INFORMATION AND CONTROL 2, 214–225 (1959)

Approximating Probability Distributions to Reduce Storage Requirements*

P. M. LEWIS II

General Electric Research Laboratory, Schenectady, New York,

and

Department of Electrical Engineering and Computer Components and Systems Group, Massachusetts Institute of Technology, Cambridge, Massachusetts

The measurement and/or storage of high order probability distributions implies exponential increases in equipment complexity. This paper considers the possibility of storing several of the lower order component distributions and using this partial information to form an approximation to the actual high order distribution.

The approximation method is based on an information measure for the "closeness" of two distributions and on the criterion of maximum entropy. Approximations consisting of products of appropriate lower order distributions are proved to be optimum under suitably restricted conditions. Two such product approximations can be compared and the better one selected without any knowledge of the actual high order distribution other than that implied by the lower order distributions.

I. INTRODUCTION

Limitations on allowable equipment complexity are an important factor in almost all systems containing computers. Indeed the performance of a great many of these systems is almost entirely determined by the size of the available machine memory and how that memory is used. In addition, it is not difficult to envision systems for which the amount of data inherent in the problem far exceeds the capacity of any foreseeable computer.

One such problem is the measurement and storage of high order discrete probability distributions. For example, the storage of an nth order binary distribution requires the use of about 2^n registers and the estima-

* The research was supported in part by the Department of the Navy, Bureau of Ships Contract NObsr 72716.

tion of the elements of such a distribution from sample functions requires the observation of an exponentially increasing number of binary symbols. Such exponential requirements (which occur frequently in information processing systems) are outside the bounds of engineering practicality.

In order to circumvent these difficulties, this paper considers the possibility of measuring and/or storing several of the lower order distributions that compose a high order distribution and using this limited information to form an approximation to the high order distribution. Stated in strictly mathematical terms, we shall be interested in the problem of approximating a high order probability distribution by some function of several of its lower order component distributions; however the related physical problem suggests that we limit ourselves to those approximations that conserve storage space. Thus the approximation should use only those component distributions that are particularly important in characterizing the high order distribution; it would be especially desirable if the approximation method itself performed this "important part" selection.

The method developed in this paper is based on an information measure for the "closeness" of two probability distributions and on the criterion of maximum entropy. Under certain conditions, the functional form of these approximations becomes particularly simple (just products of appropriate low order distributions) and the important part selection requires only a calculation of the entropies of these lower order distributions.

II. AN INFORMATION MEASURE FOR PROBABILITY DISTRIBUTIONS

We shall be concerned with nth order binary probability distributions of the form

$$P(x_1 , x_2 , \cdots , x_n)$$

with elements $P_0 , P_1 , \cdots , P_{2^n-1}$ (where P_j is the probability of the n-digit sequence that represents the number j in binary notation). We first propose a measure for the information content of such distributions.

The information contained in an nth order binary distribution is defined to be

$$\begin{aligned} I_p &= n \log 2 - H_s \\ &= n \log 2 + \sum_0^{2^n-1} P_j \log P_j \end{aligned}$$

where H_s is the entropy of the distribution.

This is a slightly different definition than that used when Information Theory is applied to communication channels. There we are concerned with the information contained in a sequence whose probability distribution is known; here we seek the information contained in the distribution.

When the distribution is known, the usual entropy measure has the property that the flatter the distribution, the more information (on the average) is contained in the sequences and the more peaked the distribution, the less information is contained in the sequences. Conversely, when the distribution is not known, the proposed measure has the property that the more the distribution is peaked, the more information it contains about the sequences. Thus, if the distribution is flat, we have very little *a priori* information about which sequence will occur and the sequences themselves give us a maximum amount of information; whereas, if the distribution is very peaked, we have a considerable amount of *a priori* knowledge about the sequences, and on the average the sequences give us little information.

