent lemmas, more or less familiar.

Let X and Y be abstract sets consisting of a finite number, α and β, of points x and y. Let p( ) be a probability distribution over X, and for each $x \in X$ let p(/x) denote a probability distribution over Y. The totality of objects thus far defined will be called a communication channel.

The situation envisaged is that X represents a set of symbols to be transmitted and Y represents the set of possible received signals. Then p(x) is the a priori probability of the transmission of a given symbol x, and p(R/x) is the probability of the received signal lying in a subset R of Y, given that x has been transmitted. Clearly, $\sum_{x \in Q} p(x)\, p(R/x)$ represents the joint probability of R and a subset Q of X, and will be written as p(Q, R). Further, $p(X, R) \equiv p(R)$ represents the absolute probability of the received signal lying in R. (The use of p for various different probabilities should not cause any confusion.)

The "information rate" of the channel "source" X is defined by $H(X) = -\sum_X p(x) \log p(x)$, where here and in the future the base of the logarithm is 2. The "reception rate" of the channel is defined by the expression

$$\sum_X \sum_Y p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \geq 0$$

If we define the "equivocation" $H(X/Y) = -\sum_X \sum_Y p(x, y) \log p(x/y)$ then the reception rate is given by H(X) - H(X/Y). The equivocation can be interpreted as the average amount of information, per symbol, lost in transmission. Indeed we see that H(X/Y) = 0 if and only if p(x/y) is either 0 or 1, for any x, y, that is, if the reception of a y uniquely specifies the transmitted symbol. When H(X/Y) = 0 the channel is called noiseless. If we interpret H(X) as the average amount of information, per symbol, required to specify a given symbol of the ensemble X, with p( ) as the only initial knowledge about X, then H(X) - H(X/Y) can be considered as the average amount, per symbol transmitted, of the information obtained by the (in general) only partial specification of the transmitted symbol by the received signal.

Let now u(v) represent a sequence of length n (where n is arbitrary but fixed) of statistically independent symbols x(y), and let the space of all sequences be denoted by U(V). In the usual manner we can define the various "product" probabilities. The n will be suppressed throughout. It is now

simple to verify the following relations:

$$\log p(u) = \sum_{i=1}^{n} \log p(x_i), \text{ where } u = \{x_1, \ldots, x_n\} \tag{1}$$

$$\log p(u/v) = \sum_{i=1}^{n} \log p(x_i/y_i), \text{ where } v = \{y_1, \ldots, y_n\} \tag{2}$$

$$H(X) = -\frac{1}{n} \sum_{U} p(u) \log p(u) \tag{3}$$

$$H(X/Y) = -\frac{1}{n} \sum_{U} \sum_{V} p(u, v) \log p(u/v) \tag{4}$$

The weak law of large numbers at once gives us the following lemma, which is fundamental for the proof of Shannon's theorem (see also section V).

LEMMA 1. For any ϵ, δ there is an $n(\epsilon, \delta)$ such that for any $n \geq n(\epsilon, \delta)$ the set of u for which the inequality $\left|H(X) + (1/n) \log p(u)\right| < \epsilon$ does not hold has $p(\ )$ probability less then δ. Similarly, but with a different $n(\epsilon, \delta)$, the set of pairs (u, v) for which the inequality $\left|H(X/Y) + (1/n) \log p(u/v)\right| < \epsilon$ does not hold has $p(\ ,\ )$ probability less than δ.

In what follows we shall need only the weaker inequalities $p(u) < 2^{-n(H(X)-\epsilon)}$ and $p(u/v) > 2^{-n(H(X/Y)+\epsilon)}$. The probability of these inequalities failing will be denoted by δ^- and δ^+, respectively.

The following lemma is required to patch up certain difficulties caused by the inequalities of lemma 1 failing to hold everywhere.

LEMMA 2. Let Z be a (u, v) set of $p(\ ,\ )$ probability greater than $1 - \delta_1$ and U_o a set of u with $p(U_o) > 1 - \delta_2$. For each $u \in U$ let A_u be the set of v's such that $(u, A_u) \in Z$. Let $U_1 \subset U_o$ be the set of $u \in U_o$ for which $p(A_u/u) \geq 1 - a$. Then $p(U_1) > 1 - \delta_2 - (\delta_1/a)$.

PROOF. Let U_2 be the set of u for which $p(A_u^c/u) > a$, where A_u^c is the complement of A_u. Then $p(u, A_u^c) \geq ap(u)$ for $u \in U_2$, and $\sum_{U_2} (u, A_u^c)$ is, by the definition of A_u, outside Z. Hence

$$\delta_1 \geq \sum_{U_2} p(u, A_u^c) \geq ap(U_2), \quad \text{or } p(U_2) \leq \frac{\delta_1}{a}$$

Thus $p(U_2 \cdot U_o) \leq \delta_1/a$ and, using $U_1 = U_o - U_o \cdot U_2$, we have

$$p(U_1) = p(U_o) - p(U_o \cdot U_2) > 1 - \delta_2 - \frac{\delta_1}{a}$$

II. We have seen that, by our definitions, the average amount of information received, per symbol transmitted, is H(X) - H(X/Y). However, in the process of transmission an amount H(X/Y) is lost, on the average. An obvious question is whether it is, in some way, possible to use the channel in such a manner that the average amount of information received, per symbol transmitted, is as near to H(X) - H(X/Y) as we please, while the information lost per symbol is, on the average, as small as we please. Shannon's theorem asserts (1), essentially, that this is possible. More precisely, let there be given a channel with rate H(X) - H(X/Y). Then for any $e > 0$ and $H < H(X) - H(X/Y)$ there is an n(e,H) such that for each $n \geq n(e,H)$ there is a family $\{u_i\}$ of message sequences (of length n) of number at least 2^{nH}, and a probability distribution on the $\{u_i\}$ such that, if only the sequences $\{u_i\}$ are transmitted, and with the given probabilities, then they can be detected with average probability of error less than e. The method of detection is that of maximum conditional probability, hence the need for specifying the transmission probability of the $\{u_i\}$. By average probability of error less than e is meant that if e_i is the fraction of the time that when u_i is sent it is misinterpreted, and p_i is u_i's transmission probability, then $\sum_i e_i p_i < e$.

A sufficient condition (2) for the above-mentioned possibility is the following:

For any $e > 0$ and $H < H(X) - H(X/Y)$ there is an n(e, H) of such value that among all sequences u of length $n \geq n(e, H)$ there is a set $\{u_i\}$, of number at least 2^{nH}, such that:

1. to each u_i there is a v-set B_i with $p(B_i/u_i) > 1 - e$
2. the B_i are disjoint.

What this says is simply that if we agree to send only the set $\{u_i\}$ and always assume that, when the received sequence lies in B_i, u_i was transmitted, then we shall misidentify the transmitted sequence less than a fraction e of the time. As it stands, however, the above is not quite complete; for, if C is the largest number such that for $H < C$ there is an n(e, H) and a set of at least 2^{nH} sequences u_i satisfying 1 and 2, C is well defined in terms of p(X/Y) alone.

However, H(X) - H(X/Y) involves p(X) in addition to p(X/Y). One might guess that C is equal to l.u.b. (H(X) - H(X/Y)) over all choices of p( ). This is indeed so, as the theorem below shows. Note the important fact that we have here a way of defining the channel capacity C without once mentioning information contents or rates. (Strictly speaking we should now consider the channel as being defined simply by p(y/x).) These remarks evidently apply equally well to Shannon's theorem, as we have stated it. We go now to the main theorem.

