Search with small sets in presence of a liar ☆

Gyula O.H. Katona

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences (HAS), P.O.B. 127,
H-1364 Budapest, Hungary

Abstract

One unknown element of an n-element set is sought by asking if it is contained in given subsets. It is supposed that the question sets are of size at most k and all the questions are decided in advance, the choice of the next question cannot depend on previous answers. At most l of the answers can be incorrect. The minimum number of such questions is determined when the order of magnitude of k is n^{α} with $\alpha < 1$. The problem can be formulated as determination of the maximum sized l-error-correcting code (of length n) in which the number of ones in a given position is at most k.

MSC: primary 90B40; secondary 94B99; 05B30; 05A05

Keywords: Search; Liar; Steiner system; Error-correcting code; Entropy

1. Introduction

Let X be a finite set of n elements, say $X = \{1, 2, \dots, n\}$, and x, $1 \leqslant x \leqslant n$ an unknown element. We want to find x by asking questions of type "is $x \in A$?" where A is a subset of X with at most k elements. However some of the answers can be false. It is supposed that the number of incorrect answers is at most l. The unknown x should be found uniquely under these informations. There are two different models, the *adaptive* model, when the choice of the next question may depend on the previous answers and the *non-adaptive* model when all the questions are decided in advance. In the present paper only the latter one is considered. Of course, we want to find the minimum number $f(n, k, l)$ of questions sufficient to find x. (An excellent survey of search problems with lies is Hill, 1995.)

In other words, our aim is to find the minimum number $f(n, k, l)$ of subsets $A \subset X, |A| \leqslant k$ of the n-element X such that the answers for the questions "is $x \in A$" uniquely determine x even if at most l of these answers are incorrect.

☆ The work was supported by the Hungarian National Foundation for Scientific Research grant numbers T016389, T016524, the European Communities (Cooperation in Science and Technology with Central and Eastern European Countries) contract number CIPACT930113.

E-mail address: ohkatona@math-inst.hu (G.O.H. Katona).

Let the characteristic vector of A be a 0,1 vector a whose jth component is 1 iff $j \in A$. Let our question sets be $A_1, A_2, \dots, A_m$, then define the $m \times n$ matrix M by the rows $a_1, a_2, \dots, a_m$. Obviously, if the unknown element x is j, then the correct answer for the ith question is "yes" iff the jth component a_{ij} of a_i is 1. The sequence of correct answers can be identified with the jth column of M. However, at most l components are changed to obtain the actual answers. It is easy to see that the element x, that is, any column can be uniquely identified from the actual answers iff these columns differ in at least $2l+1$ components, in other words their *Hamming distance* is at least $2l+1$. Therefore $f(n,k,l)$ is the minimum of such m's that there is an $m \times n$ 0,1 matrix with at most k 1's in each row and with Hamming distance at least $2l+1$ between any two columns.

Taking the complement of a set A_i, or interchanging the role of 0's and 1's in one row of M preserves the required property, therefore it can and will be supposed that $k < n/2$.

The dual problem can be formulated in terms of error-correcting codes, too. The columns of the matrix are the *code words*. Their set is an l-error-correcting code. The typical problem of coding theory is to find the largest (here n) code for given m and l. The only (in coding theory) unusual condition is that at most k of the codewords may have 1 at each fixed position.

Introduce the notation $f(n,k,0) = f(n,k)$. The value of $f(n,k)$ is more or less known (see Katona, 1966; Wegener, 1979; Luzgin, 1980). In Section 2 we give, however, some improvements.

In Section 3 we give a lower estimate, while Section 4 contains some constructions, fairly good when k is small relative to n.

The concept of *entropy* will be used in the paper. Let ξ be a random variable taking on (finitely many) distinct values with probabilities $p_1, p_2, \dots, p_N, 0 \leqslant p_i$, $\sum_{i=0}^{N} p_i = 1$. Its entropy is defined by

$$H(\xi) = -\sum_{i=0}^{N} p_i \log p_i,$$

where (and in the entire paper) log means log of base 2 and $0 \log 0$ is 0. If a pair (ξ, η) of random variables is given, then the entropy $H(\xi, \eta)$ of the pair is defined in the same way, by the (so called *joint*) probabilities of the pairs of values (ξ, η) can take on. The following inequality is well known (see e.g. Csiszár and Körner, 1981 or Feinstein, 1958):

$H(\xi, \eta) \leqslant H(\xi) + H(\eta)$ with equality iff ξ and η are independent. It can be easily extended for more random variables by induction:

$$H(\xi_1, \dots, \xi_m) \leqslant H(\xi_1) + \cdots + H(\xi_m). \tag{1.1}$$

We will use the following notation: $h(y) = -y \log y - (1-y) \log(1-y)$. This function is monotonically increasing in the interval $[0, 1/2]$, symmetric about the middle in $[0, 1]$ and $h(1/2) = 1$ (see e.g. Csiszár and Körner, 1981 or Feinstein, 1958).

2. Improvements for the case of zero lies

Let $\binom{x}{i}$ denote the polynomial $(x(x-1)\cdots(x-i+1))/i!$. The following theorem gives an algorithmic solution for $f(n,k)$.

Theorem 2.1 (Katona, 1966). *Let $2 \leqslant k < n/2$ be integers. The inequality system*

$$xk + \sum_{i=0}^{r-1} (r-i) \binom{x}{i} - rn = 0, \tag{2.1}$$

$$\sum_{i=0}^{r-1} \binom{x}{i} \leqslant n < \sum_{i=0}^{r} \binom{x}{i} \tag{2.2}$$

has a unique solution in (r,x) supposing that r is a positive integer, x is real and $r-1 \leqslant x$ holds. Then

$$f(n,k) = \lceil x \rceil.$$

The following lower and upper estimates are known.

