Note on Unique Decipherability*

E. T. JAYNES†

Summary—We consider an alphabet of a letters, used under the restrictions: 1) messages uniquely decipherable into words by use of one of the letters as a space mark, and 2) words limited to a maximum length of L letters. Although imposing these constraints simultaneously may cause a large reduction in the channel capacity of the alphabet, neither by itself causes any reduction. Accordingly, in the absence of constraints other than 1), an inequality of McMillan pertaining to uniquely decipherable messages can be made to be an equality.

Defining "semi-optimal" transmission by the condition that the mean transmission time per word is minimized for a given entropy per word, we find the attainable rate of information transmission under semi-optimal conditions. Transmission at full channel capacity is a special case of semi-optimal transmission. Some generalizations and analogies to statistical mechanics are discussed.

Introduction

THE purpose of this note is twofold: 1) to give a result related to an inequality of McMillan pertaining to unique decipherability, and 2) to illustrate the close relationship between problems of information transmission and some of the elementary problems of statistical mechanics by use of notation borrowed from the latter field. In statistical mechanics we find that in principle all thermodynamic properties of a system are determined if we can evaluate its partition function Z; or better still, $\log Z$, in its dependence on the various constraints representing experimentally imposed conditions. Similarly, many problems of information transmission under constraints are, in principle, solved if we can evaluate an appropriate partition function. For the calculation of channel capacity, this is equivalent to the method described by Shannon. The same mathematical procedure also solves a wider class of problems, in which we find the transmission rate under what are termed "semi-optimal" conditions.

Channel Capacity Under Constraints

We have an alphabet of a symbols, each of which can be transmitted in unit time. Let l_i be the length (number of letters) of word w_i, and define the partition function[1]

$$Z(\lambda) = \sum_i 2^{-\lambda l_i} \tag{1}$$

where the sum is over all words in our vocabulary. A given vocabulary (*i.e.*, a specific set of possible words) may be regarded as defining a channel. By a theorem of Shannon,[2] the capacity of this channel is the largest (actually the only) real root of $Z(\lambda) = 1$.

The notion of a "word" is meaningful only if there exists some rule by which a sequence of letters can be uniquely deciphered into words. If no such rule exists, then effectively each letter is a word. The partition function then reduces to $Z(\lambda) = a2^{-\lambda}$, and Shannon's theorem gives the well-known channel capacity (all logarithms are to the base 2)

$$C = \log a \text{ bits/symbol.} \tag{2}$$

A necessary and sufficient condition for unique decipherability into words (UD) is given by Sardinas and Patterson.[3] McMillan[4] has given two inequalities implied by UD, which had been noted before[5,6] under more restrictive conditions.[7] The first, which perhaps deserves to be called the fundamental inequality of noiseless coding theory, is in our notation

$$Z(\log a) \leq 1. \tag{3}$$

Although, as several authors have shown,[4-8] this inequality can be derived without any reference to information theory, the concepts introduced by Shannon give it a simple intuitive meaning.

Any UD coding method is a system of constraints which in some way restricts our freedom in choosing the successive letters of a message, and defines a particular channel. Eq. (3) expresses the fact that imposing these constraints can never lead to a channel with greater capacity than the value (2), which corresponds to complete freedom of choice. Thus we conjecture that (3) will be fundamental not only for UD, but also for coding systems designed for any other objective. In general, a constraint will reduce channel capacity, and a reasonable measure of the efficiency of a code is the amount of this decrease.

* Manuscript received by the PGIT, Jan. 21, 1959. This research was supported by the USAF under Contract No. AF 49 (638)-342 monitored by the AF Office of Sci. Res. of the Air Res. and Dev. Command.

† Microwave Lab. and Dept. of Phys., Stanford University, Stanford, Calif.

[1] E. Schrödinger, "Statistical Thermodynamics," Cambridge University Press, Cambridge, Eng., chap. II; 1948.

[2] C. E. Shannon, "The Mathematical Theory of Communication," University of Illinois Press, Urbana, Illinois, p. 8; 1949.

[3] A. A. Sardinas and G. W. Patterson, "A necessary and sufficient condition for unique decomposition of encoded messages," IRE Convention Record, pt. 8, pp. 104–108; 1953.

[4] B. McMillan, "Two inequalities implied by unique decipherability," IRE Trans. on Information Theory, vol. IT-2, pp. 115–116; December, 1956.

