# A Probabilistic Version of Sperner's Theorem, With Applications to the Problem of Retrieving Information From a Data Base

L. D. Baumert, R. J. McEliece, E. R. Rodemich, and H. Rumsey, Jr.
Communications Systems Research Section

*We show how the design of an optimal "merged keycode" information retrieval system involves finding the probability distribution on n-bit binary words that minimizes* $P\{X \leqslant Y_1 \cup \ldots \cup Y_r\}$ *where* $X, Y_1, \ldots, Y_r$ *are selected independently according to the given probability distribution. We then find the minimizing probability distribution in the case* $r = 1$.

## I. Introduction

In retrieving information from a large data base, such as will exist in the DSN's digital RFI Surveillance System (Ref. 1), the technique of *merged keycodes* (Ref. 2) is often useful. In this technique each record in the data base is assumed to have a certain number of attributes. Each possible attribute $A$ is assigned a binary code (whose length is normally that of one computer word) $C(A)$. If the record $R$ possesses attributes $A_1$, $A_2, \ldots,$ it is then assigned the merged keycode $C(R) = C(A_1) \cup C(A_2) \cup \ldots$, where the symbol $\cup$ denotes the logical OR operation.

If one wants to locate all records in the data base possessing a fixed set of attributes, say $B_1, B_2, \ldots, B_s$, one computes the merged keycode $D = C(B_1) \cup \ldots \cup C(B_s)$, and then tags each record $R$ such that $D \leqslant C(R)$. Clearly the set of records with the desired attributes is a subset of the tagged records. However, some of the tagged records will not have the desired attributes; such records are called *false drops*. It is obviously desirable to minimize the number of false drops, other things being equal.

Under certain circumstances, it is reasonable to model the above situation as follows. A fixed number, say $r$, binary codewords are selected independently according to a certain probability distribution $P$, which is to some extent controlled by the system design. Denote these codewords by $Y_1, Y_2, \ldots, Y_r$ – they represent the keycodes of the attributes of a randomly selected record. Let $X_1, \ldots, X_s$, $s \leqslant r$ denote the keycodes corresponding to the attributes in a random query of the data base. We assume the $X_i$ are chosen independently according to the same probability distribution $P$. The probability of false drop is then

$$P\{X_1 \cup \ldots \cup X_s \leqslant Y_1 \cup \ldots \cup Y_r\} = P\{X_1 \leqslant Y_1 \cup \ldots \cup Y_r\}^s$$

Thus we are led to ask: What is the probability distribution on $n$-bit binary words that minimizes the probability $P\{X_1 \leqslant Y_1 \cup \ldots \cup Y_r\}$ for various values of $r$?

In this article, we will solve this question for the case $r = 1$. It is hoped that the techniques developed can be brought to bear for larger values of $r$. In the remainder of this section, we will describe our main result.

Let $\Omega$ be a finite set with $n$ elements, and let $V = V(\Omega)$ denote the collection of all subsets of $\Omega$. For $x \in V$, $|x|$ will denote the cardinality of $x$. The relations of set inclusion and proper set inclusion will be denoted by $x \leqslant y$ and $x < y$, respectively. The empty set, viewed as an element of $V$, will be denoted by 0, and $\Omega$ itself, as an element of $V$, by 1.

Let $p(x)$ be a probability distribution on $V$, and let $X$ and $Y$ be elements of $V$ chosen randomly and independently according to $p$. We denote by $P\{X \leqslant Y\}$ the probability that $X$ will be a subset of $Y$. Our main result is the following.

**Theorem 1.** If $n > 1$, then for any probability distribution,

$$P\{X \leqslant Y\} \geqslant \binom{n}{\left[\frac{n}{2}\right]}^{-1}$$

Furthermore equality holds for the probability distribution defined by

$$\begin{aligned} p(x) &= \binom{n}{\left[\frac{n}{2}\right]}^{-1} && \text{if } |x| = \left[\frac{n}{2}\right] \\ &= 0 && \text{otherwise} \end{aligned}$$

In the remainder of this section, we make some remarks about Theorem 1 and its proof. The proof itself occupies Sections II–V.