Theorem 2.2 (Katona, 1966; Luzgin, 1980; Wegener, 1979).

$$\frac{\log n}{h(k/n)} \leqslant f(n,k) \leqslant \left\lceil \frac{\log n}{\log \lceil n/k \rceil} \right\rceil (\lceil n/k \rceil - 1).$$

The lower estimate was proved in Katona (1966). A somewhat weaker (than the present one) upper estimate was deduced in Katona (1966) from Theorem 2.1. However, it was not pointed out that this upper estimate was only a byproduct and some readers might have thought that Theorem 2.1 was needed mainly for proving this estimate. Luzgin (1980) and Wegener (1979) published the direct construction proving the present, somewhat improved bound. The aim of this section is to demonstrate that Theorem 2.1 provides good approximate solutions, too.

Theorem 2.3. *Let the integer $2 \leqslant R$ and the real number*

$$\kappa \geqslant \frac{R}{R!^{1/R}} \tag{2.3}$$

be fixed. Then

$$f(n, \kappa n^{1-1/R}) = \gamma n^{1/R} + \mathrm{O}(1), \tag{2.4}$$

where γ is the only real solution of the equation

$$\kappa\gamma + \frac{\gamma^R}{R!} = R + 1 \tag{2.5}$$

and $\mathrm{O}(1)$ does not depend on n, but may depend on R and κ. On the other hand, if

$$\kappa < \frac{R}{R!^{1/R}} \tag{2.6}$$

holds then

$$f(n, \kappa n^{1-1/R}) = \frac{R}{\kappa} n^{1/R} + \mathrm{O}(1) \tag{2.7}$$

is the approximate solution.

Proof. *Case* A: Suppose that (2.3) holds with strict inequality.

We will see that the pair $(R+1, \gamma n^{1/R})$ is an asymptotic solution of system (2.1), (2.2), in the below described sense.

Let

$$p_n^r(x) = \kappa n^{1-1/R} x + \sum_{i=0}^{r-1} (r-i) \binom{x}{i} - rn,$$

that is, the polynomial on the left hand side of (2.1) with our actual value of k.

It is easy to see that

$$p_n^{R+1}(\gamma n^{1/R}) = \kappa\gamma n + \frac{(\gamma n^{1/R})^R}{R!} + \mathrm{O}((\gamma n^{1/R})^{R-1}) - (R+1)n.$$

Eq. (2.5) implies

$$p_n^{R+1}(\gamma n^{1/R}) = \mathrm{O}(n^{1-1/R}). \tag{2.8}$$

Lemma 2.4. *Let $p_n^r(x)$ be a polynomial of order $r-1$ whose coefficients may depend on n and suppose that it satisfies the following conditions*:

$$(p_n^r(x))' \geqslant \kappa n^{1-1/R} \quad \textit{for } r-2 \leqslant x, \tag{2.9}$$

$$(p_n^r(x_0(n))) = 0 \quad \textit{for some } r-2 \leqslant x_0(n), \tag{2.10}$$

$$(p_n^r(x_1(n))) = \mathrm{O}(n^{1-1/R}) \quad \textit{for some } r-2 \leqslant x_1(n). \tag{2.11}$$

Then

$$|x_0(n) - x_1(n)| \leqslant \delta \tag{2.12}$$

holds where δ does not depend on n.

Proof. By Rolle's theorem

$$\frac{p_n^r(x_0(n)) - p_n^r(x_1(n))}{x_0(n) - x_1(n)}$$

is equal to $p_n^r(x)'$ for some value x in the interval determined by $x_0(n)$ and $x_1(n)$. Therefore, by (2.9), it is at least $\kappa n^{1-1/R}$. Hence, we have

$$|x_0(n) - x_1(n)| \leqslant \frac{|p_n^r(x_0(n)) - p_n^r(x_1(n))|}{\kappa n^{1-1/R}} = \frac{\mathrm{O}(n^{1-1/R})}{\kappa n^{1-1/R}}$$

by (2.10) and (2.11). The last ratio can be bounded by a constant δ. □

Let us check that our $p_n^r(x)$ satisfies the conditions of the lemma.

$$(p_n^r(x))' = \kappa n^{1-1/R} + \sum_{i=0}^{r-1}(r-i)\left(\frac{(x-1)\cdots(x-i+1)}{i!} + \cdots + \frac{x\cdots(x-i+2)}{i!}\right)$$

is monotonically increasing for $r-2 \leqslant x$, therefore it is enough to prove (2.9) at $x=r-2$, which is trivial since all the terms of the sum are non-negative at $x=r-2$. It is easy to see that $p_n^r(r-2))<0$ if n is large enough. On the other hand, the function is monotone for $r-2 \leqslant x$, therefore $p_n^r(x)=0$ has exactly one real solution $r-2<x_0(n)$. This proves the validity of condition (2.10). Condition (2.11) holds for $x_1(n)=\gamma n^{1/R}$, when $r=R+1$. The lemma can be used for $p_n^{R+1}(x)$: $|x_0(n)-x_1(n)|$ is bounded by a δ independent of n. It remained to verify that

$$\sum_{i=o}^{R}\binom{x_0(n)}{i} \leqslant n < \sum_{i=o}^{R+1}\binom{x_0(n)}{i}.$$

The stronger

$$\sum_{i=o}^{R}\binom{x_1(n)+\delta}{i} \leqslant n < \sum_{i=o}^{R+1}\binom{x_1(n)-\delta}{i}$$

will be proved, instead. The largest term of the left hand side is smaller than

$$\frac{(x_1(n)+\delta)^R}{R!} = \frac{(\gamma n^{1/R}+\delta)^R}{R!} = \frac{\gamma^R}{R!}n + \mathrm{O}(n^{1-1/R}),$$

which is smaller than 1 if $\gamma^R<R!$, and this is a consequence of the strong form of (2.3) and (2.5). We have seen that if $(R+1, x_1(n))$ is an asymptotic solution of (2.1) and (2.2) then there is an exact solution $(R+1, x_0(n))$ where $x_0(n)$ and $x_1(n)$ differ only by a δ.

