THE BASIC THEOREMS OF INFORMATION THEORY

By Brockway McMillan

Bell Telephone Laboratories

Summary. This paper describes briefly the current mathematical models upon which communication theory is based, and presents in some detail an exposition and partial critique of C. E. Shannon's treatment of one such model. It then presents a general limit theorem in the theory of discrete stochastic processes, suggested by a result of Shannon's.

1. General models of the communication problem.

1.0. *Introduction.* For the purposes of this exposition, information theory is the body of statistical mathematics which has developed, largely over the last decade, out of efforts to understand and improve the communications art. We shall not attempt a history of this development, nor any detailed justification for its existence, since either of these efforts would take us further into the technics of communication than is desirable in a short essay.

It suffices to say here that this discipline has come specifically to the attention of mathematicians and mathematical statisticians almost exclusively through the book [1] of N. Wiener and the paper [2] of C. E. Shannon.

In the remainder of this section we shall describe very broadly the kind of problem to which these two works are addressed.

1.1. *A simple model.* The simplest mathematical model of the communication problem is like the problem of parameter estimation. A parameter θ, usually ranging over a fairly abstract or at least multi-dimensional domain, represents the transmitted message. A variable, y, also fairly abstract in general, represents the received message. In realistic situations the received message is seldom a mathematically exact copy, or even an exactly predictable mutilation, of the original transmitted message. Hence, y is represented as a random variable whose distribution depends upon the parameter θ. The communication problem then is: given a sample of one value of y, to estimate the unknown θ.

There are two reasons why this model may not seem at first look to be a good one for the communication problem. One is merely that our most usual media of communication, direct acoustic transmission of voice and the written or printed word, are ones in which essentially exact transmission is possible and we are not aware of the underlying statistical nature of the problem. This is clearly a matter of degree, however, and almost anyone can find in his own experience instances in which the statistical aspect of the problem was evident.

Another apparent failing of this model is in fact real, and has led to refinement of the model. There are communication problems, mostly in technical fields, where it is realistic to assume that the recipient of y has no a priori knowledge

Received 8/9/51.

about the parameter θ. The usual situation in human experience, however, is one in which there is a great deal of a priori knowledge about the possible values of θ. There are simple experiments with mutilated text, spoken or written, which will convince one that he can, and often does, exploit his own a priori knowledge of language, speaker, and subject matter to assist in deciphering what he reads and hears. A realistic model must include this possibility.

1.2. *Stochastic transmitted message.* It was Wiener who first clearly pointed out that we may, and indeed often must, regard the transmitted message itself as a random variable drawn from a universe whose distribution function reflects our a priori knowledge of the situation. Cogent statements of this philosophy may be found both in [1] and [2]. This leads us to a model in which we have two abstract random variables, say x representing the transmitted message (replacing the parameter θ), and y the received message. There is then a joint distribution function for x and y which contains in it the complete mathematical description of the situation. One ordinarily thinks of this distribution function as being "factored" into an a priori distribution for x, representing the universe of possible messages, and a conditional distribution for y knowing x, representing for each x the universe of possible mutilations thereof.

In this second and more important model, one can still regard the communication problem as one of estimation: given the y value of a joint sample (x, y), to estimate the x value. This view is particularly appropriate in discussing the work of [1]. Here, the x and y are numerically valued time series and there is a natural numerical way to measure the deviation between the estimated and true values of x, namely, by the variance of estimate.

The statistician may alternatively wish to regard the communication problem (in either model) as one of testing hypotheses. The observed y has a distribution depending on the hypothesis "x;" the problem is to decide which x is obtained at the time of observation. This view is more appropriate to the work of [2], wherein the time series are abstract valued, and no natural measure of the "wrongness" of an incorrectly adopted hypothesis is available. In the second model, the a priori distribution for hypotheses x eliminates one kind of testing error, so that in this model there is a simple criterion of performance, namely, the total probability in the (x, y) universe of all events (x, y) in which the hypothesis adopted is correct. The reader will observe this particular criterion in sections 6 and 8.

The distinction between estimation, on the one hand, and testing among many hypotheses, on the other, is not sharp. We shall use "estimation" as a loose word to refer to the kind of model here set up for communication.

1.3. *Peculiarities of engineering applications.* Information theory is distinguished from a general study of models like these in two important respects. In the first place, as noted, the random quantities x and y of interest are, naturally, time series. Furthermore, the passage of time is explicitly recognized and the distinction between past events, which can be known, and future events which cannot be known, is carefully observed.

In the second place, the kind of question considered in information theory, particularly by Shannon, reflects the peculiar interests of communication engineers. To illustrate this, we might go back to the jointly distributed abstract variables x and y of 1.2 above and the estimation problem there stated: given a sample of one value of y, to estimate the corresponding x. Typically, a practicing statistician facing this kind of problem will find himself confronted with a given joint distribution function for the variables, or at least committed to choosing one which he thinks is representative, and his attention is directed toward such questions as the following.

a. By what criterion shall various estimates of x be compared?

b. Given the criterion, what is the best estimate of x which can be made, and how good is it?

c. How do competing methods of estimating x compare with the best?

These questions of course appear in a communication context, too. The entire effort of [1] is concentrated in this general area. It often happens, however, that the communication engineer has a freedom that the statistician seldom has, that of controlling, at least in part, the joint distribution with which he must deal. How this comes about will be discussed in a moment. We can see at once, however, that his interest in question (b) above will then extend to asking, in addition, how he can optimize his best estimate over the additional freedom he has.

1.4. *The additional freedom.* The additional freedom enjoyed by a communication engineer is like the freedom granted the designer of an experiment. Typically, technology provides the engineer with a communicating device or medium; a random variable y whose range Y represents, as above, the events which can take place at the receiving point, and a probability distribution for y which depends upon a parameter θ. As above, the range Θ of θ represents the possible events at the transmitting point. In addition, one is given a quite separate random variable x whose range X is the universe of possible messages with a probability measure appropriate thereto.

No relation is yet specified between the message x and the "stimulus" θ which is applied to the communication medium, and it is here that the extra freedom lies. Subject to limitations set by the necessary distinction between past and future, one is free to choose a mapping function $f(x)$ from X into Θ, $\theta = f(x)$. This corresponds to choosing some kind of encoding or modulation scheme transforming the original message into a form suitable for transmission.

To illustrate the effect of this, suppose that the distribution function of y has a density $\rho(\theta; y)$ with respect to some fixed underlying measure ν in the y universe, and that the distribution of x has a density $\sigma(x)$ with respect to some underlying measure μ in the x universe. Then if one fixes the relation above between x and θ, the function $\sigma(x)\rho(f(x); y)$ in $X \otimes Y$ represents the density of the resulting joint distribution of x and y relative to the product measure $\mu \otimes \nu$. It is this joint distribution with which the communication engineer works.

To the practising engineer, the most interesting theorems of Shannon's paper relate to what can be achieved by varying the encoding process repre-

sented by the function $f(x)$. The strong theorems now known are all of an asymptotic kind.

1.5. *Role of Fourier analysis.* Even the casual reader will observe in [1], and in the latter part of [2], a preoccupation with Fourier analysis. It may be well to point out that this is a kind of accident; it happens that most practical communication media are governed by linear time-invariant differential equations. Hence, the first applications of information theory have been to systems which are naturally best handled by the tools of Fourier or Laplace analysis.

2. Terminology and concepts.

2.0. *Limitation to discrete model.* We shall confine our attention to the first part of Shannon's paper [2]. This whole paper relates to the second model of the communication problem described above, with an emphasis on the kind of question discussed in 1.3. The first part of that paper is based on a fairly specific kind of model. The stochastic processes which it admits are all derived from Markov processes having finitely many states. The auxiliary devices, encoders, etc., which are admitted are defined by similar constructions. We adopt the term "finitary" to denote a restriction to these classes of objects without at this point repeating Shannon's definitions in detail. (There is a restriction, tacit in [2] but nowhere made explicitly, to devices whose graphs have the property that the terminal state of any transition is uniquely fixed when the initial state and the letter emitted are given. For the present, we take "finitary" to include this limitation.)

The central concepts of [2] may be introduced well enough here by a glossary of terms. At this purely descriptive level, we may be quite general and admit things which are not finitary.

