$$|I_{n_1,n_2}| \leq M\left\{c_0 + \left(\frac{\pi r}{18}\right)^{1/2} \tan^{3/2}\left(\frac{\pi r}{2}\right)\right\}$$

as required. Q.E.D.

ACKNOWLEDGMENT

The author thanks H. S. Piper, Jr., for many helpful discussions; his work on truncation error bounds which includes a further sharpening of the bound in the finite energy case is being prepared for publication.

REFERENCES

[1] H. D. Helms and J. B. Thomas, "Truncation error of sampling-theorem expansions," *Proc. IRE*, vol. 50, pp. 179–184, February 1962.
[2] K. Yao and J. B. Thomas, "On truncation error bounds for sampling representations of band-limited signals," *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-2, pp. 640–647, November 1966.
[3] D. Jagerman, "Bounds for truncation error of the sampling expansion," *J. SIAM*, vol. 14, pp. 714–723, July 1966.
[4] A. Papoulis, "Truncated sampling expansions," *IEEE Trans. Automatic Control*, vol. AC-12, pp. 604–605, October 1967.
[5] K. Knopp, *Theory and Application of Infinite Series*. Glasgow, Scotland: Blackie, 1928.

Approximating Discrete Probability Distributions

HARRY H. KU AND SOLOMON KULLBACK, SENIOR MEMBER, IEEE

***Abstract*—The method of minimum discrimination information estimation is applied to the problem of estimating an n-dimensional discrete probability distribution in terms of lower order marginal distributions. The procedure provides a convergent iterative algorithm. The method yields regular best asymptotically normal (RBAN) estimates. The general procedure includes as a particular case that proposed by a method using dependence trees. An example is given.**

I. INTRODUCTION

IT HAS BEEN pointed out that the problem of estimating an n-dimensional discrete probability distribution from a finite number of samples and storing the distributions in a certain limited amount of machine memory arises in designing information systems, such as communication, pattern recognition, and learning systems. In [1], [8] the problem of approximating an nth order binary distribution by a product of several of its component distributions of lower order was considered. In [2], a method to approximate optimally an n-dimensional discrete probability distribution by a set of $n - 1$ first-order dependence relationships among the n variables was presented. The procedure in [2] involves an optimization process to construct a dependence tree of maximum weight. As a matter of fact, the concept of a dependence tree does not seem to be necessary unless one is restricted to use only $n - 1$ second-order marginal distributions. Closer approximations can be obtained by a straightforward convergent iterative algorithm in terms of lower order marginal distributions. We shall see that the approximation in [2] is one of the early iterates.

Manuscript received August 21, 1968; revised December 3, 1968. This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-68-1513.

The authors are with the National Bureau of Standards, Washington, D. C. 20234, and the George Washington University, Washington, D. C. 20006.

The basic problem in the discrete case may be formulated as the statistical problem of estimating the cell entries of a multidimensional contingency table given a set of lower order marginal distributions. We may consider the complete sample table to contain all the "information" available from the particular experiment. In the process of analysis, we aim to express the sample table in a reduced number of parameters represented by the lower order marginal distributions. In other words, we are interested in knowing how much of this total information is contained in a summary consisting of sets of marginal distributions. The above interpretation is not restricted to complete sets of marginals. If the estimate computed from three out of the six second-order marginals in a four-variate table is found to be statistically "close enough" to the complete sample table, the three second-order marginals could be considered as containing essentially all the information in the four-variate table [5]. The underlying theory, properties of the estimates, proof of convergence of the iterative algorithm, and applications to contingency table analyses are given in [3]–[5]. We shall summarize the basic results and properties and apply them to the example presented in [2]. We remark that the corresponding results including appropriate modifications of the convergence proof for the case of continuous distributions are given in [7].