Case B: Suppose that (2.6) holds. Then the asymptotic solution is $(R, (R/\kappa)n^{1/R})$. Indeed,

$$p_n^R\left((R/\kappa)n^{1/R}\right) = Rn + \mathrm{O}((n^{1/R})^{R-1}) - Rn = \mathrm{O}(n^{1-1/R}).$$

The lemma can be used as before, this time for $p_n^R(x)$: $|x_0(n)-x_1(n)|$ is bounded by a δ independent of n where $x_1(n)=(R/\kappa)n^{1/R}$.

The inequalities

$$\sum_{i=0}^{R-1}\binom{x_0(n)}{i} \leqslant n < \sum_{i=0}^{R}\binom{x_0(n)}{i}$$

can be proved as in Case A.

Case C:

$$\kappa = \frac{R}{R!^{1/R}}.$$

Both $(R, R!^{1/R}n^{1/R})$ and $(R+1, R!^{1/R}n^{1/R})$ are asymptotic solutions of (2.1) and (2.2), since $\gamma = R/\kappa$ is a solution of (2.5) in this case.

A more precise notation is needed here: let $x_0^R(n)$ and $x_0^{R+1}(n)$ denote the real solution ($\geqslant R-2, R-1$, resp.) of the equations $p_n^R(x)=0$ and $p_n^{R+1}(x)=0$, respectively. By the lemma we have

$$|x_0^R(n)-x_1(n)| \quad \text{and} \quad |x_0^{R+1}(n)-x_1(n)|<\delta. \tag{2.13}$$

We want to show that either $(R, x_0^R(n))$ or $(R+1, x_0^{R+1}(n))$ is the solution of (2.1) and (2.2). The inequalities

$$\sum_{i=0}^{R-1}\binom{x_0^R(n)}{i}<n \quad \text{and} \quad n<\sum_{i=0}^{R+1}\binom{x_0^{R+1}(n)}{i} \tag{2.14}$$

can be proved as earlier. If, moreover

$$n<\sum_{i=0}^{R}\binom{x_0^R(n)}{i} \tag{2.15}$$

holds then (2.14) implies that $(R, x_0^R(n))$ is the solution, while

$$\sum_{i=0}^{R}\binom{x_0^{R+1}(n)}{i}\leqslant n \tag{2.16}$$

implies the same for $(R+1, x_0^{R+1}(n))$. An indirect way will be used to prove that either (2.15) or (2.16) must hold. Suppose

$$\sum_{i=0}^{R}\binom{x_0^R(n)}{i}\leqslant n<\sum_{i=0}^{R}\binom{x_0^{R+1}(n)}{i}, \tag{2.17}$$

$\binom{x}{i}$ is monotonically increasing for $i-1\leqslant x$, on the other hand $R-1\leqslant x_0^R(n), x_0^{R+1}(n)$ holds for large n, therefore (2.17) yields

$$x_0^R<x_0^{R+1}. \tag{2.18}$$

Consider

$$p_n^{R+1}(x)=xk+\sum_{i=0}^{R}(R+1-i)\binom{x}{i}-(R+1)n$$
$$=xk+\sum_{i=0}^{R-1}(R-i)\binom{x}{i}-Rn+\sum_{i=0}^{R}\binom{x}{i}-n=p_n^R(x)+\sum_{i=0}^{R}\binom{x}{i}-n.$$

Replacing x by $x_0^{R+1}(n)$ in the above equality

$$0=p_n^R(x_0^{R+1}(n))+\sum_{0}^{R}\binom{x_0^{R+1}(n)}{i}-n$$

is obtained. Eq. (2.17) implies $p_n^R(x_0^{R+1})\leqslant 0$ and this contradicts (2.18). □

Remark 2.5. Theorem 2.2 ensures that the order of magnitude of $f(n, \kappa n^{1-1/R})$ is $n^{1/R}$, however, it gives only that the constant is between R/κ and $R/\kappa+1$. One can easily see that the γ in Theorem 2.3 is between these numbers.

Remark 2.6. There is a somewhat incorrect statement in the proof of Lemma 5 in Katona (1966): for a given r (2.1) has exactly one non-negative real solution. The correct version is: for a given r (2.1) has at most one real solution $r-2 \leqslant x_0$. This modification does not cause too much trouble in the proof of Theorem 2.1 (that is, Theorem 3 in Katona, 1966). One must only check that Lemmas 3 and 4 lead to a solution (r, x_0) of (2.1) and (2.2) satisfying $r-1 \leqslant x_0$, starting from the trivial $m=n-1, s_0=1, s_1=n-1, s_2=\cdots=s_m=0$.

3. A lower estimate

Theorem 3.1. *$m=f(n,k,l)$ always satisfies the inequality*

$$\log n+\log\left(1+\binom{m}{1}+\cdots+\binom{m}{l}\right) \leqslant m. \tag{3.1}$$

However,

$$\frac{2nl}{n-2k} \leqslant m \tag{3.2}$$

implies the stronger

$$\log n+\log\left(1+\binom{m}{1}+\cdots+\binom{m}{l}\right) \leqslant mh\left(\frac{k}{n}+\frac{l}{m}\right). \tag{3.3}$$

Proof. Let ξ be a randomly chosen column of M with probability $1/n$. On the other hand, η is a random subset of rows of M where the probability of choosing a set is $1/(1+\binom{m}{1}+\cdots+\binom{m}{l})$ if the size of the set is at most l, otherwise 0, ξ and η are independent (η plays the role of the set of false answers.) ξ_i is the ith entry in the column ξ if $i \notin \eta$ and the complement (interchanging the role of 0 and 1) of it if $i \in \eta$.