THEOREM. For any $e > 0$ and $H < C$ there is an $n(e, H)$ such that among all sequences u of length $n \geq n(e, H)$ there is a set $\{u_i\}$, of number at least 2^{nH} such that:

1. to each u_i there is a v-set B_i, with $p(B_i/u_i) > 1 - e$
2. the B_i are disjoint.

This is not possible for any $H > C$.

PROOF. Let us note here that if we transmit the u_i with equal probability and use a result of section III (namely $P_e \leq e$) we immediately obtain the positive assertion of Shannon's theorem. We shall first indicate only the proof that the theorem cannot hold for $H > C$, which is well known. Indeed if one could take $H > C$ then, as shown in section III one would have, for n sufficiently large, the result that the information rate per symbol would exceed C. But this cannot be (3). Q. E. D. In the following we will take p( ) as that for which the value C is actually attained (4). We shall see, however, that no use of this fact is actually made in what follows, other than, of course, $C = H(X) - H(X/Y)$.

For given ϵ_1, δ_1^+, ϵ_2, δ_2^-, let $n_1(\epsilon_1, \delta_1^+)$, $n_2(\epsilon_2, \delta_2^-)$ be as in lemma 1 for $p(u/v) > 2^{-n(H(X/Y)+\epsilon_1)}$ and $p(u) < 2^{-n(H(X)-\epsilon_2)}$, respectively. Let us henceforth consider n as fixed and $n \geq \max\left(n_1(\epsilon_1, \delta_1^+), n_2(\epsilon_2, \delta_2^-)\right)$. For Z and U_o in lemma 2 we take, respectively, the sets on which the first two inequalities stated above hold. Then for any $u \in U_1$ (with a as any fixed number $< e$) and v in the corresponding A_u we have:

$$\frac{p(u/v)}{p(u)} > \frac{2^{-n(H(X/Y)+\epsilon_1)}}{2^{-n(H(X)-\epsilon_2)}} = 2^{n(C-\epsilon_1-\epsilon_2)}, \text{ or}$$

$$\frac{p(u, v)}{p(u)} > 2^{n(C-\epsilon_1-\epsilon_2)}\, p(v)$$

Summing v over A_u we have

$$\frac{p(u, A_u)}{p(u)} > 2^{n(C-\epsilon_1-\epsilon_2)} p(A_u)$$

Since $1 \geq p(A_u/u)$ we have finally

$$p(A_u) < 2^{-n(C-\epsilon_1-\epsilon_2)}$$

Let $u_1, \ldots, u_N$ be a set M of members of U such that:

a. to each u_i there is a v-set B_i with $p(B_i/u_i) > 1 - e$

b. $p(B_i) < 2^{-n(C-\epsilon_1-\epsilon_2)}$ (See footnote 5.)

c. the B_i are disjoint

d. the set M is maximal, that is, we cannot find a u_{N+1} and a B_{N+1} such that the set $u_1, \ldots, u_{N+1}$ satisfies (a) to (c).

Now for any $u \in U_1$ there is by definition an A_u such that $p(A_u/u) \geq 1 - \alpha > 1 - e$ and as we have seen above, $p(A_u) < 2^{-n(C-\epsilon_1-\epsilon_2)}$. Furthermore, for any $u \in U_1$, $A_u - A_u \cdot \sum_i B_i$ is disjoint from the B_i, and certainly

$$p\left(A_u - A_u \cdot \sum_i B_i\right) < 2^{-n(C-\epsilon_1-\epsilon_2)}$$

If u is not in M, we must therefore have

$$p\left(A_u - A_u \cdot \sum_i B_i/u\right) \leq 1 - e$$

In other words, $p\left(A_u \cdot \sum_i B_i/u\right) \geq e - \alpha$, or certainly

$$p\left(\sum_i B_i/u\right) \geq e - \alpha, \text{ for all } u \in U_1 - M \equiv U_1 - M \cdot U_1$$

Now

$$p\left(\sum_i B_i\right) = \sum_U p\left(\sum_i B_i/u\right) p(u) \geq \left\{\sum_{U_1 - M \cdot U_1} + \sum_{M \cdot U_1}\right\} p\left(\sum_i B_i/u\right) p(u)$$

$$\geq (e-\alpha)\left[1 - \beta_2^- - \frac{\delta_1^+}{\alpha} - p(M \cdot U_1)\right] + (1-e)\, p(M \cdot U_1) \geq (e-\alpha)\left[1 - \delta_2^- - \frac{\delta_1^+}{\alpha}\right]$$

if $e \leq 1/2$, since then $1 - e \geq e - \alpha$.

On the other hand, $p\left(\sum_i B_i\right) < N2^{-n(C-\epsilon_1-\epsilon_2)}$. Hence

$$N2^{-n(C-\epsilon_1-\epsilon_2)} > (e-a)\left[1 - \delta_2^- - \frac{\delta_1^+}{a}\right]$$

If $e > 1/2$ then, using $p(M \cdot U_1) < N2^{-n(H(X)-\epsilon_2)}$, we would obtain

$$N2^{-n(C-\epsilon_1-\epsilon_2)} > (e-a)\left[1 - \delta_2^- - \frac{\delta_1^+}{a} - N2^{-n(H(X)-\epsilon_2)}\right]$$

Since the treatment of both cases is identical, we will consider $e \leq 1/2$.

To complete the proof we must show that for any e and $H < C$ it is possible to choose ϵ_1, ϵ_2, δ_1^+, δ_2^-, $a < e$, and $n \geq \max\left(n_1(\epsilon_1, \delta_1^+),\ n_2(\epsilon_2, \delta_2^-)\right)$ in such a way that the above inequality requires $N \geq 2^{nH}$. Now it is clear that, if, having chosen certain fixed values for the six quantities mentioned, the inequality fails upon the insertion of a given value (say N^*) for N, then the smallest N for which the inequality holds must be greater than N^*. Let us point out that N will in general depend upon the particular maximal set considered.

We take $N^* = 2^{nH}$ and $a = e/2$. Then we can take δ_1^+, δ_2^-, and ϵ_2 so small and n so large that

$$\left[1 - \delta_2^- - \frac{\delta_1^+}{a}\right] \quad \text{is} > \frac{2}{3} \quad \text{say.}$$

We obtain finally $e/3 < 2^{-n(C-H-\epsilon_2-\epsilon_1)}$. Choosing ϵ_2 and ϵ_1 sufficiently small so that $C - H - \epsilon_2 - \epsilon_1 > 0$ we see that for sufficiently large n the inequality $e/3 < 2^{-n(C-H-\epsilon_2-\epsilon_1)}$ fails. Hence for $a = e/2$, for ϵ_1, ϵ_2, δ_1^+, δ_2^- sufficiently small the insertion of $N^* = 2^{nH}$ for N causes the inequality to fail for all n sufficiently large. Thus $N > N^* = 2^{nH}$ for such n. Q.E.D.

It is worthwhile to emphasize that the codes envisaged here, unlike those of Shannon, are uniformly good, i. e., the probability of error for the elements of a maximal set is uniformly <e. These codes are therefore error correcting, which answers in the affirmative the question as to whether the channel

capacity can be approached using such codes (6).

