Journal of Statistical Physics, Vol. 3, No. 2, 1971

Statistical Thermodynamic Formalism in the Solution of Information Theory Problems

H. Reiss[1] and C. Huang[1]

Received July 15, 1970

This is the first of several papers dealing with the application of statistical thermodynamic methodology to the solution of coding and communication theory problems. Emphasis is placed on the various "ensemble techniques" of statistical mechanics, the words or "samples" of a message taking the place of molecules in the prototype physical system. Analogs of temperature, internal energy, pressure, chemical potential, volume, entropy, etc., are developed. The isomorphism with thermodynamics is complete and these quantities transform (for example, by partial differentiation) in exactly the same way as the prototype physical quantities. The methods are nicely applicable to coding cases involving sources with memory, in which case, correlation can be discussed in terms of analog "coupling energies" between signals or words so that the store of "many-body-problem" techniques can be used. In addition, the manipulative freedom stemming from the possibility of choosing from a multiplicity of ensembles constrained by "intensive" parameters proves a distinct advantage. A concrete example dealing with the choice of a compact code for a nonextended source with memory is presented. The compact code is derived, and some discussion is given concerning the breadth of its power spectrum. In a following paper, its autocorrelation function within the framework of "pulse code modulation" is derived and transformed by Wiener theory so that the power spectrum is directly exhibited (along with the spectra for several other cases).

KEY WORDS: Information theory; coding; statistical mechanics; ensembles; thermodynamics.

Research supported under AFOSR Grant No. 70-1877. The present work is contribution No. 2643 of the Department of Chemistry, University of California—Los Angeles.

[1] Department of Chemistry, University of California, Los Angeles, California.

822/3/2-7

1. INTRODUCTION

In a recent paper,[1] the applicability of thermodynamic ideas to coding theory was explored. The paper was, however, largely didactic and confined to the simplest systems, e.g., discrete sources, without memory, coupled to noiseless channels. Information theory analogs of various thermodynamic quantities were identified, including information *temperature*, information *pressure* and *volume*, *free energy*, and *chemical potential*. Of course, the well-established *information entropy* was also discussed.

As a simple illustration of the application of such techniques, transformations among these quantities were generated using (as in real thermodynamics) the methods of partial differentiation.

The use of memoryless source led to analog situations which where the counterparts of "weakly coupled" physical systems, like ideal gases. The partition functions which appeared naturally were therefore analogs of partition functions for *molecules* rather than for whole thermodynamic systems. As such, the formalism was limited in scope.

In the present paper, we wish to generalize the method, making it applicable to situations involving memory and therefore to the analogs of strongly coupled physical systems. Furthermore, in a paper to follow, we shall apply the method to a realistic problem involving the achievement of compact, reliable communication requiring minimum bandwidth and power. This problem is described in the present paper although its detailed analysis is left for later.

Of course, we cannot expect to exceed the ideal reliable communication rate specified by the famous channel coding theorem[2]

$$I = \omega T \log_2[1 + (P/N)] \tag{1}$$

where I is the maximum information in bits which can be transmitted in time T through a channel of bandwidth ω, having average signal and noise powers P and N, respectively. In general, the greatest utility of the statistical thermodynamic approach will not lie in the establishment of general theorems such as Eq. (1) which define what *can* and *cannot* be done, but rather in the discovery of *specific means* for achieving the maximum performance allowed by such theorems. Thus, in the example to be analyzed in the following paper, specific codes are devised, aimed at achieving the communication rate permitted by Eq. (1).

Several final introductory notes: In the statistical thermodynamic approach, a variety of "ensembles" emerge in a natural manner just as in physical statistical mechanics.[3] Often, it is more convenient for mathematical manipulation to work in one ensemble rather than another. The chemist and physicist have "cultural" traditions which arm them with faith in such conceptual indeterminacy. Not the least of this faith is rooted in the fact that in those ensembles constrained by "intensive" parameters (temperature, pressure, chemical potential, etc.), the parameters admit of direct measurement and therefore possess physical meanings independent of the mathematical method itself. In information theory, it is somewhat more difficult to invest them with an equivalent measure of reality. Nevertheless, they are useful.

To the mathematician who may have a taste for rigor, attention should be called to the fact that the discussion which follows is more discursive than rigorous. This is because the formalism is indeed completely isomorphic with that of physical statistical mechanics, and the necessary limit theorems have already been carefully proved within that body of knowledge[2] so that there is no need to repeat them here. The proofs are, however, somewhat different (and perhaps less clean-cut) than those given in standard information theory. Nevertheless, they are quite parallel, for example, to manipulations via the "Chebyshev inequality" or the "weak law of large numbers."

2. ELIMINATION, WITHOUT INCREASE OF BANDWIDTH, OF REDUNDANCY IN PULSE CODE MODULATION

In order to have a concrete case in mind as we develop the formalism, consider the following problem. A continuous signal is sampled,[4] say at time intervals of $1/2\omega$, where ω is the bandwith, and the samples are converted into binary numbers for transmission by pulse code modulation (PCM)[5] over a noisy channel. In order to combat noise, check digits[6] may be added to each binary number; and to gain a measure of compactness in advance of binary coding, some procedure such as Huffman[7] or Fano[8] coding may be employed. As a final generalization, the same procedure may be used for the transmission of discrete as well as continuous messages. For example, the binary numbers, instead of representing samples of a continuous message, may correspond to the letters of the alphabet which appear in the sequence of some text being transmitted. In any event, the pulses are reconstituted into the original signal at the receiver end of the system.

The transmitted message then consists of a sequence of zeros and ones. This sequence will have a set of statistics generated by the constraints in the original message, those implicit in Huffman or Fano coding, and in the method of assigning check digits. The chance that a given digit will be zero or one depends, in some way, on the preceding digits. This correlation amounts to a redundancy which can still be squeezed out of the transmitted message. Even in the absence of correlation there may be redundancy implicit in unequal frequencies of appearance of zeros and ones. The statistics of the message can of course be determined by a suitable investigation.

Usually in PCM the pulses representing zeros and ones are of equal duration. By assigning different transmission times to different pulses, depending upon the statistics, it is possible to increase the rate of transmission. Of course, this is possible by merely shortening the duration of each pulse in scale, but for this, one pays the price of greater bandwith, a fact already evident in Eq. (1). The real question is the following: Can one, by choosing pulse transmission times of various magnitudes, decrease the mean transmission time *without* simultaneously increasing bandwidth?