(1) The restriction $n > 1$ is necessary, since with $n = 1$, the probability distribution $p(0) = p(1) = 1/2$ gives $P\{X \leqslant Y\} = 3/4$, whereas

$$\binom{n}{\left[\frac{n}{2}\right]}^{-1} = 1$$

(2) Let $Y = \{y_1, \ldots, y_M\}$ be a set of $M$ pairwise non-comparable elements of $V$, i.e., $y_1 \leqslant y_j$ iff $i = j$. If we define a probability distribution $p(x)$ by

$$\begin{aligned} p(x) &= \frac{1}{M} && \text{if } x \in Y \\ &= 0 && \text{otherwise} \end{aligned}$$

then clearly $P\{X \leqslant Y\} = P\{X = Y\} = 1/M$. Hence by Theorem 1,

$$M \leqslant \binom{n}{\left[\frac{n}{2}\right]}$$

This is Sperner's famous bound (see Ref. 3, for example) on the maximum number of non-comparable elements in $V$. Thus (for $n > 1$) Theorem 1 can be viewed as a generalization of Sperner's bound.

(3) For even values of $n$, it is possible to show that the particular probability distribution cited in the statement of Theorem 1 is the only one for which equality holds. For odd $n \geqslant 3$, the obvious alternate distribution

$$\begin{aligned} p(x) &= \binom{n}{\left[\frac{n}{2}\right]}^{-1} && \text{if } |x| = \left[\frac{n}{2}\right] + 1 \\ &= 0 && \text{otherwise} \end{aligned}$$

also achieves equality. For odd $n \geqslant 5$ this is the only other equality-achieving distribution. However, for $n = 3$ there are infinitely many extremal distributions:

$$\begin{aligned} p(x) &= a && \text{if } |x| = 1 \\ &= 1/3 - a && \text{if } |x| = 2 \\ &= 0 && \text{if } |x| = 0 \text{ or } 3, \text{ for } 0 \leqslant a \leqslant 1/3 \end{aligned}$$

(These facts can all be proved by making a careful study of our proof of Theorem 1. For the sake of brevity, however, we shall omit the details.)

(4) Our proof contains two main ideas. First, by using more-or-less standard calculus techniques, one can obtain very strong necessary conditions satisfied by any extremal probability distribution. This we do in Section II. Second, we derive in Section IV (after some preliminary material in Section III) a lemma dealing with the expected number of maximal chains through a point of $V$ which is selected according to a given probability distribution. This lemma was motivated in part by Lubell's classic proof (Ref. 3) of Sperner's theorem. Finally in Section V, these two ideas are combined to give our proof of Theorem 1.

(5) A possible alternate form of Theorem 1 would concern minimizing the probability that $X$ and $Y$ are comparable. It turns out that this problem is much easier to handle; the result is

$$P\{X \leqslant Y \quad \text{or} \quad X \geqslant Y\} \geqslant \binom{n}{\left\lfloor \frac{n}{2} \right\rfloor}^{-1}$$

for all $n$, and equality occurs only for a uniform probability distribution on the subsets of cardinality

$$\left\lfloor \frac{n}{2} \right\rfloor \quad \text{or} \quad \left\lceil \frac{n+1}{2} \right\rceil$$

However, this inequality follows already from a theorem of Motzkin and Strauss (Ref. 4), together with Sperner's original theorem.

## II. Lagrange Multipliers

The probability $Q(p) = P\{X \leqslant Y\}$ is given by the sum

$$Q(p) = \Sigma\{p(x)p(y) : x \leqslant y\} \tag{1}$$

We are asked to minimize the function $Q$ of the $2^n$ variables $(p(x), x \in V)$, subject to the following constraints:

$$\sum_{x \in V} p(x) = 1 \tag{2}$$

$$p(x) \geqslant 0, \text{all } x \in V \tag{3}$$

Suppose $p$ is a probability distribution that minimizes $Q(p)$ subject to (2) and (3), and let $G = \{x : p(x) > 0\}$. Temporarily we regard $p$ and $G$ as being fixed. Consider now the new problem of minimizing the function

$$Q_G(q) = \Sigma\{q(x)q(y) : x \leqslant y, x, y \in G\} \tag{4}$$

where $q$ is a real-valued function defined on $G$, subject to the single linear constraint

$$\sum_{x \in G} q(x) = 1 \tag{5}$$

(N.B. $q$ is *not* required to be a probability distribution.)