Our definition implies that there is a fixed amount of information inherent in a process that generates a finite set of sequences, and that some is contained in the probability distribution and the rest in the reception of the sequences. This inherent information is defined to be the maximum amount of information that could be contained in the sequences.[1]

$$H_{s_{\max}} = \log 2^n = n \log 2$$

Thus, the information contained in the probability distribution is defined to be the difference between this maximum entropy and the actual entropy of the distribution

$$I_p = H_{s_{\max}} - H_s$$

If the distribution is perfectly flat, then $H_s = H_{s_{max}}$ and $I_p = 0$, that is, no information is contained in the distribution; whereas if the distribution is perfectly peaked $H_s = 0$, the sequences give no information, and the distribution contains an amount of information equal to n log 2. In general I_p lies between 0 and n log 2.

One of the key points of the proposed definition is that zero information corresponds to a flat distribution. We assume that if the distribution is flat we have no information about the sequences; and conversely,

[1] Feinstein (1958, p. 15, Theorem 1), for example, demonstrates that n log 2 is the maximum information content.

if we have no information at all about the sequences, we assume they are equally likely. Since the possible events are unique and clearly defined to be the 2^n possible sequences n digits long, there appears to be no reason not to make this assumption in the absence of any other knowledge.

If one accepts this assumption, then he must agree to act as if it were true, at least until he hears to the contrary. If, in fact, he later finds out that the sequences *are* equally likely, we then say that he has received no information, because his new knowledge does not change the way he acts (although it may give him more confidence that his actions are correct).

III. A MEASURE FOR THE CLOSENESS OF ONE PROBABILITY DISTRIBUTION TO ANOTHER

The problem in which we are interested is that of approximating one probability distribution with another. In order intelligently to perform such an approximation, we must define some criterion, a measure of how close the approximation is to the actual distribution. Such a criterion, based on the information theory model, will be developed in this section.

By an approximation to an nth order probability distribution

$$P(x_1, \cdots, x_n)$$

with elements $P_0, \cdots, P_{2^n-1}$ we mean any other set of 2^n nonnegative numbers labeled $P_0', \cdots, P_{2^n-1}'$, such that $\sum P_j' = 1$.

Each element of this approximate or primed distribution contains an amount of information equal to

$$i_{P'} = \log 2^n P_j'$$

and the average information that would be computed by a person using the primed distribution when the unprimed distribution is the correct one is obtained by averaging the above expression using the unprimed probabilities.

$$I_{P'} = n \log 2 + \sum_{j=0}^{2^n-1} P_j \log P_j'$$

Note that $I_{P'}$ is not the information contained in the primed distribution; rather it is the information that the primed distribution gives about the unprimed distribution.

The closeness of approximation is then defined to be the difference between the information contained in the true distribution and the in-

formation contained in the approximating distribution about the true distribution

$$I_{P-P'} = I_P - I_{P'} = \sum_0^{2^n-1} P_j \log P_j - \sum_0^{2^n-1} P_j \log P_j'$$

$$= \sum_0^{2^n-1} P_j \log \frac{P_j}{P_j'}$$

This measure has the following property[2]:

$I_{P-P'}$ is always greater than or equal to zero:

$$I_{P-P'} \geqq 0$$

the equal sign only holding if the primed and the unprimed distributions are identical. Thus, the measure is always positive if the two distributions are different, is zero only if they are identical, and in all cases defines a closeness of approximation in terms of its closeness to zero.

IV. AN APPROXIMATION CRITERION

The closeness measure defined above enables one to evaluate the "goodness" of any particular approximation, but it appears of little help in finding methods for accomplishing the approximation. In general several of the lower order component distributions are known or available, and the problem is to form an approximation to the high order distribution as a function of several of these lower order distributions. The difficulty with trying to apply the closeness measure to this problem is in the determination of efficient functional forms for the approximation. In order to develop such efficient forms, we shall make a slight detour, and in fact, introduce a new approximation criterion; later we shall show how the two criteria are related.

We first introduce the idea of an extension.[3] If we have several (compatible) lower order distributions given, then any higher order distribution that reduces to these lower order ones is called an extension of those particular distributions. Thus

$P_1(x_1, x_2)$		$P_2(x_1, x_2)$	
00	$\frac{1}{8}$	00	0
01	$\frac{1}{8}$	01	$\frac{1}{4}$
10	$\frac{3}{8}$	10	$\frac{1}{2}$
11	$\frac{3}{8}$	11	$\frac{1}{4}$

[2] For a proof of this well known result see for example Feinstein (1958, p. 20).