This is the problem to which this and the following paper are addressed. In the present paper, however, we concentrate on developing methods for choosing pulse transmission times. In the succeeding paper, we compute autocorrelation functions[9] as they depend on message statistics and assigned pulse transmission times; and then by Fourier transformation derive the exact power spectrum[10] of the message and examine it for bandwith effects.

3. COMPACT CODING AND MATCHING

It is in the choice of pulse transmission times that the statistical thermodynamic methodology finds application. In the interest of generality, however, we will develop the formalism without specifically identifying the kinds of messages involved. That is, they may consist of pulses representing zeros and ones, a more extensive set of digits, or, even, of continuous waveforms. It is only necessary that certain of their statistical features (discussed below) be known.

We begin the investigation by the consideration of a very long (actually infinite) message. This is generated by a particular source operating under certain constraints; for example, under the constraints of grammar, type of language, etc. The message is itself divided into very long submessages. Hereafter, we refer to the total message as the "supermessage." The submessages are emitted by the source with different probabilities. Thus, the nth submessage will be emitted with probability P_n. The submessages are sufficiently long so that there is a very large number of them and, furthermore, so that they are substantially uncorrelated even though individual words in the message may be strongly so. The observant reader will recognize that in the usual language of information theory the submessages are "words" of a high-order "extension" of the source.[11]

We now consider a "code" or "channel"[2] whose "vocabulary" consists of the submessages emitted by the source. The constraints imposed on the supermessages which the code can "compose," aside from those of vocabulary, are not defined by language and therefore not by a set of probabilities P_n. Instead, they are represented by the specification, through the coding procedure, of the transmission time T_n, for the nth message. In addition, we may demand that the supermessage be fitted into a total time $\mathscr{T}$.

Under these conditions, the "typical" supermessage composed by the code will contain the nth submessage N_n^* times, on the average. Since there are $\mathscr{M}$ submessages, the frequency, or probability, of nth submessage in a code supermessage is

$$P_n^* = N_n^*/\mathscr{M} \tag{2}$$

There will of course be individual supermessages in which the frequency departs from the mean typified by P_n^*, but these will be "fluctuations" of very low probability. In fact, as is well known,[12] when $\mathscr{M}$ is large, messages characterized by the submessage distribution N_n^* have not only the *average* distribution but also the *most probable* one!

If the P_n of the source are different from the P_n^* defined by the code, then the source will have to work with a set of "fluctuated" supermessages of the code, i.e., with a much smaller fraction of the messages available in the code. Obviously, source and code will be very poorly matched.

Matching, however, may be achieved by changing the "spectrum" of transmission times T_n; that is, by changing the code so that $P_n^* \to P_n$.

In homely terms, the source may be viewed as an "intelligence" composing

[2] We shall use the words "code" and "channel" interchangeably.

messages out of a given vocabulary and under the constraints of language. In contrast, the code may be thought of as an "idiot" using the same vocabulary but working under another set of constraints, e.g., an assigned spectrum of transmission times and fixed $\mathscr{T}$ and $\mathscr{M}$. The question of matching is merely that of forcing the idiot operating under *its own* constraints to compose, *on the average*, the same messages as the intelligence operating under *another set* of constraints.

For our purpose, the aspect of matching of primary importance lies in the compactness (relative to the source) of the matched code. To understand this, adopt the point of view that we match a source to a code rather than the reverse. In other words, we are given a spectrum of transmission times plus $\mathscr{T}$ and $\mathscr{M}$. Then, we define classes of supermessages, each identified with a particular set of N_n. That class which contains the largest number of messages goes with $N_n{}^*$. Suppose this number of messages is Ω^*. The information I^* gained upon the receipt of one such message is measured by the logarithm of the number of alternatives.[13] Thus, we have

$$I^* = \kappa \ln \Omega^* = \log_2 \Omega^* \tag{3}$$

where

$$\kappa = \log_2 e \tag{4}$$

scales the information so that it is measured in bits even though we use natural logarithms. Since a supermessage requires time $\mathscr{T}$ for transmission, I^* is the information transmitted in that time.

Of the information distribution among the various classes of message and transmitted in time $\mathscr{T}$, that in the class defined by $N_n{}^*$ is the greatest. In passing, it is worth noting that if the total number of messages distributed in all the remaining classes is denoted by Ω', then the maximum information which can be transmitted in time $\mathscr{T}$ by the channel is

$$I_{\max} = \kappa \ln(\Omega^* + \Omega') \approx I^* \tag{5}$$

$I_{\max}$ is negligibly different from I^*. This is a common result[14] and has its root in the size of the numbers Ω^* and Ω together with the nature of the logarithm. Note, however, that it is still possible for the following to hold:

$$\ln \Omega' \gg \ln \Omega^* \tag{6}$$

so that an appreciable number of channel messages will, on the average, remain unused by a source with probabilities $P_n{}^* = N_n{}^*/\mathscr{M}$.

Thus, given a fixed transmission time $\mathscr{T}$, the maximum amount of information which can be sent through the channel in that time is given essentially by I^*. We achieve this maximum rate by using, in conjunction with the channel, a source whose probabilities of message emission are $P_n{}^*$.

Next, we invert our point of view. Now assume a fixed *source* with given probabilities P_n which emits supermessages $\mathscr{M}$ submessages long. What code or channel, defined by a spectrum of transmission times T_n, allows these supermessages to be transmitted in the smallest possible time? The probabilities P_n determine the average number of distinct supermessages which the source may construct out of

submessages, and therefore the average information in a supermessage. We are thus faced with the following problem: Given a *fixed* amount of information, what is the code which can transmit it in the shortest time?

In the previous situation, we were given a code defined by a spectrum of transmission times and a *fixed* time $\mathscr{T}$ in which to transmit supermessages constructed out of $\mathscr{M}$ submessages. In that case, the T_n were fixed and not the information. We sought to identify the source which would allow the transmission in the time $\mathscr{T}$, via this code, of the maximum information. This proved to be the one with submessage probabilities $P_n{}^*$; in other words, the source which *matched* the code. This source maximized the information rate I at

$$I = I^*/\mathscr{T} \tag{7}$$

In the present situation where we are given I^* and asked to minimize $\mathscr{T}$, it is clear that we are also maximizing I. Thus, the channel whose code is compact is again determined by matching—this time, matching of code to source!

