

A Note on Approximations to Discrete Probability Distributions*

DAVID T. BROWN†

*Department of Electrical Engineering, Massachusetts
Institute of Technology, Cambridge, Massachusetts*

An iterative method is presented which gives an optimum approximation to the joint probability distribution of a set of binary variables given the joint probability distributions of any subsets of the variables (any set of component distributions). The most significant feature of this approximation procedure is that there is no limitation to the number or type of component distributions that can be employed. Each step of the iteration gives an improved approximation, and the procedure converges to give an approximation that is the minimum information (i.e. maximum entropy) extension of the component distributions employed.

This note describes an iterative method of approximating discrete probability distributions of the type discussed by P. M. Lewis II in the last issue of this journal.¹ The iteration procedure to be presented converges to give an approximating probability distribution which, like Lewis' product approximation, is the minimum information (i.e. maximum entropy) extension of the component distributions employed. This approximation is more general than the product approximation, however, in that it can employ any set of component distributions.

Let p be the joint probability distribution $p(x_1, x_2, \dots, x_n)$ of the n binary variables $x_1, x_2, \dots, x_n$, and let p_i be a lower order or component distribution, that is, p_i is a joint probability distribution of some subset of the binary variables. The number of variables in this subset

* The note is adapted from a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Massachusetts Institute of Technology. The work was supervised by Professor P. M. Lewis II.

† Present address: IBM Product Development Laboratory, Poughkeepsie, N. Y.

¹ Some knowledge of Lewis' paper will be helpful, as many of its concepts and definitions will be used here.

defines the order of p_i . p_i can be obtained from p by summing over the variables in p but not in p_i . For example:

$$P(x_1 = 1) = P(x_1 = 1, x_2 = 0) + P(x_1 = 1, x_2 = 1)$$

In this paper, the lower case p denotes a probability distribution, while the upper case P denotes a term of this distribution with subscripts denoting component distributions and terms of component distributions.

Let the set of component distributions used in the approximation be $p_a, p_b, \dots, p_m$. There is no restriction on the size of this set or the orders of the component distributions, but obviously the approximation can be better if more and higher order distributions are used. The iteration procedure will be shown to converge to a distribution, p^r , which approximates p and is an extension of $p_a, p_b, \dots, p_m$. By this it is meant that $p_a, p_b, \dots, p_m$ are component distributions of p^r as well as of p . The optimum approximation is taken by Lewis to be that extension of the component distributions employed which contains minimum information, and it will be shown that p^r satisfies this criterion.

Initially the iteration procedure assumes a flat distribution with terms $P^0 = 2^{-n}$, thus, all sequences of the n binary variables are equally likely.² The next step is to modify p^0 so that it satisfies one of the component distributions, p_a , then a second modification causes another component distribution, p_b , to be satisfied, and so forth. When p_b is satisfied, p_a may no longer be satisfied, so the procedure will in general require each component distribution to be employed more than once before convergence is attained. Terms of the distribution at the j th step of iteration have the form:

$$P^j = P^{j-1} \frac{P_i}{P_i^{j-1}}.$$

P^{j-1} is a term at the $j - 1$ step in the iteration while P^j is the corresponding term at the j th step. P_i refers to a term in the i th component distribution (of the true distribution), while P_i^{j-1} refers to the corresponding term in the i th component distribution of the approximating distribution at the $j - 1$ step. Thus, P^{j-1} and P^j are two approximate probabilities of the same sequence of n digits while P_i^{j-1} is an approximate probability of occurrence of $k < n$ of the digits and P_i is the exact

² After a few steps of the iteration, a product approximation is achieved, so, in practice, a product approximation might be used as the initial step of the iteration to shorten the procedure.

probability of occurrence of the same k digits. For example, we might have:

$$P^j(x_1 = 1, x_2 = 0, x_3 = 1) = P^{j-1}(x_1 = 1, x_2 = 0, x_3 = 1) \frac{P(x_1 = 1, x_2 = 0)}{P^{j-1}(x_1 = 1, x_2 = 0)}.$$

The iteration procedure can now be written as follows:

$$\begin{aligned} P^0 &= 2^{-n} \\ P^1 &= P^0 \frac{P_a}{P_a^0} \\ P^2 &= P^1 \frac{P_b}{P_b^1} \\ &\vdots \\ P^{m+1} &= P^m \frac{P_m}{P_m^m} \\ P^{m+2} &= P^{m+1} \frac{P_a}{P_a^{m+1}} \\ &\vdots \end{aligned}$$

The component distributions need not be employed in any particular order, but the order used will have an effect on the rate of convergence. The following statements are proved in the appendix.

