CHAPTER 27

INTRODUCTION TO COMMUNICATION THEORY

We noted in Chapter 11 that one of the motivations behind this work was the attempt to see Gibbsian statistical mechanics and Shannon's communication theory as examples of the same line of reasoning. A generalized form of statistical mechanics appeared as soon as we introduced 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 macroscopic quantities such as pressure), and so we have little temptation to commit errors. But in communication theory the prior information consists, typically of frequencies; and this makes the probability–frequency conceptual pitfalls much more acute. For this reason it seemed best to take up communication theory only after we had seen the general connections between probability and frequency, in a variety of conceptually simpler applications.

Origins of the Theory

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, Lorentz and Poincaré, 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. In both cases 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 the problems of the field, and left nothing else for others to do.

The post–Shannon developments, with few exceptions, can be classed into efforts in two entirely different directions. On the applications side we have the Expansionists (who try to apply Shannon's ideas to other fields, as we do here), 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 then finds that nobody knows what to do with it), and the Universalist (who assures us that Shannon's work will revolutionize all intellectual activity; but is unable to offer a specific example of anything that has been changed by it).

We should not be overly critical of these efforts because, as J. R. Pierce has remarked, it is very hard to tell at first which ones make sense, which are pure nonsense, and which are the beginning of something that will in time make sense. The writer's efforts have received all three classifications from various quarters. We expect that, eventually, the ideas introduced by Shannon will be indispensable to the linguist, the geneticist, the television engineer, the neurologist, the economist. But we share with many others a feeling of disappointment that 40 years of effort along these lines has led to so little in the way of really useful advances in these fields.

During this time there has been an overabundance of vague philosophy, and of abstract mathematics; but outside of coding theory a rather embarrassing shortage of examples where specific real problems have been solved by using this theory. We believe that the reason for this is that conceptual misunderstandings, almost all of which amount to the Mind Projection Fallacy, have prevented workers from asking the right questions. In order to apply communication theory to other problems than coding, the first and hardest step is to state precisely *what is the specific problem that we want to solve*?

In almost diametric opposition to the above efforts, as far as aim was concerned, were the mathematicians, who viewed communication theory simply as a branch of pure mathematics. Characteristic of this school was 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 was to make the theorems rigorous *by the criteria of rigor then fashionable among mathematicians*, even if in so doing we limit their scope for applications. The book on information theory by A. I. Khinchin (1957) can serve as a typical example of the style prevalent in this literature.

Here again, severe criticism of these efforts is not called for. Of course, we want our principles to be subjected to the closest scrutiny the human mind can bring to bear on them; if important applications exist, the need for this is so much the greater. However, the present work is not addressed to mathematicians, but to persons concerned with real applications. So we shall dwell on this side of the story only to the extent of pointing out that the rigorized theorems are not the ones relevant to problems of the real world. Typically, they refer *only* to situations that do not exist (such as infinitely long messages) and as a result they degenerate into "nonsense theorems" which assign probability 1 to an impossible event, therefore zero to all possible events. We have no way of using such results, because our probabilities are always conditional on our knowledge of the real world. Now let's turn to some of the specific things in Shannon's papers.

The Noiseless Channel.

We deal with the transmission of information from some sender to some receiver. We shall 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 message it sends, but rather by what it *could have* sent. The usefulness of a channel lies in its readiness 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 a time t. In all cases of interest to us, this number goes eventually into an exponential increase for sufficiently large t: $W(t) \propto \exp\{Ct\}$, 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

$$C \equiv \lim_{t\to\infty} \left[\frac{1}{t} \log W(t)\right] . \tag{27–1}$$

The units in which C is measured will depend on which base we choose for our logarithms. Usually one takes base 2, in which 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 may be easier to use the base e of natural logarithms. Our channel capacities are then measured in natural units, or "nits per second." To convert, note that 1 bit = $(\ln 2) = 0.69315$ nits, or 1 nit = 1.4427 bits.

The capacity of a noiseless channel is a definite number, characteristic of the channel, which has nothing to do with human information. 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 $C = n \log_2 a$ bits/second $= n \log_e a$ 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

$$W(t) = \frac{N!}{(Nf_1)! \ldots (Nf_a)!} \tag{27–2}$$

and from Stirling's approximation, as we found in Chapter 11,

$$C = -n \sum_i f_i \log f_i \quad \text{nits/second.} \tag{27–3}$$

This attains its maximum value, equal to the previous $C = n \log a$, in the case of equal frequencies, $f_i = 1/a$. 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 matters, and this grows less rapidly than t, so it makes no difference in the limit. In view of the entropy concentration theorem of Chapter 11, this can be understood in another way: the vast majority of all *possible* messages are ones in which the letter frequencies are nearly equal.

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 case of messages in which letter A_i occurs n_i times, $i = 1, 2, \ldots, a$. The number of such messages is

$$W(n_1 \ldots n_a) = \frac{N!}{n_1! \ldots n_a!} \tag{27–4}$$

where

$$N = \sum_{i=1}^{a} n_i . \tag{27–5}$$

The total number of different messages that could have been transmitted in time t is then

$$W(t) = \sum_{n_i} W(n_1 \ldots n_a) \tag{27–6}$$

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

$$\sum_{i=1}^{a} n_i t_i \leq t. \tag{27–7}$$

The number $K(t)$ of terms in the sum (27–6) 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 meaning of $K(t)$ as the volume of a simplex.