2.1. *Sample space and measurable sets.* Let A be a finite set. We call such a set an alphabet and will have occasion to introduce further alphabets A_1, B, etc. These are all abstract finite sets. An element of A will be called a letter of A, or simply a letter when no ambiguity results.

Let I denote the set of integers: $I = (\cdots, -1, 0, 1, 2, \cdots)$.

Given an alphabet A, denote by A^I the class of infinite sequences

$$x = (\cdots, x_{-1}, x_0, x_1, x_2, \cdots)$$

where each $x_t \varepsilon A$, $t \varepsilon I$. Here x is an element of A^I, and we call x_t the letter of x at time t.

A basic set (in A^I) is a subset of A^I obtained by specifying

(i) an integer $n \geqq 1$,

(ii) a finite sequence $\alpha_0, \alpha_1, \cdots, \alpha_{n-1}$ of letters $\alpha_k \varepsilon A$.

(iii) an integer t, $-\infty < t < \infty$.

The basic set resulting from this specification consists of all sequences $x \varepsilon A^I$ such that

$$x_{t+k} = \alpha_k, \qquad 0 \leqq k \leqq n-1.$$

Let F_A be the Borel field of subsets of A^I determined by the basic sets.

2.2. *Glossary.* Our glossary now reads:

2.21. *Information source.* If μ is a probability measure defined over the Borel field F_A, the ensemble or stochastic process $[A^I, F_A, \mu]$ is an information source. Since the space A^I is fixed by the alphabet, and the Borel field F_A is always that determined by the basic sets, we can specify a source by the pair of symbols $[A, \mu]$.

2.22. *Stationary and ergodic sources.* Consider a source $[A, \mu]$. Let T be the coordinate-shift transformation defined as follows. If $x = (\cdots, x_{-1}, x_0, x_1, \cdots)$ then $Tx = (\cdots, x'_{-1}, x'_0, x'_1, \cdots)$, where $x'_t = x_{t+1}$, $t \varepsilon I$. Then T preserves membership in F_A (measurability). The source will be called stationary if (i) below holds, and ergodic if (i) and (ii) both hold.

(i) If $S \varepsilon F_A$, then $\mu(S) = \mu(TS)$.

(ii) If $S = TS$, then either $\mu(S) = 0$ or $\mu(S) = 1$.

2.23. *Transducer.* A transducer is characterized by two alphabets, A and B, and a function τ from A^I to B^I: given $x \varepsilon A^I$, $\tau(x) \varepsilon B^I$. A transducer differs from a general functional relationship in that it cannot anticipate.

If $x^{(1)} \varepsilon A^I$ and $x^{(2)} \varepsilon A^I$ and t_0 is an integer such that

$$x_t^{(1)} = x_t^{(2)} \qquad \text{for } t \leqq t_0,$$

then

$$y_t^{(1)} = y_t^{(2)} \qquad \text{for } t \leqq t_0,$$

where

$$y^{(i)} = \tau(x^{(i)}), \qquad i = 1, 2.$$

We can specify a transducer by the symbol $[A, \tau, B]$.

2.24. *Channel or communication channel.* A channel is characterized by two alphabets A and B, and a list of probability measures ν_θ defined over F_B, one for each $\theta \varepsilon A^I$. Here we have used θ to denote the "parameter" in conformance with an earlier notation.

Like a transducer, a channel cannot anticipate. That is, informally, if

$$\theta_t^{(1)} = \theta_t^{(2)} \qquad \text{for } t \leqq t_0, \tag{1}$$

we must have

$$\nu_1(S) = \nu_2(S), \tag{2}$$

where $\nu_i(S)$ denotes the value of $\nu_\theta(S)$ when $\theta = \theta^{(i)}$, for any set $S \varepsilon F_B$ which depends only on letters occuring before $t_0 + 1$. More precisely stated, (1) must imply (2) for any set $S \varepsilon F_B$ such that "$y_t^{(1)} = y_t^{(2)}$ for $t \leqq t_0$, and $y^{(1)} \varepsilon S$" implies "$y^{(2)} \varepsilon S$."

A transducer is a special case of a channel; it is a channel in which the received signal y is determined exactly by the transmitted signal θ.

We can specify a channel by the symbol $[A, \nu_\theta, B]$.

2.25. *Stationarity.* The concept of stationarity extends to channels and transducers. It suffices to define a stationary channel, a stationary transducer is a special case. Referring to the definition of a channel, this channel will be called stationary if, for any $S \varepsilon F_B$, $\nu_\theta(S) = \nu_{T\theta}(TS)$, where T is the coordinate-shift transformation.

2.26. We have so worded the definitions above that all sources are "letter generators" producing one new letter for each unit of time, and channels and transducers accept and produce one letter for each unit of time. In a careful setting of the theory, one must account for the phenomena of compression and expansion which appear when languages are translated. For example, a long business message of fairly stereotyped form, when encoded for transmission by cable, may appear in a form having many fewer letters or words than the original. There are several ways of accommodating the mathematics to this situation, but these details are unimportant in a first look at the subject and will be ignored from here on. The fact of so ignoring them does not invalidate any theorem that will be stated. It merely leaves a gap between these theorems and certain useful interpretations of them.

3. Entropy.

3.0. *Entropy.* The terms defined in Section 2, suitably hedged, are the concepts with which [2] deals. (For purposes of exposition, we have defined channels and transducers quite differently from [2]. The disparity is largely but not entirely verbal. (Cf. 10.2.)) The principal tool for their quantitative study is the concept of entropy.

Let $p_1, p_2, \cdots, p_n$ be a finite and exhaustive list of probabilities: $p_i \geqq 0$, $1 \leqq i \leqq n$, $p_1 + p_2 + \cdots + p_n = 1$. The entropy of this list is defined to be

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i = \text{Expectation } (-\log p).$$

It is by now traditional to use logs to the base 2 in this definition, but the choice of base affects the value of H only by a constant factor. We shall use the base 2.

3.1. *Marginal entropies.* To change the notation slightly, suppose that α and β run over finite index sets (alphabets) A and B, and that $p(\alpha, \beta)$ is the probability of the joint event (α, β). That is $p(\alpha, \beta) \geqq 0$, $\sum_{\alpha \varepsilon A} \sum_{\beta \varepsilon B} p(\alpha, \beta) = 1$. The entropy of this list of probabilities is denoted by $H(\alpha, \beta)$:

$$H(\alpha, \beta) = -\sum_{\alpha} \sum_{\beta} p(\alpha, \beta) \log p(\alpha, \beta).$$

We can define also two marginal entropies

$$H(\alpha) = -\sum_{\alpha} \sum_{\beta_1} p(\alpha, \beta_1) \log \left(\sum_{\beta} p(\alpha, \beta)\right),$$

$$H(\beta) = -\sum_{\beta} \sum_{\alpha_1} p(\alpha_1, \beta) \log \left(\sum_{\alpha} p(\alpha, \beta)\right),$$

and two average conditional entropies

$$H_\beta(\alpha) = H(\alpha, \beta) - H(\beta),$$
$$H_\alpha(\beta) = H(\alpha, \beta) - H(\alpha).$$

3.2. *Average conditional entropy.* These latter are called average conditional entropies because of the following formula: fix β and consider the conditional probabilities for the various $\alpha \varepsilon A$. These are $q_\beta(\alpha) = p(\alpha, \beta)/\sum_{\alpha_1} p(\alpha_1, \beta)$. The entropy of this list is

$$-\sum_\alpha q_\beta(\alpha) \log q_\beta(\alpha) = \frac{1}{r(\beta)} \sum_\alpha p(\alpha, \beta) \log p(\alpha, \beta) + \frac{1}{r(\beta)} \sum_\alpha p(\alpha, \beta) \log \sum_{\alpha_2} p(\alpha_2, \beta) \tag{1}$$

where $r(\beta)$ is defined by (2) below.

This expression is the entropy of the conditional distribution of α when it is known that a particular β has occurred. The a priori probability of this β is

$$\sum_{\alpha_1} p(\alpha_1, \beta) = r(\beta). \tag{2}$$

To average (1) over all β, we multiply it by (2) and sum over β. The result is seen to be $H_\beta(\alpha)$. This last entropy, then, is the average over all β of the entropies of the conditional distribution of α when β is known.

