Journal of Statistical Planning and Inference 11 (1985) 23–31
North-Holland

ON TWO RANDOM SEARCH PROBLEMS

András SEBŐ

Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary

Received November 1978; revised manuscript received 22 November 1983
Recommended by G.O.H. Katona

Abstract: The paper is concerned with static search on a finite set. An unknown subset of cardinality k of the finite set is to be found by testing its subsets. We investigate two problems: in the first, the number of common elements of the tested and the unknown subset is given; in the second, only the information whether the tested and the unknown subset are disjoint or not is given. Both problems correspond to problems on false coins. If the unknown subset is taken from the family of k-element sets with uniform distribution, we determine the minimum of the lengths of the strategies that find the unknown element with small error probability. The strategies are constructed by probabilistic means.

AMS Subject Classification: 94A50, 05B99.

Key words: Sequential; Static strategies; Random search; Separating systems; Random construction.

1. Introduction

The problems treated in this paper belong to a type of search problems, the common and general formulation of which may be the following:

H is a finite set, $\boldsymbol{F}$, $\boldsymbol{T}$ are systems of subsets of H, t is a mapping of pair $(\boldsymbol{F}, \boldsymbol{T})$, $F \in \boldsymbol{F}$, $T \in \boldsymbol{T}$, in the set of natural numbers. We search for an unknown set $T \in \boldsymbol{T}$ with the help of 'strategies' which are sequences of elements of $\boldsymbol{F}$; the result of the strategy $\boldsymbol{S} = (F_1, \ldots, F_N)$, $F_i \in \boldsymbol{F}$, is an N-dimensional vector $t(\boldsymbol{S}, T) = (t(F_1, T), \ldots, t(F_N, T))$; N is called the length of the strategy.

To determine T uniquely with the help of the strategy $\boldsymbol{S}$ it is necessary and sufficient to have

$$t(\boldsymbol{S}, T) \neq t(\boldsymbol{S}, T') \quad \text{for all } T \neq T'. \tag{1}$$

In this paper we only discuss problems in which F_i may depend on none of $t(F_j, T)$, $j < i$. Such strategies are called static. We could ask for the minimal length of the strategies which determine T uniquely, i.e. for which (1) holds. Instead of this, we shall consider the uniform distribution on $\boldsymbol{T}$ and define the error probability for each strategy $\boldsymbol{S}$:

$$P_e^t(S)=\frac{|\{T\in \mathcal{T}: \exists T'\in \mathcal{T},\ T'\neq T,\ t(S,T)=t(S,T')\}|}{|\mathcal{T}|}.$$

Our purpose is to find the minimal number N for which there exists a strategy S of length N with $P_e^t(S)\leq\varepsilon$, where $0\leq\varepsilon<1$ is given in advance. Denote this number by $N(H,\mathcal{F},\mathcal{T},t,\varepsilon)$.

We introduce some more notations: $H_n=\{1,2,\dots,n\}$, $V_n=2^{H_n}$, $V_n^k=\{H\in V_n: |H|=k\}$; for $A,B\in V_n$:

$$t_1(A,B)=|A\cap B|,\qquad t_2(A,B)=\begin{cases}1 & \text{if } A\cap B\neq\emptyset,\\ 0 & \text{if } A\cap B=\emptyset,\end{cases}$$

$$f(n,k,\varepsilon)=N(H_n,V_n,V_n^k,t_1,\varepsilon),$$

$$g(n,k,\varepsilon)=N(H_n,V_n,V_n^k,t_2,\varepsilon).$$

Let us notice that the minimal length of the strategies for which (1) holds is $N(H,\mathcal{F},\mathcal{T},t,0)$, i.e. $f(n,k,0)$ for $\mathcal{F}=V_n$, $\mathcal{T}=V_n^k$, $t=t_1$, and $g(n,k,0)$ for $\mathcal{F}=V_n$, $\mathcal{T}=V_n^k$, $t=t_2$.

In the present paper asymptotically exact estimates are given for $f(n,k,\varepsilon)$ and $g(n,k,\varepsilon)$ with k fixed and $n\to\infty$, $0<\varepsilon<1$. All estimates are independent of ε.