The problem in its last form is the one of primary interest. Nevertheless, it is worth noting that, under matching, the following three things are simultaneously accomplished: (1) The typical message in the "idiot" channel is the same as the typical message in the "intelligence" source. (2) The largest amount of information is transmitted in a fixed time. (3) The shortest time is required for the transmission of a fixed amount of information.

The discussion in this section has been qualitative. In the next section, we adopt a quantitative approach and introduce the statistical thermodynamic methodology.

4. COMPACT CODING AND THERMODYNAMIC EQUILIBRIUM

We now approach the quantitative aspects of the matching problem from the point of view of point (2) discussed above; namely, choosing a set of submessage probabilities P_n for a source so that the greatest amount of information is transmitted in time $\mathscr{T}$ over a channel with transmission times T_n. This means that the P_n must correspond to the frequency of appearance of the nth submessage in the average, or "typical," supermessage generated by the channel.

The class of supermessages in which there are N_n submessages of type n consists of

$$\Omega = \mathscr{M}!/\prod_n N_n! \tag{8}$$

supermessages in all. These are subject to the constraints

$$\sum_n N_n = \mathscr{M} \tag{9}$$

$$\sum_n N_n T_n = \mathscr{T} \tag{10}$$

The set of N_n which maximize Ω subject to Eqs. (9) and (10) will be those of the "typical" message, and can be obtained in a standard manner by the method of undetermined multipliers. The result is

$$N_n{}^* = e^{-T_n/\kappa\tau}/Q(\tau) \tag{11}$$

where

$$Q(\tau) = \sum_n e^{-T_n/\kappa\tau} \tag{12}$$

Here, κ is the quantity specified by Eq. (4) and τ is an undetermined multiplier. (The multiplier of course is $\kappa\tau$; we write it in this form in anticipation of later use.) The probabilities going with the typical set are then

$$P_n^* = N_n^*/\mathscr{M} = e^{-T_n/\kappa\tau}/Q(\tau) \tag{13}$$

Matching requires a source with probabilities

$$P_n = P_n^* \tag{14}$$

Until now, we have not advanced a method for defining a submessage. Any one of a number of definite criteria may be used. For example, the source may generate messages composed of discrete words. In this case, we might define each message to consist of a sequence of a fixed number, W, of words. The words in the message may, however, be highly correlated. Thus, the probability of emission of a given word may depend heavily on the words emitted previously. On the other hand, W is large enough so that the *submessages* themselves remain essentially uncorrelated.

In form, Eqs. (11)–(13) resemble those belonging to the canonical ensemble[15] employed in treatment of thermodynamic systems. In that case, P_n^* would be the probability that the system, immersed in a thermostat of temperature τ, is in a quantum state of energy T_n. $Q(\tau)$ resembles the canonical ensemble *partition function.*

That τ is indeed the analog of a temperature (we call it the *information* temperature) is demonstrated by the following argument. The mean time of transmission of a submessage is

$$T = \sum_n P_n T_n \tag{15}$$

where we have dropped the asterisk on the P_n, understanding that it still signifies the P_n^* given by Eq. (13). Since the transmission times T_n are the analogs of physical system energy levels, T is the analog of the thermodynamic internal energy. Notice that Eq. (14) can be satisfied, not only by finding a source with the proper P_n to match a channel with fixed spectrum T_n and therefore fixed P_n^*, but also by adjusting the various T_n so that the P_n^* are varied to match a source with fixed P_n. We can make the T_n depend upon one or several parameters X_i, etc., so that

$$dT_n = \sum_i (\partial T/\partial X_i)\, dX_i \tag{16}$$

so that matching can be achieved by varying the X_i. In the simplest case, there will be only one parameter X. For simplicity, we consider such a case. Then,

$$dT = \sum_n T_n\, dP_n + \sum_n P_n\, dT_n = \sum_n T_n\, dP_n - \left(-\sum_n P_n\, \partial T_n/\partial X\right) dX \tag{17}$$

If T_n from Eq. (13) is substituted into the first sum on the right of Eq. (17), we get

$$dT = \tau\, d\left\{-\kappa \sum_n P_n \ln P_n\right\} - \left\{-\sum_n P_n\, \partial T_n/\partial X\right\} dX \tag{18}$$

where we have used

$$\sum_n dP_n = 0 \tag{19}$$

Now, the term in the first curly brackets of Eq. (18) is the information per submessage when the messages are emitted with differing probabilities P_n, defined[16] by standard information theory. It is also S, the information entropy per submessage (and also has the correct form for a physical entropy of a thermodynamic system). Thus,

$$I/\mathscr{M} = S = -\kappa \sum_n P_n \ln P_n \tag{20}$$

If we define

$$\pi = -\sum_n P_n\, \partial T_n/\partial X \tag{21}$$

as the *information pressure* of a submessage, Eq. (18) may be written in the form

$$dT = \tau\, dS - \pi\, dX \tag{22}$$

When this equation is compared with the thermodynamic equation,[17] representing the combined first and second laws,

$$dU = T'\, dS - p\, dV \tag{23}$$

in which U is the internal energy, T' the temperature, S the entropy, p the pressure, and V the volume, the analog between τ and T' (as well as the other analog relationships) is immediately evident.

Notice that

$$\pi = -(\partial T/\partial X)_S \tag{24}$$

so that the information pressure measures the resistance to compression, at fixed information content S per submessage, of the average submessage transmission time. π therefore has an intuitively satisfying meaning.

Equations (13) and (14) are the matching conditions and are also the analogs of relations which would hold at thermodynamic *equilibrium* in a physical system. Matching may therefore be viewed in the following way. The source is like a *thermostat* at temperature τ in which the channel is immersed. Matching occurs when the channel is "immersed" in this "thermostat" and the channel spectrum T_n is altered, possibly by variation of the *information volume* X, so that the channel comes into "thermal" equilibrium with the source. Note that τ also measures the randomness of the source since, as $\tau \to \infty$, the various P_n become equal.

Transformations among the thermodynamic analog variables T, τ, S, π, X, etc., may be generated by partial differentiation just as in the case of thermodynamics. This process was illustrated for the simple case of a source without memory in Ref. 1.