Exact evaluation of (27–6) 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 (27–6) nor greater than $W_m K(t)$:

$$\log W_m \leq \log W(t) \leq \log W_m + a \log(Bt) \tag{27–8}$$

and so we have

$$C \equiv \lim_{t\to\infty} \frac{1}{t}\log W(t) = \lim_{t\to\infty} \frac{1}{t}\log W_m \tag{27–9}$$

i.e., to find the channel capacity, it is sufficient to maximize $\log W(n_1 \ldots n_a)$ subject to the constraint (27–7). This rather surprising fact can be understood as follows. The logarithm of $W(t)$ is given, crudely, by $\log W(t) = \log W_{max} + \log$ [number of reasonably large terms in (27–6)]. 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}. As explained by Schrödinger (1948), this same mathematical fact is the reason why, in statistical mechanics, the Darwin–Fowler method and the method of 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 Chapter 11. 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

$$\log W(n_1 \ldots n_a) \approx N\log N - \sum_{i=1}^{a} n_i \log n_i. \tag{27–10}$$

The variational problem, with λ a Lagrangian multiplier, is

$$\delta[\log W + \lambda \sum n_i t_i] = 0 \tag{27–11}$$

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

$$\begin{aligned} \delta \log W &= \delta N \log N - \delta N - \sum_i (\delta n_i \log n_i - \delta n_i) \\ &= -\sum \delta n_i \log(n_i/N). \end{aligned} \tag{27–12}$$

Therefore (27–11) reduces to

$$\sum_{i=1}^{a} [\log(n_i/N) + \lambda t_i]\delta n_i = 0 \tag{27–12}$$

with the solution

$$n_i = N e^{-\lambda t_i}. \tag{27–14}$$

To fix the value of λ we require

$$N = \sum n_i = N \sum e^{-\lambda t_i}. \tag{27–15}$$

With this choice of n_i, we find

$$\frac{1}{t}\log W_m = -\frac{1}{t}\log(n_i/N) = \frac{1}{t}\sum n_i(\lambda t_i). \tag{27–16}$$

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

$$C = \lim_{t\to\infty} \frac{1}{t}\log W(t) = \lambda. \tag{27–17}$$

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 the partition function $Z(\lambda) \equiv \sum_i e^{-\lambda t_i}$. Then the channel capacity C is the real root of*

$$Z(\lambda) = 1 \tag{27–18}$$

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 (27–15) 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

$$f_i = \frac{n_i}{N} = e^{-\lambda t_i} = e^{-C t_i}. \tag{27–19}$$

Some have concluded from this that the channel is being "used most efficiently" when we have encoded our messages so that (27–19) holds. But that would be quite mistaken 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 (27–19) tells us only that – in accordance with the entropy concentration theorem – 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 (27–74); if we impose an additional constraint requiring that the relative frequencies are given by (27–19), 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 (27–19) 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 amount which the sender can choose. In a practical communication 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; but also it has minimum perimeter for a given area. In statistical mechanics, the canonical distribution can be characterized as 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 (27–18) gives the maximum attainable W for a given transmission time, or equally well the minimum attainable transmission time for a fixed W.

As another extension of the meaning of these equations, note that we need not 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. Perhaps 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 criterion might be, for example, the amount of energy that a space probe must dissipate in transmitting a message back to earth. In this case, we could define t_i as the energy required to transmit the i'th symbol. The channel capacity given by Equation (27–18) would then be measured, not in bits per second but in bits per joule, and its reciprocal is equal to the minimum attainable number of joules needed per bit 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 the 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 the same difference equation that we used to introduce the partition function in Chapter 8:

$$W_i(t) = \sum_{jn} W_j(t - t_{inj}) \tag{27–20}$$

where the sum is over all possible sequences $S_i \to A_n \to S_j$. As before, this is a linear difference equation with constant coefficients, so its asymptotic solution must be an exponential function:

$$W_i(t) \approx B_i \exp(Ct) \tag{27–21}$$

and from the definition (27–1) it is clear that, for finite k, the coefficient C is the channel capacity. Substituting (27–21) into (27–20), we obtain

$$B_i = \sum_{j=1}^{k} Z_{ij}(C) B_j \tag{27–22}$$

where

$$Z_{ij}(\lambda) = \sum_n \exp(-\lambda t_{inj}) \tag{27–23}$$

is the "partition matrix." Compare this argument with our first derivation of a partition function in Chapter 8. If the sequence $S_i \to A_n \to S_i$ is impossible, we set $t_{inj} = \infty$. By this device we can understand the sum in (27–23) as extending over all symbols in the alphabet.

Equation (27–22) says that the matrix Z_{ij} has an eigenvalue equal to unity. Thus, the channel capacity is simply the greatest real root of $D(\lambda) = 0$, where

$$D(\lambda) \equiv \det[Z_{ij}(\lambda) - \delta_{ij}]. \tag{27–24}$$

This is one of the prettiest results given by Shannon. In the case of a single state, $k = 1$, it reduces to the previous rule, Equation (27–18).

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 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 (27–1). For some weird channels, you might find that the limit therein does not exist, in which case we cannot speak of a channel capacity, but have to characterize the channel simply by giving the function $W(t)$.