Let $B = \min \{p(x)^2 : x \in G\}$, and let $U$ denote the Euclidean neighborhood of the function $p$ (restricted to $G$) defined by

$$U = \left\{q : \sum_{x \in G} (q(x) - p(x))^2 < B\right\} \tag{6}$$

Clearly if $q \in U$, then $q(x) > 0$ for all $x \in G$. Thus by the assumed extremal property of $p$, we have

$$Q_G(q) \geqslant Q_G(p), \text{ for all } q \in U \text{ satisfying (5)} \tag{7}$$

Thus $p$ (restricted to $G$) gives a local minimum of the function $Q_G$, subject to the constraint (5), and so by the Lagrange multiplier rule there exists a constant $\lambda$ such that

$$\frac{\partial Q_G}{\partial p(x)} = \lambda, \text{ for all } x \in G \tag{8}$$

Using (1), this becomes

$$2p(x) + \sum_{y<x} p(y) + \sum_{y>x} p(y) = \lambda, \text{ for all } x \in G \tag{9}$$

Furthermore, since $P\{X \leqslant Y\} = P\{X \geqslant Y\}$ by symmetry,

$$\begin{aligned}2P\{X \leqslant Y\} &= P\{X \leqslant Y\} + P\{X \geqslant Y\} \\ &= \sum_{x \in G} p(x) \left\{\sum_{y \leqslant x} p(y) + \sum_{y \geqslant x} p(y)\right\} \\ &= \sum_{x \in G} p(x) \left\{2p(x) + \sum_{y<x} p(y) + \sum_{y>x} p(y)\right\} \\ &= \lambda \sum_{x \in G} p(x) = \lambda, \text{ by (9)}\end{aligned}$$

Thus we have identified the constant $\lambda$ in (9), and so we have

$$2p(x) + \sum_{y<x} p(y) + \sum_{y>x} p(y) = 2P\{X \leqslant Y\}, \quad \text{for all } x \in G \tag{10}$$

Equation (10) is the condition which must be satisfied by any extremal probability distribution that we will return to in Section V.

*Note*: Equation (9) follows immediately from the Kuhn-Tucker theorem of nonlinear programming (Ref. 5), and indeed one can also conclude from K-T that the left side of (9) is $\geqslant \lambda$ if $x \notin G$. We have included this elementary derivation only to make our exposition more self-contained.

## III. Preliminaries About Chains

A *chain* of length $r$ in $V$ is an $(r + 1)$-tuple $c = (x_0, x_1, \ldots, x_r)$ of elements from $V$ such that $x_0 < x_1 < \ldots < x_r$. If in addition we have $|x_{i+1}| = |x_i| + 1$, $c$ is said to be a *maximal chain* (of length $r$) from $x_0$ to $x_r$. Such a chain is said to *pass through* each of the points $x_0, x_1, \ldots, x_r$.

If $(y_0, y_1, \ldots, y_m)$ is any chain in $V$, we denote by $MC(y_0, \ldots, y_m)$ the set of all maximal chains from $y_0$ to $y_m$ which pass through each of the $y_i$'s. The number of maximal chains in $MC(y_0, \ldots, y_m)$ is denoted by $f(y_0, \ldots, y_m)$. Thus

$$f(y_0, \ldots, y_m) = |MC(y_0, \ldots, y_m)| \tag{11}$$

If $(y_0, \ldots, y_k, \ldots, y_m)$ is a chain, it is clear that every maximal chain in $MC(y_0, \ldots, y_m)$ can be decomposed uniquely into a chain from $MC(y_0, \ldots, y_k)$ followed by a chain from $MC(y_k, \ldots, y_m)$. Hence

$$f(y_0, \ldots, y_k, \ldots, y_m) = f(y_0, \ldots, y_k) f(y_k, \ldots, y_m) \tag{12}$$

and by induction it follows that

$$f(y_0, \ldots, y_m) = \prod_{i=0}^{m-1} f(y_i, y_{i+1}) \tag{13}$$

In view of (13), in order to compute $f(c)$ for a general chain $c$, it suffices to consider the case where $c = (x, y)$ consists of only two elements. This we now do.

Let $x = x_0 < x_1 < \ldots < x_r = y$ be a maximal chain from $x$ to $y$ and let $x_{i+1} - x_i = \{w_i\}$, $i = 0, 1, \ldots, r - 1$. Then $(w_0, w_1, \ldots, w_{r-1})$ is a permutation of the elements in the set $y - x$. Conversely, if $(w_0, \ldots, w_{r-1})$ is any permutation of $y - x$, and if we define $x_i = x \cup \{w_0, w_1, \ldots, w_{i-1}\}$, then $(x_0, \ldots, x_r)$ will be a maximal chain from $x_0 = x$ to $x_r = y$. Hence there is a one-to-one correspondence between chains from $x$ to $y$ and permutations of $y - x$:

$$f(x,y) = |y - x|\,! \tag{14}$$

$$f(y_0, \ldots, y_m) = \prod_{i=0}^{m-1} |y_{i+1} - y_i|\,! \tag{15}$$

As a final bit of notation, for $x \in V$ let $N(x)$ be the number of maximal chains from 0 to 1 passing through $x$. Then

$$\begin{aligned} N(x) &= f(0, x, 1) \\ &= f(0, x) f(x, 1) \\ &= |x|\,!\,(n - |x|)\,! = n!/\binom{n}{|x|} \end{aligned} \tag{16}$$