If we wish to determine how e decreases as a function of n, for fixed H, we have (7):

$$e \leqslant a + \frac{A}{B - \left(\delta_1^+/a\right)}, \text{ where } A = 2^{-n(C-H-\epsilon_1-\epsilon_2)}, \quad B = 1 - \delta_2$$

To eliminate the "floating" variable a, we proceed as follows. For $a > 0$

$$a + \frac{A}{B - \left(\delta_1^+/a\right)} \text{ achieves its minimum value at } a = \frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B}$$

and this value, namely, $\frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$, is greater than $\frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B}$

If we take

$$a = \frac{\left(A\delta_1^+\right)^{1/2} + \delta_1^+}{B} \quad \text{and} \quad e = \frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$$

then $a < e$. Hence $\frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$ is an upper bound for the minimum value of e which is possible for a given H. This expression is still a function of ϵ_1 and ϵ_2. The best possible upper bound which can be obtained in the present framework is to minimize with respect to ϵ_1 and ϵ_2. This cannot be done generally and in closed form.

Let us remark, however, that at this point we cannot say anything concerning a lower bound for e. In particular, the relation $a < e$ is a condition that is required only if we wish to make use of the framework herein considered.

III. Let us consider a channel (i. e., (S, s), (R, r), p( ) and p(/s) where s is a transmitted and r a received symbol) such that to each s there is an r-set A_s such that $p(A_s/s) \geqslant 1 - e$ and the A_s are disjoint. For each r let $p_e(r) = 1 - p(s_r/r)$ where s_r is such that $p(s_r/r) \geqslant p(s/r)$ for all $s \neq s_r$. (Then $p_e(r)$ is simply the probability that when r is received an error will be made

in identifying the symbol transmitted, assuming that whenever r is received s_r will be assumed to have been sent.) Now the inequality $\ln a \leq a - 1$ can be used to show that

$$H(S/R) \leq -P_e \log P_e - (1 - P_e) \log (1 - P_e) + P_e \log (N-1)$$

where $P_e = \sum_R p(r)\, p_e(r)$ and N is the number of symbols in S.

We now make use of the special properties of the channel considered. We have

$$P_e = \sum_R p(r)(1 - p(s_r/r)) = 1 - \sum_R p(r)\, p(s_r/r)$$

$$= 1 - \sum_S \sum_{A_s} p(r)\, p(s_r/r) - \sum_{R-\Sigma A_s} p(r)\, p(s_r/r)$$

$$\leq 1 - \sum_S \sum_{A_s} p(r)\, p(s/r) - \sum_{R-\Sigma A_s} p(r)\, p(s_o/r)$$

$$= 1 - \sum_{S-s_o} \sum_{A_s} p(r)\, p(s/r) - \sum_{R-\sum_{S-s_o} A_s} p(r)\, p(s_o/r)$$

$$= 1 - \sum_{S-s_o} p(s)\, p(A_s/s) - p(s_o)\, p\left(R - \sum_{S-s_o} A_s/s_o\right)$$

$$\leq 1 - \sum_{S-s_o} p(s)(1-e) - p(s_o)(1-e) = e \quad \text{where } s_o \text{ is any } s \text{ (8).}$$

Then $H(S/R) \leq -e \log e - (1-e) \log (1-e) + e \log (N-1)$ since for $e < 1/2$ the left side of the above inequality is an increasing function of e. (We assume of course $e < 1/2$.)

Let us consider the elements $u_1, \ldots, u_N$ of some maximal set as the fundamental symbols of a channel. Then regardless of what $p(u_i)$ is, $i = 1, \ldots, N$, the channel is of the type considered above. Hence $P_e \leq e$ (where e is as in II) and

$$H(U/V) \leq -e \log e - (1-e) \log (1-e) + e \log (N-1)$$

Here H(U/V) represents the average amount of information lost per sequence

transmitted. The average amount lost per symbol is $1/n\ H(U/V)$. Now for $N = 2^{nH}$ and $H < C$, $e = e(n) \to 0$ as $n \to \infty$. Thus $1/n\ H(U/V) \to 0$ as $n \to \infty$. In particular if we take $p(u_i) = 2^{-nH}$, then $1/n\ [H(U) - H(U/V)] \to H$ as $n \to \infty$. (This is the proof mentioned in footnote 2.)

Actually, a much stronger result will be proven, namely, that for $N = 2^{nH}$, $H < C$ (and H fixed, of course) the equivocation per sequence $H(U/V)$, goes to zero as $n \to \infty$. Since $\log (N-1) \approx nH$, a sufficient condition that $H(U/V) \to 0$ as $n \to \infty$ is that $e(n)\ n \to 0$ as $n \to \infty$.

We saw that $e \leqslant \frac{1}{B}\left[A^{1/2} + \left(\delta_1^+\right)^{1/2}\right]^2$ where $B = 1 - \delta_2$ and $A = 2^{-n(C-H-\epsilon_1-\epsilon_2)}$.

Now if we take ϵ_1, ϵ_2 sufficiently small so that $C - H - \epsilon_1 - \epsilon_2 > 0$ and $H(X) - H - \epsilon_2 > 0$, then the behavior of δ_1^+ as $n \to \infty$ is the only unknown factor in the behavior of e. If the original X consists of only x_1, x_2, and Y consists of only y_1, y_2, and if $p(x_1/y_2) = p(x_2/y_1)$, then $\log p(x/y)$ is only two-valued. If we take $\epsilon_1 = \epsilon(n)$ as vanishing, for $n \to \infty$, faster than $n^{-1/6}$, then a theorem on large deviations (9) is applicable and shows that δ_1^+, and hence e, approaches zero considerably faster than $1/n$.

We omit the details inasmuch as a proof of the general case will be given in section V.

IV. Up till now we have considered the set Y of received signals as having a finite number of elements y. One can, however, easily think of real situations where this is not the case, and where the set Y is indeed nondenumerable. Our terminology and notation will follow the supplement of (10).

We define a channel by:

1. the usual set X and a probability distribution $p(\ )$ over X
2. a set Ω of points ω
3. a Borel field F of subsets Λ of Ω
4. for each $x \in X$, a probability measure $p(\ /x)$ on F.

We define the joint probability $p(x, \Lambda) = p(x)\ p(\Lambda/x)$ and $p(\Lambda) = p(X, \Lambda) = \sum_X p(x, \Lambda)$. Since $p(x, \Lambda) \leqslant p(\ )$ for any x, Λ, we have by the Radon-Nikodym theorem

4.1 $\quad p(x, \Lambda) = \int_\Lambda p(x/\omega)\ p(d\omega)$ where $p(x/\omega)$ may be taken as $\leqslant 1$ for all x, ω.

As the notation implies, $p(x/\omega)$ plays the role of a conditional probability.

We define $H(X) = -\sum_X p(x) \log p(x)$, as before. In analogy with the finite case we define

$$4.2 \quad H(X/Y) = -\sum_X \int_\Omega \log p(x/\omega)\, p(x, d\omega)$$

To show that the integral is finite, we see first, by section 4.1 that

$$p\left(x, \{p(x/\omega) = 0\}\right) = 0$$

Furthermore, putting

$$\Lambda_i = \left\{ \frac{1}{2^{i+1}} < p(x/\omega) \leqslant \frac{1}{2^i} \right\}$$

we have, since $p(\Lambda_i) \leqslant p(\Omega) = 1$, that

$$\int_{\Lambda_i} p(x/\omega)\, p(d\omega) \leqslant \frac{p(\Lambda_i)}{2^i} \leqslant \frac{1}{2^i}$$

Hence

$$4.3 \quad p\left(x, \left\{ \frac{1}{2^{i+1}} < p(x/\omega) \leqslant \frac{1}{2^i} \right\}\right) \leqslant \frac{1}{2^i}$$

We therefore have

$$4.4 \quad -\int_\Omega \log p(x/\omega)\, p(x, d\omega) < \sum_{i=0}^{\infty} \frac{i+1}{2^i} < \infty$$

by the ratio test.