The Hamming distance between ξ and $(\xi_1, \ldots, \xi_m)$ is at most l. As the Hamming distance between the columns is at least $2l+1$ therefore the column closest to $(\xi_1, \ldots, \xi_m)$ is ξ. That is, $(\xi_1, \ldots, \xi_m)$ uniquely determines ξ.

On the other hand, the set of positions where ξ and $(\xi_1, \ldots, \xi_m)$ differ is exactly η. The conclusion is that $(\xi_1, \ldots, \xi_m)$ uniquely determines the pair (ξ, η). This implies

$$H(\xi, \eta) \leqslant \sum_{i=1}^{m} H(\xi_i) \tag{3.4}$$

by (1.1). As ξ and η are independent,

$$H(\xi, \eta)=H(\xi)+H(\eta)=\log n+\log\left(1+\binom{m}{1}+\cdots+\binom{m}{l}\right) \tag{3.5}$$

holds. As ξ_i can take on only two distinct values, 0 and 1, we have $H(\xi_i)=h(P(\xi_i)=1)$ $\leqslant 1$. This fact, (3.4) and (3.5) prove inequality (3.1) in the theorem.

To obtain a stronger estimate we need an upper estimate on $H(\xi_i)$. First the probability $P(\xi_i = 1)$ will be estimated.

$$\begin{aligned} P(\xi_i = 1) &= P(\xi_i = 1 | i \notin \eta)P(i \notin \eta) + P(\xi_i = 1 | i \in \eta)P(i \in \eta) \\ &= \frac{|A_i|}{n}P(i \notin \eta) + \frac{n - |A_i|}{n}P(i \in \eta) \\ &= \frac{|A_i|}{n}(1 - P(i \in \eta)) + \left(1 - \frac{|A_i|}{n}\right)P(i \in \eta) \\ &\leqslant \frac{|A_i|}{n} + P(i \in \eta) \leqslant \frac{k}{n} + P(i \in \eta). \end{aligned} \tag{3.6}$$

The number of possible choices of η containing i is

$$1 + \binom{m-1}{1} + \cdots + \binom{m-1}{l-1},$$

therefore

$$P(i \in \eta) = \frac{1 + \binom{m-1}{1} + \cdots + \binom{m-1}{l-1}}{1 + \binom{m}{1} + \cdots + \binom{m}{l}}.$$

This cannot exceed l/m as

$$\frac{\binom{m-1}{i-1}}{\binom{m}{i}} = \frac{i}{m} \leqslant \frac{l}{m}$$

holds for all $0 \leqslant i \leqslant l$. Hence we have $P(i \in \eta) \leqslant l/n$. (3.6) implies

$$P(\xi_i = 1) \leqslant \frac{k}{n} + \frac{l}{m}. \tag{3.7}$$

If this right hand side is at most $\frac{1}{2}$ then we can use the monotonity of $h(y)$ on the interval $[0, \frac{1}{2}]$. The inequality $k/n + l/m \leqslant 1/2$ is equivalent to (3.2). Under this condition (3.7) implies

$$H(\xi_i) \leqslant h\left(\frac{k}{n} + \frac{l}{m}\right),$$

consequently (3.4) and (3.5) lead to (3.3). □

Remark 3.2. Eq. (3.1) is nothing else but the so-called sphere packing bound (if expressed for n) well known in coding theory (see e.g. Csiszár and Körner, 1981). If $k = n/2$ then, on one hand, (3.2) cannot be satisfied, on the other hand, the condition on the size of the question sets A_i is empty. Therefore, cannot expect anything else but an inequality expressing that the columns of M form an l-error-correcting code.

Remark 3.3. Suppose that n and k tend to infinity and $k/n \to \kappa$. Eq. (3.1) implies that m also tends to infinity. Assume that $l/m \to \lambda$ (that is, the probability of a lie asymptotically does not exceed λ). If $\kappa + \lambda < 1/2$ then (3) holds for large n's, so (4) also holds. It is well known that

$$\frac{\log \sum_{i=0}^{l} \binom{m}{i}}{m} \to h(\lambda)$$

in this case, so (4) leads to

$$\frac{\log n}{h(\kappa + \lambda) - h(\lambda)} \leqslant m.$$

4. Constructions

Let Y be a set of m elements. A *Steiner system* $S(m, r, u)$ is a family of r-element subsets of Y, such that every u-element subset of Y is contained in exactly one member of the family. It is obvious that the number of members is

$$N = \frac{\binom{m}{u}}{\binom{r}{u}}. \tag{4.1}$$

Let $\mathscr{F}$ be a family of subsets of Y. The degree of the family $\mathscr{F}$ at $y \in Y$ is the number of members F of $\mathscr{F}$ such that $y \in F$. A family is *almost regular* if the difference of the degree of the family at different elements y is at most one:

$$|F\colon y_1 \in F \in \mathscr{F}| - |F\colon y_2 \in F \in \mathscr{F}| \leqslant 1 \quad (y_1, y_2 \in Y).$$

The family $\mathscr{F}$ is *building-regular* (shortly *buildreg*) if the members $F_1, F_2, \ldots, F_N$ of $\mathscr{F}$ can be listed in such a way that the family $F_1, F_2, \ldots, F_j$ is almost regular for all $j (1 \leqslant j \leqslant N)$. On the other hand, it is *storing-regular* (shortly *storereg*) if one can choose j members of F so that the family $F_{i_1}, F_{i_2}, \ldots, F_{i_j}$ is almost regular for all $j (1 \leqslant j \leqslant N)$. Obviously, if $\mathscr{F}$ is buildreg then it is storereg, too.

A Steiner system $S(m, r, u)$ is called *resolvable* if it can be decomposed into partitions of the underlying set. This can happen only when r is a divisor of m. It is easy to see that a resolvable Steiner system is buildreg, therefore it is storereg, too.

