Lecture 20

INTRODUCTION TO COMMUNICATION THEORY

At this point we have all the basic machinery of our theory developed, and have seen its application in some of the "classical" problems. We said back in the first talk that what started all this was the attempt to see statistical mechanics and communication theory as examples of the same line of reasoning. A generalized form of statistical mechanics appeared as soon as we supplemented Laplace's theory of inductive reasoning by the notion of entropy, and we ought now to be in a position to treat communication theory in a similar way.

One difference is that in statistical mechanics the prior information has nothing to do with frequencies (it consists of measured values of quantities such as pressure); while in communication theory the prior information is obtained in a different way, which makes the probability-frequency paradoxes much more acute. For this reason I thought it best to take up communication theory only after we had seen some of the general connections between probability and frequency, via the A_p distribution and the de Finetti theorem.

First the difficult matter of giving credit where credit is due. All major advances in understanding have their precursors, whose full significance is never recognized at the time. Relativity theory had them in the work of Mach, Fitzgerald, and Lorentz, to mention only the most obvious examples. Communication theory had many precursors, in the work of Gibbs, Nyquist, Hartley, Szilard, von Neumann, and Wiener. But there is no denying that the

work of Shannon (1948) represents the arrival of the main signal, just as did Einstein's of 1905. Here for the first time, ideas which had long been, so to speak, "in the air" in a vague form, are grasped and put into sharp focus.

Shannon's papers were so full of important new concepts and results that they exercised not only a stimulating effect, but also a paralyzing effect. During the first few years after their appearance, it was common to hear the opinion expressed, rather sadly, that Shannon had anticipated and solved all problems of the field, and left nothing else for others to do. Today, I think, no one entertains any such ideas, and the field has seen considerably more development.

The psst-Shannon developments, with few exceptions, can be classed into efforts in two entirely different directions. On the one hand we have the expansionists, who try to apply Shannon's ideas to other fields, as I have been doing. Others range from the entropy calculator (who works out the entropy of a television signal, the French language, a chromosome, or almost anything else you can imagine; and often finds that nobody knows what to do with the result), to the universalist (who assures us that communication theory will revolutionize all intellectual activity; but seldom offers a specific example of anything that has been changed by it).

We should not be critical of these efforts because, as J. R. Pierce has said, it is very hard to tell at present which ones make sense, which are pure nonsense, and which are the beginning of something that will in time make sense. My own efforts have received all three classifications from various quarters. I have a very strong hope, and a moderately strong belief, that the ideas introduced by Shannon will eventually be indispensable to the linguist, the geneticist, the television engineer, the neurologist, etc. But I share with many others a feeling of disappointment that twenty years

of effort along these lines has led to so little in the way of really useful advances in these fields. We have today an abundance of vague philosophy, and of abstract mathematics, but a rather embarrassing shortage of examples where specific practical problems have been solved by using communication theory.

The moral of this is, I think, that more than half the battle is in learning how to ask the right question. People who want to apply communication theory to new fields must learn that the first, and hardest, step is to state precisely what is the problem we want solved. Once we succeed in doing this, real progress comes easily. I will give some examples pertaining to statistical mechanics and decision theory in these lectures.

In almost diametric opposition to the above efforts, as far as aim is concerned, stand the mathematicians, who view communication theory simply as a branch of pure mathematics. Characteristic of this school is a belief that, before introducing a continuous probability distribution, you have to talk about set theory, Borel fields, measure theory, the Lebesgue-Stieltjes integral, and the Radon-Nikodym theorem. The important thing is to make the theorems rigorous by the criteria of rigor currently popular, even if in so doing we limit the scope of the practical theory, and/or make it unintelligible to the average scientist or engineer. The recently published books on information theory by A. Khinchin (1957) and A. Feinstein (1958) can serve as typical examples of the style prevalent in this literature.

Here again, no valid criticism of these efforts is possible. Of course, we want our principles to be subjected to the closest scrutiny one can bring to bear on them. If important applications exist, the need for this is so much the greater; fortunately, mathematicians have found the subject interesting enough to take on a not very easy task. However, the present talks are not addressed to mathematicians, but to scientists and engineers who are

interested in applications; and so I am going to dwell on this side of the story only to the extent of pointing out that the particular theorems which the mathematicians have chosen to rigorize are not always the ones relevant to real situations.

Now, in order to explain this rather cryptic remark, let's turn to some of the specific things in Shannon's papers.

20.1. The Noiseless Channel.

We deal with the transmission of information from some sender to some receiver. I will speak of them in anthropomorphic terms, such as "the man at the receiving end," although either or both might actually be machines, as in telemetry or remote control systems. Transmission takes place via some channel, which might be a telephone or telegraph circuit, a microwave link, a frequency band assigned by the FCC, the German language, the postman, the neighborhood gossip, or a chromosome. If, after having received a message, the receiver can always determine with certainty which message was intended by the sender, we say that the channel is noiseless.

It was recognized very early in the game, particularly by Nyquist and Hartley, that the capability of a channel is not described by any property of the specific messages it sends, but rather by what it could have sent. The usefulness of a channel depends on its ability to transmit any one of a large class of messages, which the sender can choose at will.

In a noiseless channel, the obvious measure of this ability is simply the maximum number, $W(t)$, of distinguishable (at the destination) messages which the channel is capable of transmitting in time t. In all cases of interest to us, this number eventually goes into an exponential increase for sufficiently large t: $W(t) \sim e^{Ct}$, so the measure of channel performance which is independent of any particular time interval is the coefficient C

of this increase. We define the channel capacity as

$$C \equiv \lim_{t\to\infty} \left[\frac{1}{t} \log W(t)\right] \tag{20-1}$$

The units in which C is measured will depend on which base we choose for our logarithms. Usually one takes the base 2, in which case C is given in "bits per second," one bit being the amount of information contained in a single binary (yes-no) decision. For easy interpretation of numerical values the bit is by far the best unit to use; but in formal operations it is easier to use the base e of natural logarithms, and I will do that in this discussion. Our channel capacities are therefore measured in natural units, or "nits per second." To convert, we note that 1 bit = $(\log_e 2)$ = 0.69315 nits, or 1 nit = 1.4427 bits.