Everything we have done in sections I, II, and III can now be carried over without change to the case defined above. A basic theorem in this connection is that we can find a finite number of disjoint sets Λ_j, $\sum_j \Lambda_j = \Omega$ such that $-\sum_X \sum_j p(x, \Lambda_j) \log p(x/\Lambda_j)$ approximates $H(X/Y)$ as closely as desired. Since we make no use of it, we shall not prove it, though it follows easily from the results given above and from standard integral approximation theorems.

V. We shall now show that $e = e(n)$ goes to zero, as $n \to \infty$, faster than $1/n$, which will complete the proof that the equivocation goes to zero

as the sequence length $n \to \infty$.

As previously mentioned, it is the behavior of δ_1^+, of lemma 1 that we must determine. The mathematical framework briefly is as follows.

We have the space $X \otimes \Omega$ of all pairs (x, ω) and a probability measure $p(\ ,\ )$ on the measurable sets of $X \otimes \Omega$. We consider the infinite product space $\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$ and the corresponding product measure

$$\prod_{i=1}^{\infty} \otimes p_i(\ ,\ ) \equiv p_\infty(\ ,\ ).$$

Let us denote a "point" of $\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$ by $(x_\infty, \omega_\infty) \equiv \{(x_1, \omega_1), (x_2, \omega_2), \ldots\}$

We define an infinite set of random variables $\{Z_i\}$, $i = 1, \ldots$ on

$$\prod_{i=1}^{\infty} \otimes (X \otimes \Omega)_i$$

by $Z_i(x_\infty, \omega_\infty) = -\log p(x_i/\omega_i)$, that is, Z_i is a function only of the i^{th} coordinate of $(x_\infty, \omega_\infty)$. Clearly the Z_i are independent and identically distributed; we shall put $E(Z_i)$ for their mean value. From section 4.4 we know that the Z_i have moments of the first order. (One can similarly show, using the fact that

$$\infty > \sum_{i=0}^{\infty} \frac{(i+1)^n}{2^i} \quad \text{for any } n > 0,$$

that they have moments of all positive orders.)

Let $S_n = \sum_{i=1}^{n} Z_i$. Then the weak law of large numbers says that for any ϵ_1, δ_1, there is an $n(\epsilon_1, \delta_1)$ such that for $n \geq n(\epsilon_1, \delta_1)$ the set of points $(x_\infty, \omega_\infty)$ on which $\left| \frac{S_n}{n} - E(Z_1) \right| \geq \epsilon_1$ has $p_\infty(\ ,\ )$ measure less than δ_1. Now, in the notation of section I, $S_n(X_\infty, \omega_\infty) = -\log p(u/v)$ where $u = \{x_1, \ldots, x_n\}$ and $v = \{\omega_1, \ldots, \omega_n\}$, while $H(X/Y) = \sum_X \int_\Omega -\log p(x/\omega)\, p(x, d\omega) = E(Z_1)$. What we have stated, then, is simply lemma 1.

Now, we are interested in obtaining an upper bound for

$$\text{Prob}\left\{\frac{S_n}{n} - E(Z_1) \geq \epsilon_1\right\}$$

More precisely we shall find sequences $\epsilon_1(n)$ and $\delta_1^+(n)$ such that, as $n \to \infty$, $\epsilon_1(n) \to 0$, $\delta_1^+(n) \to 0$ faster than $1/n$, and $n(\epsilon_1(n), \delta_1^+(n)) = n$.

Let $Z_i^{(r)} = Z_i$ whenever $Z_i < r$, and $Z_i^{(r)} = 0$ otherwise. By section 4.3,

$Z_i^{(r)}$ and Z differ on a set of probability $\leq 1/2^r$. Let $S_n^{(r)} = \sum_{i=1}^{n} Z_i^{(r)}$; then S_n and $S_n^{(r)}$ differ on a set of probability $\leq 1 - (1 - 2^{-r})^n < n/2^r$. Furthermore

$$E(Z_1) - E\left(Z_1^{(r)}\right) \leq \sum_{i=0}^{\infty} \frac{r + 1 + i}{2^{r+i}}$$

by the same argument which led to section 4. 4. We thus have:

$$\text{Prob}\left\{\frac{S_n}{n} - E(Z_1) \geq \epsilon_1(n)\right\} \leq \text{Prob}\left\{\frac{S_n^{(r)}}{n} - E(Z_1) \geq \epsilon_1(n)\right\} + \frac{n}{2^r}$$

$$\leq \text{Prob}\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\} + \frac{n}{2^r},$$

since $E(Z_1) \geq E\left(Z_1^{(r)}\right)$. In order to estimate Prob $\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\}$ we use a theorem of Feller (11) which, for our purposes, may be stated as follows:
THEOREM. Let $\{X_i\}$, $i = 1, \ldots, n$ be a set of independent, identically distributed, bounded random variables. Let $S = \sum_{i=1}^{n} X_i$ and let

$$F(x) = \text{Prob}\{S - n\,E(X_1) \leq x\}$$

Put $\sigma^2 = E([X_1 - E(X_1)]^2)$ and take $\lambda > \frac{\sup|X_1 - E(X_1)|}{\sigma n^{1/2}}$. Then if $0 < \lambda x < 1/12$ we have

$$1 - F(x\sigma n^{1/2}) = \exp[-1/2\, x^2\, Q(x)]\,[\{1 - \Phi(x)\} + \theta\lambda \exp(-1/2\, x^2)]$$

where

$$|\theta| < 9, \quad |Q(x)| \leq \frac{1}{7}\left(\frac{12\lambda x}{1 - 12\lambda x}\right) \text{ and } \Phi(x) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp[-y^2/2]\, dy$$

In order to apply this theorem, we take $r = r(n)$. Now

$$\sigma\left(Z_1^{(r)}\right) = E\left(\left[Z_1^{(r)} - E\left(Z_1^{(r)}\right)\right]^2\right)^{1/2} \to \sigma(Z_1) \quad \text{as } r \to \infty$$

Hence for suitably large n_o, $\frac{3}{2}\sigma(Z_1) > \sigma\left(Z_1^{(r)}\right) > \frac{1}{2}\sigma(Z_1)$ for $n \geq n_o$. We can

now take $\lambda \equiv \lambda(n) = \frac{2n^{-1/2}}{\sigma(Z_1)}\, r(n)$.

We henceforth consider $n \geq n_o$. We now have:

$$\text{Prob}\left\{\frac{S_n^{(r)}}{n} - E\left(Z_1^{(r)}\right) \geq \epsilon_1(n)\right\} = \text{Prob}\left\{S_n^{(r)} - n\, E\left(Z_1^{(r)}\right) \geq n\, \epsilon_1(n)\right\}$$

$$= \text{Prob}\left\{S_n^{(r)} - n\, E\left(Z_1^{(r)}\right) \geq \sigma\left(Z_1^{(r)}\right) n^{1/2} \left[n^{1/2}\, \frac{\epsilon_1(n)}{\sigma\left(Z_1^{(r)}\right)}\right]\right\}$$

$$\leq \text{Prob}\left\{S_n^{(r)} - n\, E\left(Z_1^{(r)}\right) \geq \sigma\left(Z_1^{(r)}\right) n^{1/2} \left[n^{1/2}\, \frac{2\epsilon_1(n)}{3(Z_1)}\right]\right\}$$

$$\leq \exp\left[\frac{1}{14}\, x^2 \left(\frac{12\lambda x}{1 - 12\lambda x}\right)\right] \left[\{1 - \Phi(x)\} + 9\lambda \exp\left(-\frac{x^2}{2}\right)\right]$$