[5] L. G. Kraft, "A device for quantizing, grouping and coding amplitude modulated pulses," S. M. Thesis, Dept. Elec. Eng., M.I.T., Cambridge, Mass.; 1949.

[6] B. Mandelbrot, "On recurrent noise limiting coding," *Proc. Symp. on Information Networks*, Polytechnic Inst. of Bklyn; New York, N. Y.; 1955.

[7] Given also in A. Feinstein, "Foundations of Information Theory," McGraw-Hill, New York, N. Y., pp. 17–23; 1958.

[8] M. P. Schützenberger and R. S. Marcus, "Full decodable code-word sets," IRE Trans. on Information Theory, vol. IT-5, pp. 12–15; March, 1959.

McMillan[4] considers also a strong sufficient condition for UD, called irreducibility, and shows that for fixed word lengths the strong constraint of irreducibility does not reduce channel capacity below that set by the general constraint of UD.

Irreducibility ensures UD only if the entire message is available. In applications to communication systems and to genetics,[9] the most common transmission defect is the one wherein certain parts of the message are simply lost. Although the probability that this would destroy UD for the entire balance of the message is usually very small,[10] one is led to ask for a stronger condition of UD "in the small." Without attempting a precise definition of this term, we use it in the rough sense that any reasonably long fragment of a message will still be uniquely decipherable, except for possible end-effects.

Golomb, Gordon, and Welch[9] consider channels with fixed word length k, and a structure constraint much stronger than UD, which ensures UD in the small. Among their results, they show (theorems 6 and 7) that for given k, in the limit of large alphabets even their strong constraint does not reduce capacity below the value (2).

Suppose we achieve UD by the usual method of choosing one of the letters, which we call the "space," and using it only as the terminating letter of each word; call this "spacing." Spacing is a stronger constraint than McMillan's irreducibility, but weaker than that of Golomb, Gordon, and Welch (although it still accomplishes the aim of UD in the small). We wish to find how much the channel capacity is reduced by spacing, and to show that this reduction is in fact zero, independently of the size of the alphabet, if spacing is the only constraint.

Due to the spacing constraint, the maximum number of different words of length l is not a^l, but only

$$(a-1)^{l-1}.$$

Evidently, any failure to include all of the short words in our vocabulary will have a further adverse effect on channel capacity, in addition to that imposed by spacing. Therefore we use all possible words of total length $l \leq L$, and the partition function (1) becomes

$$Z(\lambda) = \sum_{l=1}^{L} (a-1)^{l-1} 2^{-\lambda l} = \frac{1-(a-1)^L 2^{-\lambda L}}{2^\lambda - a + 1}. \tag{4}$$

Noting that for all real λ, $Z(\lambda)$ is a decreasing function, and that $Z(\lambda) \to L/(a-1)$ as $\lambda \to \log(a-1)$, it follows that if $L = (a-1)$, the exact channel capacity is $C = \log(a-1)$. If $L > (a-1)$, we find from (4) the inequalities

$$\log(a-1) < C \leq \log a. \tag{5}$$

The latter inequality in (5) is identical to (3), and it goes into an equality in the limit $L \to \infty$. But since $\log a$ is just the channel capacity we would have with an alphabet of a letters without any constraints, our assertion is proved. Stated differently, since (3) must hold for any method of achieving UD, *in the absence of other constraints, no method of achieving UD can be more efficient, as measured by channel capacity, than spacing.*

If $L < (a-1)$, then (4) leads to the inequalities

$$\left(1 - \frac{1}{L}\right) \leq \frac{C}{\log(a-1)} < 1, \tag{6}$$

the equality sign holding in the limit $a \to \infty$.

Numerical values of C obtained from (4) are given in Fig. 1. The trend for different a may seem disconcerting; one might argue that we are merely tying up one of the letters for special use, and so the loss in channel capacity could never exceed that due to removal of a single letter from the alphabet. However, the loss in capacity is in fact greatest for the large alphabets.