[3] The definition is due to Dr. J. Hartmanis of this Laboratory.

are both extensions of

$P(x_1)$	
0	$\frac{1}{4}$
1	$\frac{3}{4}$

$P(x_2)$	
0	$\frac{1}{2}$
1	$\frac{1}{2}$

In general, there is an infinite number of extensions to any given set of lower order distributions.[4] In degenerate cases, however, there may be only one possible extension. For example, the second order distributions

$P(x_1, x_2)$	$= P(x_2, x_3)$
00	0
01	$\frac{1}{2}$
10	$\frac{1}{2}$
11	0

can only be extended in one way to a third order distribution

$P(x_1, x_2, x_3)$	
000	0
001	0
010	$\frac{1}{2}$
011	0

$P(x_1, x_2, x_3)$	
100	0
101	$\frac{1}{2}$
110	0
111	0

In most cases of practical interest, this degenerate case will not happen and there will be a region of possible extensions in n-dimensional probability space.

We now propose to limit ourselves to approximations that are extensions. If we are given a set of lower order probability distributions, we shall only consider as possible approximations to the high order distribution, those functions that reduce to the given lower order distributions when properly summed or, in other words, those approximations that are extensions of these distributions. This is a considerable restriction on the generality of the approximation method, but it has at least two justifications: it is intuitively satisfying and it allows a simple and unique answer to the approximation problem in certain cases of interest.

The approximation problem can now be stated as follows. Given a set of lower order distributions, which, of all their possible extensions, should be used as an approximation to the higher order distribution from which the lower order distributions were derived. In order to decide this question, we now propose an approximation criterion:

[4] See the accompanying paper by Dr. J. Hartmanis (1959).

Of all the possible extensions, pick that one with the minimum information (maximum entropy).

Small information corresponds to randomness and large information to nonrandomness or bias. We should like our approximation to be as unbiased or random as possible (corresponding to the initial assumption that all distributions are equally likely *a priori*). Now in order for a proposed function to be an extension of a given set of probability distributions, the function must contain a certain minimum amount of information or bias; any additional information (we can argue intuitively) corresponds to additional bias on the part of the person doing the approximating, and consequently is to be avoided.

This minimum information criterion always yields a solution to the approximation problem, although sometimes the form of the solution is unwieldy. However, if we impose certain other constraints of particular importance to the storage problem, the form of the minimum information solution becomes particularly simple; in addition a close relation is found to exist between the minimum information criterion and the closeness measure defined earlier. In order to investigate this special case, we shall define a class of approximations called product approximations, and determine the conditions under which this class contains the minimum information solution.

V. PRODUCT APPROXIMATIONS

A product approximation is defined to be an approximation to a higher order distribution made up of a product of several of its lower order component distributions, such that the product is an extension of the lower order distributions. All of the product approximations to a third order distribution $P(x_1, x_2, x_3)$ are listed below

1. $P(x_1)\ P(x_2)\ P(x_3)$
2. $P(x_1, x_2)\ P(x_3)$
3. $P(x_1, x_3)\ P(x_2)$
4. $P(x_2, x_3)\ P(x_1)$
5. $P(x_1, x_2)\ P(x_3 \mid x_1)$
6. $P(x_1, x_2)\ P(x_3 \mid x_2)$
7. $P(x_1, x_3)\ P(x_2 \mid x_1)$
8. $P(x_1, x_3)\ P(x_2 \mid x_3)$
9. $P(x_2, x_3)\ P(x_1 \mid x_2)$
10. $P(x_2, x_3)\ P(x_1 \mid x_3)$

Not all products of lower order distributions are product approximations. For example,

$$P(x_1, x_2)\ P(x_2, x_3)\ P(x_1, x_3)$$

is not a product approximation since it does not reduce to $P(x_1, x_2)$ when summed over x_3 (nor does it sum to either of the other second order distributions, nor in fact, does it sum to unity).

The product approximation of an Nth order distribution contains at most a product of N terms, since it must be possible to write the terms down in a sequence such that each new term contains at least one variable (x_j) not contained in the previous terms. If the variables are then summed in reverse order back through the sequence, the unity sum property can be demonstrated. For example, one particular product expansion to a seventh order distribution can be written in sequence

$$P(x_1, x_2)\ P(x_5 \mid x_1)\ P(x_6 \mid x_5)\ P(x_3x_4 \mid x_6)\ P(x_7)$$

The unity sum property can be demonstrated by summing in sequence on x_7, x_4, x_3, x_6, x_5, x_2 and x_1.