Theorem 1. *For all* $0<\varepsilon<1$,

$$\frac{2k}{\frac{1}{2}+\log_2\pi k}\leq\liminf_{n\to\infty}\frac{f(n,k,\varepsilon)}{\log_2 n}\leq\limsup_{n\to\infty}\frac{f(n,k,\varepsilon)}{\log_2 n}\leq\frac{2k}{\log_2\pi k}.$$

Since $f(n,k,0)\geq f(n,k,\varepsilon)$ $(0\leq\varepsilon<1)$ the lower estimate is true for $f(n,k,0)$ as well, which was first proved by Diatchkov with a method utilizing information-theoretical means (Diatchkov, 1976). We shall give an elementary proof using the generalization of an idea of Moser (Erdős–Spencer, 1974).

Theorem 2.

$$\lim_{n\to\infty}\frac{g(n,k,\varepsilon)}{\log_2 n}=k$$

In the upper estimates of both theorems we use the random constuction method of Erdős and Rényi and in the proof of Theorem 2 the ideas in Diatchkov's paper (1976) which gives an upper bound for $g(n,k,0)$.

It will be useful to let the subsets of H_n correspond to n-dimensional 0–1 vectors as usual, where the vector corresponding to $A\in H_n$ is

$$v_A=(v_1,\dots,v_n),\quad v_i=\begin{cases}1 & \text{if } i\in A,\\ 0 & \text{if } i\notin A.\end{cases}$$

It is obvious that if $A\in H_n$ and $B\in H_n$, then $|A\cap B|=\langle v_A,v_B\rangle$ with $\langle\ ,\ \rangle$ the inner

product. If $S=(F_1,\dots,F_N)$, $F_i\subset H_n$, let A_S be the matrix the i-th row of which is the vector corresponding to F_i. Thus a one-to-one correspondence has been established between strategies and $N\times n$ matrices. It is clear that if the vector x corresponds to $T\in V_n^k$, then $A_S x=t_1(S,T)$. So if we define

$$P_{e,1}(A)=\frac{|\{u\in V_n^k:\ \exists v\in V_n^k,\ v\neq u,\ Av=Au\}|}{\binom{n}{k}}$$

for each matrix A, we have for all strategies S that $P_e^{t_1}(S)=P_{e,1}(A_S)$.

Similarly, if $A_S\vee x$ denotes the N-dimensional vector which has 1 in its i-th coordinates if and only if it has 1 in the i-th coordinate of at least one of the column vectors of A_S corresponding to the 1's of x, then we have

$$A_S\vee x=t_2(S,T),$$

$$P_{e,2}(A_S)=\frac{|\{u\in V_n^k:\ \exists v\in V_n^k,\ v\neq u,\ A\vee v=A\vee u\}|}{\binom{n}{k}}$$

and

$$P_e^{t_2}(S)=P_{e,2}(A_S).$$

Let the set of vectors corresponding to V_n or V_n^k inherit these notations. E.g. V_n^k is also the set of n-dimensional 0–1 vectors in which there are k 1's and $n-k$ 0's.

2. Proof of Theorem 1

We first prove the upper estimate in Theorem 1. The following statement will be proved: There is a sequence of matrices (A_n), so that A_n has rows in V_n, A_n has N_n rows and n columns, and

$$\limsup_{n\to\infty}\frac{N_n}{\log n}\leq\frac{2k}{\log \pi k}$$

with $P_e(A_n)\to 0$ $(n\to\infty)$.

We construct an $N\times n$ matrix choosing its elements either 1 or 0 independently with probability $\frac{1}{2}$. (It can easily be seen that the best estimate is expected from this distribution.) In the probability space describing this random construction, the constructed matrix will be a random variable, denote it by $L_{N\times n}$. Then $P_{e,1}(L_{N_n\times n})$ is also a random variable. We prove that the expected value $E(P_{e,1}(L_{N_n\times n}))$ converges to 0 if $N_n=(2k/\log \pi k)\log n+\omega(n)$ with $\omega(n)\to\infty$. ($\omega(n)$ is otherwise arbitrary.) This proves the statement as there is at least one $N_n\times n$ matrix A_n with $P_e(A_n)\leq E(P_{e,1}(L_{N_n\times n}))$ and $\omega(n)$ can be chosen to have a magnitude $o(\log n)$ (e.g. $\sqrt{\log n}$). Let $M_{N\times n}$ be the set of all $N\times n$ matrices.