Using

$$1 - \Phi(x) \sim \frac{1}{(2\pi)^{1/2}\, x} \exp\left(-\frac{x^2}{2}\right)$$

or

$$1 - \Phi(x) \lesssim \frac{2}{(2\pi)^{1/2}\, x} \exp\left(-\frac{x^2}{2}\right)$$

we may rewrite the above as

$$\exp\left(x^2 \left[\frac{6\lambda x}{7(1 - 12\lambda x)} - \frac{1}{2}\right]\right) \cdot \left\{9\lambda + \frac{2}{(2\pi)^{1/2}\, x}\right\}$$

Now $\lambda \equiv \lambda(n) = \frac{2n^{1/2}}{\sigma(Z_1)}\, r(n)$ and $x = n^{1/2}\, \frac{2\epsilon_1(n)}{3\sigma(Z_1)}$, while

$$\delta_1^+(n) \leq \exp\left(x^2 \left[\frac{6\lambda x}{7(1 - 12\lambda x)} - \frac{1}{2}\right]\right) \cdot \left\{9\lambda(n) + \frac{2}{(2\pi)^{1/2}\, x}\right\} + \frac{n}{2^{r(n)}}$$

It is now clear that we can pick $\epsilon_1(n)$ and $r(n)$ so that $\lambda(n) \to 0$, $x = x(n) \to \infty$, $\lambda x \to 0$ and $\delta_1^+(n) \to 0$ faster than $1/n$.

Let us point out that by using the approximation theorem of section III and thus having to deal with $-\log p(x/\Lambda_j)$, which is bounded, we can eliminate the term $n/2^{r(n)}$. This makes it likely that Feller's theorem can be proven, in our case, without the restriction that the random variables be bounded. There is in fact a remark by Feller that the boundedness condition can be replaced by the condition that Prob $\{|X_i| > n\}$ is a sufficiently rapidly decreasing function of n. But any further discussion would take us too far afield.

VI. We have, up to this point, insisted that the set X of messages be finite. We wish to relax this condition now so that the preceding work can be applied to the continuous channels considered by Shannon (1) and others. However, any attempt to simply replace finite sums by denumerable sums or integrals at once leads to serious difficulties. One can readily find simple examples for which H(X), H(X/Y) and H(X) - H(X/Y) are all infinite.

On the other hand, we may well ask what point there is in trying to work with infinite message ensembles. In any communication system there are always only a finite number of message symbols to be sent, that is, the transmitter intends to send only a finite variety of message symbols. It is quite true that, for example, an atrociously bad telegrapher, despite his intention of sending a dot, dash, or pause, will actually transmit any one of an infinite variety of waveforms only a small number of which resemble intelligible signals. But we can account for this by saying that the "channel" between the telegrapher's mind and hand is "noisy," and, what is more to the point, it is a simple matter to determine all the statistical properties that are relevant to the capacity of this "channel." The channel whose message ensemble consists of the finite number of "intentions" of the telegrapher and whose received signal ensemble is an infinite set of waveforms resulting from the telegrapher's incompetence and noise in the wire is thus of the type considered in section IV.

The case in which one is led to the consideration of so-called continuous channels is typified by the following example. In transmitting printed English via some teletype system one could represent each letter by a waveform, or each pair by a waveform, or every letter and certain pairs by a waveform, and so on. We have here an arbitrariness both in the number of message symbols and in the waveforms by which they are to be represented. It is now clear that

we should extend the definition of a channel and its capacity in order to include the case given above.

DEFINITION. Let X be a set of points x and Ω a set of points ω. Let F be a Borel field of subsets Λ of Ω, and let $p(\ /x)$ be, for each $x \in X$, a probability measure on F. For each finite subset R of X the corresponding channel and its capacity C_R is well defined by section IV. The quantity $C = \text{l.u.b. } C_R$ over all finite subsets R of X will be called the capacity of the channel $\{X, p(\ /x), \Omega\}$.

Now for any $H < C$ there is a C_R with $H < C_R \leqslant C$, so that all our previous results are immediately applicable.

We shall now show that the channel capacity defined above is, under suitable restrictions, identical with that defined by Shannon (1).

Let X be the whole real line, and Ω, ω, F, and Λ as usual. Let p(x) be a continuous probability density over X and for each $\Lambda \in F$, let $p(\Lambda/x)$ satisfy a suitable continuity condition. (See the Appendix for this and subsequent mathematical details.) Then $p(\Lambda) \equiv \int_{-\infty}^{\infty} p(x)p(\Lambda/x)\,dx$ is a probability measure. Since $p(x, \Lambda) \equiv p(x)p(\Lambda/x)$ is, for each x, absolutely continuous with respect to $p(\Lambda)$ we can define the Radon-Nikodym derivative $p(x/\omega)$ by $p(x, \Lambda) = \int_{\Lambda} p(x/\omega)p(d\omega)$. Then, with the x-integral taken as improper, we can define

$$C_p \equiv \int_{-\infty}^{\infty} dx \int_{\Omega} p(x, d\omega) \log \frac{p(x/\omega)}{p(x)} \geqslant 0$$

If we put $C_s = \text{l.u.b. } C_p$ over all continuous probability densities p(x), then C_s is Shannon's definition of the channel capacity. The demonstration of the equivalence of C, as defined above, and C_s is now essentially a matter of approximating an integral by a finite sum, as follows:

If C_s is finite, then we can find a C_p arbitrarily close to C_s; if $C_s = +\infty$ we can find C_p arbitrarily large. We can further require that p(x) shall vanish outside a suitably large interval, say [-A, A]. We can now find step-functions g(x) defined over [-A, A] that approximate p(x) uniformly to any desired degree of accuracy, and whose integral is 1. For such a step-function, C_g is well defined and approximates C_p as closely as desired by suitable choice of g(x).

Let g(x) have n steps, with area p_i, and of course $\sum_1^n p_i = 1$. By suitably

choosing positive numbers a_{ij}, integers N_i, and points x_{ij}, with x_{ij} lying in the i^{th} step of $g(x)$ and $\sum_{j=1}^{N_i} a_{ij} = p_i$, we can approximate

$$p(\Lambda) \equiv \int_{-A}^{A} g(x)\, p(\Lambda/x)\, dx \quad \text{by} \quad \sum_{i=1}^{n} \sum_{j=1}^{N_i} a_{ij}\, p(\Lambda/x_{ij})$$

and hence C_g by C_R, where $R = \{x_{ij}\}$. Thus $C \geq C_s$. On the other hand, let $R = \{x_i\}$, not as taken above. Let $p(x_i)$ be such that $H(X) - H(X/Y) = C_R$. Then the singular function $\sum_i p(x_i)\, \delta(x - x_i)$, where $\delta(\ )$ is the Dirac delta-function, can be approximated by continuous probability densities $p(x)$ such that C_p approximates C_R. Hence $C_s \geq C$, or $C = C_s$.

This can clearly be generalized to the case in which X is n-dimensional Euclidean space.

VII. We now wish to relax the condition of independence between successive transmitted symbols. Our definitions will be those of Shannon, as generalized by McMillan, whose paper (1) we now follow.

By an alphabet we mean a finite abstract set. Let A be an alphabet and I the set of all integers, positive, zero, and negative. Denote by A^I the set of all sequences $x = (\ldots, x_{-1}, x_o, x_1, \ldots)$ with $x_t \in A$, $t \in I$.

A cylinder set in A^I is a subset of A^I defined by specifying an integer $n \geq 1$, a finite sequence $a_o, \ldots, a_{n-1}$, of letters of A, and an integer t. The cylinder set corresponding to these specifications is $\left\{ x \in A^I / x_{t+k} = a_k,\ k = 0, \ldots, n-1 \right\}$. We denote by F_A the Borel field generated by the cylinder sets.