1. The approximation is at each step a unity sum distribution.
2. The approximation improves at each step according to Lewis' closeness measure.
3. The procedure converges to the minimum information extension of the component distributions employed.

Tables I and II are examples of the iteration procedure used to approximate a third order distribution when all three of its second order distributions are known. Table I approximates the distribution used as an example in Lewis' paper. It can be shown that the true distribution in this case is the only extension of all three second order component distributions, therefore, the approximation will be exact. The iteration begins with a product approximation and five steps of iteration are shown. p^6 is a considerable improvement over the product approxima-

TABLE I

Sequence $x_3 x_2 x_1$	True probability	p^2	p^3	p^4	p^5	p^6
000	0.222	0.250	0.204	0.218	0.235	0.211
001	0.111	0.083	0.109	0.115	0.095	0.107
010	0.0	0.022	0.018	0.015	0.012	0.011
011	0.111	0.089	0.115	0.096	0.102	0.115
100	0.111	0.083	0.107	0.091	0.098	0.110
101	0.0	0.028	0.024	0.020	0.016	0.015
110	0.111	0.089	0.115	0.120	0.099	0.112
111	0.334	0.356	0.310	0.325	0.343	0.319
I_{p-p^i} bits	0	0.080	0.065	0.063	0.052	0.042

$$p^2 = p(x_3 | x_2) p(x_2 x_1) \quad \text{A product approximation.}$$

$$p^3 = p^2 \frac{p(x_1 x_2)}{p^2(x_1 x_2)} \quad p^4 = p^3 \frac{p(x_2 x_3)}{p^3(x_2 x_3)}$$

$$p^5 = p^4 \frac{p(x_1 x_2)}{p^4(x_1 x_2)} \quad p^6 = p^5 \frac{p(x_1 x_3)}{p^5(x_1 x_3)}$$

TABLE II

Sequence $x_3 x_2 x_1$	True probability	p^1	p^2	p^3	p^4	p^5
000	0.300	0.200	0.267	0.291	0.287	0.284
001	0.100	0.200	0.133	0.114	0.113	0.115
010	0.100	0.100	0.100	0.109	0.112	0.115
011	0.100	0.100	0.100	0.086	0.088	0.086
100	0.100	0.100	0.133	0.114	0.117	0.116
101	0.100	0.100	0.067	0.080	0.083	0.085
110	0.100	0.100	0.100	0.086	0.083	0.085
111	0.100	0.100	0.100	0.120	0.117	0.114
I_{p-p^i} bits	0	0.076	0.026	0.014	0.013	0.012

$$\text{All } P^0 = 0.125 \quad p^1 = p^0 \frac{p(x_2 x_3)}{p^0(x_2 x_3)}$$

$$p^2 = p^1 \frac{p(x_1 x_2)}{p^1(x_1 x_2)} \quad p^3 = p^2 \frac{p(x_1 x_3)}{p^2(x_1 x_3)}$$

$$p^4 = p^3 \frac{p(x_2 x_3)}{p^3(x_2 x_3)} \quad p^5 = p^4 \frac{p(x_1 x_2)}{p^4(x_1 x_2)}$$

tion, but convergence is rather slow. Table II is an example that converges much more rapidly. A flat distribution is initially assumed, and the second step, p^2 , is identical to the best product approximation; while at the fifth step, the procedure has virtually converged. In the tables I_{p-p^j} is the measure of closeness to the true distribution.

APPENDIX

Note that summation is always over the terms of the distributions, not the subscripts or superscripts of P .

1. Proof that $\sum P^j = 1$.

The approximating distribution is at each step unity sum.

$$\sum P^j = \sum P^{j-1} \frac{P_i}{P_i^{j-1}}$$

Now partial summations may be performed on terms for which P_i and P_i^{j-1} are constant to give

$$\sum P^{j-1} \frac{P_i}{P_i^{j-1}} = \sum P_i = 1.$$

2. Proof that each step of the iteration procedure yields an improved approximation, and that the procedure will converge to an extension of the component distributions employed.