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 we take up briefly in a later Chapter. This work was an essential step in developing a way of thinking which led to 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 p_i that it will emit various messages M_i. Already we see the 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 p_i 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_i occurs in some imaginary "ensemble" of communication acts? Well, it's all right to state it that way if you want to, but it doesn't answer the question. It merely leads us to restate the question as: what defines that ensemble? How is it to be set up? Calling it by a different name doesn't help us. *What* information is that entropy $H = -\sum p_i \log p_i$ really measuring?

We take a halting first step toward answering this if we suppose that Shannon's H measures, not the information of the sender, but the ignorance of the receiver, that is removed by receipt of the message. Indeed, most later commentators make this interpretation. Yet on second thought, this does not make sense either; for Shannon proceeds to develop theorems relating H to the channel capacity C required to transmit the messages M_i. But how well a channel can transmit messages obviously depends on properties of the channel and the messages; and not at all on the state of ignorance of the receiver! You see the conceptual mess that the field has been in, for 40 years.

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 or the one at the receiving end; and Shannon offers us no explicit help on this. But implicitly, the answer seems to be clear; in view of the theorems he gives, he cannot be describing the "general philosophy" of communication between sender and receiver, as so many have supposed. He is thinking of the theory 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 Shannon is describing is that of the communication engineer when he designs the equipment.* It is his ignorance about the messages to be sent that is measured by H.

Although this viewpoint would seem perfectly natural for an engineer employed by the Bell Telephone Laboratories, as Shannon was at the time, you will not find it actually expressed in his words, or in the later literature based on the viewpoint which sees no distinction between probability and frequency. For on the frequentist view, the notion of a probability *for a person with a certain state of knowledge* simply doesn't exist, because probability is thought to be a real physical phenomenon which exists independently of human information. But the problem of choosing some probability distribution to represent the information source still does exist; it cannot be evaded. It is now clear that the whole content of the theory depends on how we do this.

We 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 a great deal of practical judgment. If our model does not correspond well to the actual situation then our theorems, however rigorous the mathematicians may have made them, can be more misleading than helpful. This is so with a vengeance in communication theory, because 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 what messages his communication system may be called upon to send. In principle, this prior knowledge could be of any sort; in particular, nothing prevents it from being semantic in nature. For example, he might know in advance that the channel will be used only to transmit stock market quotations; and not quotations from the Bible or obscene limericks. That is a perfectly valid kind of prior information, which would have definite implications for the probabilities p_i by restricting the sample space in definite, specific ways, although they might be hard to state in general mathematical terms.

We stress this point because some critics harp away incessantly on the theme that information theory does not consider semantic meaning, and hold this to be a basic defect of our whole philosophy. They could not be more mistaken; the issue of semantic meaning is not a philosophical one but a technical one. The only reason why we do not consider semantic meaning is that we do not know how to do it *as a general procedure*, although we could certainly do it "by hand" in the context of a specific, finite set of possible messages. Probably all of us have tried to restore some corrupted text by drawing upon our perception of its semantic meaning; but how do you teach a computer to do this?.

So let us assure those critics: if you will show us *a definite, usable algorithm* for assessing semantic meaning, we are most eager to incorporate this too into information theory. In fact, our present inability to do this is a serious handicap in many applications, from image restoration to pattern recognition to artificial intelligence. We need your constructive help, not your criticisms.

But in traditional Shannon–type communication theory the only kind of prior knowledge considered is "statistical" because this is amenable to mathematical treatment at once. That is, it consists of frequencies of letters, or combinations of letters, which have been observed in *past* samples of similar messages. Then a typical practical problem – indeed, the actual problem of writers of those popular text compression computer programs – is to design encoding systems which will transmit binary digits representing English text, reliably and at the maximum possible rate, given an available channel with known properties. This would be also the actual problem of designers of computer hardware like disk drives and modems, if they became a little more sophisticated. The designer will then, according to the usual viewpoint, need accurate data giving the correct frequencies of English text. Let's think about that a little.

Does the English Language have Statistical Properties?

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 many years ago when linotype machines first came into use (an inexperienced operator, who allowed his fingers to brush lightly across the keys, automatically set this in type). But already we are getting into trouble, because there does not seem to be complete agreement even as to the relative order of 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 encyclopedia might have a higher relative frequency of 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 writing 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 a tape recording of a lecture will probably be noticeably different from those one would produce if he 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 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% 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 is a thing which is not a characteristic property of the Greek language, seems rather 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 some 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 writers, 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 which occurred about 1200 B. C., some four centuries before Homer. The decipherment of Minoan Linear B script by Michel 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.

The considerations of the last two paragraphs still suggest an evolutionary development. You see that the question is very complex and far from settled; but we find it fascinating that a statistical analysis of word and syllable frequencies, representing evidence which has been there in the Iliad for some 28 Centuries for anyone who had the wit to extract it, is finally recognized as having a definite bearing on the problem.

Well, to get back to communication theory, the point we are 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 most 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 we are advocating, be seen as simple examples of maximum–entropy inference; *i.e.*, as examples of the same kind of reasoning as in statistical mechanics.

Optimum Encoding: Letter Frequencies Known.