3.3. *Properties.* Shannon [2] gives a fairly complete heuristic justification for regarding the entropy of a list of probabilities as a measure of one's a priori uncertainty as to which of the possible events will actually occur in a given trial. In the course of this demonstration, he introduces the most important mathematical properties of the H function. These are (i) its positivity, (ii) a kind of convexity property implied by the convexity of the function $-x \log x$, (iii) that composition law which permitted the identification above of the average value of (1) over β, with the earlier defined $H_\beta(\alpha)$, and (iv) $H = 0$ if and only if there is exactly one event of nonzero probability.

The convexity property (ii) mentioned above leads to the general inequality $H_\beta(\alpha) \leqq H(\alpha)$; that is, verbally, a condition (i.e. an a priori restriction on the "freedom of choice") never increases an entropy. This statement must however be taken only in the average sense in which it is stated: for any particular β, the entropy of the conditional distribution of α bears no provable relation to the marginal entropy $H(\alpha)$. It is only in the average over-all β that an inequality obtains.

4. The entropy rate of a source.

4.0. *Definition.* So far we have considered the entropy of a list of probabilities. The entropy rate of a stationary source $[A, \mu]$ is most easily defined as follows. Given $x \varepsilon A^I$, we use either of the bracket notations

(1) $$[x_t, x_{t+1}, \cdots, x_{t+n-1}], [t, t+n-1; x]$$

to denote that basic set $S \subseteq A^I$ which consists of all x' such that

$$x'_{t+h} = x_{t+h}, \qquad 0 \leqq h \leqq n-1.$$

The second notation will be used when it is to be emphasized that the basic set depends upon a particular infinite sequence x.

The possible basic sets (1), as x ranges over A^I, or, alternatively, as the x_{t+h}, $0 \leqq h \leqq n-1$, range independently over A, partition A^I into a^n measurable subsets, where a is the number of letters in the alphabet A. These subsets represent all the possible sequences of n consecutive letters. They have the respective probabilities

(2) $$\mu([x_t, x_{t+1}, \cdots, x_{t+n-1}]).$$

Our stationarity assumption makes this list of probabilities independent of t. There is then a unique number F_n, independent of t, which is the entropy of this list (2) of probabilities. We shall show presently that the limit

(3) $$\lim_{n\to\infty} \frac{1}{n} F_n$$

always exists. The value of this limit is defined to be the entropy rate of the source $[A, \mu]$.

4.1. *Interpretation.* One cannot escape the heuristic meaning of this rate; one considers the possible long sequences of text as his universe of events, and evaluates the uncertainty F_n of the outcome of a trial. This uncertainty is then prorated among the n letters. These letters represent interdependent but possibly not determinately related elementary events whose concatenation generates the universe. The result, F_n/n, represents in the limit the average uncertainty per letter generated by the source.

4.2. *Defining F_n as an integral.* We shall now prove the existence of the limit (3). The proof follows Shannon's in a different notation.

Given any $x \in A^I$, the basic set (1) defined by that x contains x. The probability (2) then may be regarded as a step function of x, equal for each x to the probability of that basic set containing x which is specified by letter values at times $t, t+1, \cdots, t+n-1$. In the same way, the definition

(4) $$f_n(x) = -\frac{1}{n} \log \mu([0, n-1; x])$$

defines a nonnegative step function of x. One verifies at once from the definition of F_n that

(5) $$\frac{1}{n} F_n = \int_{A^I} f_n(x)\, d\mu(x).$$

Regarding (2) and (4) as functions of x in this way permits us to phrase certain key problems in the language of integration theory.

4.3. *Another definition of* H. Consider now the special conditional probabilities

$$p_n(x) = \frac{\mu([x_{-n}, x_{-n+1}, \cdots, x_{-1}, x_0])}{\mu([x_{-n}, \cdots, x_{-1}])}, \qquad n \geqq 1. \tag{6}$$

Again we use the device of representing these as step functions of x. In words, $p_n(x)$ is the conditional probability of observing at time zero the letter x_0 of x, when it is known that the letters occurring at times $t = -n, -n+1, \cdots, -1$ are exactly those of x.

Define

$$\begin{aligned} g_0(x) &= f_1(x) \\ g_n(x) &= -\log p_n(x), \qquad n \geqq 1. \end{aligned} \tag{7}$$

Then $g_n(x) \geqq 0$.

One verifies by direct calculation from (6) and (7) that

$$G_n = \int_{\Delta^I} g_n(x)\, d\mu(x) \tag{8}$$

is the average conditional entropy of the next letter when n preceding letters are known. The inequality stated earlier, that adjoining a condition cannot increase an entropy, can be used to show that the G_n form a monotone sequence:

$$G_0 \geqq G_1 \geqq G_2 \geqq \cdots \geqq 0.$$

Therefore

$$\lim_{n\to\infty} G_n = H \tag{9}$$

certainly exists. The verbal interpretation of G_n, the average conditional entropy of the next letter after a long segment of text is already known, suggests that the limit H in (9) is again the average uncertainty per letter generated by the source, that is, H is the entropy rate defined in (3). The proof in 4.4 below that this is indeed so, proves the existence of the limit (3).

4.4. *Identification of two definitions.* By a direct calculation from the definitions it is found that

$$f_N(x) = \frac{1}{N} \sum_{k=0}^{N-1} g_k(T^k x). \tag{10}$$

If one integrates this and uses the assumed stationarity of μ, he obtains

$$\frac{1}{N} F_N = \frac{1}{N} (G_0 + G_1 + \cdots + G_N). \tag{11}$$

Therefore F_N/N represents the first Cesaro mean of a monotonely convergent sequence. It follows that the limit (3) exists and indeed is approached monotonely. A further consequence of (11) is that $F_N/N \geqq G_N \geqq H$.

5. The capacity of a channel.

5.0. *Channel and source.* We wish now to examine a stationary channel "driven" by a stationary source. Consider a source $[A, \mu]$, and a channel $[A, \nu_\theta, B]$. Denote by $C = A \otimes B$ the alphabet of pairs (α, β), $\alpha \in A$, $\beta \in B$. Then C^I is the class of all infinite sequences

$$(\cdots, (x_{-1}, y_{-1}), (x_0, y_0), (x_1, y_1), \cdots)$$

where $x_t \in A$, $y_t \in B$, $t \in I$. In an obvious way we can also regard $C^I = A^I \otimes B^I$, that is, as the class of paired sequences (x, y), $x \in A^I$, $y \in B^I$. It is known that the Borel field F_C is determined by the sets $X \otimes Y$ where $X \in F_A$, $Y \in F_B$. We define a measure ω for sets in F_C by the formula

$$\int_{C^I} h(x, y)\, d\omega(x, y) = \int_{A^I} d\mu(x) \int_{B^I} h(x, y)\, d\nu_x(y)$$

valid for all positive measurable $h(x, y)$. Here ν_x is the measure over F_B which is induced by the channel when the input sequence is $x \in A^I$.

The stochastic process $[C^I, F_C, \omega]$ is now a source, which we denote by $[C, \omega]$.

It is easily shown that if the original source $[A, \mu]$ and the channel are stationary, then the source $[C, \omega]$ is stationary.

5.1. *Marginal distributions.* The source $[C, \omega]$ represents the joint distribution of x and y, of input to and output from the channel. The source $[A, \mu]$ represents the marginal distribution of the input. The marginal distribution of the output is represented by the source $[B, \eta]$, where the measure η over F_B is defined by

$$\int_{B^I} k(y)\, d\eta(y) = \int_{A^I} d\mu(x) \int_{B^I} k(y)\, d\nu_x(y).$$

This marginal source is staionary if $[A, \mu]$ and $[A, \nu_\theta, B]$ are.

5.2. *Causation.* It is worth noting that the implication of causation in our language here, as we speak of a channel driven by a source, results from the fact that we consider the channel $[A, \nu_\theta, B]$ as a pregiven thing, existing independently of any particular source $[A, \mu]$; this is the typical situation in the communications art. Actually, the joint process $[C, \omega]$ is a completely symmetrical concept, as to the roles of x and y, and one may consider, at will, the conditional probabilities $\nu_x(S)$, $x \in A^I$, $S \in F_B$, the conditional probabilities of y-events, knowing x, or the conditional probabilities, say, $\bar{\mu}_y(U)$, $y \in B^I$, $U \in F_A$, of x-events, knowing y. (Indeed, given the joint process, one will find that each of these conditional probabilities ν_x, respectively $\bar{\mu}_y$, are measures for, respectively, almost all $x(\mu)$, almost all $y(\eta)$.)