It is easy to demonstrate the following relations:

$$T = \sum_n P_n T_n = \kappa\tau^2(\partial \ln Q/\partial\tau)_X \tag{25}$$

$$S = -\kappa \sum_n P_n \ln P_n = \kappa\, \partial(\tau \ln Q)_X/\partial\tau \tag{26}$$

$$\alpha = T - \tau S = -\kappa\tau \ln Q \tag{27}$$

$$\pi = -\sum_n P_n(\partial T_n/\partial X) = \kappa\tau(\partial \ln Q/\partial X)_\tau \tag{28}$$

Here α is the *information* Helmholtz free energy. Thus, as in thermodynamics, the evaluation of the partition function permits the immediate specification of the thermodynamic analog quantities.

A useful interpretation of τ is the following. If we fix the average message length T, then τ can be determined from Eq. (25). Thus, τ is in some way related to the average message length. If we define

$$\theta_n = T_n/\tau \tag{29}$$

Eq. (13) becomes

$$P_n = e^{-\theta_n/\kappa}/Q(\theta_1, \theta_2, \ldots) \tag{30}$$

so that the P_n depend only upon the reduced transmission times θ_n. The information per message therefore depends solely on these reduced times, and can be maintained constant if all the T_n are varied in proportion to τ. Thus, τ is a common scale factor, determining the average message length. At fixed information per message, one can achieve arbitrary speed of transmission (other factors such as bandwith being ignored) by merely reducing τ so that T is diminished in the same proportion. Adjustment of scale (through τ) is therefore not really a problem in information theory. The real problem consists in choosing the code which allows the transmission of the largest amount of information at fixed T and P_n, i.e., in adjusting the *relative* values of T_n (i.e., θ_n). Equation (30) accomplishes this.

Alternatively, Eq. (30), through (25), determines the smallest T for a code which can be matched to a source at temperature τ having message probabilities P_n.

5. DISCRETE SOURCE WITHOUT MEMORY

In this section, we specialize the relations of the previous section to the case treated in Ref. 1, namely, that of a source, without memory, emitting discrete words, the ith word having probability of emission p_i. If each message contains W words and we assign a spectrum of transmission times t_i to the *words*, then a message containing n_i words of type i will have a transmission time

$$T_n = \sum_i n_i t_i \tag{31}$$

Since the source has no memory, all sets of n_i are allowable subject only to the condition

$$\sum_i n_i = W \tag{32}$$

and furthermore, all permutations of words are permitted, so

$$P_n = \prod_i p_i^{n_i} \tag{33}$$

and

$$Q = q^W \tag{34}$$

where

$$q = \sum_i e^{-t_i/\kappa\tau} \tag{35}$$

where q is the partition function for a word. Using Eq. (31) in (13), noting Eq. (32), and eliminating P_n between the result and Eq. (33) gives

$$\prod_i p_i^{*n_i} = \prod_i (e^{-t_i/\kappa\tau}/q)^{n_i} \tag{36}$$

from which it is easily seen that

$$p_i^* = e^{-t_i/\kappa\tau}/q(\tau) \tag{37}$$

For matching, we adjust t_i so that

$$p_i^* = p_i \tag{38}$$

This involves solving the set of simultaneous equations, one for each i, represented by Eq. (37) with p_i^* set equal to p_i.

Note some further features of the analog. A word is like a molecule. All words are simply the *same* molecule in different time states (the analog of quantum states). When memory and correlation are not involved, the message is something like an ideal gas. We shall have more to say about this later.

6. ANOTHER ENSEMBLE

When the source has memory, one cannot achieve the most compact code by merely adjusting *word* transmission times according to Eq. (37). One can adjust *message* times in accordance with Eq. (13), but then one deals with an essentially infinite set of messages and a correspondingly large set of times. This of course is a usual problem in coding high-order extensions of a source. It is in just this case, when correlation is involved, that the statistical thermodynamic methodology is useful for the prescription of those *word* transmission times that lead to the most compact code, for it is here where the rich fund of techniques developed in connection with the physical "many-body problem" can be applied. We will solve an example of this kind later. In the present section, we discuss another ensemble.

Consider a supermessage confined to a transmission time $\mathscr{T}$ as in the previous case but from which we lift the constraint that it contain a fixed number of submessages $\mathscr{M}$. Since all of the messages possible in the presence of that constraint are still possible, together with additional ones which can now be fitted into $\mathscr{T}$, in view of the relaxing of the requirement of $\mathscr{M}$, the code which can be adapted to this situation contains an equal or greater number of messages and should be able to transmit as much or more information in the same time interval. Alternatively, it can transmit the same information in a shorter time and should be more compact. In fact, it should be the *most* compact code matched to the given source since there are no further constraints which can be removed. ($\mathscr{T}$ does not really represent a constraint since it can be varied at will, and, in any event, we think of it as going to infinity.)

In practice, how can we remove the restriction of fixed $\mathscr{M}$ in $\mathscr{T}$? One method is the following. Assume that the source, in emitting the message, breaks down intermittently in a perfectly random manner. Each failure defines the terminus of a submessage. Thus, if the message is composed of discrete words, the submessages will contain varying numbers of words; and furthermore, $\mathscr{M}$ will be indeterminate. The number of submessages going with a given set of N_n is still

$$\Omega = \left(\sum_n N_n\right)! \Big/ \pi N_n! \tag{39}$$

where $\sum_n N_n$ represents the variable $\mathscr{M}$. Now, however, the only constraint is

$$\sum_n T_n N_n = \mathscr{T} \tag{40}$$

Maximizing Ω relative to this constraint gives

$$P_n{}^* = N_n/\mathscr{M} = e^{-T_n/\kappa\tau} \tag{41}$$

Comparison of Eq. (41) with (13) shows that we are here dealing with a case in which τ has a value τ_c such that

$$Q(\tau_c) = 1 \tag{42}$$

If the source probabilities are P_n, then matching is once more achieved by adjusting the T_n such that $P_n{}^* = P_n$; that is,

$$T_n = -\kappa\tau_c \ln P_n \tag{43}$$

The mean submessage time is now

$$T_c = \sum_n P_n T_n = -\kappa\tau_c \sum P_n \ln P_n = \tau_c S \tag{44}$$

and, as indicated above, this must be the shortest mean time possible at given τ_c. If we deal with block coding,[18] message lengths are measured not by T_n but by L_n, the number (an integer) of digits in the nth message. Had we proceeded this way, Eq. (41) would have been replaced by