Suppose our alphabet consists of 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 $\sum_M$ we mean a sum over all of them. Also, define

$$N_j(M) \equiv \text{(number of times the letter } A_j \text{ appears in message } M)$$
$$N_{ij}(M) \equiv \text{(number of times the digram } A_i A_j \text{ appears in } M)$$

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_j, 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 probability $p(M)$ which E_1 assigns to the various conceivable messages. Now Mr. E_1 has no deductive proof that the letter frequencies in the 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 *expected* frequencies equal to the known past ones. Put differently, if we say that our distribution $p(M)$ "contains" certain information, we mean that that information can be extracted back out of it by the usual rule of estimation. In other words, E_1 will impose the constraints

$$\langle N_i \rangle = \sum_M N_i(M) p(M) = N f_i, \quad i = 1, 2, \ldots, a. \tag{27–25}$$

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.

We emphasize again that it makes no sense to say there exists any "physical" or "objective" probability 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, but we still want it to be transmitted as quickly and reliably as possible, whatever that message turns out to be (perhaps we know that the system will be destroyed by impact on Ganymede immediately afterward); thus there is no conceivable way in 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 $A_1 A_2$ is twice as likely as $A_2 A_3$. But from the standpoint of E_1 this could not be justified, for *as far as he knows,* 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 (27–25).

All the formalism of the maximum–entropy inference developed in Chapter 11 now becomes available to E_1. His distribution $p(M)$ will have the form

$$\log p(M) + \lambda_0 + \lambda_1 N_1(M) + \lambda_2 N_2 + \ldots + \lambda_a N_a(M) = 0 \tag{27–26}$$

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

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

where

$$z \equiv^{-\lambda_1} + \ldots + e^{-\lambda_a}. \tag{27–28}$$

From (27–25) and the general relation

$$\langle N_i \rangle = -\frac{\partial}{\partial \lambda_i} \log Z(\lambda_1 \ldots \lambda_a) \tag{27–29}$$

we find

$$\lambda_i = -\log(z f_i), \quad 1 \le i \le a \tag{27–30}$$

and, substituting back into (27–26), we find the distribution which describes E_1's state of knowledge is just the multinomial distribution:

$$p(M) = f_1^{N_1} f_2^{N_2} \ldots f_a^{N_a} \tag{27–31}$$

which is a special case of an exchangeable sequence; the probability of any particular message depends only on how many times the letters $A_1, A_2, \ldots$ appear, not on their order. The result (27–31) is correctly normalized, $\sum_M p(M) = 1$, as we see from the fact that 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 (27–31) is

$$H_1 = -\frac{1}{N} \sum_M p(M) \log p(M) = \frac{\log Z}{N} + \sum_{i=1}^{a} \lambda_i f_i = -\sum_{i=1}^{a} f_i \log f_i. \tag{27–32}$$

Having found the assignment $p(M)$, he can encode into binary digits in the most efficient way by a method found independently by Shannon (1948, Sec. 9) and R. M. Fano. 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 the messages on 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. It is left for you to prove that (1) the expected number of binary digits required to transmit a symbol is 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 $C \ge nH_1$, 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 the implication 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 unknown to E_1.

On the other hand, we contend that the probability assignment (27–26) is not an assumption at all, but the opposite. Eq. (27–26) represents, in a certain naïve sense which we shall 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 that property. Because of this, the design based on (27–26) is the safest one possible on his state of knowledge.

By that we mean the following. If, in fact, strong intersymbol correlations *do* exist unknown to E_1 (for example, Q is always followed by U), his encoding system will still be able to handle the messages perfectly well, whatever the nature of those correlations. This is what we mean by saying that the present design is the most conservative one; that it assumes *nothing* about correlations does not mean that it assumes *no* correlations and will be in trouble if correlations are in fact present. On the contrary, it means that it is prepared in advance *for whatever kind of correlations might exist*; they will not cause any deterioration in performance. We stress this point because it was not noted by Shannon, and it does not seem to be comprehended in the more recent literature.

But if he had been given this additional information about some particular kind of correlations, he could have used it to arrive at a new 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 correlations in the messages were suddenly to change, this new encoding system would likely become worse than the one just found.

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}, $1 \le i \le a$, $1 \le j \le a$, which represent the expected relative frequencies of the digrams $A_i A_j$. E_2 will assign message probabilities $p(M)$ so as to agree with his state of knowledge,

$$\langle N_{ij} \rangle = \sum_M N_{ij}(M) p(M) = (N-1) f_{ij} \tag{27–33}$$

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 probability distribution over messages $p(M)$ which has maximum entropy subject to these constraints. The problem is solved if he can evaluate the partition function

$$Z(\lambda_{ij}) = \sum_M \exp\left[-\sum_{ij=1}^{a} \lambda_{ij} N_{ij}(M) \right] . \tag{27–34}$$