With a little practice, appropriate product approximations can be written down by inspection.

As a class, product expansions suffer from the disadvantage of using only part of the available information. Thus even if one knows all three of the second order distributions that compose a third order distribution, he can only use two of them in forming a product approximation. In a more general case, there are $n(n-1)/2$ second order distributions in an Nth order distribution of which only $N-1$ can be used in the product approximation.

However, this apparent disadvantage of the product extension as a general approximation method becomes considerably less important when the storage problem is considered. One of the principal aspects of this problem is the throwing away of unimportant data and the selection of that information which is most important for the proposed application. Thus we are only interested in using part of the information (the important part) and the product approximation shows us how to use this partial information to obtain an approximation to the total information.

We shall now demonstrate that under certain conditions, the product approximation is the best way to use the partial information that we do have: namely, we shall prove the following result:

Given a set of lower order probabilities P_a, P_b, $\cdots$, P_n such that the product

$$P' = P_aP_b \cdots P_n$$

is a product approximation, then this product approximation contains

the smallest information of the entire class of extensions of P_a, P_b, $\cdots$, P_n.

The proof is straightforward; consider any other extension of P_a, P_b, $\cdots$, P_n; call this other extension distribution P''. According to the well known result in information theory referred to earlier

$$\sum P_j'' \log P_j'' \geqq \sum P_j'' \log P_j'$$

The expression on the right can be written

$$\begin{aligned}\sum P_j'' \log P_j' &= \sum P_j'' \log P_a P_b \cdots P_n \\ &= \sum P_j'' \log P_a \\ &\quad + \sum P_j'' \log P_b + \cdots + \sum P_j'' \log P_n\end{aligned}$$

Since the P'' distribution is also an extension of $P_a P_b \cdots P_n$ the terms on the right can be partially summed to obtain

$$\sum P_j'' \log P_j' = \sum P_a \log P_a + \sum P_b \log P_b + \cdots + \sum P_n \log P_n$$

But this is exactly (the negative of) the entropy of the P' distribution

$$\sum P_j'' \log P_j' = \sum P_j' \log P_j'$$

Thus the original expression becomes

$$\sum P_j'' \log P_j'' \geq \sum P_j' \log P_j'$$

which proves the result.

Thus we have shown that under the above restrictions, the product expansion is the best (minimum information) approximation.

For example, if we are given only two of the second order components of a third order distribution, the best approximation is their appropriate product. However, if we are given all three second order components, they do not form a product extension and the conditions of the result no longer apply. One could solve the minimum information optimization problem for this case, and a complicated function of all three probabilities would result. However, if one desires to retain the simplicity of the product approximations, he could consider all three of the possible product expansions obtained by neglecting one of the given second order distributions and, using the closeness measure, select the best. This would not yield quite as good an approximation as using all three distributions, but it offers the advantage of simplicity, and in addition represents the desired "important data" selection.

VI. CLOSENESS MEASURE APPLIED TO PRODUCT EXPANSIONS

The closeness measure takes a particularly simple form when applied to product expansions. Since the only term in the information measure

$$I_{p-p'} = \sum P_i \log P_i - \sum P_i \log P_i'$$

that depends on the approximating distribution is the last one, our first efforts will be directed toward evaluating the function

$$-\sum P_i \log P_i'$$

The approximating distribution P_i' consists of a product of lower order probabilities of P_i which we can symbolize as

$$P' = P_a P_b \cdots P_n$$

The expression we are trying to evaluate becomes

$$-\sum P_i \log P_a P_b \cdots P_n$$

and, on expanding the log,

$$-(\sum P_i \log P_a + \sum P_i \log P_b + \cdots + \sum P_i \log P_n)$$

Since the lower order distributions are components of the actual distribution, a partial summation of each particular term yields (for example)

$$\sum P_i \log P_a = \sum P_a \log P_a = -H_a$$

where H_a is the entropy of the P_a distribution. Then, the above expression becomes

$$-\sum P_i \log P_i' = (H_a + H_b + \cdots + H_n)$$

which, when substituted into the equation for the closeness measure, gives

$$\begin{aligned} I_{p-p'} &= (H_a + H_b + \cdots + H_n) - H_p \\ &= I_p - (I_a + I_b + \cdots + I_n) \end{aligned}$$