$$E(P_{e,1}(L_{N\times n})) = \sum_{\substack{X \\ X\in M_{N\times n}}} \frac{1}{2^{nN}} \sum_{\substack{u\in V_n^k:\, \exists v\in V_n^k \\ v\neq u,\ Xv=Xu}} \frac{1}{\binom{n}{k}}$$

$$= \sum_{u\in V_n^k} \frac{1}{\binom{n}{k}} \sum_{\substack{X\in M_{N\times n} \\ \exists v\neq v_0,\ Xv=Xv_0}} \frac{1}{2^{nN}} \tag{2}$$

where $v_0 = (1, \ldots, 1, 0, \ldots, 0)^T$ (k ones, $n-k$ zeroes). As

$$\sum_{\substack{X\in M_{N\times n} \\ \exists v\neq u,\ Xv=Xv}} \frac{1}{2^{nN}}$$

has the same value for all $u \in V_n^k$, the right-hand side of (2) is equal to

$$\sum_{\substack{X\in M_{N\times n} \\ \exists v\neq v_0,\ Xv=Xv_0}} \frac{1}{2^{nN}}.$$

Thus we only have to compute the number of matrices for which there exists a $v = V_n^k$ with $Xv = Xv_0$. Since Xu, $u \in V_n^k$, is the sum of columns in X corresponding to the 1's of u, $Xv = Xv_0$ is equivalent to $Xv' = Xv_0'$, where $v' = v - u(v, v_0)$, $v_0' = v_0 - u(v_1, v_0)$ and $u(v_1, v_0)$ is the vector the i-th coordinate of which is 1 if and only if it is 1 in both v ad v_0. Hence it follows that the number of $N\times n$ matrices has to be found which have i columns among the first k and i among the last $n-k$ columns, such that the two groups of i columns have equal sums ($1 \le i \le k$). This number is majorized by

$$\sum_{i=1}^{k} \binom{k}{i}\binom{n-k}{i}\left[\sum_{l=0}^{i}\binom{i}{l}^2 2^{n-2i}\right]^N.$$

Using

$$\sum_{l=0}^{i}\binom{i}{l}^2 = \binom{2i}{i} \le \frac{2^{2i}}{\sqrt{\pi i}}$$

which follows from the inequality

$$\left(\frac{n}{e}\right)^n \sqrt{2\pi n}\, e^{1/(12n+1)} \le n! \le \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\, e^{1/12n}$$

(Robbins, 1955), and using (2), we have

$$E(P_{e,1}(L_{N\times n})) \le \frac{1}{2^{nN}} \sum_{i=1}^{k} \binom{k}{i}\binom{n-k}{i} \frac{2^{nN}}{(\pi i)^{N/2}} \le \sum_{i=1}^{k} \binom{k}{i} \frac{n^i}{(\pi i)^{N/2}}.$$

It is enough to demand of each member of the sum to tend to 0, i.e. $2^{i\log n - (N_n/2)\log \pi i} \to 0$, which holds if

$$N_n = \frac{2k}{\log \pi k} \log n + \omega(n) \quad (\omega(n) \to \infty).$$

Proof of the lower estimate in Theorem 1. The proof is based on the fact that the number of elements of the set $A_n V_n^k = \{A_n v : v \in V_n^k\}$, with A_n an $N \times n$ matrix, can be estimated from above as a function of N. So, if N is small, we can arrive at the conclusion that there are many vectors v in V_n^k for which $A_n v$ is the image of another vector as well; consequently $P_e(A_n)$ is large:

$$P_e(A_n) \geq 1 - \frac{|A_n V_n^k|}{|V_n^k|}.$$

Let us take arbitrarily two numbers, a, b, $a \neq b$ (depending on n and k) instead of 0 and 1; V_n^k can also be represented as a set of vectors in which k coordinates are a and $n-k$ are b. From here on in this section, V_n^k will denote this set of vectors. a en b will be chosen so that a large part of the set $A_n V_n^k$ will be within a small sphere, and all its elements will be on a unit lattice in the n-dimensional space. From this it will follow that $A_n V_n^k$ can possess only few elements.