This can be done by solving the combinatorial problem of the number of different messages with given $\{N_{ij}\}$, or by observing that (27–34) can be written in the form of a matrix product:

$$Z = \sum_{ij=1}^{a} \left(Q^{N-1}\right)_{ij} \tag{27–35}$$

where the matrix Q is defined by

$$Q_{ij} \equiv e^{-\lambda_{ij}} . \tag{27–36}$$

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 (27–34). The modified partition function then becomes a trace:

$$Z' = \mathrm{Tr}(Q^N) = \sum_{k=1}^{a} q_k^N \tag{27–37}$$

where the q_k are the roots of $|Q_{ij} - q\delta_{ij}| = 0$. 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 $N \to \infty$,

$$\lim \frac{1}{N} \log Z = \lim \frac{1}{N} \log Z' = \log q_{max} \tag{27–38}$$

where q_{max} is the greatest eigenvalue of Q. The probability of a particular message is now a special case of (27–82):

$$p(M) = \frac{1}{Z} \exp[-\sum \lambda_{ij} N_{ij}(M)] \tag{27–39}$$

which yields the entropy as a special case of (27–82):

$$S = -\sum_M p(M) \log p(M) = \log Z + \sum_{ij} \lambda_{ij} \langle N_{ij} \rangle . \tag{27–40}$$

In view of (27–33) and (27–38), Mr. E_2's entropy per symbol reduces, in the limit $N \to \infty$, to

$$H_2 = \frac{S}{N} = \log q_{max} + \sum_{ij} \lambda_{ij} f_{ij} \tag{27–41}$$

or, since $\sum_{ij} f_{ij} = 1$, we can write (27–41) as

$$H_2 = \sum_{ij} f_{ij}(\log q_{max} + \lambda_{ij}) = \sum_{ij} f_{ij} \log\left(\frac{q_{max}}{Q_{ij}}\right). \tag{27–42}$$

Thus to calculate the entropy we do not need q_{max} as a function of the λ_{ij} (which would be impractical analytically 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:

$$D(q) \equiv \det(Q_{ij} - q\delta_{ij}) \tag{27–43}$$

and note, for later purposes, some well-known properties of determinants. The first is

$$D(q)\delta_{ik} = \sum_{j=1}^{a} M_{ij}(Q_{kj} - q\delta_{kj}) = \sum_j M_{ij}Q_{kj} - qM_{ik} \tag{27–44}$$

and similarly,

$$D(q)\delta_{ik} = \sum_j M_{ji}Q_{jk} - qM_{ki} \tag{27–45}$$

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 ith row and jth column of the matrix $(Q - q1)$. If q is any eigenvalue of Q, the expression (27–44) vanishes 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

$$M_{ik}M_{jl} - M_{il}M_{jk} = 0, \quad \text{if } D(q) = 0. \tag{27–46}$$

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

$$M_{ij} = a_i b_j, \quad \text{if } D(q) = 0. \tag{27–47}$$

Substitution into (27–44) and (27–47) then shows that the quantities b_j form the *right eigenvectors* of Q, while a_i is a *left eigenvector:*

$$\sum_j Q_{kj}b_j = qb_k, \quad \text{if } D(q) = 0 \tag{27–48}$$

$$\sum_i a_i Q_{ik} = a_k q, \quad \text{if } D(q) = 0. \tag{27–49}$$

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 (27–43), this gives

$$\frac{dD}{d\lambda_{kl}} = \frac{\partial D}{\partial \lambda_{kl}} + \frac{\partial D}{\partial q}\frac{\partial q}{\partial \lambda_{kl}} = -M_{kl}Q_{kl} - \frac{\partial q}{\partial \lambda_{kl}}\mathrm{Tr}(M) = 0 \tag{27–50}$$

Using this relation, the condition (27–33) fixing the Lagrangian multipliers λ_{ij} in terms of the prescribed digram frequencies f_{ij}, become

$$f_{ij} = -\frac{\partial}{\partial \lambda_{ij}} \log q_{max} = \frac{M_{ij}Q_{ij}}{q_{max}\mathrm{Tr}(M)}. \tag{27–51}$$

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

$$f_i = \sum_{j=1}^{a} f_{ij} = \frac{M_{ii}}{\mathrm{Tr}(M)} \tag{27–52}$$

where we have used the fact that (27–44) vanishes for $q = q_{max}$, $i = k$. Thus, from (27–51) and (27–52), the ratio needed in computing the entropy per symbol is

$$\frac{Q_{ij}}{q_{max}} = \frac{f_{ij}}{f_i}\frac{M_{ii}}{M_{ij}} = \frac{f_{ij}}{f_i}\frac{b_i}{b_j} \tag{27–53}$$

where we have used (27–47). Substituting this into (27–42), we find that the terms involving b_i and b_j cancel out, and E_2's entropy per symbol is just

$$H_2 = -\sum_{ij} f_{ij} \log\left(\frac{f_{ij}}{f_i}\right) = -\sum_{ij} f_{ij}\log f_{ij} + \sum_i f_i \log f_i . \tag{27–54}$$

This is never greater than E_1's H_1, for from (27–82), (27–54),

$$H_2 - H_1 = \sum_{ij} f_{ij} \log\left(\frac{f_i f_j}{f_{ij}}\right) \le \sum_{ij} f_{ij}\left[\frac{f_i f_j}{f_{ij}} - 1\right] = 0$$

where we used the fact that $\log x \le x - 1$ in $0 \le x < \infty$, with equality if and only if $x = 1$. Therefore,

$$H_2 \le H_1 \tag{27–55}$$

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

$$N_{ij}(M) = \delta(i,i_1)\delta(j,i_2) + \delta(i,i_2)\delta(i,i_3) + \cdots + \delta(i,i_{N-1})\delta(j,i_N) \tag{27–56}$$