Let $S(m, r, r - l)$ be a storereg Steiner system, and suppose that

$$n \leqslant \frac{\binom{m}{r-1}}{\binom{r}{r-1}}. \tag{4.2}$$

Define the matrix M with the help of $S(m,r,r-l)$: the columns of the matrix are the characteristic (zero–one) vectors of the first n members of $S(m,r,r-l)$, in the order of the definition of a storereg family. It is easy to see that the Hamming distance between any two columns is at least $2l+1$. (Actually, it is at least $2l+2$.) On the other hand, the number of ones in any given row is at most $\lceil nr/m \rceil$. That is, if n satisfies (4.2), k satisfies

$$\left\lceil \frac{nr}{m} \right\rceil \leqslant k \tag{4.3}$$

then

$$f(n,k,l) \leqslant m \tag{4.4}$$

holds.

Lemma 4.1. *Suppose that n,k,l are given and*

$$k \leqslant \frac{nr}{(nr!/l!)^{1/(r-l)} + r - l} + 1 \tag{4.5}$$

and

$$\left\lceil \frac{nr}{k-1} \right\rceil \leqslant m \tag{4.6}$$

hold with some integer $l<r$. Then (4.2) *and* (4.3) *also hold.*

Proof. Suppose that (4.6) holds. Then we have

$$\left\lceil \frac{nr}{m} \right\rceil \leqslant \left\lceil \frac{nr}{\lceil nr/(k-1) \rceil} \right\rceil \leqslant \left\lceil \frac{nr}{nr/(k-1)} \right\rceil = k-1<k.$$

On the other hand, (4.5) implies the following inequality:

$$\frac{nr!}{l!} \leqslant \left(\frac{nr}{k-1} - (r-l) \right)^{r-l}.$$

Hence, by using (4.6), we obtain

$$n \leqslant \left(\frac{nr}{k-1} - (r-l) \right)^{r-l} \frac{l!}{r!} \leqslant (m-(r-l))^{r-l} \frac{l!}{r!}$$

$$\leqslant m(m-1)\dots(m-(r-l)+1)\frac{l!}{r!} = \frac{\begin{pmatrix} m \\ r-1 \end{pmatrix}}{\begin{pmatrix} r \\ r-1 \end{pmatrix}}. \qquad \square$$

This lemma and (4.4) result in

Theorem 4.2. *If n,k and l satisfy* (4.5) *with some positive integer $r(l<r)$ and there is a storereg Steiner family $S(m,r,r-l)$ where m satisfies* (4.6) *then*

$$f(n,k,l) \leqslant m.$$

Of course, the ideal case is when $m = \lceil nr/(k-1) \rceil$, otherwise one should choose the smallest possible m.

The following theorem is an asymptotic consequence of the above results.

Theorem 4.3. *Suppose that $l < R$ are fixed, n tends to infinity,*

$$k \sim \kappa n^{1-1/(R-l)}, \tag{4.7}$$

then

$$\frac{R-2l}{\kappa} \leqslant \liminf \frac{f(n,k,l)}{n^{1/(R-l)}} \tag{4.8}$$

holds. On the other hand, if

$$\kappa < R\left(\frac{l!}{R!}\right)^{1/(R-l)} \tag{4.9}$$

and there is an infinite series of such pairs (n,k) which satisfy (4.7) *with some $R(l<R)$ and there is a storereg Steiner family $S(m,R,R-l)$ where*

$$\left\lceil \frac{nR}{k-1} \right\rceil \leqslant m \sim \frac{R}{\kappa} n^{1/(R-l)},$$

then

$$\liminf \frac{f(n,k,l)}{n^{1/(R-l)}} \leqslant \frac{R}{\kappa} \tag{4.10}$$

completes (4.9).

Proof. A weakened form of (3.3) will be used to prove (4.8):

$$\log n \leqslant mh\left(\frac{k}{n} + \frac{l}{m}\right). \tag{4.11}$$

It will be shown that if

$$m \leqslant \left(\frac{R-2l}{\kappa} - \varepsilon\right) n^{1/(R-l)} \tag{4.12}$$

holds with some $0 < \varepsilon$ (and k satisfies (4.7)) then (4.11) cannot hold.

The trivial inequality

$$-2x \leqslant \log(1-x) \quad (0 \leqslant x \leqslant \tfrac{1}{2})$$

implies

$$h(x) = -x\log x - (1-x)\log(1-x) \leqslant -x\log x + 2x(1-x). \tag{4.13}$$

This inequality will be applied for $x = k/n + l/m$ where k and m are determined by (4.7) and (4.12), respectively, l is fixed.

It is easy to see that

$$x \geqslant n^{-1/(R-l)}\left(\kappa + \frac{l}{(R-2l)/\kappa - \varepsilon}\right) + \mathrm{o}(n^{-1/(R-l)}),$$

therefore

$$-x\log x + 2x(1-x)$$
$$\leqslant \frac{1}{R-l}\left(\kappa + \frac{l}{(R-2l)/\kappa - \varepsilon}\right) n^{-1/(R-l)}\log n + \mathrm{o}(n^{-1/(R-l)}\log n)$$

holds. By (4.13) this is an upper estimate on $h(x)$ too. Eqs. (4.11) and (4.12) lead to a contradiction since

$$\left(\frac{R-2l}{\kappa} - \varepsilon\right)\frac{1}{R-l}\left(\kappa + \frac{l}{(R-2l)/\kappa - \varepsilon}\right) < 1$$

can be easily checked. The first part of the theorem is proved.