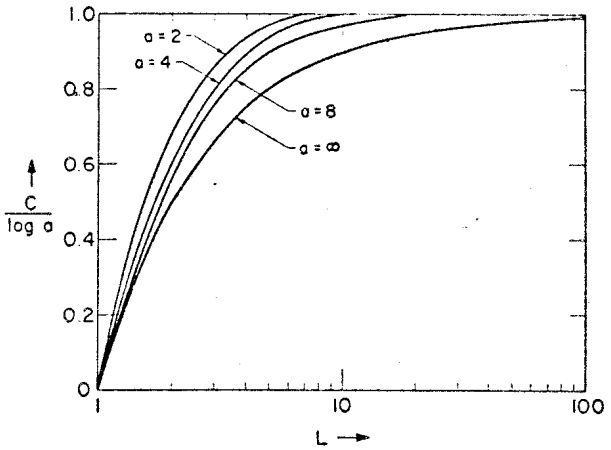

Fig. 1—Reduction in channel capacity due to restriction on maximum word length, for various alphabet sizes.

This situation may be understood as follows. Under no restrictions on word length, $L \to \infty$, we show in the next section that operation at full channel capacity requires a mean word length $\langle l \rangle = a$. The space then occurs with the same relative frequency, $(1/a)$, that it would have if it were not assigned any special function. This is the reason why UD, by itself, need not restrict transmission rate; the space is being used just as efficiently as any other symbol. However, when a becomes large, the effect of fixed L is to force $\langle l \rangle < a$. It is this tying up of channel time by too frequent repetition of the space which actually causes all the decrease in channel capacity, and explains the lower position, in Fig. 1, of the curves for large a.

Semi-Optimal Transmission

If the word w_i occurs with probability,

$$p_i = \frac{2^{-\lambda l_i}}{Z(\lambda)}, \tag{7}$$

the rate of transmission (entropy per word) is maximized

[9] S. W. Golomb, Basil Gordon, and L. R. Welch, "Comma-free codes," *Canadian Jour. Math.*, vol. 10, no. 2, pp. 202-209; 1958.

[10] M. P. Schützenberger, "On an application of semi-group methods to some problems in coding," IRE Trans. on Information Theory, vol. IT-2, pp. 47-60; September, 1956.

for a given mean length of word; equivalently, the average transmission time per word is minimized for a given entropy per word. The average length and entropy per word, under these semi-optimal conditions, are given by[11]

$$\langle l \rangle = -\frac{\partial}{\partial \lambda} \log Z(\lambda) \tag{8}$$

$$S = \log Z(\lambda) + \lambda \langle l \rangle, \tag{9}$$

which are parametric equations connecting the quantities of interest. From them we can construct the "operating characteristic" of the channel, in which we plot the time rate of transmission,

$$H = \frac{S}{\langle l \rangle} \text{ bits/symbol}, \tag{10}$$

as a function of the average word length $\langle l \rangle$. A family of operating characteristics, computed from (4) in the case of no restriction on maximum word length, *i.e.*, from the partition function

$$Z(\lambda) = (2^{\lambda} - a + 1)^{-1}, \tag{11}$$

is given in Fig. 2 for various alphabet sizes. The range of attainable operating conditions consists of all points lying below the curve.

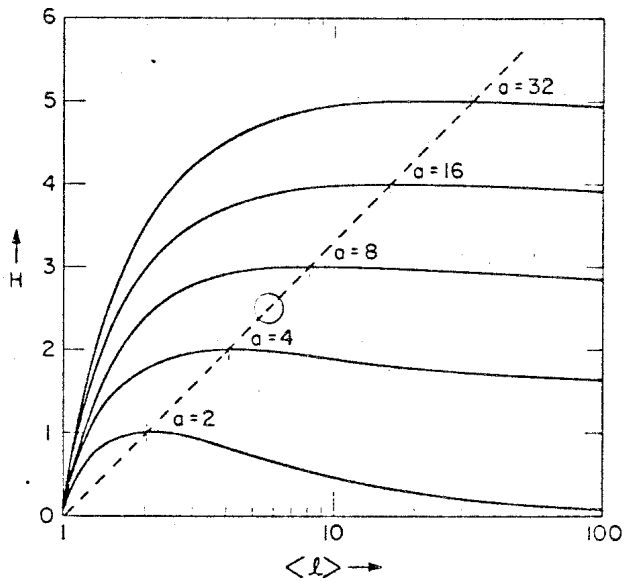

Fig. 2—Maximum attainable transmission rate as a function of average word length, for various alphabet sizes. For each curve, operation at full channel capacity occurs at the intersection with the dotted line.