The capacity of a noiseless channel is a definite number, characteristic of the channel, which contains no subjective features. Thus, if a noiseless channel can transmit n symbols per second, chosen in any order from an alphabet of a letters, we have $W(t) = a^{nt}$, or C = n log a nits/second. Any constraint on the possible sequences of letters can only lower this number. For example, if the alphabet is $A_1, A_2, \ldots, A_a$, and it is required that in a long message of N = nt symbols the letter A_i must occur with relative frequency f_i, then the number of possible messages in time t is only

$$W(t) = \frac{N!}{(Nf_1)! \ \ldots \ (Nf_a)!} \tag{20-2}$$

and from Stirling's approximation, we find, as in Eq. (10-17),

$$C = -n \sum_i f_i \log f_i \quad \text{nits/second.} \tag{20-3}$$

This attains its maximum value, equal to the previous C = n log a, in the case of equal frequencies, $f_i = a^{-1}$. Thus we have the interesting result that a constraint requiring all letters to occur with equal frequencies does not decrease channel capacity at all. It does, of course, decrease the number

W(t) by an enormous factor; but the decrease in log W is what counts, and this grows less rapidly than t, so it makes no difference in the limit.

Suppose now that symbol A_i has transmission time t_i, but there is no other constraint on the allowable sequences of letters. What is the channel capacity? Well, consider first the class of messages in which letter A_i occurs n_i times, i = 1, 2, ..., a. The number of such messages is

$$W(n_1 \ldots n_a) = \frac{N!}{n_1! \ldots n_a!} \tag{20-4}$$

where

$$N = \sum_{i=1}^{a} n_i \quad . \tag{20-5}$$

The total number of different messages that can be transmitted in time t is then

$$W(t) = \sum_{n_i} W(n_1 \ldots n_a) \tag{20-6}$$

where we sum over all choices of $(n_1 \ldots n_a)$ compatible with $n_i \geq 0$ and

$$\sum_{i=1}^{a} n_i t_i \leq t \; . \tag{20-7}$$

The number K(t) of terms in the sum (20-6) satisfies $K(t) \leq (Bt)^a$ for some $B < \infty$. This is seen most easily by imagining the n_i as coordinates in an a-dimensional space and noting the geometrical interpretation of (20-7).

Exact evaluation of (20-6) would be quite an unpleasant job. But it's only the limiting value that we care about right now, and we can get out of the hard work by the following trick. Note that W(t) cannot be less than the greatest term $W_m = W_{max}(n_1 \ldots n_a)$ in (20-6) nor greater than $W_m K(t)$:

$$\log W_m \leq \log W(t) \leq \log W_m + a \log (Bt) \tag{20-8}$$

and so we have

$$C \equiv \lim_{t \to \infty} \frac{1}{t} \log W(t) = \lim_{t \to \infty} \frac{1}{t} \log W_m \tag{20-9}$$

i.e., to find the channel capacity, it is sufficient to maximize $\log W(n_1 \ldots n_a)$ subject to the constraint (20-7). This rather surprising fact can be understood as follows. The logarithm of W(t) is given, rather crudely, by

$$\log W(t) = \log W_{max} + \log [\text{number of reasonably large terms in (20-6)}]$$

Even though the number of large terms tends to infinity as t^a, this is not rapid enough to make any difference in comparison with the exponential increase of W_{max}. This same mathematical fact is the reason why, in statistical mechanics, the Darwin-Fowler method and the method of the most probable distribution lead to the same results in the limit of large systems.

We can solve the problem of maximizing $\log W(n_1 \ldots n_a)$ by the same Lagrange multiplier argument used in Lecture 10, Section (10.6). The problem is not quite the same, however, because now N is also to be varied in finding the maximum.

Using the Stirling approximation, which is valid for large n_i, we have as before

$$\log W(n_1 \ldots n_a) \simeq N \log N - \sum_{i=1}^{a} n_i \log n_i \qquad (20\text{-}10)$$

The variational problem with λ a Lagrangian multiplier, is

$$\delta[\log W + \lambda \sum n_i t_i] = 0 \qquad (20\text{-}11)$$

but since $\delta N = \sum \delta n_i$, we have

$$\delta \log W = \delta N \log N - \delta N - \sum_i (\delta n_i \log n_i - \delta n_i)$$

$$= - \sum \delta n_i \log\left(\frac{n_i}{N}\right) \qquad (20\text{-}12)$$

Therefore (20-11) reduces to

$$\sum_{i=1}^{a} \left[\log\left(\frac{n_i}{N}\right) + \lambda t_i\right] \delta n_i = 0$$

with the solution

$$n_i = N\, e^{-\lambda t_i} \tag{20-13}$$

To fix the value of λ we require

$$N = \sum n_i = N \sum e^{-\lambda t_i} \tag{20-14}$$

With this choice of n_i, we find

$$\frac{1}{t} \log W_m = -\frac{1}{t} \sum n_i \log\left(\frac{n_i}{N}\right) = \frac{1}{t} \sum n_i (\lambda t_i) \tag{20-15}$$

In the limit, $t^{-1} \sum n_i t_i \to 1$, and we find

$$C = \lim_{t\to\infty} \frac{1}{t} \log W(t) = \lambda \ . \tag{20-16}$$

So, our final result can be stated very simply:

To calculate the capacity of a noiseless channel in which symbol A_i has transmission time t_i and which has no other constraints on the possible messages, define a partition function

$$Z(\lambda) = \sum_i e^{-\lambda t_i} \tag{20-17}$$

Then the channel capacity C is the real root of

$$Z(\lambda) = 1. \tag{20-18}$$

You see already a very strong resemblance to the reasoning and the formalism of statistical mechanics, in spite of the fact that we have not yet said anything about probability. From (20-14) we see that $W(n_1 \ldots n_a)$ is maximized when the relative frequency of symbol A_i is given by the canonical distribution

$$f_i = \frac{n_i}{N} = e^{-\lambda t_i} = e^{-Ct_i} \tag{20-19}$$

Should we conclude from this that the channel is being "used most efficiently"

when we have encoded our messages so that (20-19) holds? No, that wouldn't be quite the right way of putting it. Because, of course, in time t the channel will actually transmit one message and only one; and this remains true regardless of what relative frequencies we use. Equation (20-19) tells us only that the overwhelming majority of all possible messages that the channel could have transmitted in time t are ones where the relative frequencies are canonical.

