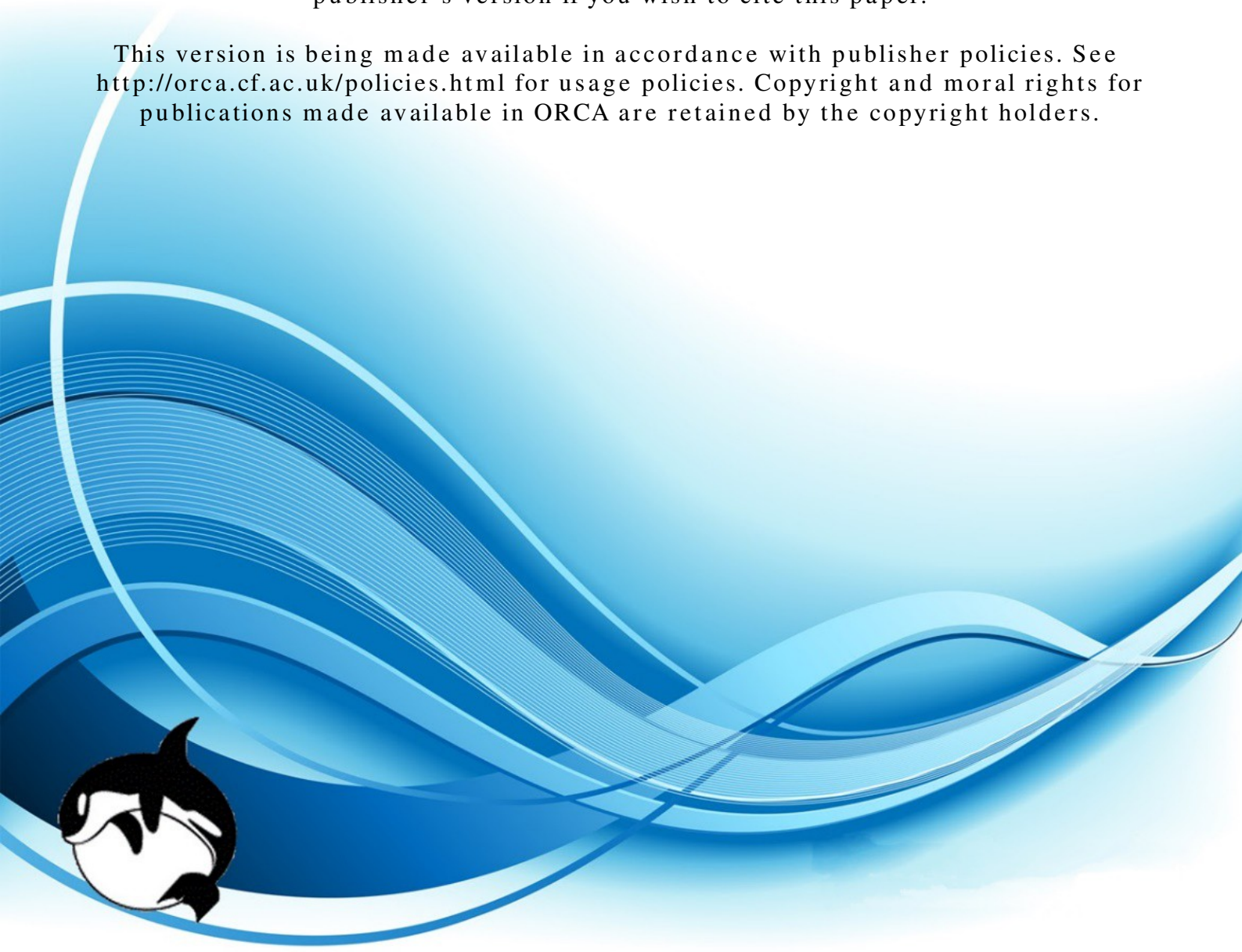

Nonadaptive group testing with lies: Probabilistic existence theorems

Anatoly Zhigljavsky,
School of Mathematics, Cardiff University,
Senghennydd Road, Cardiff CF24 4YH, UK
e-mail ZhigljavskyAA@cardiff.ac.uk

Abstract

We consider a wide range of combinatorial group testing problems with lies including binary, additive and multiaccess channel group testing problems. We derive upper bounds for the number of tests in the optimal nonadaptive algorithms. The derivation is probabilistic and is therefore non-constructive; it does not provide the way of constructing optimal algorithms. In the asymptotic setting, we show that the leading term for the number of tests does not depend on the number of lies and it is thus the same as for the zero-lie case. However, the other terms in the asymptotic upper bounds depend on the number of lies and substantially influence the upper bounds in the non-asymptotic situation.

Key words: Probabilistic method, Existence theorems, Group testing, Significant factors, Multiaccess channel, Searching with lies.

AMS classification: 62C10, 90Cxx, 28Dxx.