Thus the closeness measure can be expressed as the information in the actual distribution minus the sum of the informations in the (product) approximating distributions. In particular the second term in the closeness measure, which can be interpreted as the information given by the approximating distribution about the actual distribution, can be cal-

culated as the sum (and difference) of the entropies of the approximating distributions and *can thus be determined without knowing the actual higher order distribution.* This is a very basic mutual property of the product approximation and the information measure and is one of the main justifications for the definitions made in their development.

The use of this property greatly simplifies the comparison of two or more proposed product approximations, in order to select that approximation for which the closeness measure is smallest. Since I_p (the information in the actual distribution) is the same for all approximations, the best approximation is the one for which the sum of the informations in the approximating distributions is greatest. Thus, two or more proposed approximations can be compared and the best one selected without any knowledge of the actual distribution beyond that given by the approximations.

This last statement, surprising at first, is easily explainable. The process of comparison consists of selecting that approximation containing the greatest amount of correlation. Thus, if x_1 and x_{100} are less correlated than x_1 and x_2, then even without knowing the true distribution, it is clear that one would rather use the term $P(x_1, x_2)$ than the term $P(x_1, x_{100})$ in the approximation.

Furthermore, there are other types of approximation that have this property. If one has the choice of approximating a given function by using a finite number of terms from either of two different orthogonal

TABLE I

Sequence 3 2 1	True Probability $P(x_3x_2x_1)$	Approximations $P(x_3)P(x_2)$ $P(x_1)$	$P(x_3)P(x_2x_1)$	$P(x_3 \mid x_1)$ $P(x_2x_1)$	$P(x_3 \mid x_2)$ $P(x_2x_1)$
0 0 0	0.222	0.088	0.148	0.176	0.250
0 0 1	0.111	0.110	0.049	0.044	0.083
0 1 0	0	0.110	0.049	0.055	0.022
0 1 1	0.111	0.137	0.198	0.176	0.089
1 0 0	0.111	0.110	0.185	0.176	0.083
1 0 1	0	0.137	0.062	0.066	0.028
1 1 0	0.111	0.137	0.062	0.056	0.089
1 1 1	0.333	0.171	0.247	0.264	0.355
$I_{p-p'}$ (bits)	0	0.575	0.323	0.309	0.080

expansions, he can tell which is the better approximation, in the minimum mean square error sense, without knowing the true function, by just picking that expansion having the greatest mean square value.

This selection property relates the mathematical concepts of approximating probability distributions to the engineering problem of reducing equipment complexity, as was originally discussed in the introduction.

Example 1. Elementary results (Feinstein, 1958, Chapt. 2) in information theory can be used, in certain cases, to show that of two terms that are interchangeable in a product approximation, one always gives a better approximation. For example:

$$P(x_1 , x_2) \text{ better than } P(x_1)\, P(x_2)$$

$$P(x_1 \mid x_2) \text{ better than } P(x_1)$$

$$P(x_1 \mid x_2 , x_3) \text{ better than } P(x_1 \mid x_2)$$

These are properties that any meaningful measure would be expected to possess.

Example 2. *Approximation Characteristics.* In Table I are listed a third order probability distribution together with four product approximations, in order of increased goodness. At the bottom of each column is the approximation measure $I_{p-p'}$.

ACKNOWLEDGMENT

This research was done in close association with Dr. J. Hartmanis of the General Electric Research Laboratory whose work on a closely related topic appears in an accompanying paper. Dr. Hartmanis' criticisms and discussions have contributed significantly to the ideas and concepts of this paper.

RECEIVED: March 5, 1959; revised April 3, 1959.

REFERENCES

FEINSTEIN, A. (1958). "Foundations of Information Theory." McGraw-Hill, New York.

HARTMANIS, J. (1959). The application of some basic inequalities for entropy. *Inform. and Control* **2,** 199 (1959).