On the other hand we have a generalization of the remark following (20-3); if we impose an additional constraint requiring that the relative frequencies are given by (20-19), which might be regarded as defining a new channel, the channel capacity would not be decreased. But any constraint requiring that all possible messages have letter frequencies different from (20-19) will decrease channel capacity.

There are many other ways of interpreting these equations. For example, in our above arguments we supposed that the total time of transmission is fixed and we wanted to maximize the number W of possible messages among which the sender can choose. In a practical communications system, the situation is usually the other way around; we know in advance the extent of choice which we demand in the messages which might be sent over the channel, so that W is fixed. We then ask for the condition that the total transmission time of the message be minimized subject to a fixed W.

It is well known that variational problems can be transformed into several different forms, the same mathematical result giving the solution to many different problems. A circle has maximum area for a given perimeter; and also it has minimum perimeter for a given area. In statistical mechanics, the canonical distribution can be characterized as the one with maximum entropy for a given expectation of energy; or equally well as the one with minimum expectation of energy for a given entropy. Similarly, the channel capacity

found from (20-18) gives the maximum attainable W for a given transmission time, while its reciprocal is equal to the minimum attainable transmission time for a fixed W.

As another extension of the meaning of these equations, note that we don't have to interpret the quantity t_i as a time; it can stand equally well for the "cost," as measured by any criterion, of transmitting the i'th symbol. For example, it might be that the total length of time the channel is in operation is of no importance, because the apparatus has to sit there in readiness whether it is being used or not. The real economic criterion might be the total amount of choice W_b of different messages which the apparatus is capable of transmitting before breaking down, for a given installation cost. The lifetime of the apparatus might be limited by the total number of times a certain relay has to open and close. In this case, we could define t_i as the number of times this relay must operate in the course of transmitting the i'th symbol. The channel capacity given by Equation (20-18) would then be measured, not in nits per second, but in "nits per relay operation," and its reciprocal is equal to the minimum attainable number of relay operations per nit of transmitted information.

A more complicated type of noiseless channel, also considered by Shannon, is one where the channel has a memory; it may be in any one of a set of "states," $\{S_1 \ldots S_k\}$ and the possible future symbols, or their transmission times, depend on the present state. For example, suppose that if the channel is in state S_i, it can transmit symbol A_n, which leaves the channel in state S_j, the corresponding transmission time being t_{inj}. Surprisingly, the calculation of channel capacity in this case is quite easy.

Let $W_i(t)$ be the total number of different messages the channel can transmit in time t, starting from state S_i. Breaking down $W_i(t)$ into several terms according to the first symbol transmitted, we have

$$W_i(t) = \sum_{jn} W_j(t - t_{inj}) \qquad (20\text{-}20)$$

where the sum is over all possible sequences $S_i \rightarrow A_n \rightarrow S_j$. This is a linear difference equation with constant coefficients, so its assymptotic solution must be an exponential function:

$$W_i(t) \approx B_i \exp(Ct) \qquad (20\text{-}21)$$

and from the definition (20-1) it is clear that, for finite k, the coefficient C is the channel capacity. Substituting (20-21) into (20-20), we obtain

$$B_i = \sum_{j=1}^{k} Z_{ij}(C)\, B_j \qquad (20\text{-}22)$$

where

$$Z_{ij}(\lambda) = \sum_{n} \exp(-\lambda t_{inj}) \qquad (20\text{-}23)$$

is the "partition matrix." If the sequence $S_i \rightarrow A_n \rightarrow S_j$ is impossible, we set $t_{inj} = \infty$. By this device we can understand the sum in (20-23) as extending over all symbols in the alphabet.

Equation (20-22) says that the matrix Z_{ij} has an eigenvalue equal to unity. Thus, the channel capacity is the greatest real root of $D(\lambda) = 0$, where

$$D(\lambda) \equiv \det[Z_{ij}(\lambda) - \delta_{ij}] \ . \qquad (20\text{-}24)$$

In the case of a single state, k = 1, this reduces to the previous rule, Equation (20-18).

The problems solved above are, of course, only especially simple ones. By inventing channels with more complicated types of constraints on the allowable sequences (i.e. with a long memory), you can generate mathematical problems as involved as you please. But it would still be just the mathematics--as long as the channel is noiseless, there would be no difficulties of principle. In each case you simply have to count up the possibilities and apply the definition (20-1). For some weird channels, you might find that the limit

therein does not exist, in which case we can't speak of a channel capacity, but have to characterize the channel simply by giving the function W(t).

20.2. The Information Source.

When we take the next step and consider the information source feeding our channel, fundamentally new problems arise. There are mathematical problems aplenty, but there are also more basic conceptual problems, which have to be considered before we can state which mathematical problems are the significant ones.

It was Professor Norbert Wiener who first suggested the enormously fruitful idea of representing an information source in probability terms. He applied this to some problems of filter design, which I will take up briefly in a later lecture. This work was an essential step in developing a way of thinking which led to modern communication theory.

It is perhaps difficult nowadays for us to realize what a big step this was. Previously, communication engineers had considered an information source simply as a man with a message to send; for their purposes an information source could be characterized simply by describing that message. But Wiener suggested instead that an information source be characterized by giving the probabilities that it will emit various messages. Already we can see some conceptual difficulties faced by a frequency theory of probability--the man at the sending end presumably knows perfectly well which message he is going to send. What, then, could we possibly mean by speaking of the probability that he will send something? There is nothing analogous to "chance" operating here.