1 Introduction

We consider the following class of search problems. Assume that there are n elements $x_1, \dots, x_n$. Out of them, either t or $\le t$ of the so-called target elements have to be found by means of certain tests (t must be much smaller than n). The tests can be made for the groups $X = (x_{i_1}, \dots, x_{i_s})$ of s elements of the set $\mathbf{X} = \{x_1, \dots, x_n\}$; these groups will be called test sets. The results of the tests are expressed through the so-called test function f defined below in (1).

A search algorithm is a collection of the test sets. We only consider nonadaptive algorithms which are specified before the actual tests start. An algorithm $\mathcal{D}_N = \{X_1, \dots, X_N\}$ is called L-lie separating if all the target elements can be determined from the results of the N tests at $X_i \in \mathcal{D}_N$ provided that up to L mistakes (lies) in the test results are possible. An algorithm $\mathcal{D}_N$ is optimal if its length N is minimal.

The purpose of the paper is the derivation of the upper bounds for the length of optimal algorithms. In other words, we prove theorems that ensure the existence of search algorithms with a length smaller than the derived upper bounds. The method for the derivation of these bounds is probabilistic and is not constructive; it does not provide the procedures of constructing the optimal algorithms.

Let T be an unknown collection of the target elements; we shall call T the target set. We consider a class of test functions f which for a target set T and a test set X are defined as

$$f(X,T)=f_K(X,T)=\min\{K, |X \cap T|\}, \tag{1}$$

where $|\cdot|$ stands for the number of elements in a discrete set and K is some integer. This corresponds to the so-called K-channel model, see e.g. Du and Hwang (2000), Section 10.4. The following three special cases of (1) are very well known.

In the *binary* search $K = 1$; that is,

$$f(X,T)=\begin{cases} 0 & \text{if } X \cap T = \emptyset, \\ 1 & \text{if } X \cap T \neq \emptyset\,. \end{cases}$$

In the *additive* model $K = \infty$ and therefore $f(X,T) = |X \cap T|$; for this model, after testing a group X we receive the number of target elements in X. In the *multiaccess channel* model $K = 2$ and therefore $f(X,T) = \min\{2, |X \cap T|\}$. The binary search model is by far the most popular in search theory and applications.

In the case of no lies (L=0) the problems we are dealing with are often considered as the problems of the combinatorial group testing, see Du and Hwang (2000). To give a rough idea about the activity in the area and the results we obtain, assume that the observations are error–free, n is large and t is small relative to n. It is then often possible to prove existence of the group testing algorithms (and sometimes even to construct such algorithms) that provide exact determination of all target elements in

$$N(t,n) \;=\; C_t \ln n + o(\ln n) \quad (n \to \infty) \tag{2}$$

tests, for some constant C_t. Asymptotic problems deal with minimizing the constant C_t and, possibly, the other terms in $o(\ln n)$ in (2).

The main result of the present paper is Theorem 3 in Section 4 which implies that for a wide range of the L-lie search problems the constant C_t in (2) does not depend on the number of lies L and is the same as for the 0-lie case, see (26). Note, however, that the other terms in $o(\ln n)$ do depend on L and make a significant impact on the upper bounds, unless the value of n is astronomically large. In particular, if t is known and fixed then

$$\begin{aligned} N(t,n) &= C_t \ln n + O(1)\,, \quad n \to \infty, && \text{no lies}\,; \\ N_L(t,n) &= C_t \left(\ln n + \tfrac{2L}{t+1} \ln\ln n\right) + O(1)\,, \quad n \to \infty, && \le L \text{ lies}\,. \end{aligned}$$

Similar effect is observed in the case when a weak separation is required (note that in this case the constant C_t is smaller than in the case of strong, or full, separation). In the case of weak separation, only a $(1-\gamma)$-part of all target sets has to be separated, where γ is either fixed or tends to zero as $n \to \infty$.

Group testing is a well established area attracting attention of specialists in optimum algorithm, combinatorics, information theory and discrete search. The paper Dorfman (1943), devoted to sequential procedures of blood testing for detection of syphilitic men, is usually considered as the first work on the theory of group testing. A state-of-art in the field is well presented in the monographs Du and Hwang (2000, 2006). We do not consider an important problem of finding efficient algorithms. We only refer to Du and Hwang (2006), Ghosh and Avila (1985), Katona (1966), Katona and Srivastava (1983), Macula (1997) and Patel (1987) as a sample of works dealing with algorithm construction schemes in important specific cases including the case of the binary model with two and, more generally, t defectives.

Existence theorems constitute an important part of the theory of discrete mathematics, see e.g. Ahlswede (1987), Alon, Spencer and Erdős (1992). In the field of group testing, the corresponding activity has been originated in the seminal work Rényi (1965) and has been successfully continued by many authors, numerous examples are given in Du and Hwang (2000, 2006).