Note that as a function of $x$, $N(x)$ achieves its minimum value when $|x| = n/2$ for $n$ even, and $|x| = (n \pm 1)/2$ for $n$ odd. Thus if we define $\alpha(n) = [n/2]!\,(n - [n/2])\,!$, we have

$$N(x) \geqslant \alpha(n), \text{ all } x \in V, \tag{17}$$

with equality iff $x = \lfloor n/2 \rfloor$ or $n - \lceil n/2 \rceil$.

## IV. A Basic Lemma

Let $G$ be a subset of $V$. If $c = (x_0, \ldots, x_r)$ is a chain in $V$, and at least one element of $c$ lies in $G$, we define $L_G(c)$ ("the last element of $c$ lying in $G$"), as follows:

$$L_G(c) = x_k, \text{ where } k = \max\{i: x_i \in G\} \tag{18}$$

If no element of $c$ lies in $G$, $L_G(c)$ is undefined.

Further, we define for each $x \in V$,

$$N_G(x) = |\{c \in MC(0, 1): L_G(c) = x\}| \tag{19}$$

Thus $N_G(x)$ is the total number of maximal chains from 0 to 1 whose last element in $G$ is $x$. If now for each $x \epsilon V$ we define $g(x)$ by

$$g(x) = |\{c \epsilon MC(x, 1): L_G(c) = x\}| \tag{20}$$

i.e., $g(x)$ is the number of maximal chains from $x$ to 1 whose last element in $G$ is $x$ itself, it follows that

$$N_G(x) = f(0,x)\, g(x) \tag{21}$$

If $p$ is a probability distribution on $V$, and if $X$ is an element of $V$ chosen randomly according to $p$, the number of maximal chains from 0 to 1 passing through $X$, $N(X)$ is a random variable whose expectation is given by $E(N(X)) = \Sigma\{p(x)N(x)\colon\ x \epsilon V\}$. The following lemma gives another formula for $E(N(X))$ which is crucial in our proof of Theorem 1.

**Lemma 1**

Let $p$ be a probability distribution on $V$, and let $G = \{x \epsilon V\colon p(x) > 0\}$. Then

$$\sum_{x \epsilon V} p(x)N(x) = \sum_{y \epsilon G} N_G(y) \sum_{x \leqslant y} p(x) \binom{|y|}{|x|}^{-1}$$

**Proof**

Using the fact that $p(x) = 0$ if $x \notin G$, and (16), we have

$$\sum_{x \epsilon V} p(x)N(x) = \sum_{x \epsilon G} p(x)\, f(0, x)\, f(x, 1) \tag{22}$$

We now classify the chains in $MC(x, 1)$ according to their last element in $G$. If $y \geqslant x$, then the number of maximal chains from $x$ to 1 whose last element in $G$ is $y$ is $f(x,y)\, g(y)$. Thus

$$f(x, 1) = \sum_{y \geqslant x} f(x, y) g(y) \tag{23}$$

Replacing $f(x, 1)$ in (22) by the sum (23), and interchanging the order of summation, we get

$$\sum_{x \epsilon V} p(x)N(x) = \sum_{y \epsilon G} g(y) \sum_{x \leqslant y} p(x) f(0, x) f(x, y)$$

$$= \sum_{y \epsilon G} f(0, y) g(y) \sum_{x \leqslant y} p(x) \frac{f(0, x) f(x, y)}{f(0, y)}$$

But by (21), $f(0, y)g(y) = N_G(y)$. And by (14),

$$\frac{f(0,x) f(x,y)}{f(0,y)} = \frac{|x|!\, |y - x|!}{|y|!} = \binom{|y|}{|x|}^{-1} .$$

This proves Lemma 1.