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

$$\langle f_{ij}\rangle = \frac{\langle N_{ij}\rangle}{N-1} = \frac{1}{N-1}\sum_M p(M) N_{ij}(M) = f_i f_j \tag{27–57}$$

using for $p(M)$ the distribution (27–82) of E_1. In fact, the solutions found by E_1 and E_2 are identical if $f_{ij} = f_i f_j$, for then we have from (27–51), (27–52), and (27–47),

$$Q_{ij} = e^{-\lambda_{ij}} = q_{max}\sqrt{f_i f_j}. \tag{27–58}$$

Using (27–38), (27–56), and (27–58), we find that E_2's distribution (27–39) reduces to (27–82). This is a rather nontrivial example of what we noted in Chapter 11, Eq. (11–74).

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 product rule

$$p(AB|I) = p(A|BI)p(B|I)$$

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

$$p(A|BI) = \frac{p(AB|I)}{p(B|I)} \tag{27–59}$$

a relation which in conventional theory, which never mentions prior information I, is taken as the *definition* of a conditional probability (*i.e.*, the ratio of two "absolute" probabilities). In our case, let I stand for the general statement of the problem leading to the solution (27–39), and let

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

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

Then $p(AB|I)$ is the same as $p(M)$ in (27–39). Using (27–56), this reduces to

$$p(AB|I) = p(i_1 \ldots i_N|I) = Z^{-1} Q_{i_1 i_2} Q_{i_2 i_3} \cdots Q_{i_{N-1} i_N} \tag{27–60}$$

and in

$$p(B|I) = \sum_{i_m=1}^{a} \cdots \sum_{i_N=1}^{a} p(i_1 \ldots i_N|I) \tag{27–61}$$

the sum generates a power of the matrix Q, just as in the partition function (27–35). Writing, for brevity, $i_{m-1} = i$, $i_m = j$, $i_N = k$, and

$$R \equiv \frac{1}{Z} Q_{i_1 i_2} \cdots Q_{i_{m-2} i_{m-1}}, \tag{27–62}$$

we have

$$p(B|I) = R \sum_{k=1}^{a} (Q^{N+m+1})_{ik} = R \sum_{jk=1}^{a} Q_{ij} (Q^{N-m})_{jk} \tag{27–63}$$

and so

$$p(A|BI) = \frac{Q_{ij} Q_{i_m i_{m+1}} \cdots Q_{i_{N-1} i_N}}{\sum_{k=1}^{a} (Q^{N-m+1})_{ik}} \tag{27–64}$$

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 preceding symbol A_i, and not on any other details of B. This property defines a *generalized Markov Chain.* There is a huge literature dealing with them; it is perhaps the most thoroughly worked out branch of probability theory, and we used a rudimentary form of it in calculating the conditional sampling distributions in Chapter 3. The basic tool, from which essentially all else follows, is the matrix p_{ij} of "elementary transition probabilities." This is the probability $p_{ij} = p(A_j|A_i I)$ that the next symbol will be A_j, given that the last one was A_i. Summing (27–64) over $i_{m+1} \ldots i_N$, we find that for a chain of length N the transition probabilities are

$$p_{ij}^{(N)} = p(A_j|A_i I) = \frac{Q_{ij} - T_j}{\sum_k Q_{ik} T_k} \tag{27–65}$$

where

$$T_j \equiv \sum_{k=1}^{a} (Q^{N-m})_{jk}. \tag{27–66}$$

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 (27–65) for $N \to \infty$ that is ever considered. This example shows that prior knowledge of the length of the chain 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 $|q_1| \geq |q_q| \geq \ldots \geq |q_a|$. There exists a nonsingular matrix A such that AQA^{-1} takes the canonical "superdiagonal" form:

$$AQA^{-1} = \overline{Q} = \begin{pmatrix} C_1 & 0 & 0 & \cdots \\ 0 & C_2 & 0 & \cdots \\ 0 & 0 & C_3 & \cdots \\ \vdots & \vdots & \vdots & C_m \end{pmatrix} \tag{27–67}$$

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

$$C_i = \begin{pmatrix} q_i & 1 & 0 & 0 & \cdots \\ 0 & q_i & 1 & 0 & \cdots \\ 0 & 0 & q_i & 1 & \cdots \\ 0 & 0 & 0 & q_i & 1 \\ \vdots & \vdots & \vdots & 0 & q_i \end{pmatrix} \quad \text{or,} \quad C_i = \begin{pmatrix} q_i & & & \\ & q_i & & \\ & & \ddots & \\ & & & q_i \end{pmatrix}. \tag{27–68}$$