In recent years a number of papers have been published on the problem of constructing optimal algorithms for finding one, two or three defectives in search with lies, see e.g. Katona (2002), Porat and Rothschild (2008) as well as survey papers Deppe (2006) and Chen (2008). The author is not aware of any results on existence theorems for group testing algorithms in the presence of lies, except for an earlier paper of the author Zhigljavsky (2003), where some preliminary results have been reported.

The paper is organized as follows. In Section 2 we consider the group testing problems from the general point of view of discrete search and prove a general existence theorem. In Section 3 we demonstrate that for many interesting group testing problems the upper bounds can typically be written in the form $N_L(n) = \min\left\{k \ge 1 : \sum_i q_{i,n} r_{i,n}^k < 1\right\}$ for suitable coefficients $q_{i,n}$ and $r_{i,n}$. We also provide explicit formulae for these coefficients. In Section 4 we demonstrate how to derive asymptotic expressions for the upper bounds $N_L(n)$ from the asymptotic expressions for the upper bounds $N_0(n)$. As a consequence, known asymptotic expressions for $N_0(n)$ imply asymptotic expressions for $N_L(n)$.

2 General Existence Theorem

2.1 Discrete search problems

For the main case of the error–free tests, discrete search problems can often be determined as quadruples $\{\mathcal{T}, \mathcal{X}, f, \mathcal{Y}\}$, where $\mathcal{T} = \{T\}$ is a target field, that is an ordered collection of all possible *targets* T, $\mathcal{X} = \{X\}$ is a test field, that is a collection of all possible *test sets* X, and $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ is a *test function* mapping $\mathcal{X} \times \mathcal{T}$ to some space $\mathcal{Y}$, see O'Geran et al (1991) for details. A value $f(X, T)$ for fixed $X \in \mathcal{X}$ and $T \in \mathcal{T}$ is regarded as test or experimental result at X when the unknown target is T. We only consider *solvable* search problems, where every $T \in \mathcal{T}$ can be separated (found) by means of the test results at all $X \in \mathcal{X}$.

A *nonadaptive algorithm* $\mathcal{D}_N$ of length N is a collection $\mathcal{D}_N = \{X_1, \dots, X_N\}$ of test sets, which are chosen before the tests start. We shall not consider adaptive (sequential) algorithms and will omit the word 'nonadaptive' while referring to the algorithms.

For a pair of targets $T, T' \in \mathcal{T}$, we say that $X \in \mathcal{X}$ separates T and T' if $f(X, T) \neq f(X, T')$. We say that an algorithm $\mathcal{D}_N = \{X_1, \dots, X_N\}$ separates T in $\mathcal{T}$ if for any $T' \in \mathcal{T}$, such that $T' \neq T$, there is $X \in \mathcal{D}_N$ which separates the pair (T, T'). An algorithm $\mathcal{D}_N$ is *separating* if it separates all T in $\mathcal{T}$. (These algorithms are also called *strongly separating*; this distinguishes them from the weakly separating algorithms which provide separability for the majority of targets only rather than for all of them.)

In an L-lie search problem some test results can be wrong (that is, differ from $f(X, T)$ but still belong to $\mathcal{Y}$) but the total number of wrong answers is bounded by a given number $L \geq 0$.

If a 0-lie search problem is solvable then the corresponding L-lie search problem is solvable as well. Indeed, to provide a separating algorithm for the L-lie problem one may take a separating algorithm for the ordinary 0-lie search problem and repeat all the tests $2L + 1$ times. Analogously to the 0-lie case, separating algorithms should provide unique identifiability of all $T \in \mathcal{T}$.

An important fact is that if a non-sequential algorithm $\mathcal{D}_N = \{X_1, \dots, X_N\}$ is applied to a general L-lie search problem, then one can guarantee that the target can be uniquely defined if and only if the two vectors $F_T = (f(X_1, T), \dots, f(X_N, T))$ and $F_{T'} = (f(X_1, T'), \dots, f(X_N, T'))$ differ in at least $2L + 1$ components; here (T, T') is any pair of different targets in $\mathcal{T}$.

This fact can be expressed in terms of the Hamming distance between the vectors F_T and $F_{T'}$; recall that the Hamming distance between two vectors $F = (f_1, \dots, f_N)$ and $F' = (f_1^{'}, \dots, f_N^{'})$ is the number of components of F and F' that are different, that is,

$$d_H(F, F^{'}) = \text{ the number of } i \ (1 \leq i \leq N) \text{ such that } f_i \neq f_i^{'} .$$

Specifically, for a general L-lie search problem $\{\mathcal{T}, \mathcal{X}, f, \mathcal{Y}\}$ an algorithm $\mathcal{D}_N = \{X_1, \dots, X_N\}$ is separating if and only if for any $T, T' \in \mathcal{T}$, $T \neq T'$

$$d_H(F_T, F_{T'}) \geq 2L + 1 \,, \tag{3}$$

where $F_T = (f(X_1, T), \dots, f(X_N, T))$, $F_{T'} = \big(f(X_1, T^{'}), \dots, f(X_N, T^{'})\big)$ and $d_H(\cdot, \cdot)$ is the Hamming distance in $\mathcal{Y}^N$.