## V. Proof of Theorem 1

We are now in a position to give a short proof of Theorem 1. The idea of the proof is to estimate the expected number of maximal chains through a randomly selected point of $V$ in two ways. On one hand, this expectation is certainly at least $\alpha(n)$ by (17). On the other hand, using the machinery we have developed in Sections II–IV, we will show that this expectation (at least for an extremal probability distribution) is at most $P\{X \leqslant Y\} \cdot n!$ The resulting bound, $P\{X \leqslant Y\} \geqslant \alpha(n)/n!$, is the bound of Theorem 1. Let us now see how this proof goes in detail.

Let $p$ be a probability distribution that minimizes $Q(p) = P\{X \leqslant Y\}$, and let $G = \{x \epsilon V\colon p(x) > 0\}$. Then by Lemma 1,

$$\sum_{x \epsilon V} p(x)N(x) = \sum_{y \epsilon G} N_G(y) \sum_{x \leqslant y} p(x) \binom{|y|}{|x|}^{-1} \tag{24}$$

By (17), $N(x) \geqslant \alpha(n)$ for all $x$. Hence

$$\sum_{x \epsilon V} p(x)N(x) \geqslant \alpha(n) \tag{25}$$

Let us rewrite the inner sum in (24) in the following way:

$$\sum_{x\leqslant y} p(x)\binom{|y|}{|x|}^{-1} = p(y) + \sum_{x<y} p(x)\binom{|y|}{|x|}^{-1} \qquad (26)$$

We now claim that if $n > 1$, and $x < y$,

$$p(x)\binom{|y|}{|x|}^{-1} \leqslant \frac{1}{2}p(x) \qquad (27)$$

If $x \neq 0$, (27) is obvious since then the binomial coefficient will be $\geqslant 2$. If $x = 0$, the binomial coefficient is 1, but (27) is true anyway because $p(0) = 0$. This can be seen as follows. Let $x = 0$ in (10); we get $P\{X \leqslant Y\} = 1/2 + 1/2\, p(0)$. Hence if $p(0) > 0$, then $P\{X \leqslant Y\} > 1/2$. But since $n > 1$ we can select two non-comparable elements $x_1$ and $x_2$ and define a probability distribution $q$ by setting $q(x_1) = q(x_2) = 1/2$. For this probability distribution we clearly have $P\{X \leqslant Y\} = 1/2$. This shows that no distribution that minimizes $P\{X \leqslant Y\}$ for $n > 1$ can have $p(0) > 0$, and this completes the proof of (27).

Combining (26) and (27), we get

$$\sum_{x\leqslant y} p(x)\binom{|y|}{|x|}^{-1} \leqslant p(y) + \frac{1}{2}\sum_{x<y} p(x)$$

$$\leqslant p(y) + \frac{1}{2}\sum_{x<y} p(x) + \frac{1}{2}\sum_{x>y} p(x) \qquad (28)$$

We now apply (10) to (28) and conclude that

$$\sum_{x\leqslant y} p(x)\binom{|y|}{|x|}^{-1} \leqslant P\{X \leqslant Y\}, \text{ if } y \in G \qquad (29)$$

Combining (24), (25), and (29), we get

$$\alpha(n) \leqslant P\{X \leqslant Y\} \sum_{y\in G} N_G(y) \qquad (30)$$

Finally we observe that each maximal chain from 0 to 1 is counted at most once in the sum (30). (If $c$ is such a chain and if $L_G(c) = y$, it is counted by the term $N_G(y)$.) Since the total number of such chains is $n!$ (see Eq. 14), we get $\Sigma N_G(y) \leqslant n!$ and hence, finally,

$$P\{X \leqslant Y\} \geqslant \frac{\alpha(n)}{n!} = \binom{n}{\left[\frac{n}{2}\right]}^{-1} \qquad (31)$$

This completes the proof of Theorem 1.

# References

1. Levitt, B. K., "Analysis of a Discrete Spectrum Analyzer for the Detection of Radio Frequency Interference," in *DSN Progress Report* 42-38, April 15, 1977, pp. 83-98.
2. Knuth, D. E., *Sorting and Searching*, Vol. 3 in *The Art of Computer Programming*, Reading: Addison-Wesley, 1973.
3. Lubell, D., "A Short Proof of Sperner's Lemma," *J. Combinatorial Theory*, 1 (1966), p. 299.
4. Motzkin, T. S., and E. G. Strauss, "Maxima for Graphs and a New Proof of a Theorem of Turan," *Canadian J. Math.* 17 (1965), pp. 535-540.
5. Cottle, R. N., and Lemke, C. E., eds., *Nonlinear Programming*, SIAM-AMS Proceedings, Vol. 9, Providence: American Math. Society, 1976.