II. MINIMUM DISCRIMINATION INFORMATION ESTIMATION

For convenience we discuss the results in terms of a four-variate distribution (the example is also four-variate) but the general results are easily apparent. Let $\pi(\mathbf{x})$, $\mathbf{x} = (i, j, k, l)$, $i = 0, 1, \cdots, r - 1$, $j = 0, 1, \cdots, s - 1$, $k = 0, 1, \cdots, t - 1$, $l = 0, 1, \cdots, u - 1$, be a discrete

probability distribution that may be specified by hypothesis, given by observations, or derived by some estimation procedure. Let $p(\mathbf{x})$ be a member of a class of discrete probability distributions having a specified set of some lower order marginal distributions. The minimum discrimination information estimate is $p^*(\mathbf{x})$, where $p^*(\mathbf{x})$ minimizes the discrimination information

$$I(p : \pi) = \sum_{\mathbf{x}} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\pi(\mathbf{x})} \tag{1}$$

over the class of $p(\mathbf{x})$ distributions, that is, $p^*(\mathbf{x})$ is that member of the class of $p(\mathbf{x})$ distributions that is "closest" to $\pi(\mathbf{x})$ in terms of the measure of divergence (1). It has been shown [5, th. 2.3] that

$$I(p : \pi) = I(p^* : \pi) + I(p : p^*) \tag{2}$$

where $p(\mathbf{x})$ is any member of the class, and [3] that $p^*(\mathbf{x})$ is unique and a regular best asymptotically normal (RBAN) estimate, that is, it has the statistical properties of a maximum likelihood estimate.

We remark that if $\pi(\mathbf{x})$ is a member of the class of $p(\mathbf{x})$ distributions then $p^*(\mathbf{x}) = \pi(\mathbf{x})$ and $I(p^* : \pi) = 0$. A reasonable criterion for the choice of $\pi(\mathbf{x})$ in the present context, is to assume that dependence, or interactions, among the various variables are in fact fully represented by the lower order marginals. Equivalently, because of the form of the $p^*(\mathbf{x})$ distribution which we examine shortly, one may say that the estimates of the cell entries represent a "generalized" independence distribution subject to restraints imposed by the lower order marginal distributions. This assumption is tacitly made in the dependence tree scheme proposed in [2], and is explicitly stated here. Since the uniform distribution is a general form of independence, we may take $\pi(\mathbf{x})$ to be the uniform distribution $\pi_0(\mathbf{x}) = 1/rstu$ and compute $p_1^*(\mathbf{x})$ when the discrete distribution is to be approximated by first-order marginals, compute $p_2^*(\mathbf{x})$ when the discrete distribution is to be approximated by second-order marginals, etc. We remark that since higher order marginals determine lower order marginals we may also compute $p_2^*(\mathbf{x})$ with $\pi(\mathbf{x}) = p_1^*(\mathbf{x})$, $p_3^*(\mathbf{x})$ with $\pi(\mathbf{x}) = p_1^*(\mathbf{x})$ or $p_2^*(\mathbf{x})$, etc.

We shall now indicate the form of the p^* distribution under various sets of marginal restraints. The basis for the statements is given in [3] and further discussed and illustrated in [5].

When the first-order marginals

$$p(i),\ p(j),\ p(k),\ p(l) \tag{3}$$

are given, where for example,

$$p(i) = \sum_{j,k,l} p(i, j, k, l) \tag{4}$$

we denote the minimizing distribution by p_1^*, and it has the form

$$p_1^*(i, j, k, l) = a(i)b(j)c(k)d(l)\pi(i, j, k, l) \tag{5}$$

where $a(i), \cdots, d(l)$ are determined to satisfy the marginal restraints. When the π distribution satisfies the criterion of independence, that is,

$$\pi(i, j, k, l) = \pi(i)\pi(j)\pi(k)\pi(l),$$

then

$$p_1^*(i, j, k, l) = p(i)p(j)p(k)p(l). \tag{6}$$

When the second-order marginals

$$p(i, j),\ p(i, k),\ p(i, l),\ p(j, k),\ p(j, l),\ p(k, l) \tag{7}$$

are given, where for example,

$$p(i, j) = \sum_{k,l} p(i, j, k, l), \tag{8}$$

we denote the minimizing distribution by p_2^*, and it has the form

$$p_2^*(i, j, k, l) = a(i, j)b(i, k)c(i, l) \cdot d(j, k)e(j, l)f(k, l)\pi(i, j, k, l) \tag{9}$$

where $a(i, j), \cdots, f(k, l)$ are determined to satisfy the marginal restraints

$$p(i, j) = a(i, j) \sum_{k,l} b(i, k)c(i, l) \cdots f(k, l)\pi(i, j, k, l)$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \tag{10}$$
$$p(k, l) = f(k, l) \sum_{i,j} a(i, j)b(i, k) \cdots e(j, l)\pi(i, j, k, l).$$