It happens that in most applications the ν_x are pregiven, and the $\bar{\mu}_y$ derivative.

5.3. *Channel capacity.* To use Shannon's notation, let $H(x, y)$ denote the entropy rate of the source $[C, \omega]$, $H(x)$ the entropy rate of the marginal source $[A, \mu]$, and $H(y)$ that of the marginal source $[B, \eta]$. The quantity $R = H(x) + H(y) - H(x, y)$ is defined to be the transmission rate achieved by the source

$[A, \mu]$ over the channel $[A, \nu_\theta, B]$. The supremum or least upper bound of these rates, as μ is allowed to vary, is defined to be the capacity of that channel.

5.4. *Interpretation.* An intuitive interpretation of the rate $H(x) + H(y) - H(x, y)$ can be obtained if we assume that the quantity $H_x(y) = H(x, y) - H(x)$ can be given the same verbal interpretation when the x and y are stochastic processes that it was given earlier when the random quantities involved were drawn from finite populations. That it can, in the same limiting sense that the entropy concept has been carried over to stochastic processes, is easy to show. Foregoing this demonstration, we observe that $R = H(y) - H_x(y)$; that is, the rate of transmission R is the marginal rate of the output, $H(y)$, diminished by that amount of uncertainty at the output which arises from the average uncertainty of y even when x is known, that is, by $H_x(y)$, the average conditional entropy of y when x is known. In this verbal way, at least, R represents that portion of the "randomness" or average uncertainty of each output letter which is not assignable to the randomness created by the channel itself.

Another observation here is also pertinent. Because of the symmetry of R in x and y (which is more than a mere consequence of the notation!) we also have $R = H(x) - H_y(x)$. This shows R as the rate of the original source diminished by the average uncertainty as to the input x when the output y is known.

6. The fundamental theorem.

6.0. *As justifying the theory.* So far, we have introduced a list of what is hoped are natural-seeming concepts, and have stated a few mathematical results to help justify the rather picturesque language used in introducing them. The concepts themselves can only be justified as objects worthy of mathematical attention by the existence of theorems relating them. There is one such theorem, the so-called fundamental theorem for a noisy channel ([2], Theorem 11), which in itself performs this task completely. We shall quote this theorem and sketch its proof. This will complete our general exposition and lead us to our general limit theorem.

6.1. The Theorem. The fundamental theorem relates to this question. Suppose we are given a stationary channel with input alphabet A, and a stationary ergodic source with alphabet A_1. We are permitted to insert a stationary transducer $[A_1, \tau, A]$ between the source and channel, to create in effect, a new stationary channel with input alphabet A_1. With this freedom, what is the optimum transmission rate which can be achieved between source and output?

For the class of finitary sources, channels, and transducers, admitted in the model used in [2], this question is answered by Shannon's theorem: Let the given channel have capacity C and the given source have rate H. Then if $H < C$, for any $\epsilon > 0$ there exists a transducer such that a rate $R > H - \epsilon$ can be achieved. If $H \geqq C$, there exists similarly a transducer such that $C \geqq R > C - \epsilon$. No rate greater than C can be achieved.

Actually, Shannon's proof of this theorem proves the following more complete result.

THEOREM. *Let the given channel have capacity C and the given source have rate H. If $H < C$, then, given any $\epsilon > 0$, there exists an integer $n(\epsilon)$ and a transducer (depending on ϵ) such that when $n(\epsilon)$ consecutive received letters are known, the corresponding n transmitted letters can be identified correctly with probability at least $1 - \epsilon$. If $H > C$ no such transducer exists.*

This statement is perhaps more satisfying to a statistician, in that the logarithmic quantities H and C appear only in the hypotheses. The conclusion is then given in terms of the criterion of performance suggested in 1.2.

6.2. *Interpretation.* In the vernacular, this theorem asserts that if a channel has adequate capacity C, an infinitesimal margin being mathematically adequate, then virtually perfect transmission of the material from the source can be achieved, but not otherwise. Here, of course, we have used "virtually perfect" to describe transmission at a rate

$$R = H - \epsilon_1 \geqq H - \epsilon. \tag{1}$$

The sense in which this is to be interpreted as virtually perfect transmission is, of course, an asymptotic one and refers to the rate at which certain probabilities decay as the amount of available received text increases.

Engineering experience has been that the presence in the channel of perturbations, noise, in the engineer's language, always degrades the exactitude of transmission. Our verbal interpretation above leads us to expect that this need not always be the case; that perfect transmission can sometimes be achieved in spite of noise. This practical conclusion runs so counter to naive experience that it has been publicly challenged on occasion. What is overlooked by the challengers is, of course, that "perfect transmission" is here defined quantitatively in terms of the capabilities of the channel or medium, perfection can be possible only when transmission proceeds at a slow enough rate. When it is pointed out that merely by repeating each message sufficiently often one can achieve virtually perfect transmission at a very slow rate, the challenger usually withdraws. In doing so, however, he is again misled, for in most cases the device of repeating messages for accuracy does not by any means exploit the actual capacity of the channel.

Historically, engineers have always faced the problem of *bulk* in their messages, that is, the problem of transmitting rapidly or efficiently in order to make a given facility as useful as possible. The problem of noise has also plagued them, and in many contexts it was realized that some kind of exchange was possible, for example, noise could be eliminated by slower or less "efficient" transmission. Shannon's theorem has given a general and precise statement of the asymptotic manner in which this exchange takes place.

The statistician will recognize the exchange between bulk and noise as akin to the more or less general exchange between sample size and validity or significance.

7. The asymptotic equipartition property.

7.0. *A Basic Lemma.* The theorem quoted in 6.1 is termed fundamental in

[2] because it answers a question which is clearly fundamental in the communications art, and because it defines the applicability of the central concept of channel capacity. Many of the later results in [2] then concern the calculation of capacities for practical or interesting channels.

The proof of this fundamental theorem rests directly on a lemma (Theorem 3 of [2]) which itself is a basic limit theorem in the theory of stochastic processes. As a mathematical theorem, this lemma requires very little of the specialized imagery of communication theory for its understanding. A mathematician, therefore, is likely to regard it as the more fundamental element. A generalization of it is the one contribution of the present paper.

7.1. *Shannon's form.* The basic limit theorem, as given in Theorem 3 of [2], asserts that the text from an ergodic finitary source possesses what we shall call an asymptotic equipartition property. The basic sets

$$[x_0, x_1, \cdots, x_{n-1}], \tag{1}$$

as x ranges over A^I describe a partition of A^I, as we noted earlier: a partition into a^n events, each one of which is the occurrence of a particular string of n letters. Shannon's Theorem 3 asserts that, if H is the rate of a finitary ergodic source, then, given $\epsilon > 0$ and $\delta > 0$, there exists an $n_0(\epsilon, \delta)$ such that, given any $n \geqq n_0$, the basic sets (1) above can be divided into two classes:

(i) a class whose union has μ-measure less than ϵ,

(ii) a class each member E of which has a measure $\mu(E)$ such that $| H + 1/n \log \mu(E) | < \delta$.

That is, this theorem asserts the possibility of dividing the long segments of text from a finitary source into a class of roughly equally probable segments plus a residual class of small total probability.

7.2. *Stronger form.* Let us introduce here the step functions $f_n(x)$ defined in (4) of Section 4: $f_n(x) = -1/n \log \mu([0, n-1; x])$. In terms of these, the possibility of dividing the long segments of text into the categories (i) and (ii) above is easily seen to be equivalent to the assertion that the sequence $(f_n(x))$ converges in probability to the constant H.

We shall say that a source $[A, \mu]$ has the asymptotic equipartition property, AEP, if the sequence $(f_n(x))$ converges in probability to a constant.

Shannon's Theorem 3 then asserts that a finitary ergodic source has the AEP. We shall improve this in Section 9 to read as follows.