The equation represented in Fig. 2 is found by eliminating λ from the above equations:

$$H = \log \langle l \rangle + \frac{\langle l \rangle - 1}{\langle l \rangle} \log \left[\frac{a - 1}{\langle l \rangle - 1} \right]. \tag{12}$$

[11] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620-630, May 15, 1957; vol. 108, pp. 171-190, October 15, 1957.

It is in the limit $\langle l \rangle \to \infty$ that we are effectively removing one letter from the alphabet, and so $H \to \log (a - 1)$.

From (8), (9), and (10) the condition for maximum transmission rate is

$$\frac{dH}{d\langle l \rangle} = -\frac{\log Z}{\langle l \rangle^2} = 0, \tag{13}$$

or $\log Z(\lambda) = 0$. Under these conditions, we find $H = \lambda = C$, the channel capacity; thus (8) and (9) provide a simple alternative derivation of Shannon's rule for calculating channel capacity.[2]

In the case of the partition function (11), operation at full channel capacity occurs when $H = \lambda = C = \log a$; and (8) then gives $\langle l \rangle = a$, as previously noted.

The case $a = 32$ corresponds to the English language, if we consider the space and any five punctuation marks as included in the alphabet. The average word length in English is far less than 32 symbols, and varies with the source. One thousand consecutive words from Shannon's fundamental paper[2] had a mean length (including the space) of 5.9 symbols; while a similar analysis of James Michener's "Sayonara" gave a mean length of only 5.4. From Fig. 2, we find that because of Michener's tendency to use short words, unique decipherability by spacing costs him 0.3 bit per symbol in information content, while Shannon's loss was only 0.25.

The many additional constraints in English cause the actual transmission rate to fall considerably below the semi-optimal rate. Taking Shannon's estimate[12] of the redundancy of English as about 50 per cent, the actual operating region of English text would be given roughly by the circle in Fig. 2. From this we see that 1) only about 15 per cent of the redundancy is due to use of short words, and 2) the same rate of information transmission and the same mean word length to which we are accustomed could be achieved with an alphabet of only 6 symbols (5 letters and a space), if used at maximum efficiency.

Generalization

The above relations are easily generalized to the case where the transmission time is different for different symbols, and where we have other types of constraints.

[12] Shannon, *loc. cit.*, p. 26. See also C. E. Shannon, "Prediction and Entropy of Printed English," *Bell Sys. Tech J.*, vol. 30, pp. 50-64; January, 1951. Here the estimated redundancy is increased to about 75 per cent, from experiments in which human subjects attempted to restore missing parts of English text. However, the ability to do this may depend on semantic as well as purely statistical factors; and in any event the only properties which could be utilized for encoding efficiently into a smaller alphabet, are known frequencies. Estimates based on measured letter and word frequencies remain not much greater then 50 per cent, the value used above. Even these measurements suffer from fundamental ambiguities, some of which were pointed out by G. A. Barnard, "Statistical calculation of word entropies for four western languages," IRE Trans. on Information Theory, vol. IT-1, pp. 49-53; March, 1955. Fundamentally, of course, it is meaningless to say that there exists one and only one "true" redundancy for English text; one can speak only of the redundancy corresponding to certain specified statistical information.

For example, let the k'th letter of the alphabet have transmission time t_k, and denote the space by the a'th letter. Then in (1) we interpret l_i as the total transmission time of word w_i, and its evaluation is again elementary, with the result

$$Z(\lambda) = \frac{2^{-\lambda t_a}}{1 - Q(\lambda)}, \tag{14}$$

where

$$Q(\lambda) \equiv \sum_{k=1}^{a-1} 2^{-\lambda t_k}.$$

From this the channel capacity and transmission rate under semi-optimal conditions may be found. Constraints of the form that certain combinations of letters do not occur may lead to involved mathematical problems, but not to any new difficulties of principle. Shannon's Theorem 1 shows[2] that one common type of constraint is included if we generalize the partition function to a "partition matrix," operation at full channel capacity then occurring when the greatest eigenvalue of this matrix is unity.

Of course, the quantity t_k above need not be interpreted as a time. It can equally well stand for the "cost," as measured on any basis, of transmitting the k'th symbol. The theory then gives us the method of transmitting which is most economical with respect to this cost assignment; semi-optimal transmission minimizes the average cost per word for a given entropy per word.

Relation to Statistical Mechanics

The partition function (11), with the number 2 replaced by e, is almost identical to the one arising in quantum statistical mechanics, describing a harmonic oscillator. The operating characteristic in Fig. 2, for the case $a = 2$, represents nothing more than an unconventional way of plotting the Einstein specific heat function of a harmonic oscillator.