$$P_n{}^* = \exp(-L_n/\kappa\tau_c') \tag{45}$$

with τ_c' now dimensionless. Equation (44) would become

$$L_c = \tau_c' S \tag{46}$$

The L_n in Eq. (45), when the P_n^* are fixed, still depend on the choice of τ_c'. Alternatively, Eq. (42), now written as

$$\sum_n \exp(-L_n/\kappa\tau_c') = 1 \tag{47}$$

does not determine τ_c' until the L_n are fixed. The capacity C of the code is given by the information transmitted per digit,

$$C = S/L_c = 1/\tau_c' \tag{48}$$

Obviously, the most compact code is the one with the largest capacity. Clearly, the maximum information per digit is achieved when we have absolutely no preinformation concerning which digit is likely to appear. In this case, if there are r digits, the information per digit would be $\kappa \ln r$, so that the value of τ_c' for the most compact code should be given by

$$1/\tau_c' = C_{\max} = \kappa \ln r \qquad \text{or} \qquad 1/r = \exp(-1/\kappa\tau_c') \tag{49}$$

Choosing this value of τ_c' in Eq. (47), it becomes

$$\sum_n r^{-L_n} = 1 \tag{50}$$

which is the "equality" part of the famous McMillan inequality.[19] Introduction of Eq. (49) into (46) gives, for the *most compact block* code,

$$L_c = S/(\kappa \ln r) = S/\log_2 r \tag{51}$$

which is Shannon's first theorem,[20] or the source coding theorem.

The special value of τ_c' prescribed by Eq. (49) arises strictly in the case of block coding and is connected with the fact that our "scale" is constrained by the integer requirements on the various L_n. No such unique determination of τ_c is available for the continuous case to which Eq. (44) applies. Here, as mentioned earlier, τ and τ_c are simply scale factors which, as will be seen later, cancel out of most problems.

In the block coding case, maximally compact codes can only be achieved by coding very high-order extensions of the source. In the continuous case without the integer restriction, word transmission times t_i can be chosen so that the coding of low-order extensions achieves compactness.

Although $Q(\tau)$ has the outward form of an analog of the partition function in the canonical ensemble, it is not the *strict* analog if words are considered the analogs of molecules, unless the number of words in each submessage is fixed, say at W. Further insight into this analog question can be obtained by considering a supermessage composed of $\mathscr{M}$ submessages and $\mathscr{W}$ words confined to a time interval $\mathscr{T}$.

Here again, we achieve this condition in practice by letting the source fail intermittently in a random manner. In addition, we consider only those supermessages which happen, in time $\mathscr{T}$, to contain $\mathscr{W}$ words.

The constraints are now

$$\sum_n N_n = \mathscr{M} \tag{52}$$

$$\sum_n T_n N_n = \mathscr{T} \tag{53}$$

$$\sum_n N_n W_n = \mathscr{W} \tag{54}$$

where W_n is the number of words in the nth submessage. The N_n for the typical supermessage are obtained by maximizing

$$\Omega = \mathscr{M}!/\prod_n N_n! \tag{55}$$

with the result

$$P_n{}^* = N_n/\mathscr{M} = e^{\mu W_n/\kappa\tau} e^{-T_n/\kappa\tau}/Z \tag{56}$$

where μ is another undetermined multiplier and

$$Z = \sum_n e^{\mu W_n/\kappa\tau} e^{-T_n/\kappa\tau} \tag{57}$$

If we gather all messages having W words together, the sum in Eq. (57) may be expressed more compactly as

$$Z = \sum_W e^{\mu W/\kappa\tau} Q(T, W, X) \tag{58}$$

where $Q(T, W, X)$ is indeed the canonical ensemble analog partition function for a submessage with a fixed number W of words. Z will be recognized, immediately, as the analog *grand ensemble* partition function for an "open" submessage constrained to a "chemical potential" μ.[21a,b]

It may be shown[22] in a straightforward manner that

$$T = \kappa\tau^2(\partial \ln Z/\partial\tau)_{\mu,X} \tag{59}$$

$$\langle W\rangle = \kappa\tau(\partial \ln Z/\partial\mu)_{\tau,X} \tag{60}$$

$$S = \kappa\{\partial(\tau \ln Z)/\partial\tau\}_{\mu,X} \tag{61}$$

$$\pi X = \kappa\tau \ln Z \tag{62}$$

Here, $\langle W\rangle$ is the *average* word content of a submessage, just as T is the *average* transmission time. The "intensive" parameters τ and μ are constrained, while T and W are allowed to fluctuate. Although these parameters are not directly measurable as in the case of physical systems, they are nevertheless useful concepts for the purpose of mathematical manipulation. For example, it often is easier to evaluate a sum in

one ensemble where it may contain an infinite number of terms than in another where it may be finite. To give a simple illustration, we derive the "equation of state" for a submessage generated by a source without memory.

Introducing Eq. (34) into (58) gives

$$Z = \sum_{W=1}^{W=\infty} \{qe^{\mu/\kappa\tau}\}^W = qe^{\mu/\kappa\tau}/(1 - qe^{\mu/\kappa\tau}) \tag{63}$$

Substitution of this result into Eq. (60) yields

$$1 - 1/\langle W \rangle = qe^{\mu/\kappa\tau} \tag{64}$$

and substitution back into Eq. (63) gives

$$Z = \langle W \rangle - 1 \tag{65}$$

so that, according to Eq. (62),

$$\pi X = \kappa\tau \ln\{\langle W \rangle - 1\} = \kappa\tau \ln\langle W \rangle \tag{66}$$

where unity can be ignored in comparison with $\langle W \rangle$.

Since π measures the resistance of a submessage to compression at fixed information content, we see that this resistance rises less than linearly with word content, i.e., the message becomes more compressible with increased W.

Equations (56) and (57) reduce to Eqs. (13) and (12), respectively, if μ is set equal to zero. Thus, we see that $Q(\tau)$ derived originally is in fact a *grand* partition function for a grand ensemble with *information chemical* potential equal to zero. In order for it to correspond to the canonical ensemble, it would be necessary to fix W rather than allow it to fluctuate.