An information source $[A, \mu]$ consists of an alphabet A and a probability measure μ defined on F_A. Let T be defined by $T(\ldots, x_{-1}, x_o, x_1, x_2, \ldots) = (\ldots, x'_{-1}, x'_o, x'_1 \ldots)$ where $x'_t = x_{t+1}$. Then $[A, \mu]$ will be called stationary if, for $S \in F_A$, $\mu(S) = \mu(TS)$ (clearly T preserves measurability) and will be called ergodic if it is stationary and $S = TS$ implies that $\mu(S) = 1$ or 0.

By a channel we mean the system consisting of:

1. a finite alphabet A and an abstract space B.
2. a Borel field of subsets of B, designated by β, with $B \in \beta$
3. the Borel field of subsets of $B^I \equiv \prod_{-\infty}^{\infty} \otimes B_i$ (where $B_i = B$) which we define in the usual way, $\prod_{-\infty}^{\infty} \otimes \beta$, and designate F_β.
4. a function ν_x which is, for each $x \in A^I$, a probability measure on F_β,

and which has the property that if $x_t^1 = x_t^2$ for $t \leq n$, then $\nu_{x^1}(S) = \nu_{x^2}(S)$ for any $S \in F_\beta$ of the form $S = S_1 \otimes S_2$, where $S_1 \in \prod_{-\infty}^{n} \otimes \beta$ and $S_2 = \prod_{n+1}^{\infty} \otimes B$.

Consider a stationary channel whose input A is a stationary source $[A, \mu]$. Let $C^I = A^I \otimes B^I$ and $F_C = F_A \otimes F_\beta$. We can define a probability measure on F_C by $p(R, S) \equiv p(R \otimes S) = \int_R \nu_x(S)\, d\mu(x)$ for $R \in F_A$, $S \in F_\beta$, assuming certain measurability conditions for $\nu_x(S)$. It is then possible to define the information rate of the channel source, the equivocation of the channel, and the channel capacity in a manner analogous to that of section I. Assuming that $\mu(\ )$ and $p(\ ,\ )$ are ergodic, McMillan proves lemma 1 of section I in this more general framework. Hence the proof of section III remains completely valid, except for the demonstration that the theorem cannot hold for $H > C$.

The difficulty that we wish to discuss arises in the interpretation of $p(\ /u)$. A glance at McMillan's definitions shows that $p(B/u)$ no longer can be interpreted as "the probability of receiving a sequence lying in B, given that the sequence u was sent." This direct causal interpretation is valid only for $\nu_x(\ )$. But the result of the theorem of section II is the existance of a set u_i and disjoint sets B_i such that $p(B_i/u_i) > 1 - e$. Under what conditions can we derive from this an analogous statement for $\nu_{u_i}(B_i)$?

Suppose that for a given integer N we are given, for each sequence $x_1, \ldots, x_{N+1}$ of message symbols, a probability measure $\nu(\ /x_1, \ldots, x_{N+1})$ on the Borel field β of received signals (not sequences of signals). We envisage here the situation in which the received signal depends not only upon the transmitted symbol x_{N+1} but also upon the preceding N symbols which were transmitted.

If $u = \{x_1, \ldots, x_n\}$ then

$$p(\ /u) \equiv \sum_{[x_{-N+1}, \ldots, x_o]} \frac{p(x_{-N+1}, \ldots, x_n)}{p(x_1, \ldots, x_n)} \times [\nu(\ /x_{n-N}, \ldots, x_n) \otimes \ldots \otimes \nu(\ /x_{-N+1}, \ldots, x_1)]$$

Let us write the bracket term, which is a probability measure on received sequences of length n, as $\nu_n(\ /x_{-N+1}, \ldots, x_n)$. Now if $p(B_i/u_i) > 1 - e$, then,

since

$$\sum_{[x_{-N+1}, \ldots, x_o]} \frac{p(x_{-N+1}, \ldots, x_n)}{p(x_1, \ldots, x_n)} = 1$$

there must be at least one sequence $\{x_{-N+1}, \ldots, x_n\}$ for which

$$\nu_n(B_i/x_{-N+1}, \ldots, x_n) > 1 - e$$

A minor point still remains: we had 2^{nH} sequences u_i and we now have the same number of sequences, but of length n + N. In other words, we are transmitting at a rate H' = (n/n+N) H. But since N is fixed we can make H' as near as we choose to H by taking n sufficiently large; hence we can still transmit at a rate as close as desired to the channel capacity.

It is evident that by imposing suitable restrictions on $\nu_x(\ )$ we can do the same sort of thing in a more general context. These restrictions would amount to saying that the channel characteristics are sufficiently insensitive to the remote past history of the channel.

In this connection some interesting mathematical questions arise. If we define the capacity following McMillan for the $\nu(\ /x_1, \ldots, x_{N+1})$ as above, is the capacity actually achieved? It seems reasonable that it is, and that the channel source that attains the capacity will automatically be of the mixing type (see ref. 12, p. 36, Def. 11.1; also p. 57) and hence ergodic. Because of the special form of $\nu_x(\ )$ it easily follows that the joint probability measure would likewise be of mixing type and hence ergodic.

The question of whether or not the equivocation vanishes in this more general setup is also unsettled. Presumably one might be able to extend Feller's theorem to the case of nonindependent random variables that approach independence, or perhaps actually attain independence when far enough apart. To my knowledge nothing of this sort appears in the literature.

Finally there is the question of whether or not, in the more general cases, the assertion that for H > C the main theorem cannot hold is still true. While this seems likely, at least in the case of a channel with finite memory, it is to my knowledge unproven.

APPENDIX

It is our purpose here to supply various proofs that were omitted in the body of the work.

1. $H(X) - H(X/Y)$ is a continuous function of the $p(x_i)$, $i = 1, \ldots, a$.

PROOF. $H(X)$ is clearly continuous. To show the same for $H(X/Y)$ we need only show that for each i, $-p(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i)$ is a continuous function of $p(x_1), \ldots, p(x_a)$. Now

$$p(x_i/\omega) = \frac{p(x_i, d\omega)}{p(d\omega)} = p(x_i) \frac{p(d\omega/x_i)}{p(d\omega)}$$

But since $\sum_i p(\Lambda/x_i) \geq p(\Lambda)$, we have (see ref. 13, p. 133)

$$\frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} = \frac{p(d\omega/x_i)}{p(d\omega)} \cdot \frac{p(d\omega)}{\sum_i p(d\omega/x_i)}$$

$$= \frac{p(d\omega/x_i)}{p(d\omega)} \cdot \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}$$

almost everywhere with respect to $\sum_i p(\ /x_i)$ and hence, certainly, almost everywhere with respect to each $p(\ /x_i)$. Thus

$$\frac{p(d\omega/x_i)}{p(d\omega)} = \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Bigg/ \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}$$

almost everywhere with respect to $p(\ )$. The dependence on the $p(x_i)$ is now explicitly continuous, so that each $p(x_i/\omega)$ is a continuous function of $p(x_1), \ldots, p(x_a)$ almost everywhere with respect to each $p(\ /x_i)$. We now wish to show that $-p(x_i) \int_\Omega \log p(x_i/\omega)\, p/d\omega(x_i)$ is a continuous function of the $p(x_i)$.