Lewis takes $I_{p-p^j} = \sum P \log(P/P^j)$ as a measure of the closeness of p^j to p . This function is positive and decreases as the approximation, p^j , improves.

Expanding the logarithm and making a substitution for P^j gives

$$I_{p-p^j} = \sum P \log P - \sum P \log P^{j-1} \frac{P_i}{P_i^{j-1}}.$$

Now expand the second logarithm.

$$I_{p-p^j} = \sum P \log P - \sum P \log P^{j-1} - \sum P \log P_i + \sum P \log P_i^{j-1}$$

The first two terms are the definition of $I_{p-p^{j-1}}$. The second two terms can be partially summed in a manner similar to that in proof 1 to give

$$I_{p-p^j} = I_{p-p^{j-1}} - \sum P_i \log P_i + \sum P_i \log P_i^{j-1}.$$

But in general

$$\sum P \log P \geq \sum P \log P'.$$

Applying this general result to the above sums yields

$$I_{p-p^i} \leq I_{p-p^{i-1}}$$

proving that the approximation improves at each step.

Equivalently,

$$\sum P \log P^j \geq \sum P \log P^{j-1}.$$

The equality holds only if $p^j = p^{j-1}$, but $p^j = p^{j-1}$ only if $p_i = p_i^{j-1}$. Thus I_{p-p^i} steadily decreases if the component distributions are not satisfied, but $I_{p-p^i} \geq 0$, therefore convergence of p^j to p^r which is an extension of $p_a, p_b, \dots p_m$ is assured.

3. Proof that p^r , the limiting value of the iteration, is the minimum information extension of $p_a, p_b, \dots p_m$, the component distributions employed.

Terms of the distribution p^r , which is the final value of the iteration, can be written

$$P^r = 2^{-n} \frac{P_a}{P_a^0} \frac{P_b}{P_b^1} \frac{P_c}{P_c^2} \dots$$

Now we can write

$$\begin{aligned} \sum P^r \log P^r &= -n \log 2 + \sum P^r \log \frac{P_a}{P_a^0} + \sum P^r \log \frac{P_b}{P_b^1} \\ &\quad + \sum P^r \log \frac{P_c}{P_c^2} + \dots \end{aligned}$$

Here, the logarithm has been expanded and summation performed on the first term. Because p^r is shown in proof 2 to be an extension of $p_a, p_b, p_c, \dots$, partial summations may be performed in a manner similar to that in the previous proofs to give

$$\begin{aligned} \sum P^r \log P^r &= -n \log 2 + \sum P_a \log \frac{P_a}{P_a^0} + \sum P_b \log \frac{P_b}{P_b^1} \\ &\quad + \sum P_c \log \frac{P_c}{P_c^2} \dots \end{aligned}$$

It is to be shown that of the entire set of extensions to $p_a, p_b, \dots p_m$;

p^r is that extension with minimum information. Let p^* be any extension of $p_a, p_b, \dots, p_m$.

$$\begin{aligned} \sum P^* \log P^r &= -n \log 2 + \sum P^* \log \frac{P_a}{P_a^0} + \sum P^* \log \frac{P_b}{P_b^1} \\ &\quad + \sum P^* \log \frac{P_c}{P_c^2} + \dots \end{aligned}$$

and partial summation can be performed because p^* is an extension.

$$\begin{aligned} \sum P^* \log P^r &= -n \log 2 + \sum P_a \log \frac{P_a}{P_a^0} + \sum P_b \log \frac{P_b}{P_b^1} \\ &\quad + \sum P_c \log \frac{P_c}{P_c^2} + \dots \end{aligned}$$

Therefore,

$$\sum P^* \log P^r = \sum P^r \log P^r$$

But

$$\sum P^* \log P^r \leq \sum P^* \log P^*$$

This is the same general inequality used in proof 2.

So

$$\sum P^r \log P^r \leq \sum P^* \log P^*$$

As the information in p^r is taken as $n \log 2 + \sum P^r \log P^r$ and the information in any extension is $n \log 2 + \sum P^* \log P^*$, the proof that p^r is the minimum information extension is complete.

RECEIVED: May 1, 1959.

REFERENCE

LEWIS, P. M. II (1959). Approximating probability distributions to reduce storage requirements. *Inform. and Control* **2**, 214-225.