Since in deriving Eqs. (12) and (13) the constraint, Eq. (54), was not used, we see that $\mu = 0$ corresponds to the case in which no condition is placed on the total number of words permitted in the supermessage.

The most compact code is still achieved when $Q(\tau) = 1$. This corresponds to the case in which the constraints, Eq. (52), as well as Eq. (54), are absent.

7. SOURCE WITH CORRELATION

In this section, we return to the problem of pulse code modulation (PCM) described in Section 2 and apply the methodology of the preceding text to its solution. We treat a very simple example in order to illustrate the method, but more complex ones can be handled with no essential increase in difficulty. To refresh the reader's memory, we are given a source which emits messages consisting of sequences of zeros and ones. This source is in fact some "sampled," possibly continuous, signal, the samples having been coded as binary numbers, with check digits if necessary. The statistics of this source are known and, for simplicity, we assume that correlation, or memory, extends only to the previous digit. In the present example, the term "word," as used in the preceding text, is taken to mean "digit."

We now define the following: p_{11} is the probability that a one will be emitted, given that the previous digit was a one; p_{00} is the probability that a zero will be emitted, given that the previous digit was a zero; p_{01} is the probability that a one will be emitted, given that the previous digit was a zero; p_{10} is the probability that a zero will be emitted, given that the previous digit was a one. In addition, we define P_1 as the *a priori* probability that a one is emitted, P_0 as the *a priori* probability that a zero is emitted, and P_{11}, P_{00}, P_{01}, P_{10} as the respective probabilities that the subscript pairs are observed.

Clearly, the following relations hold:

$$\begin{aligned} P_1 &= p_{11} + p_{01}, \qquad P_0 = p_{10} + p_{00} \\ P_{11} &= P_1 p_{11}, \qquad P_{00} = P_0 p_{00} \\ P_{01} &= P_0 p_{01}, \qquad P_{10} = P_1 p_{10} \end{aligned} \tag{67}$$

The problem is to choose word transmission times t_{11}, t_{00}, t_{01}, and t_{10} which, on the basis of the above statistics, lead to the shortest mean time of transmission per word. Because of the correlation, we cannot simply use a variant of Eq. (37) [with $q(\tau) = 1$] to get something like

$$t_{10} = -\kappa\tau \ln p_{10}, \quad \text{etc.} \tag{68}$$

Furthermore, when we discover the proper values for t_{11}, t_{00}, t_{01}, and t_{10}, we must investigate whether we have to pay for their use with increased bandwidth.

The problem is simplified considerably by inversion; that is, we assume a code or channel in which t_{11}, t_{00}, t_{01}, and t_{10} are *specified*, and then find a source with word probabilities to match it. Thus, if n_{11}, n_{00}, n_{01}, and n_{10} represent the numbers of 11, 00, 01, and 10 pairs in the message, these will be allowed to "fluctuate" along with the numbers of ones and zeros while the "intensive" paremeters τ and μ are held fixed. In this way, no restrictions are placed either on the n's or on the sums in the relevant partition functions, which can then be evaluated in closed form.

It is then possible to calculate the *average* values $\langle n_{11} \rangle$, etc., at fixed τ and μ. We thus arrive at relations between the $\langle n \rangle$'s and t's which can be inverted to give the t's as functions of $\langle n \rangle$'s. The $\langle n \rangle$'s are then chosen to be those of the *source* and the corresponding t's will, from what has been said in previous sections, be those yielding the compact code for given τ and μ. To get the most compact code, in accordance with Section 7, μ is set equal to zero and τ is determined by the requirement Eq. (42).

Thus, even though μ and τ are not directly measurable, their introduction removes certain bothersome restrictions from the process of mathematical manipulation, and so they play important roles exhibiting some advantages of the statistical thermodynamic methodology.

The most convenient ensemble for the present problem is the grand ensemble introduced in Section 7. If the nth submessage contains n_{11}, n_{00}, n_{01}, and n_{10} of the respective pairs, then

$$T_n = n_{11}t_{11} + n_{00}t_{00} + n_{01}t_{01} + n_{10}t_{10} \tag{69}$$

Note that T_n can vary with the n's.

In the same submessage, the number of *words* will be

$$W_n = n_{11} + n_{00} + n_{01} + n_{10} \tag{70}$$

Using Eqs. (69) and (70) in (57) gives

$$Z = \sum_{\text{(all possibilities)}} (\gamma s_{11})^{n_{11}} (\gamma s_{00})^{n_{00}} (\gamma s_{01})^{n_{01}} (\gamma s_{10})^{n_{10}} \tag{71}$$

where

$$\begin{aligned} \gamma &= e^{\mu/\kappa\tau} \\ s_{ij} &= e^{-t_{ij}/\kappa\tau}, \qquad i = 0,1, \quad j = 0,1 \end{aligned} \tag{72}$$

The sum is over all message possibilities. This means over all pair numbers n_{ij}, the pairs appearing in *all* possible permutations. The "all possibilities" means every conceivable message of every conceivable length! This illustrates the great freedom from restriction gained through the introduction of τ and μ.

We already know [see Eq. (56)] that the terms in the sum (71), when normalized by $1/Z$, measure the probability of the message with the specified n_{ij}. Thus, $\langle n_{11} \rangle$, for example, is given by

$$\begin{aligned} n_{11} &= 1/Z \sum_{\text{(all possibilities)}} n_{11} (\gamma s_{11})^{n_{11}} (\gamma s_{00})^{n_{00}} (\gamma s_{01})^{n_{01}} (\gamma s_{10})^{n_{10}} \\ &= s_{11} \{\partial \ln Z / \partial s_{11}\}_{\gamma, s_{00}, s_{01}, s_{10}} \end{aligned} \tag{73}$$

Similar relations hold for the other n's, so that

$$\begin{aligned} \langle n_{00} \rangle &= s_{00}\, \partial \ln Z / \partial s_{00} \\ \langle n_{01} \rangle &= s_{01}\, \partial \ln Z / \partial s_{01} \\ \langle n_{10} \rangle &= s_{10}\, \partial \ln Z\, \partial s_{10} \end{aligned} \tag{74}$$

where for simplicity we have omitted the subscripted variables held constant in the partial differentiations.