To this end let $\{p_j(x_1), \ldots, p_j(x_a)\}$, $j = 1, \ldots$ be a convergent sequence of points in a-dimensional Euclidean space R_a, with limit $\{p_o(x_1), \ldots, p_o(x_a)\}$. Then we have $\lim_{j\to\infty} p_j(x_i/\omega) = p_o(x_i/\omega)$ almost everywhere with respect to each $p(\ /x)$. We must now show that

$$-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \to -p_o(x_i) \int_\Omega \log p_o(x_i/\omega)\, p(d\omega/x_i).$$

Suppose, first, that $p_o(x_i) \neq 0$. Now from section IV we have

$$\int_\Omega - \log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i) < \infty$$

whenever $p(x_i) \neq 0$. Take $p(x_i) = p(x_2) = \ldots p(x_a) = 1/a$. Then

$$\int_\Omega - \log a \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} p(d\omega/x_i) < \infty \quad \text{or clearly}$$

$$\int_\Omega - \log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} p(d\omega/x_i) < \infty. \quad \text{But}$$

$$- \log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \geqslant - \log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\}, \quad j = 1, 2, \ldots$$

Since the last term is also bounded below by $\log p_j(x_i)$, then by reference 14, p. 110, we have

$$\lim_{j \to \infty} \int_\Omega - \log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i)$$

$$= \int_\Omega - \log \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} p(d\omega/x_i)$$

Since $p_o(x_i) \neq 0$, $-p_j(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i) = p_j(x_i)$

$$\int_\Omega - \log \left[p_j(x_i) \cdot \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_j(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} \right] p(d\omega/x_i) \to p_o(x_i)$$

$$\int_\Omega - \log \left[p_o(x_i) \left\{ \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \Big/ \frac{\sum_i p_o(x_i)\, p(d\omega/x_i)}{\sum_i p(d\omega/x_i)} \right\} \right] p(d\omega/x_i)$$

$$= -p_o(x_i) \int_\Omega \log p(x_i/\omega)\, p(d\omega/x_i)$$

If $p_o(x_i) = 0$, we can clearly assume $p_j(x_i) \neq 0$, since we have to show that $-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \to 0$. As before we have

$$-\log \frac{p_j(x_i/\omega)}{p_j(x_i)} \leqslant -\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}, \text{ therefore}$$

$$-p_j(x_i) \int_\Omega \log p_j(x_i/\omega)\, p(d\omega/x_i) \leqslant p_j(x_i) \int_\Omega -\log \frac{p(d\omega/x_i)}{\sum_i p(d\omega/x_i)}\, p(d\omega/x_i)$$

$$+ p_j(x_i) \log \frac{1}{p_j(x_i)} \to 0 \quad \text{as } p_j(x_i) \to 0 \text{ (i.e., as } j \to \infty).$$

2. We wish here to rigorize the discussion of section VI.

We assume that $p(\ /x)$ satisfies the following continuity condition: For any finite closed interval I and any ϵ there is a $\delta(I, \epsilon)$ such that

$$\left| \frac{p(\Lambda/x_2)}{p(\Lambda/x_2)} - 1 \right| < \epsilon \quad \text{for} \quad |x_1 - x_2| \leqslant \delta \quad \text{and } x_1, x_2 \in I,$$

whenever $p(\Lambda/x_2) \neq 0$. It follows that if, for $x_1 \in I$, $p(\Lambda/x_1) = 0$, then for $x_2 \in I$ and $|x_1 - x_2| < \delta$, $p(\Lambda/x_2) = 0$. (Indeed, since $\{x/p(\Lambda/x) = 0\}$ is evidently both open and closed, for any Λ, $p(\Lambda/x)$ either vanishes everywhere or nowhere.) That $p(\Lambda) \equiv \int_{-\infty}^{\infty} p(x)\, p(\Lambda/x)\, dx$ is a probability measure is a simple consequence of reference 14, p. 112, Theorem B. Since $p(x)\, p(\Lambda/x)$ is continuous, $p(\Lambda)$ can vanish only if $p(x)\, p(\Lambda/x)$ is zero for all x. Hence, for all x, $p(x)\, p(\Lambda/x)$ is absolutely continuous with respect to $p(\Lambda)$.

We can sharpen this result as follows: Let I be a closed interval over which $p(x) \neq 0$. Then for a given ϵ we can find a δ such that $p(x_1) > p(x_2)/2$ and $p(\Lambda/x_1) \geqslant (1-\epsilon)\, p(\Lambda/x_2)$, for $x_1, x_2 \in I$ and $|x_1 - x_2| < \delta$. We thus have

$$\int_{-\infty}^{\infty} p(x)\, p(\Lambda/x)\, dx \geqslant 2\delta \frac{p(x_2)}{2} p(\Lambda/x_2)(1-\epsilon) = \delta\, p(x_2)(1-\epsilon)\, p(\Lambda/x_2)$$

Thus for any $x_2 \in I$,

$$p(x_2)\, p(\Lambda/x_2) \leqslant \frac{1}{(1-\epsilon)\delta}\, p(\Lambda) \equiv k(x_2)\, p(\Lambda)$$

which defines $k(x_2) < \infty$. As in section IV, we can easily show that

$$-\infty < -\int_\Omega \log p(x/\omega)\, p(x, d\omega) < \infty \quad \text{for all } x. \quad \text{Now}$$

$$\int_\Omega p(x, d\omega) \log \frac{p(x)}{p(x/\omega)} \leqslant \int_\Omega p(x, d\omega) \left\{ \frac{p(x)}{p(x/\omega)} - 1 \right\} \log e$$

$$= \int_\Omega p(x/\omega)\, p(d\omega) \left\{ \frac{p(x)}{p(x/\omega)} - 1 \right\} \log e = 0,$$

the next to last equality being justified by reference 14, p. 133. Therefore, if $\int_\Omega p(x, d\omega) \log \frac{p(x/\omega)}{p(x)}$ is, say, continuous in x, then

$$\lim_{a \to \infty} \int_{-a}^{a} \int_\Omega p(x, d\omega) \log \frac{p(x/\omega)}{p(x)}\, dx$$

is meaningful and is either positive or equal to $+\infty$.

We shall now show that $\int_\Omega p(x, d\omega) \log \frac{p(x/\omega)}{p(x)}$ is indeed continuous. To this end let x_i be a convergent sequence of real numbers with limit x_o. We shall show that $p(x_i/\omega) \to p(x_o/\omega)$ almost everywhere with respect to $p(x_o, \ )$. (Since for $p(x_o) = 0$ this assertion is trivially true, we assume that $p(x_o \neq 0)$.) Let $A_{in}^+ = \{p(x_i/\omega) - p(x_o/\omega) > 1/n\}$ and $A_{in}^- = \{p(x_i/\omega) - p(x_o/\omega) < -1/n\}$. Now

$$p(x_i)\, p(A_{in}^+/x_i) - p(x_o)\, p(A_{in}^+/x_o) = \int_{A_{in}^+} (p(x_i/\omega) - p(x_o/\omega))\, p(d\omega) > 1/n\, p(A_{in}^+).$$

There is clearly no loss in generality in assuming $p(x_i) \neq 0$. Then

$$p\left(A_{in}^+/x_i\right) - p\left(A_{in}^+/x_o\right) > \frac{p(x_o)}{k(x_o)\, n\, p(x_i)}\, p\left(A_{in}^+/x_o\right) + \frac{p(x_o) - p(x_i)}{p(x_i)}\, p\left(A_{in}^+/x_o\right)$$

Now $\frac{p(k_o)}{k(x_o)\, n\, p(x_i)} + \frac{p(x_o) - p(x_i)}{p(x_i)}$ is positive and bounded away from zero for all i sufficiently large. By the continuity condition on $p(\ /x)$ we therefore have $p\left(A_{in}^+/x_o\right) = 0$ for $i > i(n)$ suitably chosen. We get a similar result for $p\left(A_{in}^-/x_o\right)$. Let A_n^+ be the set of points ω which lie in infinitely many A_{in}^+, and similarly for A_n^-. Then $p\left(A_n^\pm/x_o\right) = 0$, and so,

$$p\left(\sum_n A_n^+ + \sum_n A_n^-/x_o\right) = p\left(x_o, \sum_n \left(A_n^+ + A_n^-\right)\right) = 0$$

But for any $\omega \in \Omega - \sum_n A_n^+ - \sum_n A_n^-$, $p(x_i/\omega) \to p(x_o/\omega)$, which was to be shown.