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

$$Q^n = A\overline{Q}^n A^{-1} \tag{27–69}$$

and as $n \to \infty$, the elements of $\overline{Q}^n$ 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 $\overline{Q}$, we have

$$\lim_{N\to\infty} \left[\frac{T_j}{q_1^{N-m}}\right] = A_{j1} \sum_{k=1}^{a} (A^{-1})_{1k}, \tag{27–70}$$

$$\lim_{N\to\infty} \left[\frac{T_j}{\sum_k Q_{ik} T_k}\right] = \frac{A_{j1}}{q_1 A_{i1}}, \tag{27–71}$$

and the limiting transition probabilities are

$$p_{ij}^{(\infty)} = \frac{Q_{ij}}{q_1}\frac{A_{j1}}{A_{i1}} = \frac{Q_{ij}}{q_1}\frac{M_{ij}}{M_{ii}} \tag{27–72}$$

where we have used the fact that the elements A_{j1} $(j = 1, 2, \ldots, a)$ from an eigenvector of Q with eigenvalue $q_1 = q_{max}$, so that, referring to (27–47), (27–82), $A_{j1} = Kb_j$ where K is some constant. Using (27–51), (27–52), we have finally,

$$p_{ij}^{(\infty)} = \frac{f_{ij}}{f_i}. \tag{27–73}$$

From this long calculation we learn many things. In the first place, for a sequence of finite length (the only kind that actually exists), the exact solution has intricate fine details that depend on the length. This, of course, could not be learned by those who try to jump directly into an infinite set at the beginning of a problem. Secondly, it is interesting that standard matrix theory was adequate to solve the problem completely. Finally, in the limit of infinitely long sequences the exact solution of the maximum–entropy problem does indeed go into the familiar Markov chain theory. This gives us a deeper insight into the basis of, and possible limitations on, Markov chain analysis.

Exercise 22-1. *The exact meaning of this last statement might be unclear; in a classical Markov chain the transition probabilities two steps down the chain would be given by the square of the one–step matrix p_{ij}, three steps by the cube of that matrix, and so on. But our solution determines those* *multi–step probabilities by summing (27–64) over the appropriate indices, which is not obviously the same thing. Investigate this and determine whether the maximum entropy multi–step probabilities are the same as the classical Markov ones, or whether they become the same in some limit.*

We see that the Maximum–Entropy Principle suffices to determine explicit solutions to problems of optimal encoding for noiseless channels. Of course, as we consider more complicated constraints (trigram frequencies, etc.), pencil-and-paper methods of solution will become impossibly difficult (there is no 'standard matrix theory' for them), and to the best of our knowledge we must resort to computers.

Now Shannon's ostensibly strongest theorem concerns the limit as $n \to \infty$ of the problem with n-gram frequencies given; his $H \equiv lim\ H_n$ is held to be the "true" entropy of the English language, which determines the "true" minimum channel capacity required to transmit it. We do not question this as a valid mathematical theorem, but from our discussion above it is clear that such a theorem can have no relevance to the real world, because there is no such thing as a "true" n-gram frequency for English even when $n = 1$.

Indeed, even if such frequencies did exist, think for a moment about how one would determine them. Even if we do not distinguish between capital and small letters and include no decimal digits or punctuation marks in our alphabet, there are $26^{10} = 1.41 \times 10^{14}$ ten–grams whose frequencies are to be measured and recorded. To store them all on paper at 1000 entries per sheet would require a stack of paper about 7000 miles high.

************************* MORE COMING! NOISY CHANNELS! *******************

When we turn to the noisy channel, fundamentally new features appear, and even the basic theorems given by Shannon have been called into question by mathematicians. But they did not have the tool of probability theory as logic; so let us see whether we are now in a better position to deal with these problems.

The Noisy Channel

From a physical standpoint, the basic idea is that the transmitted messages are corrupted en route, in an unpredictable and uncontrollable way, so that a given sent message can produce any one of several!

************************* MORE COMING! *******************

Fixing a Noisy Channel

Let us examine the simplest nontrivial case, where the noise acts independently (without memory) on each separate letter transmitted. Suppose that each letter has independently the probability ϵ of being transmitted incorrectly. Then in a message of N letters the probability that there are r errors is the binomial

$$p(r) = \binom{N}{r} \epsilon^r (1-\epsilon)^{N-r} \qquad (27\text{–}74)$$

and the expected number of errors is $\langle r \rangle = N\epsilon$. Then if $N\epsilon << 1$, we might consider the communication system satisfactory for most purposes. However, it may be essential that the message be transmitted without any error at all (as in sending a computer code instruction to a satellite in orbit). The field of fancy error–correcting codes has a large literature and much sophisticated theory; but a very popular and simple procedure is the checksum.

Suppose, as is usually the case in computer practice, that our "alphabet" consists of $2^8 = 256$ different characters sent as eight–bit binary numbers, called 'bytes'. At the end of the message one transmits one more byte, which is numerically the sum $(mod\,256)$ of the N previous ones. The receiver recalculates this sum from the first N bytes received, and compares it with the transmitted checksum. If they agree, then it is virtually certain that the transmission was error–free (if there is an error, then there must be at least two errors which just happened to cancel each other out in the checksum, and the probability of this is astronomically small, far less than ϵ). If they disagree, then it is certain that there was a transmission error, so the receiver sends back a "please repeat" signal to the transmitter, and the process is repeated until error-free transmission is achieved.

Let us see just how good the checksum procedure is according to probability theory. Write, for brevity,

$$q \equiv (1-\epsilon)^{N+1}. \tag{27–75}$$

Then to achieve error–free transmission, there is

probability q that it will require $(N+1)$ symbols transmitted.
probability $(1-q)q$ that $2(N+1)$ symbols will be required.
probability $(1-q)^2q$ that $3(N+1)$ symbols will be required,
$\cdots$ and so on.