2.2 Existence theorem

In this section we formulate and prove a general existence theorem (Theorem 1). Note that in the case of error-free tests, some versions of Theorem 1 are known in literature, see Rényi (1965), O'Geran et al (1991), Zhigljavsky and Zabalkanskaya (1996), Dyachkov and Rykov (1983), Zhigljavsky (2003). In the case of L-lie search problems a rather complete version of Theorem 1 was sketched in Zhigljavsky (2003). However, for the sake of completeness we provide the full proof of this theorem.

Theorem 1. (Existence theorem for a general L-lie search problem)
Let $\{\mathcal{T}, \mathcal{X}, f, \mathcal{Y}\}$ be a solvable L-lie search problem with $|\mathcal{T}| > 1$, $\mathcal{R}$ be a probability distribution on $\mathcal{X}$ and for $T_i, T_j \in \mathcal{T}$ we set

$$p_{ij} = \Pr\{f(X, T_i) = f(X, T_j)\}\,,$$

where X is random and distributed according to $\mathcal{R}$. Then there exists a separating algorithm $\mathcal{D}_N = \{X_1, \ldots, X_N\}$ with length

$$N \le N_L(n) = \min\left\{ k = W+1, W+2, \ldots : \sum_{i=2}^{|\mathcal{T}|} \sum_{j=1}^{i-1} \sum_{w=0}^{W} \binom{k}{w} (p_{ij})^{k-w} (1-p_{ij})^{w} < 1 \right\}, \quad (4)$$

where $W = 2L$.

Proof. For a given algorithm $\mathcal{D}_N = \{X_1, \ldots, X_N\}$, consider the matrix

$$\mathcal{A}_N = \|f(X_i, T_j)\|_{i,j=1}^{N,|\mathcal{T}|}$$

with rows and columns corresponding to the test sets X_i and targets T_j, respectively. Let a_j $(j = 1, \ldots, |\mathcal{T}|)$ be the columns of the matrix $\mathcal{A}_N$. According to (3) the algorithm $\mathcal{D}_N$ is separating if and only if $d_H(a_i, a_j) \ge 2L + 1$ for all the pairs of columns (a_i, a_j) with $i \ne j$. Note that this implies $N \ge 2L + 1$.

Assume now that $X_1, \ldots, X_N$ in $\mathcal{D}_N$ are random, independent and distributed according to $\mathcal{R}$. Then for any fixed pair (i, j) such that $i \ne j$ $(i, j = 1, \ldots, |\mathcal{T}|)$, the Hamming distance $d_H(a_i, a_j)$ between the columns a_i and a_j of the matrix $\mathcal{A}_N$ is random and has the Binomial distribution with parameters N and p_{ij}; that is, for any integer w $(0 \le w \le N)$ we have

$$\Pr\{d_H(a_i, a_j) = w\} = \binom{N}{w} (p_{ij})^{N-w} (1-p_{ij})^{w}\,. \quad (5)$$

Let us introduce the events

$$E_{ij} = \{d_H(a_i, a_j) \le 2L \text{ for the pair } (T_i, T_j) \in \mathcal{T} \times \mathcal{T}\}$$

and the event

$$E = \{d_H(a_i, a_j) \le W \text{ for at least one pair } (T_i, T_j) \in \mathcal{T} \times \mathcal{T},\ i \ne j\} = \bigcup_{1 \le i < j \le |\mathcal{T}|} E_{ij}\,,$$

where $W = 2L$. In accordance with (5), we have for the probabilities of the events E_{ij}:

$$\Pr\{E_{ij}\} = \sum_{w=0}^{W} \binom{N}{w} (p_{ij})^{N-w} (1-p_{ij})^{w} \ . \tag{6}$$

Estimating the probability of a union of events E_{ij} by the sum of individual probabilities, we obtain the following estimate:

$$\Pr\{E\} = \Pr\left\{ \bigcup_{1\le j<i\le|\mathcal{T}|} E_{ij} \right\} \le \sum_{1\le j<i\le|\mathcal{T}|} \Pr\{E_{ij}\}\,. \tag{7}$$

Using the necessary and sufficient condition of separation (3), the formula (6) and the estimator (7), we obtain

$$\begin{aligned}
\Pr\{\mathcal{D}_N \text{ is separating}\} &= \Pr\{d_H(a_i,a_j) \ge W+1 \text{ for all pairs } (T_i,T_j)\in\mathcal{T}\times\mathcal{T},\ i\ne j\} \\
&= 1-\Pr\{d_H(a_i,a_j)\le W \text{ for at least one pair } (T_i,T_j)\in\mathcal{T}\times\mathcal{T},\ i\ne j\} \\
= 1-\Pr\{E\} \ge 1-\sum_{i=2}^{|\mathcal{T}|}\sum_{j=1}^{i-1}\Pr\{E_{ij}\} &= 1-\sum_{i=2}^{|\mathcal{T}|}\sum_{j=1}^{i-1}\sum_{w=0}^{W}\binom{N}{w}(p_{ij})^{N-w}(1-p_{ij})^{w}
\end{aligned}$$