When the third-order marginals

$$p(i, j, k),\ p(i, j, l),\ p(i, k, l),\ p(j, k, l) \tag{11}$$

are given, where for example,

$$p(i, j, k) = \sum_{l} p(i, j, k, l), \tag{12}$$

we denote the minimizing distribution by p_3^*, and it has the form

$$p_3^*(i, j, k, l) = a(i, j, k)b(i, j, l) \cdot c(i, k, l)\ d(j, k, l)\pi(i, j, k, l) \tag{13}$$

where $a(i, j, k), \cdots, d(j, k, l)$ are determined to satisfy the marginal restraints

$$p(i, j, k) = a(i, j, k) \cdot \sum_{l} b(i, j, l)c(i, k, l)d(j, k, l)\pi(i, j, k, l)$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \tag{14}$$
$$p(j, k, l) = d(j, k, l) \cdot \sum_{i} a(i, j, k)b(i, j, l)c(i, k, l)\pi(i, j, k, l).$$

When the subset

$$p(i, j),\ p(i, l),\ p(j, k) \tag{15}$$

of the second-order marginals is given (see [2] and the example) we denote the minimizing distribution by p_a^*, and it has the form

$$p_a^*(i, j, k, l) = a(i, j)b(i, l)c(j, k)\pi(i, j, k, l) \tag{16}$$

where $a(i, j)$, $b(i, l)$, $c(j, k)$ are determined to satisfy the marginal restraints

$$\begin{aligned} p(i, j) &= a(i, j) \sum_k \sum_l b(i, l)c(j, k)\pi(i, j, k, l) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p(j, k) &= c(j, k) \sum_i \sum_l a(i, j)b(i, l)\pi(i, j, k, l). \end{aligned} \tag{17}$$

Note that the subset (15) implies the first-order marginals.

The p^* distribution may be determined by a convergent iterative procedure successively satisfying the marginal restraints. The proof that the iterative algorithm converges to the p^* distribution is given in [3] for the discrete case and in [7] for the continuous case. For the p_2^* distribution the iteration cycles through

$$\begin{cases} p^{(6n+1)}(i, j, k, l) = \dfrac{p(i, j)}{p^{(6n)}(i, j)} p^{(6n)}(i, j, k, l) \\ p^{(6n+2)}(i, j, k, l) = \dfrac{p(i, k)}{p^{(6n+1)}(i, k)} p^{(6n+1)}(i, j, k, l), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p^{(6n+6)}(i, j, k, l) = \dfrac{p(k, l)}{p^{(6n+5)}(k, l)} p^{(6n+5)}(i, j, k, l); \end{cases} \tag{18}$$

for the p_3^* distribution, the iteration cycles through

$$\begin{cases} p^{(4n+1)}(i, j, k, l) = \dfrac{p(i, j, k)}{p^{(4n)}(i, j, k)} p^{(4n)}(i, j, k, l) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p^{(4n+4)}(i, j, k, l) = \dfrac{p(j, k, l)}{p^{(4n+3)}(j, k, l)} p^{(4n+3)}(i, j, k, l); \end{cases} \tag{19}$$

for the p_a^* distribution the iteration cycles through

$$\begin{cases} p^{(3n+1)}(i, j, k, l) = \dfrac{p(i, j)}{p^{(3n)}(i, j)} p^{(3n)}(i, j, k, l) \\ p^{(3n+2)}(i, j, k, l) = \dfrac{p(i, l)}{p^{(3n+1)}(i, l)} p^{(3n+1)}(i, j, k, l) \\ p^{(3n+3)}(i, j, k, l) = \dfrac{p(j, k)}{p^{(3n+2)}(j, k)} p^{(3n+2)}(i, j, k, l) \end{cases} \tag{20}$$

with $p^{(0)}(i, j, k, l) = \pi(i, j, k, l)$.