The expected length of transmission to achieve error–free operation is then the sum

$$\langle L\rangle = (N+1)q[1 + 2(1-q) + 3(1-q)^2 + 4(1-q)^3 + \cdots]. \tag{27–76}$$

Since $|1-q| < 1$, the series converges to $1/q^2$, and so

$$\langle L\rangle = \frac{N+1}{(1-\epsilon)^{N+1}} \simeq N e^{N\epsilon}, \tag{27–77}$$

the approximation holding reasonably well if $N >> 1$. But if the message is so long that $N\epsilon >> 1$, this procedure fails; there is almost no chance that we could transmit it without error in any feasible time.

But now an ingenious device comes to the rescue, and shows how much a little probability theory can help us to achieve exactitude. Let us break the long message into m shorter blocks of length $n = N/m$, and transmit each block with its own checksum. From (27–77) the expected total transmission length is now

$$\langle L\rangle = m\,\frac{n+1}{(1-\epsilon)^{n+1}} = N\frac{n+1}{n(1-\epsilon)^{n+1}} \tag{27–78}$$

It is evident that if the blocks are too long, then we shall have to repeat too many of them; if they are too short, then we shall waste transmission time sending many unnecessary checksums. Thus there should be an optimal block length, which minimizes (27–78). Providentially, this turns out to be independent of N; varying n, (27–78) reaches a minimum when

$$1 + n(n+1)\log(1-\epsilon) = 0, \qquad \text{or} \qquad (1-\epsilon)^{n+1} = \exp(-1/n) \tag{27–79}$$

For all practical purposes, then, the optimal block length is

$$(n)_{opt} = \frac{1}{\sqrt{\epsilon}} \tag{27–80}$$

and the minimum achievable expected length is

$$\langle L\rangle_{min} = N\cdot\frac{n+1}{n}\cdot\exp\left(\frac{1}{n}\right) \simeq N(1+2\sqrt{\epsilon}) \tag{27–81}$$

By breaking a long message into blocks, we have made an enormous improvement. If $\epsilon \simeq 10^{-4}$, then it would be impractical to send an error–free message of length $N = 100,000$ bytes in a single block; for one expects about 10 errors in each transmission. The expected transmission length would be about $22,000N$ bytes, signifying that we would have to repeat the message, on the average, about 22,000 times before achieving one error–free result. But the optimal block length is about $n \simeq 100$, and by using this the expected length is reduced to $\langle L \rangle = 1.020N$. This signifies that we are sending 1000 blocks, of which each has one extra byte (which accounts for the factor $(n+1)/n \simeq 1+\sqrt{\epsilon}$) and about 10 will probably need to be repeated (which corresponds to the factor $\exp(1/n) \simeq 1 + \sqrt{\epsilon}$). But the minimum in (27–78) is very broad; if $40 \le n \le 250$, we have $\langle L \rangle \le 1.030N$. If $\epsilon = 10^{-6}$, then the block technique allows us to transmit error–free messages of any length with virtually no penalty in transmission time [$\langle L \rangle \simeq 1.002N$ if n is anywhere near 1000].

To the best of our knowledge, the block technique is an intuitive *ad hockery*, not derived uniquely from any optimality criterion; yet it is so simple to use and comes so close to the best that could ever be hoped for [$\langle L \rangle = N$], that there is hardly any incentive to seek anything better.

In the early days of microcomputers, messages were sent to and from disks in block lengths of 128 or 256 bytes, which would be optimal if the error probability for each byte were of the order $\epsilon \simeq 10^{-5}$. At the time of writing (1991) they are being sent instead in blocks of 1024 to 4096 bytes, suggesting that disk reading and writing is now reliable to error probabilities of the order of 10^{-8} or better. Of course, it is conservative design to use block lengths somewhat shorter than the above optimal value, to hedge against deterioration in performance as the equipment wears out and the error rate increases.

But let us note a point of philosophy; in this discussion, have we abandoned our stance of probability theory as logic, and reverted to frequency definitions? Not at all! It is perfectly true that *if* the error probability ϵ is indeed an "objectively real" frequency of errors measured in some class of repetitions of all this, then our $\langle L \rangle_{min}$ is equally well the objectively real minimum achievable average transmission length *over that same class of repetitions.*

But there are few cases where this is really known to be true; such experiments are costly in time and resources. In the real world, they are never completed before the design becomes frozen and the manufactured product delivered to the customers. Indeed, reliability experiments on highly reliable systems can never be really completed at all, because in the time it requires to do them, our state of knowledge and technical capabilities will change, making the original purpose of the test irrelevant.

Our present point is that probability theory as logic works as well, in the following sense, whether our probabilities are or are not known to be real frequencies. As we saw in Chapter 8, it is an elementary derivable consequence of probability theory as logic that our probabilities are the best estimates of those frequencies that we can make on the information we have.

Then, whatever the evidence on which that probability assignment ϵ was based, the above equations still describe the most rational design *that could have been made, here and now, on the information we had.* As noted, this remains true even if we know in advance that only a single message is ever going to be sent over our communication system. Thus probability theory as logic has a wider range of applications, even in situations where one sometimes pretends that he is using a frequency definition for psychological reasons.

******************* A LITTLE MORE HERE! ****************************