THEOREM. *For any source $[A, \mu]$, the sequence $(f_n(x))$ converges in L^1 mean (μ). If $[A, \mu]$ is ergodic, and has rate H, this sequence converges in L^1 mean to the constant H.*

Since L^1 convergence here implies convergence in probability, (a fact easily proved,) we have the

COROLLARY. *Every ergodic source has the AEP.*

These are the limit theorems mentioned in the Summary. As we shall see in Section 8, they permit extending Shannon's fundamental theorem, 6.1, to other than finitary sources.

7.3. *Interpretation.* Returning to 7.1 and the description there of the AEP, we see that most of the probability must be accounted for by the aggregate (ii) of "likely" long sequences. That is, if the source has the AEP, there are, for large enough n, 2^{nH} likely basic sets $[x_0, \cdots, x_{n-1}]$, roughly equally probable, accounting in the aggregate for all but a small fraction of the total probability.

7.4. *Another corollary.* The proof in [2] of the fundamental theorem uses also a consequence of the AE property. We examine this consequence briefly.

Consider a stationary source $[A, \mu]$ and a stationary channel $[A, \nu_\theta, B]$. Suppose that both $[A, \mu]$, and the joint process $[C, \omega]$ which results when this source drives the channel, have the AEP. We can write

$$(2) \quad \begin{aligned} &-\frac{1}{n}\log \omega([0, n-1; x] \otimes [0, n-1; y]) \\ &\quad = -\frac{1}{n}\log \frac{\omega([0, n-1; x) \otimes [0, n-1; y)]}{\mu([0, n-1; x])} - \frac{1}{n}\log \mu([0, n-1; x]). \end{aligned}$$

Our hypothesis that the joint process has the AEP now implies that the left member of this equation converges in measure to a constant, namely the entropy rate of the joint process, $H(x, y)$. (Here the notation is misleading. In $H(x, y)$, the x and y are labels merely. Equation (2) involves x and y as specific variables.) Also by hypothesis the second term on the right converges in probability to $H(x)$, the entropy rate of $[A, \mu]$. It follows then that the first term on the right converges in probability also to a constant, which constant must then be $H_x(y)$, by 5.4.

Now the first term on the right of (2) is

$$(3) \qquad -\frac{1}{n}\log \bar{\mu}_{n,x}([0, n-1; y]),$$

where the argument of the logarithm is the conditional probability of $[0, n-1; y]$ knowing that $[0, n-1; x]$ has occurred. We have therefore proved the following.

COROLLARY. *If* $[A, \mu]$ *and* $[C, \omega]$ *have the AEP, then the functions* (3) *converge in probability to a constant.*

8. Proof of the fundamental theorem.

8.0. *Introduction.* For simplicity, we do not examine the question of ergodicity and consider only the most interesting of the cases cited in the statement of the theorem (6.1), that in which we are given a finitary source $[A_1, \mu]$ of entropy rate H and a finitary channel $[A, \nu_\theta, B]$ of capacity $C > H$, both stationary. Our problem is then, given $\epsilon > 0$, to exhibit an $n(\epsilon)$ and a finitary transducer $[A_1, \tau, A]$ such that, when the given source drives the channel through this transducer it is possible at the receiver, given $n(\epsilon)$ consecutive received letters, to identify the corresponding $n(\epsilon)$ transmitted letters correctly with a probability exceeding $1 - \epsilon$. Here the probability is not conditional (i.e., not given the received letters) but in the universe of joint events at transmitter and receiver.

We shall review Shannon's argument. He does not supply detailed epsilontics here, and we shall not either. Generally, the manner in which they could be supplied is evident enough, though at one point we must consider a detail. (My efforts to make them simple have so far failed, however.)

8.1. *The "likely" events.* The channel $[A, \nu_\theta, B]$ has capacity $C = H + 2\gamma$, say, where $\gamma > 0$. There, therefore, exists a source, say $[A, \mu^*]$, which achieves over this channel a rate

$$R^* \geqq C - \gamma = H + \gamma. \tag{1}$$

We will use asterisks to denote quantities referring to this source. Let H^* be the entropy rate of $[A, \mu^*]$, and let $[C, \omega^*]$ denote the joint process of input (x) and output (y) when $[A, \mu^*]$ drives the channel. For a simpler notation let $K^* = H_y^*(x)$, the average conditional entropy of input to $[C, \omega^*]$ when output is known. Then by definition (5.3)

$$R^* = H^* - K^*. \tag{2}$$

We now invoke the AEP for the processes $[A_1, \mu]$, $[A, \mu^*]$, and $[C, \omega^*]$. For large n there are roughly 2^{nH} equally likely basic sets $[0, n-1; w]$ from $[A_1, \mu]$, call these the likely outputs of $[A_1, \mu]$. Similarly there are roughly 2^{nH^*} equally likely basic sets $[0, n-1; x]$ from $[A, \mu^*]$, the likely outputs of $[A, \mu^*]$. Furthermore, consider the possible basic sets $[0, n-1; y]$ at the output of the channel. With the exception of an aggregate of these of small total probability in $[C, \omega^*]$, the conditional probabilities in $[C, \omega^*]$ of the $[0, n-1; x]$, knowing $[0, n-1; y]$, are such that roughly, there are 2^{nK^*} equally likely $[0, n-1; x]$ for each $[0, n-1; y]$, call these the likely antecedents to $[0, n-1; y]$.

In each of these definitions the "likely" objects in sum exhaust most of the probability. In particular, the likely antecedents of $[0, n-1; y]$ exhaust most of the a posteriori probability in $[C, \omega^*]$ of the basic sets $[0, n-1; x]$ when $[0, n-1; y]$ is known. Let us use the word "package" to mean "the aggregate of likely antecedents to a given $[0, n-1; y]$."

8.2. *Marked basic sets.* The nub of Shannon's proof lies in the fact that the packages are so small that it is easy to find 2^{nH} of them which are disjoint. Indeed, suppose one designates, "marks," 2^{nH} of the likely basic sets $[0, n-1; x]$ from $[A, \mu^*]$, doing so at random. Then the probability that a particular $[0, n-1; x]$ be marked in this process is $2^{n(H-H^*)}$. Consider the 2^{nK^*} likely antecedents of some $[0, n-1; y]$. The conditional probability that two or more of these get marked, knowing that one of them is marked, is of the order of

$$2^{nK^*} \cdot 2^{n(H-H^*)} = 2^{n(H-R^*)} \leqq 2^{-n\gamma},$$

by (1) and (2). This probability may be made small by choosing a large n.

8.3. *Distinguishability a posteriori of marked inputs.* Conceptually, we now have this situation: some 2^{nH} basic sets $[0, n-1; x]$ have been specially marked. Given a $[0, n-1; y]$, the received message, and knowing in addition that a marked basic set $[0, n-1; x]$ has been transmitted (has occurred) there is but

a small conditional probability, in the joint universe of $[C, \omega^*]$ and of random markings, that either of the following events has occurred.

(i) The actual $[0, n-1; x]$ which occurred is not a likely antecedent of $[0, n-1; y]$;

(ii) The actual $[0, n-1; x]$ which occurred is a likely antecedent of $[0, n-1; y]$, but there are other marked $[0, n-1; x]$ in the same package.

There is now virtual certainty in the joint universe of $[C, \omega^*]$ and random markings that the actual $[0, n-1; x]$ is a unique marked likely antecedent of $[0, n-1; y]$ when we know $[0, n-1; y]$ a priori, and that a marked $[0, n-1; x]$ is transmitted. That is, by making a marking at random, one is almost certain to have chosen a limited vocabulary of 2^{nH} basic sets $[0, n-1; x]$ which are almost certain to be distinguishable a posteriori, knowing $[0, n-1; y]$.

8.4. *The transducer.* The next step is deceptively simple. One shows easily that, given a marking, a finitary transducer can be described which maps the 2^{nH} likely $[0, n-1; w]$ from $[A_1, \mu]$ on to the marked $[0, n-1; x]$ from $[A, \mu^*]$. When one drives this transducer from $[A_1, \mu]$, the likely output basic sets $[0, n-1; x]$ are just those which were marked. Therefore, when one operates the channel from $[A_1, \mu]$ through this transducer he has essentially only the vocabulary of marked basic sets appearing at the input to the channel. Let us call the resulting joint process of input x to, and output y from, the channel the source $[C, \omega]$. This source itself depends on the marking.