We already know [Eq. (60)] that

$$\langle W \rangle = \kappa\tau\, \partial \ln Z / \partial \mu = \gamma\, \partial \ln Z / \partial \gamma \tag{75}$$

Thus, we may calculate

$$P_{11}^{*} = \langle n_1 \rangle / \langle W \rangle \tag{76}$$

and the remaining similar probabilities if $Z(\gamma, s_{11}, s_{00}, s_{01}, s_{10})$ can be evaluated.

The problem in its present form is a variant of the Ising model problem for a ferromagnet,[23] which has received much attention in physics. Accordingly, for the evaluation of Z we may use one of the several methods developed for that case. A convenient procedure is the so-called "matrix method."[24] This is applied most easily to our case if the form of Z, given by Eq. (58), is used, in which case we must first evaluate $Q(T, W, X)$.

The messages of W words over which the sum in Q goes may be divided into two classes; those ending in one and zero, respectively. The part of the sum which includes only those ending in one will be denoted by $Q_1^{(W)}$ and the part over those ending in zero by $Q_0^{(W)}$. These quantities may be generated from the corresponding quantities for messages possessing $W - 1$ words by the relations

$$\begin{aligned} Q_1^{(W)} &= s_{11}Q_1^{(W-1)} + s_{01}Q_0^{(W-1)} \\ Q_0^{(W)} &= s_{10}Q_1^{(W-1)} + s_{00}Q_0^{(W-1)} \end{aligned} \tag{77}$$

which are easily explained. For example, if we have $Q_1^{(W-1)}$ and add a one to all of the messages to which it corresponds, we will first of all have messages of length W ending in one. The addition of a one to messages of length $W - 1$, all ending in one, multiplies, by a factor $e^{-t_{11}/\kappa\tau} = s_{11}$, each term in partition function sum $Q_1^{(W-1)}$ which corresponds to those messages. Thus, the first term on the right of the first of Eqs. (77) is $s_{11}Q_1^{(W-1)}$. The remainder of Eq. (77) is derived in the same way. Now,

$$Q(T, W, X) = Q_1^{(W)} + Q_0^{(W)} \tag{78}$$

and it is easily shown[24] that when W is large

$$\ln Q = \ln(Q_1^{(W)} + Q_0^{(W)}) = W \ln \lambda \tag{79}$$

where λ is the largest eigenvalue of the matrix of coefficients on the right of Eq. (77). Thus, for all practical purposes, we may write

$$Q(T, W, X) = \lambda^W \tag{80}$$

The secular equation is

$$\begin{vmatrix} (s_{11} - \lambda) & s_{01} \\ s_{10} & (s_{00} - \lambda) \end{vmatrix} = 0 \tag{81}$$

of which the largest root is

$$\lambda = \tfrac{1}{2}(s_{11} + s_{00}) + \{\tfrac{1}{4}(s_{11} + s_{00})^2 + s_{10}s_{01} - s_{11}s_{00}\}^{1/2} \tag{82}$$

According to Eq. (58), then, using Eq. (78),

$$Z = \sum_{W=1}^{W=\infty} (\gamma\lambda)^W = \gamma\lambda/(1 - \gamma\lambda) \tag{83}$$

Using Eq. (75),

$$\langle W \rangle = \gamma\, \partial \ln Z/\partial\gamma = 1/(1 - \gamma\lambda) \tag{84}$$

while from Eqs. (73) and (74) we obtain, as a typical relation,

$$\langle n_{ij} \rangle = \frac{1}{1 - \gamma\lambda} \frac{s_{ij}}{\lambda} \frac{\partial\lambda}{\partial s_{ij}} = \langle W \rangle \frac{s_{ij}}{\lambda} \frac{\partial\lambda}{\partial s_{ij}} \tag{85}$$

where in the last member we have used Eq. (84).

Although $\langle n_{ij} \rangle$ and $\langle W \rangle$ depend individually on γ, and therefore on μ, we see that

$$P_{ij}^* = \frac{\langle n_{ij} \rangle}{\langle W \rangle} = \frac{s_{ij}}{\lambda} \frac{\partial \lambda}{\partial s_{ij}} \tag{86}$$

does not, so that it is not even necessary to set $\mu = 0$ before the proper P_{ij}^* is determined. Making use of Eq. (82), we get

$$\partial\lambda/\partial s_{11} = \{\tfrac{1}{2} + \tfrac{1}{4}(s_{11} - s_{00})[\tfrac{1}{4}(s_{11} + s_{00})^2 + s_{10}s_{01} - s_{11}s_{00}]^{-1/2}\} \tag{87}$$

$$\partial\lambda/\partial s_{00} = \{\tfrac{1}{2} - \tfrac{1}{4}(s_{11} - s_{00})[\tfrac{1}{4}(s_{11} + s_{00})^2 + s_{10}s_{01} - s_{11}s_{00}]^{-1/2}\} \tag{88}$$

$$\partial\lambda/\partial s_{01} = \tfrac{1}{2}s_{10}[\tfrac{1}{4}(s_{11} - s_{00})^2 + s_{10}s_{01} - s_{11}s_{00}]^{-1/2} \tag{89}$$

$$\partial\lambda/\partial s_{10} = \tfrac{1}{2}s_{01}[\tfrac{1}{4}(s_{11} - s_{00})^2 + s_{10}s_{01} - s_{11}s_{00}]^{-1/2} \tag{90}$$

These relations, together with Eq. (82) inserted in Eq. (86), relate the s_{ij} (and therefore t_{ij}) to the $P_{ij} = P_{ij}^*$ (under matching) of the source, and thus determine the compact code.