Assume that N is large enough to provide the positivity of the right-hand side in the last inequality; this is satisfied, for example, for $N = N_L(n)$ defined in (4). For such N the probability that a random algorithm $\mathcal{D}_N$ of length N is separating is positive and the discreteness of $\mathcal{T}$ immediately implies the existence of a deterministic separating algorithm with this length. □

The inequality (7) and therefore the upper bound (4) seem to be crude. This may be true if either n is small or all p_{ij} are (approximately) equal. However, the bound (4) seems to be reasonably sharp in many difficult problems including the problems discussed below. This can be explained by the fact that the values p_{ij} for most pairs (T_i, T_j) are relatively small and therefore the value of the sum in the right–hand side of (4) is basically determined by the terms corresponding to very few pairs (i, j). Theorem 2 establishes an asymptotic version of this fact.

Note that one can consider a version of the general L-lie search problem where all wrong answers are the same; that is, the wrong results are equal to some value $y \in \mathcal{Y}$ which can be obtained by correct answers as well. This problem is a little simpler than the general L-lie problem and in this problem it is enough to ensure that $d_H(F_T, F_{T'}) \ge L+1$, rather than (3), to guarantee the strong separability of an algorithm. For this problem, the upper bound $N_L(n)$ of (4) can be reduced to a sharper bound $N_{L/2}(n)$.

2.3 Specification of the problems considered in the paper

In the problems we consider $\mathcal{T}$ is either $\mathcal{P}_n^t$ or $\mathcal{P}_n^{\le t}$ and $\mathcal{X} = \mathcal{P}_n^s$, where $1 \le t \le n$, $1 \le s \le n$, $\mathcal{P}_n^k = \{\{x_{i_1}, \ldots, x_{i_k}\}, \quad 1 \le i_1 < \ldots i_k \le n\}$ is the collection of all sets containing

exactly k elements x_i, and $\mathcal{P}_n^{\le k} = \bigcup_{j=0}^{k} \mathcal{P}_n^j$ is the collection of all sets containing containing not more than k elements.

The probability distribution $\mathcal{P}$ is uniform on $\mathcal{X} = \mathcal{P}_n^s$; the test function f belongs to the general class (1) with some K; in the asymptotic considerations we only consider $K = 1, 2$ and ∞. The set $\mathcal{Y}$ is uniquely defined by (1); it is $\mathcal{Y} = \{0, 1, \dots, K\}$.

Since $\mathcal{X} = \mathcal{P}_n^s$ and the probability distribution $\mathcal{P}$ is uniform on $\mathcal{X}$, we can write $p_{ij} = k_{ij}/|\mathcal{P}_n^s|$, where $|\mathcal{P}_n^s| = \binom{n}{s}$ and $k_{ij} = k(T_i, T_j)$ is the number of $X \in \mathcal{X} = \mathcal{P}_n^s$ such that $f(X, T_i) = f(X, T_j)$; that is,

$$k_{ij} = k(T_i, T_j) = \mid \{X \in \mathcal{P}_n^s : \quad f(X, T_i) = f(X, T_j)\} \mid . \tag{8}$$

In accordance with O'Geran et al (1991) and Zhigljavsky (2003) the numbers k_{ij} will be called *Rényi coefficients.*

Several other randomization schemes (that is, ways of defining the measure $\mathcal{P}$) in search problems are known. For example, a randomization with $\mathcal{X} = \mathcal{P}_n^{\le n}$ and a random inclusion of elements of $\mathbf{X}$ into X with fixed probability (to be optimised at a later stage) is known to work quite well in some group testing problems, see Du and Hwang (2000, 2006), Dyachkov and Rykov (1983). We however believe (our belief is based on extensive numerical evidence) that the scheme we consider allows to achieve better bounds, and it is also more appealing from the practical point of view, see Zhigljavsky (2003) for a discussion. One of the attractive features of the present scheme is that all the probabilistic statements can be formulated as equivalent combinatorial ones.

Note finally that there is a simple way, see Zhigljavsky (2003), of generalizing all the results of the present paper to the case $\mathcal{X} = \mathcal{P}_n^{\le s}$ or, more generally, $\cup_{s \in S} \mathcal{P}_n^s$, where S is any subset of $\{1, \dots, n\}$, but we did not find reasons why this could be advantageous over the simpler case $\mathcal{X} = \mathcal{P}_n^s$.