Suppose now that (n,k) and $S(m,R,R-l)$ are chosen according to the second part of the theorem and prove (4.10). It is easy to see that (4.5) hold with $r=R$. Therefore Theorem 4.2 can be applied:

$$f(n,k,l) \leqslant m \sim \frac{R}{\kappa} n^{1/(R-l)}. \qquad \square$$

If $k < c\sqrt{n}$ we have a somewhat improved construction. Suppose that $l+1$ divides m and Y is an m-element set. The family $N(m,l+2,2) = \{A_1,\ldots,A_{m/(l+1)},B_1,\ldots,B_N\}$ of subsets of Y is called a *nearly Kirkman system* if the A's have size $l+1$, $\{A_1,\ldots,A_{m/(l+1)}\}$ forms a partition of Y, the B's have size $l+2$ and every 2-element subset of Y is contained in exactly one member of the family. A nearly Kirkman system is *resolvable* (*buildreg*, *storereg*) if the subsystem of B's is resolvable (builddreg, storereg). Very little is known about the existence of such systems. However, the case of $l=1$ was settled, mainly in Kotzig and Rosa (1974). Papers Baker and Wilson (1977), Brouwer (1978), Rees and Stinton (1987) completed it with the small cases.

Theorem 4.4. (Kotzig and Rosa, 1974; Baker and Wilson, 1977; Brouwer, 1978; Rees and Stinton, 1987). *If* 6 *divides m then there is a resolvable nearly Kirkman triple system.*

These nearly Kirkman systems can be analogously used for our purposes, like the Steiner systems earlier. Suppose

$$l+1|m, \quad \left\lceil \frac{n(l+2)}{k+1/(l+1)} \right\rceil \leqslant m. \tag{4.14}$$

The columns of the $m\times n$ matrix M will be the characteristic vectors of the first n members of $N(m,l+2,2)$, starting with the A's and then in order of the definition of the storereg family. It is easy to see that the Hamming distance between any two columns is at least $2l+1$, we have to verify only that the number of 1's in every row is at most k.

This is trivially true when $n \leqslant m/(l+1)$ and $1 \leqslant k$. Therefore $n > m/(l+1)$ can be supposed. Then each column, except the first $m/(l+1)$ ones, contains exactly $l+2$ 1's. The total number of 1's is $n(l+2) - m/(l+1)$. The storereg property implies that the number of 1's in a row is at most

$$\left\lceil \frac{n(l+2) - m/(l+1)}{m} \right\rceil = \left\lceil \frac{n(l+2)}{m} - \frac{1}{l+1} \right\rceil . \tag{4.15}$$

By (4.14) we have

$$\frac{n(l+2)}{m} \leqslant k + \frac{1}{l+1}$$

and hence (4.15) is at most k, proving the other important property of M.

Of course, this construction works only when n does not exceed the number $m/(l+1) + N$ of the members of $N(m, l+2, 2)$. N can be determined from the equality

$$\binom{m}{2} = \frac{m}{l+1}\binom{l+1}{2} + N\binom{l+2}{2}.$$

Therefore, we have to prove the following inequality:

$$n \leqslant \frac{m}{l+1} + N = \frac{m^2 + m}{(l+2)(l+1)}. \tag{4.16}$$

We claim that this holds when

$$k \leqslant \sqrt{\frac{l+2}{l+1}\sqrt{n}} - \frac{1}{l+1}.$$

Indeed, the latter inequality implies

$$\left(k + \frac{1}{l+1}\right)^2 \leqslant \frac{l+2}{l+1} n.$$

Hence, by (4.14) $n(l+2)(l+1) \leqslant m^2$ is obtained, proving (4.16).

It is somewhat surprising that this construction is the best possible under the given condition on k.

Theorem 4.5. *If* $k < \sqrt{[(l+2)/(l+1)]\sqrt{n}} - 1/(l+1)$,

$$m = \left\lceil \frac{n(l+2)}{k + 1/(l+1)} \right\rceil .$$

is divisible by $l+1$ *and there is a nearly Kirkman system* $N(m, l+2, 2)$ *then*

$$f(n,k,l) = \left\lceil \frac{n(l+2)}{k + 1/(l+1)} \right\rceil .$$

Proof. Let M be an $m \times n$ 0,1 matrix, in which the Hamming distances of columns are at least $2l+1$ and the number of 1's in each row is at most k. Denote the number of columns containing i 1's by s_i $(0 \leqslant i \leqslant m)$. We give now a lower estimate on the total number of 1's in M:

$$\sum_{i=0}^{m} is_i = \sum_{i=0}^{l+2} is_i + \sum_{i=l+3}^{m} is_i \geqslant \sum_{i=0}^{l+2} (l+2)s_i - \sum_{i=0}^{l+2} (l+2-i)s_i + \sum_{i=l+3}^{m} (l+2)s_i$$

$$= \sum_{i=0}^{m} (l+2)s_i - \sum_{i=0}^{l+2} (l+2-i)s_i = n(l+2) - \sum_{i=0}^{l+2} (l+2-i)s_i.$$

Comparing it with the trivial upper estimate km, the following inequality is obtained:

$$n(l+2) - \sum_{i=0}^{l+2} (l+2-i)s_i \leqslant km. \tag{4.17}$$

It is easy to see that

$$\sum_{i=0}^{l} s_i \leqslant 1 \tag{4.18}$$

holds by the distance condition for the columns.

Some cases will be distinguished.

Case A: $\sum_{i=0}^{l} s_i = 0$. Two columns containing exactly $l+1$ 1's cannot have a common 1, therefore $s_{l+1} \leqslant m/(l+1)$. Eq. (4.17) becomes

$$n(l+2) - \frac{m}{l+1} \leqslant km, \tag{4.19}$$

proving the statement for this case.