To illustrate the process in greater detail, consider the simple case in which

$$P_{11} = P_{00} = \alpha \tag{91}$$

and

$$P_{01} = P_{10} = \beta \tag{92}$$

Since

$$P_{11} + P_{00} + P_{01} + P_{10} = 1 \tag{93}$$

we have

$$\alpha + \beta = \tfrac{1}{2} \tag{94}$$

From Eqs. (67) and (94),

$$p_{11} = p_{00} = 2\alpha \tag{95}$$

$$p_{01} = p_{10} = 2\beta \tag{96}$$

and

$$P_1 = P_0 = \tfrac{1}{2} \tag{97}$$

Symmetry now requires

$$s_{11} = s_{00} = s_a \tag{98}$$

$$s_{01} = s_{10} = s_b \tag{99}$$

Introducing these equations into Eqs. (87)–(90) and the results into (86) with

$$\lambda = 1 \tag{100}$$

in accordance with the combined requirements of Eqs. (42) and (80), yields

$$\alpha = P_{11} = P_{00} = s_a/2 \tag{101}$$

$$\beta = P_{01} = P_{10} = s_b/2 \tag{102}$$

Using Eq. (94) we find

$$s_a = 2\alpha \tag{103}$$

$$s_b = 1 - 2\alpha \tag{104}$$

From Eq. (72),

$$t_{11} = t_{00} = -\kappa\tau \ln s_a \tag{105}$$

$$t_{01} = t_{10} = -\kappa\tau \ln s_b \tag{106}$$

and therefore,

$$t_{11}/t_{01} = t_{00}/t_{10} = (\ln 2\alpha)/\ln(1 - 2\alpha) \tag{107}$$

Thus, except for an arbitrary scale factor, all the times leading to a compact code are determined in terms of α, the parameter which determines the correlation. The compact mean transmission time per word is

$$\langle t \rangle = 2t_{11}\{\alpha + [\ln(1 - 2\alpha)/\ln 2\alpha]\} \tag{108}$$

where t_{11} is taken as the scale.

According to Eq. (108), the value of α that gives the smallest mean time per word (aside from scale) is $\alpha = 0$, for which the mean time is itself zero. This is perfectly consistent since $\alpha = 0$ implies $P_{11} = P_{00} = 0$, which can only be true for the *single* message of the type

$$1010101010... \tag{109}$$

in which one and zero alternate. Since only one message is possible, we already know what it is; and zero transmission time is therefore required!

8. CONCLUDING REMARKS

Although, as shown in the last section, $\langle t \rangle$ can be made small by the proper choice of word transmission times, it is still necessary to determine how much, if any, increased bandwidth is required for the achievement of this more rapid communication. In order to examine this question in detail, it is necessary to calculate the power spectrum of a signal resembling Fig. 1, but adjusted for different pulse lengths.

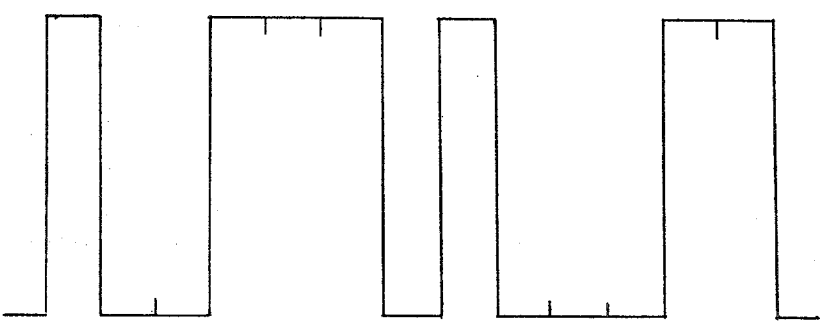

Fig. 1. Binary PCM signal with pulses representing zeros and ones of equal length.

This is a separate and difficult problem which is approached most expeditiously by first computing the autocorrelation function and applying Wiener theory.[9] This will be done in a subsequent paper.

It should be noted that in the method of Section 7, t_{ij} is the analog of an interaction or coupling energy in the Ising model. In fact, throughout the entire analog, time is the analog of energy. The problem treated in Section 7 contains *nearest-neighbor* interactions only. There is no increased difficulty in principle when longer-range interactions are involved, although there may be increased tedium. However, certain *approximate* methods have been developed in connection with the *physical* problem which could be used without alteration in the information theory case. In fact, one's whole intuition gained from acculturation with the physical problem should be transferable.

Another difficulty, of course, is the acquisition of knowledge concerning the actual statistics of the message. Here also, experience with physical systems should prove useful.

REFERENCES

1. H. Reiss, *J. Stat. Phys.* **1**:107 (1969).
2. C. E. Shannon, *Proc. IRE* **37**:10 (1949).
3. T. L. Hill, *Introduction to Statistical Thermodynamics* (Addison-Wesley, New York, 1960), Chapter 1.
4. D. Sakrison, *Communication Theory: Transmission of Wave Forms and Digital Information* (John Wiley and Sons, New York, 1968), p. 59.
5. A. M. Rosie, *Information and Communication Theory* (Blackie and Son, London, 1966), p. 148.
6. R. G. Gallager, *Information Theory and Reliable Communication* (John Wiley and Sons, New York, 1968), Chapter 6.
7. N. Abramson, *Information Theory and Coding* (McGraw-Hill, New York, 1963), p. 77.
8. A. M. Rosie, *Information and Communication Theory* (Blackie and Son, London 1966), p. 81.
9. *Ibid.*, p. 44.
10. *Ibid.*, p. 41.
11. N. Abramson, *Information and Communication Theory* (McGraw-Hill, New York, 1963), p. 29.
12. R. H. Fowler, *Statistical Mechanics* (Cambridge, New York, 1929), p. 27.
13. R. G. Gallager, *Information Theory and Reliable Communication* (John Wiley and Sons, New York, 1968), Chapter 2.
14. T. L. Hill, *Introduction to Statistical Mechanics* (Addison-Wesley, Reading, Massachussets, 1960), p. 478.
15. T. L. Hill, *Introduction to Statistical Thermodynamics* (Addison-Wesley, Reading, Massachussets, 1960), p. 6.
16. N. Abramson, *Information Theory and Coding* (McGraw-Hill, New York, 1963), Chapter 2.
17. H. Reiss, *Methods of Thermodynamics* (Blaisdell, New York, 1965), p. 79.
18. N. Abramson, *Information Theory and Coding* (McGraw-Hill, New York, 1963), p. 46.
19. *Ibid.*, p. 59.
20. *Ibid.*, p. 68.

21a. H. Reiss, *Methods of Thermodynamics* (Blaisdell, New York, 1965), p. 113.

21b. T. L. Hill, *Introduction to Statistical Thermodynamics* (Addison-Wesley, Reading, Massachusetts, 1960).

22. *Ibid.*, p. 25.
23. *Ibid.*, Chapter 14.
24. N. Davidson, *Statistical Mechanics* (McGraw-Hill, New York, 1962), pp. 378–393.