By the probability of a message, do we mean the frequency with which he sends that particular message? The question is absurd--a sane man sends a given message at most once, and most messages never. Do we mean the frequency

with which the message M occurs in some imaginary "ensemble" of communication acts? Well, it's all right to state it this way if you want to, but it doesn't answer the question. It merely leads us to re-state the question as: what do we mean by the ensemble? How is it to be set up? Calling it by a different name doesn't help us.

Right at this point we have to state clearly what is the specific problem we want solved. A probability distribution is a means of describing a state of knowledge. But whose state of knowledge do we want to talk about? Evidently, not the man at the sending end. Is it the man at the receiving end? Well, that might be relevant to the problem I have in mind. But basically, since I am talking to scientists and engineers, I want to consider communication theory, not as describing the "general philosophy" of communication between sender and receiver, but as something of practical value to an engineer whose job is to design the technical equipment in the communication system. In other words, the state of knowledge we want to describe is that of the communication engineer when he designs the equipment.

This consideration is something you will not find in the previous literature based on the viewpoint which sees no distinction between probability and frequency; on this view, the notion of a probability for a person with a certain state of knowledge simply doesn't exist. Nevertheless, from any viewpoint, the problem of choosing some probability distribution to represent the information source does exist. It cannot be evaded, and the whole content of the theory depends on how we do this.

I have already emphasized several times that in probability theory we never solve an actual problem of practice. We solve only some abstract mathematical model of the real problem. Setting up this model requires not only mathematical ability, but also practical judgment. If our model does not correspond well to the actual situation, our theorems, however rigorous,

may be more misleading than helpful.

This is so with a vengeance in communication theory because, as I will show in this lecture, not only the quantitative details, but even the qualitative nature of the theorems that can be proved, depend on which probability model we use to represent an information source.

The purpose of this probability model is to describe the communication engineer's prior knowledge about the messages to be sent. In principle, this prior knowledge could be of any sort; but in "traditional" communication theory the only kind of prior knowledge considered consists of frequencies of letters, or combinations of letters, which have been observed in past samples of similar messages. A typical practical problem is to design equipment which will transmit English text at a given rate, while using the smallest possible channel capacity. The engineer will then, according to the usual viewpoint, need accurate data giving the correct frequencies of English text. Let's think about that a little more.

Suppose we try to characterize the English language, for purposes of communication theory, by specifying the relative frequencies of various letters, or combinations of letters. Now we all know that there is a great deal of truth in statements such as "the letter E occurs more frequently than the letter Z." Long before the days of communication theory, many people made obvious common-sense use of this knowledge. One of the earliest examples is the design of the Morse telegraphic code, in which the most frequently used letters are represented by the shortest codes--the exact prototype of what Shannon formalized and made precise a century later.

The design of our standard typewriter keyboard makes considerable use of knowledge of letter frequencies. This knowledge was used in a much more direct and drastic way by Ottmar Mergenthaler, whose immortal phrase

ETAOIN SHRDLU

was a common sight in the newspapers not so many years ago. But already we are getting into trouble, because there does not seem to be complete agreement even as to the relative order of the twelve most common letters in English, let alone the numerical values of their relative frequencies. For example, according to Pratt (1942) the above phrase should read

ETANOR ISHDLF

while Tribus (1961) gives it as

ETOANI RSHDLC

As we go into the less frequently used letters, the situation becomes still more chaotic.

Of course, we readily see the reason for these differences. People who have obtained different values for the relative frequencies of letters in English have consulted different samples of English text. It is obvious enough that the last volume of an encyclopaedia might have a higher relative frequency for the letter Z than the first volume. There is no reason to expect that letter frequencies would be the same in, say, a textbook on organic chemistry, a treatise on the history of Egypt, and a modern American novel. The writing of educated people would reveal systematic differences in word frequencies from the writings of people who had never gone beyond grade school. Even within a much narrower field, we would expect to find significant differences in letter and word frequencies in the writings of James Michener and Ernest Hemingway. The letter frequencies in the transcript of the tape recording of this lecture will probably be noticeably different from those I would produce if I sat down and wrote out the lecture verbatim.

The fact that statistical properties of a language vary with the author and circumstances of writing is so clear that it has become a useful research tool. A recent doctoral thesis in classics submitted to Columbia University by James T. McDonough (1961) contains a computer-run statistical analysis of

Homer's Iliad. Classicists have long debated whether all parts of the Iliad were written by the same man, and indeed whether Homer is an actual historical person. The analysis showed stylistic patterns consistent throughout the work. For example, 40.4 per cent of the 15,693 lines end on a word with one short syllable followed by two long ones, and a word of this structure never once appears in the middle of a line. Such consistency in a thing which is not a characteristic property of the Greek language, seems very strong evidence that the Iliad was written by a single person in a relatively short period of time, and it was not, as had been supposed by many nineteenth century classicists, the result of an evolutionary process over several centuries.

Of course, the evolutionary theory is not demolished by this evidence alone. If the Iliad was sung, we must suppose that the music had the very monotonous rhythmic pattern of primitive music, which persisted to a large extent as late as Bach and Haydn. Characteristic word patterns may have been forced on the composers, by the nature of the music. Archaeologists tell us that the siege of Troy, described in the Iliad, is not a myth but an historical fact, occurring about 1200 B. C., some four centuries before Homer. The decipherment of Minoan Linear B script by Michael Ventris in 1952 established that Greek existed already as a spoken language in the Aegean area several centuries before the siege of Troy; but the introduction of the Phoenician alphabet, which made possible a written Greek language in the modern sense, occurred only about the time of Homer. You see that the question is very complex and far from settled; but I find it fascinating that a statistical analysis of word and syllable frequencies, representing evidence which has been there in the Iliad for some twenty-eight centuries for anyone who had the wit to extract it, is now recognized as having a definite bearing on the problem. Undoubtedly, this is only the beginning of this type of analysis.