We set $a = (k/n) - 1$ and $b = (k/n)$. If the i-th row of A_n has m 1's, then the i-th coordinate of all elements of $A_n V_n^k$ will be of the form

$$v\left(\frac{k}{n} - 1\right) + (m - v)\frac{k}{n}$$

for some v ($v = 1, 2, \ldots, k$). Since the difference between two numbers of this form can only be $1, 2, \ldots, k$, $A_n V_n^k$ lies on a unit lattice in the n-dimensional space. Let $\zeta = (\zeta_1, \ldots, \zeta_n)$ be a random variable with $P(\zeta = v) = 1/\binom{n}{k}$ for all $v \in V_n^k$. To prove that a large part of $A_n V_n^k$ is within a small sphere, we shall use Markov's inequality, and therefore $E(\|A_n \cdot (\zeta_1, \zeta_2, \ldots, \zeta_n)\|^2)$ has to be estimated ($\| \ \|$ is the Eucledian norm). If A_n has $m(i)$ 1's in its i-th row ($i = 1, 2, \ldots, N_n$) then

$$A_n \cdot (\zeta_1, \ldots, \zeta_n) = (\zeta_i + \cdots + \zeta_{i_{m(1)}}, \ldots, \zeta_{j_1} + \cdots + \zeta_{j_{m(N_n)}}). \tag{3}$$

This shows that for the estimation of $E(\|A_n(\zeta_1, \ldots, \zeta_n)\|^2)$ we need the following lemma:

Lemma.

$$E((\zeta_1 + \cdots + \zeta_m)^2) \leq \frac{k(n-k)}{4(n-1)} \quad (m = 1, 2, \ldots, n)$$

and equality holds if and only if $m = \frac{1}{2}n$.

Proof.

$$E((\zeta_1 + \cdots + \zeta_n)^2) = (ka + (n-k)b)^2 = 0$$

and by symmetry

$$E((\zeta_1 + \cdots + \zeta_n)^2) = 2\binom{n}{2}E(\zeta_1 \zeta_2) + nE(\zeta_1^2).$$

From this we get $E(\zeta_1\zeta_2) = -E(\zeta_1^2)/(n-1)$. Clearly

$$E(\zeta_1^2 + \cdots + \zeta_n^2) = \frac{k(k-n)^2}{n^2} + (n-k)\frac{k^2}{n^2} = \frac{k(n-k)}{n},$$

and hence

$$E(\zeta_1^2) = \frac{k(n-k)}{n^2}, \qquad E(\zeta_1\zeta_2) = -\frac{k(n-k)}{(n-1)n^2}.$$

We have

$$E((\zeta_1 + \cdots + \zeta_m)^2) = 2\binom{m}{2}E(\zeta_1\zeta_2) + mE(\zeta_1^2) = \frac{k(n-k)}{n^2(n-1)}\, m(n-m).$$

This formula has its maximum at $m = \frac{1}{2}n$, and the proof of the lemma is completed.

Equation (3) and the lemma prove the inequality

$$E(|A_n \cdot (\zeta_1, \ldots, \zeta_m)\|^2) \le N_n \frac{k(n-k)}{4(n-1)}.$$

Then, by the Markov inequality and passing to the complementary event,

$$P(\|A_n \cdot (\zeta_1, \ldots, \zeta_n)\|^2) \le KN_n \frac{k(n-k)}{4(n-1)} \le 1 - \frac{1}{K} = \frac{K-1}{K} \tag{4}$$

for all $K > 1$. As $A_n \cdot (\zeta_1, \ldots, \zeta_n)$ has its values in $A_n V_n^k$, i.e. on a unit lattice, (4) means that at least $|A_n V_n^k|(K-1)/K$ elements of $A_n V_n^k$ lie in the sphere of radius