When the p_1^* distribution is given as in (6) by using the relation (2) (see [5]), we have

$$\begin{cases} I(p : p_a^*) = I(p_2^* : p_a^*) + I(p : p_2^*) \\ I(p : p_1^*) = I(p_a^* : p_1^*) + I(p : p_a^*) \\ I(p : p_1^*) = I(p_2^* : p_1^*) + I(p : p_2^*) \\ I(p : p_2^*) = I(p_3^* : p_2^*) + I(p : p_3^*) \end{cases} \tag{21}$$

and in the iterations (18)–(20), we may take $p^{(0)}(i, j, k, l) = p_1^*(i, j, k, l)$, and in (19) we may also take $p^{(0)}(ijkl) = p_2^*(ijkl)$. Since the discrimination information values in (21) are ≥ 0,

$$I(p : p_1^*) \geq I(p : p_a^*) \geq I(p : p_2^*) \geq I(p : p_3^*) \tag{22}$$

with equality in the first, second, and third pair, respectively, if and only if $p_1^* = p_a^*$, $p_a^* = p_2^*$, $p_2^* = p_3^*$ (see [6, pp. 14–18]).

Using the value of $I(p : p^*)$ as a measure of the goodness of the approximation p^* to p, the inequalities in (22) provide the ordering (in increasing goodness) $p_1^*, p_a^*, p_2^*, p_3^*$.

III. Example

In Table I is given the binary probability distribution $p(i, j, k, l)$ used in [2] and also the p_1^* distribution computed as in (6). The optimization procedure in [2] is selected to use the branches of the dependence tree corresponding to the second-order marginals $p(j, k)$, $p(i, j)$, $p(i, l)$.

If we now follow the iteration (20), with $p^{(0)}(i, j, k, l) = p_1^*(i, j, k, l)$, and starting the cycles with the sequence of the marginals as selected in [2], we get

$$\begin{cases} p^{(1)}(i, j, k, l) = \dfrac{p(j, k)}{p(j)p(k)} p(i)p(j)p(k)p(l) \\ \qquad = p(j, k)p(i)p(l) \\ p^{(2)}(i, j, k, l) = \dfrac{p(i, j)}{p(i)p(j)} p(j, k)p(i)p(l) \\ \qquad = \dfrac{p(i, j)p(j, k)p(l)}{p(j)} \\ p^{(3)}(i, j, k, l) = \dfrac{p(i, l)}{p(i)p(l)} \cdot \dfrac{p(i, j)p(j, k)p(l)}{p(j)} \\ \qquad = p(j, k)p(i, j)p(i, l)/p(i)p(j), \end{cases} \tag{23}$$

and since, as may be verified,

$$p^{(3)}(j, k) = p(j, k),\ p^{(3)}(i, j) = p(i, j),$$
$$p^{(3)}(i, l) = p(i, l), \tag{24}$$

further iteration merely reproduces the $p^{(3)}$ distribution so that $p_a^* = p^{(3)}$. Note that the order here is different from that in (20) but the reader may verify that $p^{(3)}(i, j, k, l)$ is the same in (20) and (23). The optimization procedure in [2] yields as the optimum approximation

$$\begin{aligned} p_a(i, j, k, l) &= p(i)p(j \mid i)p(k \mid j)p(l \mid i) \\ &= p(j, k)p(i, j)p(i, l)/p(i)p(j). \end{aligned} \tag{25}$$