3 Computation of upper bounds: non-asymptotic case

To compute the upper bound (4) we need to compute the whole set of the Rényi coefficients (8); that is, the set $\{k_{ij},\ (T_i, T_j) \in \mathcal{T} \times \mathcal{T}\}$. As shown in Zhigljavsky (2003), in the problems we consider the set $\mathcal{T} \times \mathcal{T}$ can be partitioned into a few subsets, these subsets are defined in (9), where the Rényi coefficients are equal and have a closed-form representation through the binomial coefficients, see (13). In this section we summarize and specify the results of Zhigljavsky (2003) for the setups we consider.

Let us introduce the multinomial coefficient

$$\binom{n}{n_1\ n_2 \dots n_k} = \frac{n!}{n_1! n_2! \dots n_k!} \quad \text{for } n_r \ge 0,\ \sum_{r=1}^{k} n_r = n$$

and assume

$$\binom{n}{n_1\ n_2 \dots n_k} = 0 \quad \text{if} \ \ \min\{n_1, \dots, n_k\} < 0 \, .$$

Let $0 \le p \le m \le l \le n$, $p < l$. Denote

$$\mathcal{T}(n,l,m,p) = \{(T,T') \in \mathcal{P}_n^{\le n} \times \mathcal{P}_n^{\le n} : \ |T| = m, \ |T'| = l, \ |T \cap T'| = p\} .$$

Note that the condition $p < l$ guarantees that $T \ne T'$ for all pairs $(T,T') \in \mathcal{T}(n,l,m,p)$.

Easy counting arguments, see Zhigljavsky and Zabalkanskaya (1996), allow to compute the number of different non-ordered pairs in $\mathcal{T}(n,l,m,p)$, which is

$$Q_{n,l,m,p} = |\mathcal{T}(n,l,m,p)| = \begin{cases} \binom{n}{p \ m-p \ l-p \ n-l-m+p} & \text{if } m < l \\ \\ \frac{1}{2}\binom{n}{p \ m-p \ m-p \ n-2m+p} & \text{if } m = l \, . \end{cases} \tag{9}$$

Let (T,T') be any pair in $\mathcal{P}_n^{\le n} \times \mathcal{P}_n^{\le n}$ such that $T \ne T'$. For a given test field $\mathcal{X}$, we define

$$\mathcal{X}_{uvr}(T,T') = \{X \in \mathcal{X} : \ |X \cap (T \backslash T')| = u, \ |X \cap (T' \backslash T)| = v, \ |X \cap T \cap T'| = r\} \tag{10}$$

where u, v, r are some nonnegative integers. The pair (T,T') belongs to some $\mathcal{T}(n,l,m,p)$ with l,m,p such that $0 \le p \le m \le l \le n$ and $p < l$. The sets $\mathcal{X}_{uvr}(T,T')$ can be non-empty only if $0 \le u \le l-p, \ 0 \le v \le m-p, \ 0 \le r \le p$. Joining these restrictions on the parameters u,v,r with the restrictions on p,m and l in the definition of the sets $\mathcal{T}(n,l,m,p)$, we obtain the combined parameter restriction

$$0 \le p \le m \le l \le n, \ p < l, \ 0 \le u \le l-p, \ 0 \le v \le m-p, \ 0 \le r \le p \, . \tag{11}$$

The test field $\mathcal{X} = \mathcal{P}_n^s$ is balanced in the sense that the number $|\mathcal{X}_{uvr}(T,T')|$ does not depend on the choice of the pair $(T,T') \in \mathcal{T}(n,l,m,p)$ for any set of integers u,v,r,p,m,l satisfying (11). This number is equal to

$$R_{n,l,m,p,u,v,r,s} = |\mathcal{X}_{uvr}(T,T')| = \binom{p}{r}\binom{l-p}{u}\binom{m-p}{v}\binom{n-l-m+p}{s-r-u-v} \tag{12}$$

with $\binom{b}{a} = 0$ for $a < 0$ and $a > b$.

Theorem 3.3 in Zhigljavsky (2003) implies that for the general case of K-channel model with test function (1), $\mathcal{X} = \mathcal{P}_n^s$ and $(T_i,T_j) \in \mathcal{T}(n,l,m,p)$ with $0 \le p \le m \le l \le n$, $p < l$, the value of the Rényi coefficient k_{ij} does not depend on the choice of the pair $(T_i,T_j) \in \mathcal{T}(n,l,m,p)$ and equals $k_{ij} = K(n,l,m,p,s) =$

$$= \sum_{r=0}^{p}\sum_{u=0}^{m-p} R_{n,l,m,p,u,u,r,s} + \sum_{r=0}^{p}\sum_{u=q}^{l-p}\sum_{v=u+1}^{m-p} R_{n,l,m,p,u,v,r,s} + \sum_{r=0}^{p}\sum_{v=q}^{m-p}\sum_{u=v+1}^{l-p} R_{n,l,m,p,u,v,r,s} \tag{13}$$

with $q = \max\{0, K-r\}$.