If the probabilities sketched in 8.3 can be relied on for this new situation, it is evident that we have described a transducer, depending on a random marking, which, when $[0, n-1; y]$ is given, permits the correct identification of the $[0, n-1; x]$ which occurred in all but a set of cases of small probability (a posteriori, knowing $[0, n-1; y]$) in the joint universe of random markings and events in $[C, \omega]$. We can assume that for all likely $[0, n-1; x]$ the input $[0, n-1; w]$ which produced it is unique. Then the average, over the joint universe of markings and events in $[C, \omega]$, of the probability that the actual $[0, n-1; w]$ which occurred is not the one determined by this procedure is small. By the Tchebycheff inequality, then, all but a small fraction of the markings will describe transducers which make the probability of misidentifying the actual $[0, n-1; w]$ *simultaneously* small for all but a small fraction of the $[0, n-1; y]$.

8.5. *Critique.* This argument shows that it is somehow easy to describe a transducer which will make the probability of error small. There is, however, a gap in the argument. The probabilities calculated in 8.3 were based on $[C, \omega^*]$. In 8.4 we used these as though they applied to any $[C, \omega]$ which might arise when a marking had been made. If they are both ergodic, and this we are tacitly assuming, ω and ω^* are either identical or else each assigns unit probability to a null set of the other. (This is almost trivial to prove. To my knowledge it was first explicity noted by G. W. Brown.) The probabilities in 8.3 are based on relations which hold only almost everywhere in $[C, \omega^*]$, and therefore, possibly, at most on a null set in $[C, \omega]$. This point is not touched on in [2].

In 10.1 we shall show that finitary channels have a kind of continuity which

permits passage from $[C, \omega^*]$ to $[C, \omega]$, when n is large enough, without serious modification of the probabilities. Shannon's argument is then valid, though incomplete, for finitary channels. Indeed, it is valid for any channel having this kind of continuity, but I have not yet found a satisfying formulation of the property or isolation of the class.*

9. The limit theorem.

9.0. *Introduction* This section is devoted principally to the proof of the theorem quoted in 7.2, which has as a corollary that every ergodic source has the AEP. We recall the definitions of 2.1 and 2.2, and use the following notation.

Given any fixed $x \in A^I$, The symbols $[t, t+n-1; x]$, $[x_t, \cdots, x_{t+n-1}]$ denote that basic set which consists of all $x' \in A^I$ such that $x'_{t+k} = x_{t+k}$, $k = 0, 1, \cdots, n-1$.

Given a source $[A, \mu]$, the symbols $\int f(x)\, d\mu(x)$, $\int f\, d\mu$, denote integration over the space A^I. Integration over a measurable subset $S \subseteq A^I$ is denoted by one of $\int_S f(x)\, d\mu(x)$, $\int_S f\, d\mu$.

Following [3], we append "(μ)" to a statement which holds almost everywhere with respect to μ, or to a statement involving mean convergence relative to μ.

9.1. *The Theorem.* Given the source $[A, \mu]$ we define the following step functions of $x \in A^I$.

$$
\begin{aligned}
p_n(x) &= \frac{\mu([-n, 0; x])}{\mu([-n, -1; x])}, && n \geqq 1, \\
p_0(x &= \mu([x_0]); \\
g_n(x) &= -\log p_n(x), && n \geqq 0; \\
f_n(x) &= -\frac{1}{n} \log \mu(0, n-1; x]), && n \geqq 1.
\end{aligned} \tag{1}
$$

The function $p_n(x)$ is the conditional probability that the letter which occurs at time $t = 0$ is x_0 when it is known that the letters between time $t = -n$ and $t = -1$ are also those of the infinite sequence x. The definitions of $g_n(x)$ and $f_n(x)$ need no comment. They are related by the important and easily verified formula

$$
f_N(x) = \frac{1}{N} \sum_{k=1}^{N-1} g_k(T^k x). \tag{2}
$$

What is now to be proved is the

THEOREM. *For any source $[A, \mu]$, the sequence $(f_n(x))$ converges in L^1 mean (μ). If $[A, \mu]$ is ergodic, the limit of this sequence is almost everywhere constant and equal to H, the information rate of $[A, \mu]$.*

9.2. PROOF. The proof of this theorem requires the following intermediate results.

(i) The sequence $(p_n(x))$ converges almost everywhere (μ).

(ii) Each $g_n(x) \in L^1(\mu)$, and the sequence $(g_n(x))$ converges in L^1 mean (μ).

These will be established in 9.3 and 9.4, respectively. Granted the second of them, the theorem to be proved follows easily, as we now show.

We have $g_n(x) \in L^1$, and $\lim_n \int | g_n - g | d\mu = 0$ for some function $g \in L^1$. Then the mean ergodic theorem (e.g., [4], equation 2.42) implies that $\sum_{k=0}^{N-1} g(T^k x)/N$ converge in L^1 mean to an invariant function $h(x) = h(Tx)$. When μ is ergodic $h(x) = H$, a constant, almost everywhere.

By (2) of 9.1

$$\int | f_N - h | d\mu \leqq \int \left| \frac{1}{N} \sum_{k=0}^{N-1} [g_k(T^k x) - g(T^k x)] \right| d\mu(x) + \int \left| \frac{1}{N} \sum_{k=0}^{N-1} g(T^k x) - h(x) \right| d\mu(x)$$

$$\leqq \frac{1}{N} \sum_{k=0}^{N-1} \int | g_k(x) - g(x) | d\mu(x) + \int \left| \frac{1}{N} \sum_{k=0}^{N-1} g(T^k x) - h(x) \right| d\mu(x).$$

In the second inequality we use the stationarity of μ to obtain the first term. This term represents the first Cesaro mean of a sequence which by hypothesis has zero as a limit, hence it has also the limit zero. The second term also goes to zero as $N \to \infty$, by the mean ergodic theorem. We conclude then that $f_n \to h$ in L^1 mean, and that $f_n \to H$ in L^1 mean when μ is ergodic. We identify this constant H with the entropy rate of $[A, \mu]$ in 9.4.

9.3. *First Lemma.* We now prove that the sequence $(p_n(x))$ converges almost everywhere (μ). For any given set $D \in F_A$ define, in analogy with 9.1, (1)

$$p_n(x, D) = \frac{\mu([-n, -1; x] \cap D)}{\mu([-n, -1; x])}, \qquad n \geqq 1;$$

this is the conditional probability of D knowing $x_{-n}, x_{-n+1}, \cdots, x_{-1}$. It is a result of Doob [5] that such a sequence of conditional probabilities is a martingale (positive and bounded) and converges almost everywhere.

Given $\alpha \in A$, let D_α denote the basic set of all $x \in A^I$ with $x_0 = \alpha$. Given any $x \in A^I$, the value of $p_n(x)$ is one of the numbers $p_n(x, D_\alpha)$ obtained as α ranges over the finite set A. Therefore

$$| p_n(x) - p_m(x) | \leqq \sum_{\alpha \in A} | p_n(x, D_\alpha) - p_m(x, D_\alpha) |, \tag{3}$$

because the left member is, for each x, one of the summands on the right.

Except for x in a certain null set (μ), each term on the right of (3) converges to zero as m and n go to infinity, by the result of Doob quoted above. By (3), then, the sequence $p_n(x)$ converges almost everywhere (μ), say to $p(x)$.

It follows at once that the sequence $g_n(x) = -\log p_n(x)$ converges almost everywhere to $g(x) = -\log p(x)$, if we admit convergence to $+\infty$.

9.4. *Second Lemma.* We must now show that $g_n(x) \varepsilon L^1$ and that the sequence $g_n(x)$ converges in mean (μ). The integrability of the $g_n(x)$ is simple to establish directly, but will follow automatically from stronger results which are needed later. We need a uniform bound for the contribution of the "unbounded part" of g_n to the value of $\int g_n \, d\mu$. We shall therefore show that

$$\int_{A_{n,L}} g_n \, du \leqq O(L2^{-L}) \tag{4}$$

uniformly in n, where $A_{n,L}$ is the set of x's such that $g_n(x) \geqq L$.