Its numerical values are given in Table I, and we note that

$$p^{(3)}(i, j, k, l) = p_a(i, j, k, l) = p_a^*(i, j, k, l).$$

Using the iterative relationships, and following the procedure used in proving the convergence of the iterative algorithm in [3], it may be shown that

$$I(p : p^{(n)}) = I(p^{(n+1)} : p^{(n)}) + I(p : p^{(n+1)}) \tag{26}$$

$$I(p : p_a) = I(p_2^* : p_a) + I(p : p_2^*) \tag{27}$$

where p_a is any of the iterates in the procedure leading to p_2^*, and consequently

$$I(p : p_a) \geq I(p : p_2^*) \tag{28}$$

TABLE I

i	j	k	l	$p(i, j, k, l)$	$p_1^*(i, j, k, l)$	$p_a(i, j, k, l)$	$p_2^*(i, j, k, l)$
0	0	0	0	0.100	0.04556	0.130	0.09977
0	0	0	1	0.100	0.04556	0.104	0.10000
0	0	1	0	0.050	0.05569	0.037	0.04958
0	0	1	1	0.050	0.05569	0.030	0.04927
0	1	0	0	0.000	0.05569	0.015	0.00051
0	1	0	1	0.000	0.05569	0.012	0.00026
0	1	1	0	0.100	0.06806	0.068	0.10011
0	1	1	1	0.050	0.06806	0.054	0.05035
1	0	0	0	0.050	0.05569	0.053	0.05027
1	0	0	1	0.100	0.05569	0.064	0.09996
1	0	1	0	0.000	0.06806	0.015	0.00039
1	0	1	1	0.000	0.06806	0.018	0.00076
1	1	0	0	0.050	0.06806	0.033	0.04945
1	1	0	1	0.050	0.06806	0.040	0.04978
1	1	1	0	0.150	0.08319	0.149	0.14992
1	1	1	1	0.150	0.08319	0.178	0.14962
				1.000	1.00000	1.000	1.00000

with equality if and only if

$$p_a(i, j, k, l) = p_2^*(i, j, k, l).$$

Carrying out the iteration (18) until there is agreement to at least 0.0014 in the second-order marginals we get values tabulated as p_2^* in Table I. Recomputing certain results given in [2] we list in Table II decompositions corresponding to (26) and (27).

IV. Remarks

The various relationships given are valid either for theoretical distributions or observed distributions. It would be of interest to compare the results of applying the p_2^* distribution to the pattern recognition problem described in [2] using their optimal approximation.

In this discussion we have not considered the possible hypotheses and tests of significance which arise in the statistical applications to contingency table analysis as discussed in [3], [4], [5].

TABLE II

$I(p : p_1^*)$	0.36867
$I(p^{(1)} : p_1^*)$	0.18899
$I(p : p^{(1)})$	0.17968
$I(p^{(2)} : p^{(1)})$	0.07943
$I(p : p^{(2)})$	0.10025
$I(p^{(3)} : p^{(2)})$	0.00506
$I(p : p^{(3)})$	0.09519
$I(p_2^* : p^{(3)})$	0.08539
$I(p : p_2^*)$	0.00980

Acknowledgment

The authors wish to thank Y. Molk of George Washington University and Mrs. R. Varner of the National Bureau of Standards for the programming and computations.

References

[1] D. T. Brown, "A note on approximations to discrete probability distributions," *Inform. and Control*, vol. 2, pp. 386–392, December 1959.
[2] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Information Theory*, vol. IT-14, pp. 462–467, May 1968.
[3] C. T. Ireland and S. Kullback, "Contingency tables with given marginals," *Biometrika*, vol. 55, pp. 179–188, March 1968.
[4] ——, "Minimum discrimination information estimation," *Biometrics*, vol. 24, pp. 707–713, September 1968.
[5] H. H. Ku and S. Kullback, "Interaction in multidimensional contingency tables: an information theoretic approach," *J. Research NBS*, vol. 72B, pp. 159–199, September 1968.
[6] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959. New York: Dover, 1968.
[7] ——, "Probability densities with given marginals," *Ann. Math. Stat.*, vol. 39, pp. 1236–1243, August 1968.
[8] P. M. Lewis, "Approximating probability distributions to reduce storage requirement," *Inform. and Control*, vol. 2, pp. 214–225, September 1959.