As a consequence, for the general K-channel model (1) the upper bound (4) for the length of the optimal separating algorithm can be written as

$$N_L(n) = \min\left\{k \ge W+1 : \sum_{l,m}\sum_{p \le m}\sum_{w=0}^{W}\binom{k}{w} Q_{n,l,m,p}\left(p_{n,l,m,p,s}\right)^{k-w}\left(1-p_{n,l,m,p,s}\right)^{w} < 1\right\} \tag{14}$$

with $W = 2L$,

$$p_{n,l,m,p,s} = K(n,l,m,p,s)/\binom{n}{s},$$

$K(n,l,m,p,s)$ and $Q_{n,l,m,p}$ defined in (13) and (9), respectively, and $p_{n,m,m,m,s} = 0$ for all s and m. In (14), the first summation is taken over m, l such that $0 \le m \le l \le t$ for the case $\mathcal{T} = \mathcal{P}_n^{\le t}$; for the case $\mathcal{T} = \mathcal{P}_n^t$ the first summation disappears and $m = l = t$.

In the case of zero lies, $L = W = 0$ and the third sum in (14) contains only one term (with $w = 0$). In this case, (14) simplifies to

$$N_0(n) = \min\left\{k \ge 1 : \sum_{l,m} \sum_{p \le m} Q_{n,l,m,p} \left(p_{n,l,m,p,s}\right)^k < 1\right\}, \tag{15}$$

which is equivalent to formula (6) in Zhigljavsky (2003).

In the three particular cases ($K = 1, 2$ and ∞), formulae (13), (14) and (15), can be simplified, see Theorems 4.2, 4.3 and 4.4 in Zhigljavsky (2003).

For $K = 1$ (binary model), we obtain from (13) and (12)

$$K(n,l,m,p,s) = \binom{n}{s} - \binom{n-l}{s} - \binom{n-m}{s} + 2\binom{n-l-m+p}{s},$$

for $K = \infty$ (additive model):

$$K(n,l,m,p,s) = \sum_{u=0}^{p} \binom{l-p}{u}\binom{m-p}{u}\binom{n-l-m+2p}{s-2u},$$

and for $K = 2$ (multiaccess channel model):

$$K(n,l,m,p,s) = \binom{n}{s} - \binom{n-l}{s} - \binom{n-m}{s} - l\binom{n-l}{s-1} - m\binom{n-m}{s-1}$$
$$+2\binom{n-l-m+p}{s} + (l+m)\binom{n-l-m+p}{s-1} + 2(l-p)(m-p)\binom{n-l-m+p}{s-2}.$$

4 Asymptotic bounds

The asymptotic behaviour of $N_0(n)$ defined through (15), has been investigated in Zhigljavsky (2003). Below we show how to modify the asymptotic expressions for $N_0(n)$ to the case $L \ge 0$. For the derivation of the main result we shall need some bounds for the values of the Lambert W-function.

4.1 Asymptotic behaviour of the Lambert W-function

Consider the equation

$$\frac{e^x}{x} = z \tag{16}$$

with respect to x. For $z < e = 2.71828...$ this equation does not have a real solution; for $z > e$ there are two real solutions. Denote these solutions $x(z)$ and $\tilde{x}(z)$ with $0 < \tilde{x}(z) < 1$

and $x(z) > 1$. We shall be interested only in the solution $x(z)$, the largest solution of the equation (16). Below we shall need an asymptotic behaviour of $x(z)$ as $z \to \infty$.

We can express

$$x(z) = -\mathbf{W}_{-1}\left(-1/z\right) , \tag{17}$$

where $\mathbf{W}_{-1}(\cdot)$ is the lower branch of the Lambert W-function, see, for example, Corless et al (1996) and Barry et al (2000). Note that the function $\mathbf{W}_{-1}(t)$ has real values only for $-1/e \le t < 0$.

Known approximations for $\mathbf{W}_{-1}(t)$ (see Barry et al (2000) for a survey) do not imply simple bounds and clear asymptotic for $x(z)$ as $z \to \infty$. Rather than using the approximations and bounds known in the literature, we shall use the bounds derived in the following lemma.

Lemma 1. *Let $x(z)$ be the largest solution of the equation* (16)*. Then for all $z > e = 2.71828...$ we have*

$$\ln z + \ln\ln z < x(z) < \ln z + \ln\ln z + c_{\max} \tag{18}$$

with $c_{\max} = 0.4587$.