As before, let x_i be a convergent sequence with limit x_o.

a. Let us assume first that $p(x_o) \neq 0$. Now

$$\left| \int_\Omega -\log p(x_o/\omega)\, p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i) \right|$$

$$= \left| \int_\Omega \left[-\log p(x_o/\omega) + \log p(x_i/\omega)\right] p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i) \right.$$

$$\left. + \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_o) \right| \leq \left| \int_\Omega \left[-\log p(x_o/\omega) + \log p(x_i/\omega)\right] p(d\omega/x_o) \right|$$

$$+ \left| \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_o) - \int_\Omega -\log p(x_i/\omega)\, p(d\omega/x_i) \right|$$

To show that the first of the last two terms goes to zero, we remark, first, that since $p(x_o) \neq 0$ and $p(\Lambda/x_o) \leq (1+a)\, p(\Lambda/x_i)$ for any Λ, for i suitably large, it follows, as in section IV, that

$$\int_{\{-\log p(x_i/\omega) \geq R\}} -\log p(x_i/\omega)\, p(d\omega/x_o)$$

is uniformly bounded for all i, where we use the previously shown result that $p(x)\, p(\Lambda/x) \leq k(x)\, p(\Lambda) \leq M\, p(\Lambda)$ for M suitably chosen and x in a closed interval containing x_o. It is now a simple exercise, by using reference 14, p. 110, to justify the interchange of limit and integration, so that the term in question vanishes as $i \to \infty$. The relation $p(\ /x_o) \leq (1+a)\, p(\ /x_i) \leq (1+a)^2\, p(\ /x_o)$, with $a \to 0$ as $i \to \infty$, at once shows that the second term likewise vanishes as $i \to \infty$.

b. Now suppose that $p(x_o) = 0$. Then by definition we take

$$\int_\Omega -\log p(x_o/\omega)\, p(x_o, d\omega) = 0$$

If p(x) is identically zero in some neighborhood of x_o there is nothing to be proven. We can then assume that $p(x_i) \neq 0$. For a closed interval containing x_o and the x_i, we have, for $|x_i - x_j|$ sufficiently small (or equivalently, i and

j sufficiently large) that $p(\ /x_i) \geq (1-\epsilon)\, p(\ /x_j)$. Thus

$$p(x_i/\omega) \geq p(x_i) \frac{p(d\omega/x_j)}{p(d\omega)} (1-\epsilon). \quad \text{Hence}$$

$$-\log p(x_i/\omega) \leq -\log [p(x_i)(1-\epsilon)] - \log \frac{p(d\omega/x_j)}{p(d\omega)}$$

for fixed j and any i, both sufficiently large. Further, since $p(x_j) \neq 0$, $p(x_j)\, p(\Lambda/x_j) = M\, p(\Lambda)$ for suitable M. Hence

$$p(x_i)\, p(\Lambda/x_i) \leq \frac{1}{1-\epsilon} \frac{p(x_i)}{p(x_j)} M\, p(\Lambda)$$

Since $p(x_i) \to 0$, we have, for sufficiently large i, $p(x_i)\, p(\Lambda/x_i) \leq p(\Lambda)$, so that $p(x_i/\omega) \leq 1$ or $-\log p(x_i/\omega) \geq 0$. Therefore

$$\int_\Omega -\log p(x_i/\omega)\, p(x_i, d\omega) \leq -p(x_i) \log [p(x_i)(1-\epsilon)]$$

$$+ p(x_i) \int_\Omega -\log \frac{p(d\omega/x_j)}{p(d\omega)} p(d\omega/x_i)$$

As i approaches ∞, the last integral approaches

$$\int_\Omega -\log \frac{p(d\omega/x_j)}{p(d\omega)} p(d\omega/x_o), \quad \text{which is} < \infty,$$

using arguments as in section IV. Since $p(x_i) \to 0$, we have, finally,

$$\int_\Omega -\log p(x_i/\omega)\, p(x_i, d\omega) \to 0 \quad \text{as } i \to \infty.$$

References and Footnotes

1. C. E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27, 379-423, 623-656; also B. McMillan, The basic theorems of information theory, Ann. Math. Stat. 24, 196-219.

2. That it is indeed sufficient will be shown in section III.

3. R. M. Fano, Lecture notes on statistical theory of information, Massachusetts Institute of Technology, spring, 1952. This statement asserts that if the channel is considered as transmitting sequence by sequence its capacity per symbol is still bounded by C. Using the fact that the reception rate per symbol may be written as

$$\frac{H(V) - H(V/U)}{n}$$

the statement follows upon noticing that H(V/U) depends only upon single-received-symbol probabilities and that H(V) is a maximum when those probabilities are independent. The expression H(V) - H(V/U) then reduces to a sum of single-symbol channel rates, from which the assertion follows at once.

4. It is not difficult to see that H(X) - H(X/Y) is a continuous function of the "variables" $r_i = p(x_i)$, $i = 1, \ldots, a$. This is true also in the context of section IV (c.f. Appendix). Since the set of points in a-dimensional cartesian space R_a defined by $r_i \geqslant 0$ and $\sum_{i=1}^{a} r_i = 1$ is a closed set, H(X) - H(X/Y) attains a maximum value. This point is, however, not critical, for, given $H < C$ we can certainly find $p(\ )$ such that $H < H(X) - H(X/Y) < C$ and then use H(X) - H(X/Y) in place of C.

5. This condition appears to be superfluous. It is, however, strongly indicated by the immediately preceding result and is, in fact, essential for the proof.

6. E. M. Gilbert, A comparison of signalling alphabets, Bell System Tech. J. 31, in particular p. 506.

7. Up to here, the possibility that certain quantities are not integers can be seen not to invalidate any of the various inequalities. In what follows, the modifications needed to account for this possibility are obvious and insignificant and are therefore omitted.

8. Word-wise, this string of inequalities states simply: (a) that in order to minimize the probability of misidentifying the transmitted s we should guess the s with greatest conditional probability as the one actually transmitted; (b) if instead of the above recipe, we assume that s was sent, whenever $r \epsilon A_s$ is received, for all s except s_o, and that in all other circumstances we shall assume s_o to have been sent, then the probability of error is less than e; (c) hence, since P_e is the error obtained by the best method of guessing, $P_e \leqslant e$.

9. See reference 13, pp. 144-5. This was pointed out by Professor R. M. Fano.

10. J. L. Doob, Stochastic Processes (John Wiley and Sons, Inc., New York, 1953).

11. W. Feller, Generalization of a probability limit theorem of Cramer, Trans. Amer. Math. Soc. 54, 361-372 (1943).

12. E. Hopf, Ergodentheorie (Julius Springer, Berlin, 1937).

13. W. Feller, An Introduction to Probability Theory (John Wiley and Sons, Inc., New York, 1950).

14. P. R. Halmos, Measure Theory (D. Van Nostrand, New York, 1950).