Each of the soluble problems of statistical mechanics also provides the solution to a certain problem of information transmission under constraints, and the mathematical analogy may be set up in other ways than the one indicated here. In the above type of analogy, noted before by Mandelbrot,[6] the mean word length corresponds to the thermodynamic energy function, the parameter λ to the reciprocal temperature. Thus the transmission rate H corresponds, not to the thermodynamic entropy function, but to the ratio (entropy)/(energy).

It is interesting that such a fundamental notion as channel capacity has no thermodynamic analog. In thermodynamics the absolute value of the entropy has no meaning; only entropy differences can be measured in experiments. Consequently the condition that H is maximized, equivalent to the statement that the Helmholz free energy function vanishes ($A \equiv E - TS = 0$), corresponds to no condition which could be detected experimentally.

Generalization of the above analysis to the case where the different words are no longer statistically independent is also straightforward, and corresponds to the transition in statistical mechanics from the Maxwell-Boltzman "molecular" viewpoint, to the Gibbsian "global" viewpoint.[1] We mention two examples of the correspondence which then exists.

The partition function of the linear Ising chain,[13] with an easy generalization, provides also an explicit solution to the problem of encoding a message into binary digits, in a way which is optimal from the standpoint of a person who knows the digram frequencies of the source, but has no other statistical information. The corresponding solution for trigrams would be of considerable interest in connection with the theory of ferromagnetism.

The two-dimensional Ising model of ferromagnetism,[13] the partition function of which was first obtained by Onsager, gives the solution to a problem in which a message in binary digits has strong correlations between adjacent symbols, and also between the n'th and the $(n + M)$'th, where M is a large fixed number. Its most striking feature is a logarithmic singularity, signifying physically a phase transition (ferromagnetic Curie point). Translated into communication theory, it can be said that at a certain critical strength of the intersymbol correlations, as measured by the parameter λ, there occurs a sudden collapse of transmission rate to a very low value, $dH/d\lambda$ becoming infinite at a single point.[14]

Conclusion

Much of what we have said has already been pointed out by others.[6,15] However, the basic mathematical identity of these two fields has had, thus far, very little influence on the development of either. There is an inevitable difference in detail, because the applications are so different; but we should at least develop a certain area of common language, so that a worker in one field can decide quickly whether work in the other has a bearing on his problems.

We suggest that one way of doing this is to recognize that the partition function, for many decades the standard avenue through which calculations in statistical mechanics are "channeled," is equally fundamental to communication theory. Even within communication theory, there are advantages to be had by adopting this terminology

[13] G. F. Newell and E. W. Montroll, "On the theory of the Ising model of ferromagnetism," *Rev. Mod. Phys.*, vol. 25, pp. 353-389; April, 1953.

[14] This type of message structure strongly resembles that occurring in certain styles of music, where strong correlations appear after an interval of 2^n bars, n being a small integer. This phenomenon of collapse in transmission rate then has some amusing implications, which we leave for the reader to develop.

[15] A referee kindly informs me that the following reference also contains material along the lines discussed here: Apostel, Mandelbrot, and Morf, "Linguistic statistique macroscopique" in "Logique, Langage et Theorie de L'information," Presses Universitaires de France, Paris, France, pp. 1-78; 1957.

and notation as standard. For example, expressions of the form $\sum D^{-n_i}$, which occur repeatedly in coding theory, are really partition functions. The "rather algebraic" nature of this theory derives in part from the fact that often only one value of D is considered. If we generalize by setting $D = 2^{\lambda}$, with λ a continuously variable parameter, we have a true partition function, which has analytical properties useful in deriving theorems; indeed, this is just what McMillan[4] has done. A partition function $Z(\lambda)$ is, of course, the same as a generating function of the variable $t = 2^{-\lambda}$. However, from a general standpoint the partition function is a more powerful analytical tool because it remains single-valued under conditions where the generating function would develop an infinite number of Riemann surfaces.

The way in which the partition function varies for different values of λ often tells one the effect of some departure from ideal conditions. Thus, in the problem treated above, we see from inspection of Fig. 2 that in the case of small alphabets it is essential to encode in such a way that the mean word length is held close to the optimal value; while in a large alphabet the mean word length can vary widely with very little effect on attainable transmission rate.