Well, to get back to communication theory, the point I am making is simply this: it is utterly wrong to say that there exists one and only one "true" set of letter or word frequencies for English text. If we use a mathematical model which presupposes the existence of such uniquely defined frequencies, we might easily end up proving things which, while perfectly valid as mathematical theorems, are worse than useless to an engineer who is faced with the job of actually designing a communication system to transmit English text efficiently.

But suppose our engineer does have extensive frequency data, and no other prior knowledge. How is he to make use of this in describing the information source? Many of the standard results of communication theory can, from the viewpoint I am advocating, be seen as simple examples of maximum-entropy inference; i.e. as examples of the same kind of reasoning as in statistical mechanics. To understand this was my original goal, discussed in Lecture 1.

20.3. Optimum Encoding: Letter Frequencies Known.

Suppose our alphabet consists of a different symbols $A_1, A_2, \ldots, A_a$, and we denote a general symbol by A_i, A_j, etc. Any message of N symbols then has the form $A_{i_1} A_{i_2} \ldots A_{i_N}$. We denote this message by M, which is a shorthand expression for the set of indices: $M = \{i_1\ i_2 \ldots i_N\}$. The number of conceivable messages is a^N. By $\sum_M$ I mean a sum over all of them. Also, define

$N_j(M) \equiv$ (number of times the letter A_j appears in the message M)

$N_{ij}(M) \equiv$ (number of times the digram $A_i A_j$ appears in M),

and so on.

Consider first an engineer E_1, who has a set of numbers $(f_1 \ldots f_a)$ giving the relative frequencies of the letters A_i, as observed in past samples of

messages, but has no other prior knowledge. What communication system represents rational design on the basis of this much information, and what channel capacity does E_1 require in order to transmit messages at a given rate of n symbols per second? To answer this, we need the probabilities p(M) which E_1 assigns to the various conceivable messages. Now Mr. E_1 has no deductive proof that the letter frequencies in future messages will be equal to the f_i observed in the past. On the other hand, his state of knowledge affords no grounds for supposing that the frequency of A_i will be greater than f_i rather than less, or vice versa. So he is going to suppose that frequencies in the future will be more or less the same as in the past, but he is not going to be too dogmatic about it. He can do this by requiring of the distribution p(M) only that it yield <u>expected</u> frequencies equal to the past ones. In other words,

$$\langle N_i \rangle = \sum_M N_i(M)\ p(M) = Nf_i, \qquad i = 1, 2, \ldots, a \qquad (20\text{-}25)$$

Of course, p(M) is not uniquely determined by these constraints, and so E_1 must at this point make a free choice of some distribution.

Let me emphasize again that it makes no sense to say that there exists any "physical" or "objective" distribution p(M) for this problem. This becomes especially clear if we suppose that only a single message is ever going to be sent over the communication system; thus there is no conceivable procedure by which p(M) could be measured as a frequency. But this would in no way affect the problem of engineering design which we are considering.

In choosing a distribution p(M), it would be perfectly possible for E_1 to assume some message structure involving more than single letters. For example, he might suppose that the digram A_1A_2 is more likely than A_3A_4. But from the standpoint of E_1 this could not be justified, for <u>as far as he knows</u>, a design based on any such assumption is as likely to hurt as to help. From E_1's standpoint, rational conservative design consists just in carefully

avoiding any such assumptions. This means, in short, that E_1 should choose the distribution p(M) by maximum entropy consistent with (20-25).

All the formalism of the maximum-entropy inference developed in Lecture 10 now becomes available to E_1. His distribution p(M) will have the form

$$\log p(M) + \lambda_0 + \lambda_1 N_1(M) + \lambda_2 N_2(M) + \ldots + \lambda_a N_a(M) = 0 \qquad (20\text{-}26)$$

and in order to evaluate the Lagrangian multipliers λ_i, he will use the partition function

$$Z(\lambda_1 \ldots \lambda_a) = \sum_M \exp[-\lambda_1 N_1(M) - \ldots - \lambda_a N_a(M)] = z^N \qquad (20\text{-}27)$$

where

$$z = e^{-\lambda_1} + \ldots + e^{-\lambda_a} \; . \qquad (20\text{-}28)$$

From (20-25) and the general relation

$$\langle N_i \rangle = - \frac{\partial}{\partial \lambda_i} \log Z(\lambda_1 \ldots \lambda_a) \qquad (20\text{-}29)$$

we find

$$\lambda_i = - \log(z f_i) \; , \qquad 1 \leq i \leq a \qquad (20\text{-}30)$$

and, substituting back into (20-26), we find the distribution which describes E_1's state of knowledge is just the multinomial distribution:

$$p(M) = f_1^{N_1} f_2^{N_2} \ldots f_a^{N_a} \qquad (20\text{-}26a)$$

which is a special case of an exchangeable sequence; the probability of any particular message depends only on how many times the letters A_1, A_2, ... appear, not on their order. The number of different messages possible for specified N_i is just the multinomial coefficient

$$\frac{N!}{N_1! \ldots N_a!} \; .$$

The entropy per symbol of the distribution (20-26a) is

$$H_1 = \frac{S}{N} = -\frac{1}{N}\sum_{M} p(M)\ \log\ p(M) = \frac{\log Z}{N} + \sum_{i=1}^{a} \lambda_i f_i$$

$$= -\sum_{i=1}^{a} f_i\ \log\ f_i \qquad (20\text{-}31)$$

Having found the assignment p(M), he can encode into binary digits in the most efficient way by a method found independently by R. M. Fano and C. E. Shannon (1948, Sec. 9). Arrange the messages in order of decreasing probability, and by a cut separate them into two classes so the total probability of all messages to the left of the cut is as nearly as possible equal to the probability of messages to the right. If a given message falls in the left class, the first binary digit in its code is 0; if in the right, 1. By a similar division of these classes into subclasses with as nearly as possible a total probability of 1/4, we determine the second binary digit, etc. I leave it for you to prove that (1) the expected number of binary digits required to transmit the message is numerically equal to H_1, when expressed in bits, and (2) in order to transmit at a rate of n of the original message symbols per second, E_1 requires a channel capacity $C \geq nH_1$, a result first given by Shannon.