$$\left(KN_n \frac{k(n-k)}{4(n-1)}\right)^{1/2}.$$

But in the N-dimensional space in the sphere of radius R there are 'not much more' elements of any cubic lattice than its volume T_R^N; more precisely,

$$\frac{K-1}{K}\, |A_n V_n^k| \le DT_{R^*}^N(1 + o(1)) \tag{5}$$

is true with some constant D for all $K > 1$, where

$$R^* = \left(KN_n \frac{k(n-k)}{4(n-1)}\right)^{1/2}.$$

We compute DT_R^N:

$$DT_R^N(1 + o(1)) = D\frac{2\pi^{N/2}}{\Gamma(\frac{1}{2}N)}\frac{R}{N}(1 + o(1)) = \frac{2C}{N\sqrt{\pi N}}\frac{(2\pi e)^{N/2}}{N^{N/2}} R^N(1 + o(1))$$

$$\le \left(\frac{2\pi e}{N}\right)^{N/2} R^N \tag{6}$$

(if $N \ge N_0$ but we are not interested in small N's; here we used the Stirling for-

mula). Let us insert (6) in (5):

$$\frac{K-1}{K}|A_n V_n^k| \le \left(\frac{2\pi e}{N_n}\right)^{N_n/2}\left(KN_n \frac{k(n-k)}{4(n-1)}\right)^{N_n/2}$$

$$= \left(\frac{Ke}{2}\pi k \frac{n-k}{n-1}\right)^{N_n/2} \le \left(\frac{Ke}{2}\pi k\right)^{N_n/2} \quad (K>1). \tag{7}$$

We write $1-P_e(A_n) \le |A_n V_n^k|/\binom{n}{k}$ in (7) to have an estimate for $P_e(An)$:

$$P_e(A_n) \ge 1 - \frac{K}{K-1}\frac{(\frac{1}{2}Ke\pi k)^{N_n/2}}{\binom{n}{k}} \ge 1 - \frac{K}{K-1}\, 2^{(N_n/2)\log_2(Ke/2)\pi k - k\log_2 n} \tag{8}$$

(for all $K>1$). If

$$N_n = \frac{2k}{\frac{1}{2}+\log_2 \pi k}\log n\left(\frac{2k}{\log_2 \sqrt{2}\pi k}\log_2 n\right),$$

then, with arbitrary constant $1<K<2\sqrt{2}/e$, (8) gets the form

$$P_e(A_n) \ge 1 - C_1 2^{-C_2 \log_2 n}$$

where $C_1>0$, $C_2>0$. Thus $P_e(A_n)\to 1$ as $n\to\infty$. □

3. Proof of Theorem 2

Using the heuristic fact that the best strategy is the one with biggest Shannon entropy, it is not difficult to see that the best upper estimate on $g(n,k,\varepsilon)$ can be expected from the following random construction: choose each element of A_n to be 0 with probability $(\frac{1}{2})^{1/k}$ and 1 with probability $1-(\frac{1}{2})^{1/k}$ independently (Diatchkov, 1976). Denote by $D_{N_n\times n}$ the random matrix determined by this construction. Similarly to the proof of the upper estimate of Theorem 1 (namely to (2)), we can write

$$E(P_{e,2}(B_{N_n\times n})) = \sum_{\substack{X\in M_{N_n\times n}\\ \exists v\neq v_0,\ Xv=Xv_0}} P(X) \le \sum_{r=0}^{k-1}\sum_{X\in M^r_{N_n\times n}} P(X) \tag{9}$$

with the same v_0 as in (2), and where

$$M^r_{N_n\times n} = \{X\in M_{N_n\times n}: \exists v\neq v_0,\ \langle v, v_0\rangle = r,\ X_n v = X_n v_0\}.$$