Case B: $s_j = 1$ for some j $(0 \leqslant j \leqslant l)$. A sharpening of (4.17) will be used in this case. It can be supposed, without loss of generality, that the first column contains j 1's and they stand in the first j rows. Let s_{uv} denote the number of columns containing u 1's in the first j rows and v 1's in the rest of the rows. First, give an upper bound on the number of 1's in the matrix formed by the last $n-1$ columns of M:

$$\sum (u+v)s_{uv} \leqslant (k-1)j + k(m-j) = km - j. \tag{4.20}$$

(To be precise, one should write s^*_{uv} for the submatrix.) Let us give a lower estimate on the left hand side, analogous to (4.17):

$$n(l+2) - \sum_{u+v<l+2} (l+2-u-v)s_{uv} \leqslant \sum (u+v)s_{uv}.$$

Combine this inequality with (4.20).

$$n(l+2) - \sum_{u+v<l+2} (l+2-u-v)s_{uv} \leqslant km. \tag{4.21}$$

Case BA: $j = l$, that is, $s_l = 1$. It is sufficient to consider the term with $u+v = l+1$ on the left hand side of (4.21). All other terms are 0. However, s_{uv} is zero for all $v < l+1$ because of the distance condition between the columns. For the same reason $s_{0,l+1} \leqslant (m-j)/(l+1)$. Eq. (4.21) leads to

$$n(l+2) - \frac{m-l}{l+1} \leqslant km.$$

This inequlity is stronger than (4.19), proving the statement in this case.

Case BB: $j<l$. All terms in the sum on the left hand side of (4.21) are zero, including $s_{0,l+1}$. The so obtained inequality

$$n(l+2) \leqslant km$$

implies (4.19), completing the proof. □

The above constructions are good when k is relatively small. The next construction is trivial and rather week, but works for all values of the parameters. Let $C(m^*, 2l+1)$ a *binary code*, that is, a set of n binary sequences of length m^*, with pairwise Hamming distance at least $2l+1$. Consider them as columns of a matrix. It "almost" satisfies the conditions of our problem, the "only" missing condition is that the number of 1's may exceed k. Cutting the matrix into parts of width k and making these $\lceil n/k \rceil$ parts "disjoint" by "pulling" the matrix vertically $\lceil n/k \rceil$-times longer, the so obtained matrix will satisfy all the conditions if $m^* \lceil n/k \rceil \leqslant m$. This trivial construction can be somewhat improved if we exploit the fact that the large Hamming distance ensures a large number of 0's.

Theorem 4.6. *Let $A(m^*, 2l+1)$ be the maximum number of codewords in a binary code of length m^* with pairwise Hamming distance at least $2l+1$. Then*

$$f(n,k,l) \leqslant \frac{nm^*}{2k} + \frac{1}{2k}\sqrt{n^2(m^*)^2 - 2m^*n(n-1)(2l+1)} + m^*, \tag{4.22}$$

where m^* is the smallest integer such that

$$n \leqslant A(m^*, 2l+1). \tag{4.23}$$

Proof. By (4.23) there exists an $m^* \times n$ 0,1 matrix M such that the pairwise Hamming distance of its columns are at least $2l+1$. Denote the number of 1's in the ith row by a_i $(1 \leqslant i \leqslant m^*)$. Replace the first row of M by $\lceil a_1/k \rceil$ new rows, obtained by cutting the first row into parts containing at most k 1's, that is, the ith of these new rows contains a 1 in the jth position iff the first row of M contains a 1 here and it is the $(i-1)k+1$st or … or ikth 1 in the first row. Denote the so obtained matrix by M'. It is obvious that the pairwise Hamming distance between the columns of M' is at least $2l+1$ again.

Repeating this step with all of the rows of M we arrive at a matrix M_1 having at most

$$\frac{\sum_{i=1}^{m^*} a_i}{k} + m^* \tag{4.24}$$

rows, n columns and the Hamming distance is not less than before.

Counting the number of (0,1)-pairs in the same rows of M, the following inequality can be obtained:

$$\binom{n}{2}(2l+1) \leqslant \sum_{i=1}^{m^*} a_i(n-a_i).$$

Apply the Cauchy–Schwarz inequality for the right hand side:

$$\sum_{i=1}^{m^*} a_i n - \sum_{i=1}^{m^*} a_i^2 \leqslant n \sum_{i=1}^{m^*} a_i - \frac{(\sum_{i=1}^{m^*} a_i)^2}{m^*}.$$

The last two inequalities lead to

$$\binom{n}{2}(2l+1) \leqslant n \sum_{i=1}^{m^*} a_i - \frac{(\sum_{i=1}^{m^*} a_i)^2}{m^*}.$$

Solve this quadratic ineqality for $\sum_{i=1}^{m^*} a_i$:

$$\sum_{i=1}^{m^*} a_i \leqslant \frac{1}{2} n m^* + \frac{1}{2}\sqrt{n^2(m^*)^2 - 2m^* n(n-1)(2l+1)}.$$

Substituting this in (4.24), the right hand side of (4.22) is obtained. □

Remark 4.7. Suppose that n and k tend to infinity and $k/n \to \kappa$, moreover l/m tends to λ. The asymptotic Varshamov–Gilbert bound (see e.g. MacWilliams and Sloane, 1977) gives

$$1 - h(2\lambda^*) \leqslant \lim \frac{\log A(m^*, 2l+1)}{m^*}$$

when $l/m^* \to \lambda^* \leqslant 1/4$. Use the minimality of m^*:

$$A(m^* - 1, 2l+1) < n.$$

These two inequalities imply

$$m^* \leqslant \frac{\log n}{1 - h(2\lambda^*)} + \mathrm{o}(\log n).$$

Use a weakened version of (4.22):

$$m \leqslant \frac{nm^*}{2k} + m^*. \tag{4.25}$$

Then

$$m \leqslant \frac{\log n}{1 - h(2\lambda^*)}\left(\frac{1}{2\kappa} + 1\right) + \mathrm{o}(\log n) \tag{4.26}$$

is obtained. Eq. (4.25) also implies

$$\frac{m}{m^*} \leqslant \frac{n}{2k} + 1 \to \frac{1}{2\kappa} + 1.$$

Hence we have

$$\lambda = \lim \frac{l}{m} = \lim \frac{l}{m^*} \lim \frac{m^*}{m} \geqslant \lambda^* \frac{1}{1/2\kappa + 1},$$

that is,

$$\lambda^* \leqslant \lambda\left(\frac{1}{2\kappa} + 1\right).$$

Substitute this in (4.26):

$$f(n,k,l) \leqslant \frac{\log n}{1 - h(2\lambda(1/2\kappa + 1))}\left(\frac{1}{2\kappa} + 1\right) + o(\log n) \tag{4.27}$$

if $\lambda(1/2\kappa + 1) \leqslant 1/4$. Using (4.22) in its full power, a somewhat stronger, but more complicated form can be obtained.