The preceding mathematical steps are so well-known that they might be called trivial. However, the rationale which we have given them differs essentially from that of conventional treatments, and in that difference lies the main point of this section. Conventionally, one would use the frequency definition of probability, and say that E_1's probability assignment p(M) is the one resulting from the assumption that there are no intersymbol influences. Such a manner of speaking carries a connotation that the assumption might or might not be correct, and that its correctness must be demonstrated if the resulting design is to be justified; i.e. that the resulting

encoding rules might not be satisfactory if there are in fact intersymbol influences.

On the other hand, I contend that the probability assignment (20-26) is not an assumption at all, but the exact opposite. Eq. (20-26) represents, in a certain naive sense which I want to come back to later, the complete absence of any assumption on the part of E_1, beyond specification of expected single-letter frequencies, and it is uniquely determined by this property. The design based on (20-26) is the safest one possible on his state of knowledge. By that I mean the following. If, in fact, strong intersymbol influences do exist unknown to E_1, his encoding system will still be able to handle the messages perfectly well. If he had been given this additional information about intersymbol influences, he could have used it to arrive at an encoding system which would be still more efficient (i.e. would require a smaller channel capacity), as long as messages with only the specified type of correlation were transmitted. But if the type of intersymbol influence in the messages were suddenly to change, this new encoding system would likely become worse than the original one.

20.4. Better Encoding From Knowledge of Digram Frequencies.

Here is a rather long mathematical derivation which has, however, useful applications outside the particular problem at hand. Consider a second engineer, E_2. He has a set of numbers f_{ij}, $1 \leq i \leq a$, $1 \leq j \leq a$, which represent the expected relative frequencies of the digrams A_iA_j. E_2 will assign message probabilities $p(M)$ so as to agree with his state of knowledge,

$$\langle N_{ij} \rangle = \sum_M N_{ij}(M)\ p(M) = (N-1)\ f_{ij} \qquad (20\text{-}32)$$

and in order to avoid any further assumptions which are as likely to hurt as to help as far as he knows, he will determine the distribution $p(M)$ which has maximum entropy subject to this constraint. The problem is solved if he

can evaluate the partition function

$$Z(\lambda_{ij}) = \sum_{M} \exp\left[- \sum_{i,j=1}^{a} \lambda_{ij} N_{ij}(M) \right] \quad . \tag{20-33}$$

This can be done by solving the combinatorial problem of the number of different messages with given N_{ij}, or by observing that (20-33) can be written in the form of a matrix product:

$$Z = \sum_{i,j=1}^{a} \left(Q^{N-1} \right)_{ij} \tag{20-34}$$

where the matrix Q is defined by

$$Q_{ij} = e^{-\lambda_{ij}} \tag{20-35}$$

The result can be simplified formally if we suppose that the message $A_{i_1} \ldots A_{i_N}$ is always terminated by repetition of the first symbol A_{i_1}, so that it becomes $A_{i_1} \ldots A_{i_N} A_{i_1}$. The digram $A_{i_N} A_{i_1}$ is added to the message and an extra factor $\exp(-\lambda_{ij})$ appears in (20-33). The modified partition function then becomes a trace:

$$Z' = \mathrm{Tr}(Q^N) = \sum_{k=1}^{a} q_k^{\,N} \tag{20-36}$$

where the q_k are the roots of $\left| Q_{ij} - q\delta_{ij} \right| = 0$. This simplification would be termed "use of periodic boundary conditions" by the physicist. Clearly, the modification leads to no difference in the limit of long messages; as $N \to \infty$,

$$\lim \frac{1}{N} \log Z = \lim \frac{1}{N} \log Z' = \log q_{\max} \tag{20-37}$$

where $q_{\max}$ is the greatest eigenvalue of Q.

The probability of a particular message is now a special case of (10-28):

$$p(M) = \frac{1}{Z} \exp\left[- \sum \lambda_{ij} N_{ij}(M) \right] \tag{20-38}$$

which yields the entropy as a special case of (10-34):

$$S = - \sum_{M} p(M) \log p(M)$$

$$= \log Z + \sum_{ij} \lambda_{ij} \langle N_{ij} \rangle \tag{20-39}$$

In view of (20-32) and (20-37), Mr. E_2's entropy per symbol reduces, in the limit $N \to \infty$, to

$$H_2 = \frac{S}{N} = \log q_{max} + \sum_{ij} \lambda_{ij} f_{ij} \tag{20-40}$$

or, since $\sum_{ij} f_{ij} = 1$, we can write (20-40) as

$$H_2 = \sum_{ij} f_{ij} (\log q_{max} + \lambda_{ij})$$

$$= \sum_{ij} f_{ij} \log \left(\frac{q_{max}}{Q_{ij}} \right) \tag{20-41}$$

Thus, to calculate the entropy we do not need q_{max} as a function of the λ_{ij} (which would be impractical for a > 3), but we need find only the ratio q_{max}/Q_{ij} as a function of the f_{ij}.

To do this, we first introduce the characteristic polynomial of the matrix Q:

$$D(q) \equiv \det(Q_{ij} - q\delta_{ij}) \tag{20-42}$$

and note, for later purposes, some well-known properties of determinants (Bocher, 1907, pp. 31-33). The first is

$$D(q)\delta_{ik} = \sum_{j=1}^{a} M_{ij} (Q_{kj} - q\delta_{kj})$$

$$= \sum_{j} M_{ij} Q_{kj} - qM_{ik} \tag{20-43a}$$

and similarly,

$$D(q)\delta_{ik} = \sum_{j} M_{ji} Q_{jk} - qM_{ki} \tag{20-43b}$$

in which M_{ij} is the cofactor of $(Q_{ij} - q\delta_{ij})$ in the determinant D(q); i.e. $(-)^{i+j} M_{ij}$ is the determinant of the matrix formed by striking out the i'th row and j'th column of the matrix (Q - ql). If q is any eigenvalue of Q,

the expressions (20-43) vanish for all choices of i and k.