Let $E_{n,K}$ be the set of x's where

$$K \leqq g_n(x) < K + 1. \tag{5}$$

Let B denote a typical basic set $[-n, -1; x]$. Given $\alpha \varepsilon A$, let D_α, as before, be the basic set of all x such that $x_0 = \alpha$. By its definition, $g_n(x)$ is constant over each $B\Lambda D_\alpha$, in fact, it has there the value $-\log[\mu(B\Lambda D_\alpha)/\mu(B)]$. Hence $g_n(x)$ is measurable.

Let a be the number of letters in the alphabet A. There are altogether finitely many, namely a^{n+1}, sets $B\Lambda D_\alpha$ covering A^I. Since $g_n(x) \geqq 0$ everywhere, we have

$$A^I = \bigcup_B \bigcup_{K=0}^{\infty} B\Lambda E_{n,K}.$$

For fixed n, K, let D^K range over those D_α such that $B\Lambda D_{n,K} \neq \phi$. Then the step character of $g_n(x)$ implies that $B\Lambda E_{n,K} = \bigcup_{D^K} B\Lambda D^K$. Therefore

$$\int_{B\Lambda E_{n,K}} g_n \, d\mu = \sum_{D^K} \int_{B\Lambda D^K} g_n \, d\mu \tag{6}$$

and, furthermore, over any $B\Lambda D^K$, (5) holds. Therefore $-\log[\mu(B\Lambda D^K)/\mu(B)] \geqq K$, or $\mu(B\Lambda D^K) \leqq 2^{-K}\mu(B)$. From (6), then,

$$\int_{B\Lambda E_{n,K}} g_n \, d\mu < \sum_{D^K} (K+1)2^{-K}\mu(B) < a(K+1)2^{-K}\mu(B).$$

We have then that

$$\int_{E_{n,K}} g_n \, d\mu = \sum_B \int_{B\Lambda E_{n,K}} g_n \, d\mu < (K+1)2^{-K}, \tag{7}$$

since $\sum \mu(B) = 1$. The right member of (7) is the Kth term of a convergent series and is independent of n. Since $A_{n,L} = \bigcup_{K \geqq L} E_{n,K}$, (4) follows at once.

That $g_n \varepsilon L^1$ follows by summing (7) over all $K \geqq 0$. This summation gives a uniform bound, say β, for $\int g_n \, d\mu$. Define $g_n^L(x) = \inf(g_n(x), L)$, $g^L(x) = \inf(g(x),$

L). Then $\lim_n g_n^L(x) = g^L(x), (\mu)$, and this convergence is dominated by the integrable function L. Hence

$$\lim_n \int | g_n^L - g^L | \, d\mu = 0, \tag{8}$$

and $g^L \varepsilon L^1$. Furthermore

$$\int g^L \, d\mu = \lim_n \int g_n^L \, d\mu \leqq \limsup_n \int g_n \, d\mu \leqq \beta. \tag{9}$$

By (9) and the definition of the left-hand side,

$$\int g \, d\mu = \lim_{L\to\infty} \int g^L \, d\mu \leqq \beta$$

whence $g \varepsilon L^1$. Furthermore

$$\lim_{L\to\infty} \int | g - g^L | \, d\mu = \lim_L \int (g - g^L) \, d\mu = 0. \tag{10}$$

We have now

$$\int | g_n - g | \, d\mu \leqq \int | g_n - g_n^L | \, d\mu + \int | g_n^L - g^L | \, d\mu + \int | g^L - g | \, d\mu.$$

The first term on the right is dominated by

$$\int_{A_{n,L}} g_n \, d\mu$$

where $A_{n,L}$ is the set over which $g_n(x) \geqq L$. By (4) and (8) therefore,

$$0 \leqq \limsup_n \int | g_n - g | \, d\mu \leqq 0(L2^{-L}) + \int | g^L - g | \, d\mu.$$

We let $L \to \infty$ and use (10) to conclude that $\lim_n \int | g_n - g | \, d\mu = 0$.

This establishes the mean convergence of the sequence $(g_n(x))$.

It was shown in 4.3 that the entropy rate of $[A, \mu]$ is $\lim_{n\to\infty} \int g_n \, d\mu$. From what we have just shown, $\lim_{n\to\infty} \int g_n \, d\mu = \int g \, d\mu$. In 9.2, $h(x)$ is defined as the limit of $1/N \sum_{k=0}^{N-1} g(T^k x)$ and we know by the ergodic theorem then that $\int h \, d\mu = \int g \, d\mu$. When μ is ergodic $h(x) = H$, a constant, almost everywhere. Therefore

$$H = \int H \, d\mu = \int h \, d\mu = \int g \, d\mu = \lim \int g_n \, d\mu.$$

This identifies H with the entropy rate of $[A, \mu]$.

10. Finitary devices.

10.0. *Sources.* Shannon's Markov-like sources, which we have here called finitary, are defined by a construction equivalent (in a sense to be made precise later) to one now to be described.

Consider a Markov process with finitely many states, enumerated $1, 2, \cdots, S$. Let p_{ij} be the probability of the transition from state j to state i, and ξ_i the stationary probability of occupancy of state i, so that

$$\xi_i = \sum_{j=1}^{S} p_{ij}\xi_j , \qquad 1 \leqq i \leqq S.$$

Let B be the alphabet whose letters are the symbols $1, 2, \cdots, S$. We may suppose that the Markov process makes a transition at each time $\tau = t + \frac{1}{2}$, $t = 0, \pm 1, \pm 2, \cdots$. We define the stationary information source $[B, \nu]$ by the rule that the letter which occurs at time $t \varepsilon I$ is the name of the state in which the Markov process is at that time. The p_{ij} and the ξ_i are enough to define this source. A source defined in this way will be called a finite Markov source.

Let A be an arbitrary alphabet and let φ be a function from B to A: $\alpha = \varphi(\beta)$, $\beta \varepsilon B$, $\alpha \varepsilon A$. Given $y \varepsilon B^I$, say $y = (\cdots, y_{-1}, y_0, y_1, \cdots)$, we define $x = \Phi(y) \varepsilon A^I$ by $x = (\cdots, x_{-1}, x_0, x_1, \cdots)$, where $x_t = \varphi(y_t)$, $t \varepsilon I$. Let μ be the measure over F_A defined by this construction. In the notation of [3], $\mu = \nu\Phi^{-1}$. The source $[A, \mu]$ we will call a *projection* of $[B, \nu]$. The notion of projection clearly applies even when $[B, \nu]$ is not Markov.

An arbitrary source $[A, \mu]$ will be called finitary if it is a projection of some finite Markov source $[B, \nu]$. Shannon's sources are all of this kind in the sense that, given any source of his, there is a projection of a finite Markov source which produces the same ensemble of text, and conversely.

Consider now an $[A, \mu]$, a projection by φ of $[B, \nu]$. Given $\alpha \varepsilon A$, let $\varphi^{-1}(\alpha)$ denote that subset of B consisting of all β such that $\varphi(\beta) = \alpha$. We will call $[A, \mu]$ *unifilar* if for each $\alpha \varepsilon A$ and each state $i \varepsilon B$ there is at most one transition from i to $\varphi^{-1}(\alpha)$ which has nonzero probability.

The definition of [2], paragraph 7, and certain related results, are tacitly restricted to finitary and unifilar sources. The word "finitary" as we use it in discussing the proof of the fundamental theorem (Section 8) may, however, be interpreted in the wider sense defined above: being a projection of a finite Markov process.

10.1. *Channels.* We now frame a definition of "finitary channel" consonant with that just given for finitary sources. A finitary channel is specified by:

(i) An input alphabet A.

(ii) An output alphabet B.

(iii) A finite set $D = (1, 2, \cdots, K)$ of states. We treat D as an alphabet.

(iv) A set of Markov transition matrices, $\| q_{ij}(\alpha) \|$, one matrix for each $\alpha \varepsilon A$. Each element $q_{ij}(\alpha)$ represents a conditional probability of transition from state $j \varepsilon D$ to state $i \varepsilon D$ knowing that the input letter is α. We have $\Sigma_i\, q_{ij}(\alpha) = 1$ for each j and α.

(v) A function ψ from D to B. The output letter from the channel is $\psi(i)$ whenever the transition is to the state $i \varepsilon D$.