5. Further remarks and questions

1. Let us illustrate the difference among the resolvability, buildreg and storereg properties on the family of all r-element subsets of an m-element set. The family can be resolvable only when r divides m. The celebrated theorem of Baranyai (1975) states that it is really resolvable in this case.

It is easy to see that the first $m/\gcd(m,r)$ sets of a buildreg family covers X exactly $r/\gcd(m,r)$ times, and so on, the system splits into $\binom{m}{r}\gcd(m,r)/m$ subfamilies, each of which covers X $r/\gcd(m,r)$ times, and they can be obtained from each other by permuting the elements of X. This is an equivalent formulation of the definition of buidreg property for the present case. It was only conjectured by Baranyai and the present author (see e.g. Katona, 1991) that the family of all r-element sets of an m-element set possesses the latter property.

However, the storereg property of the same family is proved as an easy lemma in Katona (1966).

2. Suppose that a family is *regular*, that is, every element of the underlying set is contained in the same number of subsets. It is easy to see that this does not imply even the storereg property even for the case $r = 2$: the vertex-disjoint union of two odd cycles with $j = n/2$ gives a counterexample. A finite projective geometry is another obvious counterexample.

Problem 5.1. *Find a sufficient condition for a regular family to make it storereg.*

3. The storereg property is not indispensable for Theorem 4.3 What is really needed is that the degrees of the elements in the subfamily differ in a relatively small number. This serves as a motivation for the following problem. Let $\mathscr{F} = \{F_1, F_2, \ldots, F_N\}$ be a family of r-element subsets of an m-element set. Let $\rho(\mathscr{G})$ denote the difference between the maximum and minimum degrees of the family $\mathscr{G}$ at the elements. $\sigma(\mathscr{F}, j)$ is defined as the minimum of $\rho(\mathscr{G})$ for all j-member subfamilies $\mathscr{G}$ of $\mathscr{F}$. Further, let $\sigma(\mathscr{F}) = \max_{1 \leqslant j \leqslant N} \sigma(\mathscr{F}, j)$.

Problem 5.2. *Prove that there is a Steiner system $S(m,r,u)$ whose σ is small, where "small" is a function of m, r and u.*

On the other hand, it would also be interesting to determine $\max \sigma(\mathscr{F})$ for different classes of families.

4. Our results are good only in the case when k is small, namely, if its order of magnitude is n^{α} where $\alpha < 1$. If k is near to n, then the problem becomes "real" coding theory. The lower estimate in Remark 3.3 can be probably improved by the linear programming method, as it was suggested by Linial (1998). The good constructions must be similar to the codes known from the theory of error-correcting codes.

Problem 5.3. *Find good lower and upper estimates on $f(n,k,l)$ when $k = \kappa n$ and l is fixed.*

Problem 5.4. *Find good lower and upper estimates on $f(n,k,l)$ when $k = \kappa n$ and $l = \lambda m$. (The last condition might make more sense in the form $l \leqslant \lambda m$.)*

Remarks 3.3 and 4.7 show that the order of magnitude in this case is constant times $\log n$, however both the lower and upper estimates can be improved, (4.27) is especially weak and is valid only for the case when λ is less than half of κ.

References

Baker, R.D., Wilson, R.M., 1977. Nearly Kirkman triple systems. Utilitas Math. 11, 289–296.

Baranyai, Zs., 1975. On the factorization of the complete uniform hypergraph. In: Hajnal, A., Rado, R., Sós, V.T., (Eds.), Infinite and Finite Sets, Vol. I, Colloq. Keszthely, 1973, North-Holland, Amsterdam; Colloq. Math. Soc. Janos. Bolyai 10, 91–108.

Brouwer, A.E., 1978. Two new nearly Kirkman triple systems. Utilitas Math. 13, 311–314.

Csiszár, I., Körner, J., 1981. Information Theory. Akadémiai Kiadó, Budapest.

Feinstein, A., 1958. Foundations of Information Theory. McGraw-Hill, New York.

Hill, R., 1995. Searching with lies. In: Surveys in Combinatorics, Stirling. London Mathematical Society Lecture Note Series, Vol. 218. Cambridge University Press, Cambridge, pp. 41–70.

Katona, G., 1966. On separating systems of a finite set. J. Combin. Theory 1, 174–194.

Katona, G.O.H., 1991. Rényi and the combinatorial search problems. Studia Sci. Math. Hungar. 1, 363–378.

Kotzig, A., Rosa, A., 1974. Nearly Kirkman systems. Proceedings of the Fifth S.-E. Conference, Boca Raton; Congr. Numer. 10, 607–614.

Linial, N., 1998. Personal communication.

Luzgin, V.N., 1980. Separating systems of partitions of a finite set. Combinatorial Analysis, Vol. 5. Moskov. Gos. Univ., Moscow, pp. 39–45. (in Russian).

MacWilliams, F.J., Sloane, N.J.A., 1977. The Theory of Error-Correcting Codes. North-Holland, Amsterdam, New York, Oxford.

Rees, R., Stinton, D.R., 1987. On resolvable group-divisible designs with block size 3. Ars Combin. 23, 107–120.

Wegener, I., 1979. Separating systems, whose elements are sets of at most k elements. Discrete. Math. 28, 219–222.