The second identity applies only when q is an eigenvalue of Q. In this case, all minors of the matrix M are known to vanish. In particular, the second order minors are

$$M_{ik}M_{jl} - M_{il}M_{jk} = 0 \quad , \text{ if } D(q) = 0. \tag{20-44a}$$

This implies that the ratios (M_{ik}/M_{jk}) and (M_{ki}/M_{kj}) are independent of k; i.e. that M_{ij} must have the form

$$M_{ij} = a_i b_j \quad , \quad \text{if } D(q) = 0 \ . \tag{20-44b}$$

Substitution into (20-43a) and (20-43b) then shows that the quantities b_j form a <u>right eigenvector</u> of Q, while a_i is a <u>left eigenvector</u>:

$$\sum_j Q_{kj} b_j = q b_k \quad , \quad \text{if } D(q) = 0 \tag{20-43c}$$

$$\sum_i a_i Q_{ik} = q a_k \quad , \quad \text{if } D(q) = 0 \tag{20-43d}$$

Suppose now that any eigenvalue q of Q is expressed as an explicit function $q(\lambda_{11},\lambda_{12},\ldots,\lambda_{aa})$ of the Lagrangian multipliers λ_{ij}. Then, varying a particular λ_{kl} while keeping the other λ_{ij} fixed, q will vary so as to keep D(q) identically zero. By the rule for differentiating the determinant (20-42), this gives

$$\frac{dD}{d\lambda_{kl}} = \frac{\partial D}{\partial \lambda_{kl}} + \frac{\partial D}{\partial q}\frac{\partial q}{\partial \lambda_{kl}}$$

$$= - M_{kl} Q_{kl} - \frac{\partial q}{\partial \lambda_{kl}} \mathrm{Tr}(M) = 0 \tag{20-45}$$

where

$$\mathrm{Tr}(M) \equiv \sum_{i=1}^{a} M_{ii} \tag{20-46}$$

is the trace, or diagonal sum, of the matrix M.

Using this relation, the condition (20-32) fixing the Lagrangian multipliers λ_{ij} in terms of the prescribed digram frequencies f_{ij}, becomes

$$f_{ij} = -\frac{\partial}{\partial \lambda_{ij}} \log q_{max} = \frac{M_{ij} Q_{ij}}{q_{max} Tr(M)} \tag{20-47}$$

The single-letter frequencies are proportional to the diagonal elements of M:

$$f_i = \sum_{j=1}^{a} f_{ij} = \frac{M_{ii}}{Tr(M)} \tag{20-48}$$

where we have used the fact that (20-43a) vanishes for $q = q_{max}$, $i = k$. Thus, from (20-47) and (20-48), the ratio needed in computing the entropy per symbol is

$$\frac{Q_{ij}}{q_{max}} = \frac{f_{ij}}{f_i} \frac{M_{ii}}{M_{ij}} = \frac{f_{ij}}{f_i} \frac{b_i}{b_j} \tag{20-49}$$

where we have used (20-44b). Substituting this into (20-41), we find that the terms involving b_i and b_j cancel out, and E_2's entropy per symbol is just

$$\begin{aligned} H_2 &= -\sum_{ij} f_{ij} \log \left(\frac{f_{ij}}{f_i} \right) \\ &= -\sum_{ij} f_{ij} \log f_{ij} + \sum_i f_i \log f_i \end{aligned} \tag{20-50}$$

This is never greater than E_1's H_1, for from (20-31), (20-50),

$$\begin{aligned} H_2 - H_1 &= \sum_{ij} f_{ij} \log \left(\frac{f_i f_j}{f_{ij}} \right) \\ &\leq \sum_{ij} f_{ij} \left[\frac{f_i f_j}{f_{ij}} - 1 \right] = 0 \end{aligned}$$

where we used the fact that $\log x \leq x - 1$ with equality if and only if $x = 1$. Therefore,

$$H_2 \leq H_1 \tag{20-51}$$

with equality if and only if $f_{ij} = f_i f_j$, in which case E_2's extra information was only what E_1 would have inferred. To see this, note that in the message

$M = \{i_1 \ldots i_N\}$, the number of times the digram $A_i A_j$ occurs is

$$N_{ij}(M) = \delta(i,i_1)\delta(j,i_2) + \delta(i,i_2)\delta(j,i_3) + \ldots + \delta(i,i_{N-1})\delta(j,i_N) \qquad (20\text{-}52)$$

and so, if we ask E_1 to estimate the frequency of digram $A_i A_j$, by the criterion of minimizing the expected square of the error, he will make the estimate

$$\langle f_{ij} \rangle = \frac{\langle N_{ij} \rangle}{N-1} = \frac{1}{N-1} \sum_M p(M)\, N_{ij}(M) = f_i f_j \qquad (20\text{-}53)$$

using for $p(M)$ the distribution (20-26a) of E_1. In fact, the distributions $p(M)$ found by E_1 and E_2 are identical if $f_{ij} = f_i f_j$, for then we have from (20-47), (20-48), and (20-44b),

$$Q_{ij} = e^{-\lambda_{ij}} = q_{max}^{\sqrt{f_i f_j}} \quad . \qquad (20\text{-}54)$$

Using (20-37), (20-52), and (20-54), we find that E_2's distribution (20-38) reduces to (20-26a). This is a rather nontrivial example of what we noted in Lecture 10, Eq. (10-76).