Consider a stationary source $[A, \mu]$ driving the channel so specified. Let λ_j be the stationary probability of finding the channel in state $j \varepsilon D$, (if such a probability exists). Then the probability that the letter α be presented to the channel and that the channel make a transition to state $i \varepsilon D$ is

$$\sum_j \mu([\alpha]) q_{ij}(\alpha) \lambda_j . \tag{1}$$

Stationarity of the system requires now that the sum of these numbers over all $\alpha \varepsilon A$ be λ_i. That is, the vector $(\lambda_i, i \varepsilon D)$ must be invariant under left multiplication by the Markov matrix

$$Q = \| \sum_{\alpha \varepsilon A} \mu([\alpha])\ q_{ij}(\alpha) \| .$$

At least one such invariant probability vector exists. If, for example, each matrix $\| q_{ij}(\alpha) \|$ has a unique such invariant vector, then in general the λ_i will also be unique and they will be continuous functions of the letter frequencies of the source.

Given the λ_i above, the joint probability that letters $\alpha_1, \cdots, \alpha_n$ be presented to the channel and that the corresponding sequence of states of the channel be $i_1, i_2, \cdots, i_n$ is similar to the expression (1):

$$\mu([\alpha_1, \cdots, \alpha_n]) \sum_{j \varepsilon D} q_{i_n i_{n-1}}(\alpha_n) \cdots q_{i_2 i_1}(\alpha_2) q_{i_1 j}(\alpha_1) \lambda_j . \tag{2}$$

The joint probability of input letters $[\alpha_1, \cdots, \alpha_n]$ and output letters $[\beta_1, \cdots, \beta_n]$ is found by summing (2) for

(1) all i_1 in $\psi^{-1}(\beta_1)$,

(2) all i_2 in $\psi^{-1}(\beta_2)$,

$\cdot$

$\cdot$

$\cdot$

(n) all i_n in $\psi^{-1}(\beta_n)$.

The conditional probability of $[\beta_1, \cdots, \beta_n]$ knowing $[\alpha_1, \cdots, \alpha_n]$ is then

$$\sum_n \cdots \sum_1 \sum_{j \varepsilon D} q_{i_n i_{n-1}}(\alpha_n) \cdots q_{i_1 j}(\alpha_1) \lambda_j , \tag{3}$$

where $\sum_k$ denotes the summation of i_k over $\psi^{-1}(\beta_k)$. The expression (3) depends on $[\alpha_1, \cdots, \alpha_n]$ and $[\beta_1, \cdots, \beta_n]$, and not otherwise upon past history. It is independent of the source except for the continuous dependence of the λ_j upon the letter frequencies. This continuity is sufficient for the proof in Section 8, since it is easy there to guarantee that the source $[A, \mu^*]$ and the source which results from putting $[A_1, \mu]$ through the transducer there defined have virtually the same letter frequencies.

10.2. *A discrepancy.* The purist will observe that in 2.24 we defined a channel as a set of conditional probability measures ν_θ over outputs, where θ represents the input sequence. The construction in 10.1 is not obviously of this kind, since the measures ν_θ there obtained might well depend not only on θ but also on the particular source-ensemble in mind at the moment. We will not clarify the point here. Some pedagogic license was used in 2.24, and it is simpler to enlarge the notion of channel beyond that defined there than to try to reconcile the two definitions.

11. A useful theorem. Let Ω be an abstract countable set of elements ω. Let A be a finite set, an alphabet. Let μ be a probability measure over a Borel field containing all sets $S \otimes W$, where $S \in F_A$ and $W \subseteq \Omega$. Define the measures μ_ω over F_A by $\mu_\omega(S) = \mu(S \otimes \omega)/\mu(A^I \otimes \omega)$. This definition is valid for almost every ω. Define $\bar{\mu}$ over F_A by $\bar{\mu}(S) = \mu(S \otimes \Omega)$. Suppose that the source $[A, \bar{\mu}]$ is ergodic and has rate H. Suppose that

$$\int [-\log \mu(A^I \otimes \omega)]\, d\mu(x \otimes \omega) < \infty. \tag{1}$$

Then the functions $f_n(x, \omega) = -(\log \mu([0, n-1; x] \otimes \omega))/n$ converge in L^1 mean to H relative to μ. Considering ω as a parameter, for almost every ω the functions $f_n(x, \omega)$ converge in L^1 mean to H relative to μ_ω.

PROOF. Since $\mu([0, n-1; x] \otimes \omega) \leqq \bar{\mu}([0, n-1; x])$ we have

$$f_n(x, \omega) \geqq -\frac{1}{n} \log \bar{\mu}([0, n-1; x]) = g_n(x), \tag{2}$$

where the second equality sign defines $g_n(x)$. Fix n and consider the countable list of events $[0, n-1; x] \otimes \omega$. By the composition law (3.2), extended to infinite sums, the entropy of this list of events is the sum of the entropy of the $[0, n-1; x]$ and the conditional entropy of ω knowing $[0, n-1; x]$:

$$H([0, n-1; x] \otimes \omega) = H([0, n-1; x]) + H_{x,n}(\omega). \tag{3}$$

Now the convexity law (3.3) implies that the average conditional entropy $H_{x,n}(\omega)$ is always less than the unconditional entropy of ω, which latter is the integral asserted to be finite in (1). Hence there is a finite K such that for all n

$$H_{x,n}(\omega) \leqq K. \tag{4}$$

From (2), (3), (4) and the definitions of $\bar{\mu}$ and the entropies,

$$\int |f_n(x, \omega) - g_n(x)|\, d\mu(x \otimes \omega) \int f_n(x, \omega)\, d\mu(x \otimes \omega) - \int g_n(x)\, d\bar{\mu}(x)$$

$$= \frac{1}{n} H([0, n-1; x] \otimes \omega - \frac{1}{n} H([0, n-1; x]) \leqq \frac{1}{n} K.$$

Therefore

$$\int | f_n - H | d\mu \leqq \int | f_n - g_n | d\mu + \int | g_n - H | d\bar{\mu} \leqq \frac{1}{n} K + \int | g_n - H | d\bar{\mu}.$$

Since by hypothesis and (9.1) g_n tends to H in L^1 mean $(\bar{\mu})$, the first conclusion of the theorem follows.

For the second conclusion of the theorem, we note that

$$\int | f_n(x, \omega) - H | d\mu(x \otimes \omega) = \sum_{\omega} \mu(A^I \otimes \omega) \int | f_n(x, \omega) - H | d\mu_\omega(x).$$

Since the left-hand side has limit zero, every term on the right for which $\mu(A^I \otimes \omega) \neq 0$ must have limit zero.

As an application of this theorem let $[B, \nu]$ be a stationary source. Let $[A, \bar{\mu}]$ be a projection by φ of $[B, \nu]$. Let Ω coincide with the alphabet B and for $S \in F_A$ define μ by $\mu(S \otimes \omega) = \nu(\Phi^{-1}(S) \cap D_\omega)$ where D_ω is the set of $y \in B^I$ such that $y_{-1} = \omega$. The theorem then implies (if $[A, \bar{\mu}]$ is ergodic) that the rate of $[A, \mu]$ may be calculated by considering conditional probabilities knowing that the letter ω occurred at time -1. When $[B, \nu]$ is finite and Markov, this often leads to simplified calculations.

As another application, consider a fixed countable partition of A^I into sets $S_\omega \in F_A$. Given an ergodic source $[A, \bar{\mu}]$, define $\mu(S \otimes \omega)$ by $\mu(S \otimes \omega) = \bar{\mu}(S \cap S_\omega)$. The theorem then implies that the entropy rate of $[A, \bar{\mu}]$ can be calculated using only partitions which refine the given one.

REFERENCES

[1] NORBERT WIENER, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley and Sons, 1949.

[2] C. E. SHANNON, "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, pp. 379–423, pp. 623–656.

[3] P. R. HALMOS, *Measure Theory*, D. Van Nostrand, 1950.

[4] NORBERT WIENER, "The ergodic theorem," *Duke Mathematical Journal*, Vol. 5, pp. 1–18.

[5] J. L. DOOB, *Stochastic Processes*, John Wiley and Sons, 1953.