It can be simply verified that for any fixed r,

$$\sum_{X\in M^r_{N_n\times n}} P(X) \le \binom{k}{r}\binom{n-k}{k-r} q_r^{N_n} \tag{10}$$

where q_r is the probability of the event that $B_{1\times n}$ is a $1\times n$ matrix for which there exists a vector $v=(0,\dots,0,1,0,\dots,0,1,0,\dots,0,1,0,\dots,0)$ with 1's at positions $i_1, i_2, \dots, i_k$, where $i_1<i_2<\cdots<i_k$, $i_r\le k<i_{r+1}$, that satisfies the equation $L_{1\times n}v = L_{1\times n}v_0$. We have

$$q_r = \left[\left(\frac{1}{2}\right)^{1/k}\right]^k \left[\left(\frac{1}{2}\right)^{1/k}\right]^{k-r} + \left[\left(\frac{1}{2}\right)^{1/k}\right]^r \left\{1 - \left[\left(\frac{1}{2}\right)^{1/k}\right]^{k-r}\right\}^2 + 1 - \left[\left(\frac{1}{2}\right)^{1/k}\right]^r = \left(\frac{1}{2}\right)^{(k-r)/k}.$$

Here the first member is the probability of the event that in $B_{1\times n}$ the $1, 2, \ldots, k, i_{r+1}, i_{r+2}, \ldots, i_k$-th coordinates are 0, the third member is the probability of the event that at least one coordinate of the $i_1, i_2, \ldots, i_r$-th coordinate is 1, and the second is the probability of the event that among the $i_1, \ldots, i_r$-th places there are only 0's but there is at least one 1 among the $1, 2, \ldots, k$-th coordinates, and at least one 1 among the $i_1, \ldots, i_k$-th coordinates. We put $q_r = (\frac{1}{2})^{(k-r)/k}$ in (10) and substitute it in (9) to obtain the estimate

$$E(P_{e,2}(B_{N_n\times n})) \le \sum_{r=0}^{k-1} \binom{k}{r}\binom{n-k}{n-r}\left(\frac{1}{2}\right)^{N_n(k-r)/k}.$$

As $\binom{n-k}{k-r} \le n^{k-r}$ it follows, setting $N_n = k\log_2 n + \omega(n)$ (with $\omega(n)\to\infty$ but $\omega(n)$ otherwise arbitrary), that

$$E(P_{e,2}(L_{N_n\times n}))\to 0 \quad (n\to\infty).$$

Hence

$$g(n,k,\varepsilon) \le k\log_2 n(1+o(1)).$$

The lower estimate is trivial as t_2 can have only two values and therefore $2^{N_n}/\binom{n}{k} \ge 1 - P_{e,2}(A_n)$. So $1 - P_{e,2}(A_n)\to 1$ implies $N_n \ge k\log_2 n(1+o(1))$. □

4. Comments

Although the estimates for $f(n,k,\varepsilon)$ and $g(n,k,\varepsilon)$, $0<\varepsilon<1$, could be proved quite simply, the estimation for $f(n,k,0)$ and $g(n,k,0)$ seems much more difficult. The upper estimates given for them are multiples of the lower estimates (Diatchkov, 1976, 1977, and Rényi, 1965). There are no exact estimates even for $k=2$.

Other problems arise when taking systems of subsets different from ours for F, T or t. Rényi (1965) and Katona (1966) have examined $N(H_n, V_n^k, V_1, t_1, 0)$ and Katona (1966) practically solves the problem for this case.

The cases $\varepsilon = 0$ need combinatorial and algebraic methods but the validity and simplicity of the exact results for $0<\varepsilon<1$ have some information-theoretical reasons, and maybe a more general statement of the character of coding theorems is in the background. However, there are no results in this direction.

Acknowledgement

The author wishes to thank Professors A.G. Diatchkov and G.O.H. Katona for their help.

References

Erdős, P. and A. Rényi (1963). On two problems of information theory. *Publ. Math. Inst. Hungar. Acad. of Sci.* 8, 241–254.

Diatchkov, A.G. (1976). On a problem of false coins. *Colloquium in Combinatorics*, Keszthely.

Diatchkov, A.G. (1977). Theoria poiska i planirovanie eksperimentov. Type-script.

Rényi, A. (1965). On the theory of random search. *Bull. Amer. Math. Soc.* 71, 809–823.

Katona, G.O.H. (1966). On separating systems of finite sets. *J. Combinat. Theory* 1, 174–194.

Erdős, P. and J. Spencer (1974). *Probabilistic Methods in Combinatorics*. Academic Press, Budapest, 100–101.

Robbins, D. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* 62, 26–29.