20.5. Relation to a Stochastic Model.

The quantities introduced above acquire a deeper meaning in terms of the following problem. Suppose that part of the message has been received, what can Mr. E_2 then say about the remainder of the message? This is answered by recalling our Rule 1:

$$(AB|X) = (A|BX)(B|X)$$

or, the conditional probability of A, given B, is

$$(A|BX) = \frac{(AB|X)}{(B|X)} \qquad (20\text{-}55)$$

a relation which in conventional theory, which does not use X, is taken as the definition of a conditional probability (i.e., as a ratio of two "absolute" probabilities). In our case, let X stand for the general statement of the

problem leading to the solution (20-38), and let

$$B \equiv \text{"The first } (m-1) \text{ symbols are } \{i_1 i_2 \ldots i_{m-1}\}."$$

$$A \equiv \text{"The remainder of the message is } \{i_m \ldots i_N\}."$$

Then $(AB|X)$ is the same as $p(M)$ in (20-38). Using (20-52), this reduces to

$$(AB|X) = (i_1 \ldots i_N|X) = Z^{-1} Q_{i_1 i_2} Q_{i_2 i_3} \ldots Q_{i_{N-1} i_N} \tag{20-56}$$

and in

$$(B|X) = \sum_{i_m=1}^{a} \ldots \sum_{i_N=1}^{a} (i_1 \ldots i_N|X) \tag{20-57}$$

the sum generates a power of the matrix Q, just as in the partition function (20-34). Writing, for brevity, $i_{m-1} = i$, $i_m = j$, $i_N = k$, and

$$R \equiv \frac{1}{Z} Q_{i_1 i_2} \ldots Q_{i_{m-2} i_{m-1}} , \tag{20-58}$$

we have

$$(B|X) = R \sum_{k=1}^{a} (Q^{N-m+1})_{ik} = R \sum_{j,k=1}^{a} Q_{ij} (Q^{N-m})_{jk} \tag{20-59}$$

and so

$$(A|BX) = \frac{Q_{ij} Q_{i_m i_{m+1}} \ldots Q_{i_{N-1} i_N}}{\sum_{k=1}^{a} (Q^{N-m+1})_{ik}} \tag{20-60}$$

since all the Q's contained in R cancel out, we see that the probabilities for the remainder $\{i_m \ldots i_N\}$ of the message depend only on the immediately ppeceding symbol A_i, and not on any other details of B. This property defines a Markov Chain. There is a huge literature dealing with them; it is perhaps the most thoroughly worked out branch of probability theory. The basic tool, from which essentially all else follows, is the matrix p_{ij} of "elementary transistion probabilities." This is the probability $p_{ij} = (A_j|A_i X)$ that the next symbol will be A_j, given that the last one was A_i. Summing (20-60) over $i_{m+1} \ldots i_N$, we find

$$P_{ij}^{(N)} = (A_j | A_i X) = \frac{Q_{ij} - T_j}{\sum_j Q_{ij} T_j} \tag{20-61}$$

where

$$T_j \equiv \sum_{k=1}^{a} (Q^{N-m})_{jk} \tag{20-62}$$

The fact that T_j depends on N and m is an interesting feature. Usually, one considers from the start a chain indefinitely prolonged, and so it is only the limit of (20-61) for $N \to \infty$ that is ever considered. This example shows that prior knowledge of how long the chain is going to be can affect the transition probabilities; however, the limiting case is clearly of greatest interest.

To find this limit we need a little more matrix theory. The equation $D(q) = \det(Q_{ij} - q\delta_{ij}) = 0$ has a roots $(q_1 q_2 \ldots q_a)$, not necessarily all different, or real. Label them so that $|q_1| \geq |q_2| \geq \ldots \geq |q_a|$. There exists a nonsingular matrix A such that $A\,Q\,A^{-1}$ takes the canonical "superdiagonal" form:

$$A\,Q\,A^{-1} = \bar{Q} = \begin{pmatrix} C_1 & 0 & 0 & \ldots \\ 0 & C_2 & 0 & \ldots \\ 0 & 0 & C_3 & \ldots \\ \vdots & \vdots & \vdots & \\ \cdot & \cdot & \cdot & C_m \end{pmatrix} \tag{20-63}$$

where the C_i are sub-matrices which can have either of the forms

$$C_i = \begin{pmatrix} q_i & 1 & 0 & 0 & \ldots & & \\ 0 & q_i & 1 & 0 & \ldots & & \\ 0 & 0 & q_i & 1 & \ldots & & \\ \cdot & \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & q_i & 1 \\ \cdot & \cdot & \cdot & \cdot & & & \\ & & & & & 0 & q_i \end{pmatrix} \quad \text{or,} \quad C_i = \begin{pmatrix} q_i & & & \\ & q_i & & \\ & & \ddots & \\ & & & q_i \end{pmatrix} \tag{20-64}$$

The result of raising Q to the n'th power is

$$Q^n = A \bar{Q}^n A^{-1} \tag{20-65}$$

and as $n \to \infty$, the elements of $\bar{Q}^n$ arising from the greatest eigenvalue $q_{max} = q_1$ become arbitrarily large compared to all others. If q_1 is nondegenerate, so that it appears only in the first row and column of $\bar{Q}$, we have

$$\lim_{N\to\infty} \frac{T_j}{q_1^{N-m}} = A_{j1} \sum_{k=1}^{a} (A^{-1})_{1k} \quad , \tag{20-66}$$

$$\lim_{N\to\infty} \frac{T_j}{\sum_j Q_{ij} T_j} = \frac{A_{j1}}{q_1 A_{i1}} \quad , \tag{20-67}$$

and the limiting transition probabilities are

$$p_{ij}^{(\infty)} = \frac{Q_{ij}}{q_1} \frac{A_{j1}}{A_{i1}} = \frac{Q_{ij}}{q_1} \frac{M_{ij}}{M_{ii}} \tag{20-68}$$

where we have used the fact that the elements A_{j1} (j = 1, 2, ..., a) form an eigenvector of Q with eigenvalue $q_1 = q_{max}$, so that, referring to (20-44b), (20-44c), $A_{j1} = Kb_j$ where K is some constant. Using (20-47), (20-48), we have finally,

$$p_{ij}^{(\infty)} = \frac{f_{ij}}{f_i} \tag{20-69}$$

which is just what would be taken, on the frequency theory, as the <u>definition</u> of the transition probability.