Proof. Set $y = \ln z + \ln\ln z$. Then

$$\frac{e^y}{y} = \frac{z \ln z}{\ln z + \ln\ln z} < z \quad \text{for all} \;\; z > e$$

and

$$\frac{e^{y+c}}{y+c} = \frac{e^c z \ln z}{\ln z + \ln\ln z + c} < z$$

for all $c > 0$. The largest value of c such that the equation

$$\frac{e^c z \ln z}{\ln z + \ln\ln z + c} = z$$

has a solution in $\{z, c\}$ with $z > e$ and $c > 0$ is $c = 0.4586751453870819...$ Any value larger than this guarantees the right-hand side inequality in (18) for all $z > e$. □

Note that we can reduce the value of the constant $c_{\max}$ (down to any positive value, of course) in the right-hand side of the inequality (18) on the expense of the reduction of the interval (z_*, ∞) for z, where this inequality holds. For example,

$z_* \cong 23.1$ for $c = 0.4$, $z_* \cong 124$ for $c = \frac{1}{3}$, $z_* \cong 4247$ for $c = \frac{1}{4}$ and $z_* \cong 199249$ for $c = \frac{1}{5}$.

We can see that z_* rapidly grows as $c_{\max}$ decreases.

Note finally that since $x(z) = -\mathbf{W}_{-1}(-1/z)$, see (17), the inequality (18) is essentially the inequality for the Lambert W-function $\mathbf{W}_{-1}(t)$.

4.2 Asymptotics for N

The case of the general group testing problem with no lies and the test function (1) has been considered in Zhigljavsky (2003). Theorem 5.1 in this paper implies that the equation for $N = N_0(n)$ (asymptotically, as $n \to \infty$ and $t/n \to 0$) becomes

$$c_* n^\alpha r^N = 1\,, \tag{19}$$

where $c_* > 0$, $\alpha > 0$ and $0 < r < 1$ are some constants which are uniquely determined by the parameters that define the group testing problem.

In the next statement we show how to generalize the asymptotic expressions for $N_0(n)$ derived in Zhigljavsky (2003) to the case $L \geq 0$.

Theorem 2. *Consider an L-lie group testing problem where $\mathcal{T}$ is either $\mathcal{P}_n^t$ or $\mathcal{P}_n^{\leq t}$, $\mathcal{X} = \mathcal{P}_n^s$, $n \to \infty$, $t/n \to 0$ and $s/n \to \lambda$ as $n \to \infty$ with $0 < \lambda < 1$. Let $N_L(n)$ and $N_0(n)$ be defined by (14) and (15), correspondingly. Assume that the asymptotic (as $n \to \infty$) expression for $N_0(n)$ is determined as the solution to the equation (19) with some constants $c_* > 0$, $\alpha > 0$ and $0 < r < 1$. Then the asymptotic expression for $N = N_L(n)$ is determined as the solution to the equation*

$$\frac{c_*}{W!}\left(\frac{1-r}{r}\right)^W N^W n^\alpha r^N = 1\,, \tag{20}$$

where the coefficients c_*, α and r are the same as in (19) and $W = 2L$.

Proof. Consider the non-asymptotic expression (14) for $N_L(n)$. Let us first demonstrate that asymptotically (as $n \to \infty$) all the terms with $w < W$ are dominated by the corresponding terms with $w = W$. Obviously, $N_L(n) \to \infty$ as $n \to \infty$ since $N_L(n) \geq N_0(n) \to \infty$. Consider the ratio $J_k(w, W) = J_k(w)/J_k(W)$ of the corresponding terms in (14), where

$$J_k(w) = \binom{k}{w} Q_{n,l,m,p} \left(p_{n,l,m,p,s}\right)^{k-w} \left(1-p_{n,l,m,p,s}\right)^w\,,$$

$w < W$ and $k \to \infty$. All probabilities $p_{n,l,m,p,s}$ are uniformly (with respect to n) bounded away from 1 (see formula (45) in Zhigljavsky (2003)) and therefore there exists a constant C such that

$$\left(\frac{p_{n,l,m,p,s}}{1-p_{n,l,m,p,s}}\right) \leq C \quad \forall n, l, m, p, s.$$

This implies

$$J_k(w, W) = \binom{k}{w} \Big/ \binom{k}{W} \left(\frac{p_{n,l,m,p,s}}{1-p_{n,l,m,p,s}}\right)^{W-w} \leq \text{const}\, \frac{(k-W)!}{(k-w)!} \to 0 \;\text{ as }\; k \to \infty\,,$$

where const$= C^{W-w} W!/w!$.

Therefore, asymptotically (as $n \to \infty$) we can ignore all the terms in (14) with $w < W$ and define $N_L(n)$ by

$$N_L(n) = \min\left\{k \geq W+1 :\; \binom{k}{W} \sum_{l,m} \sum_{p \leq m} Q_{n,l,m,p} \left(p_{n,l,m,p,s}\right)^{k-W} \left(1-p_{n,l,m,p,s}\right)^W < 1\right\} \tag{21}$$

The combination of indices l, m, p that define the asymptotically dominating term in the expression (15) for the 0-lie case and determine the coefficients c_*, α and r in the asymptotic equation (19) remain the same; they also define the dominating terms in (21). The ratio of these dominating terms for the 0-lie and L-lies cases is

$$\binom{k}{W}\left(\frac{1-p_{n,l,m,p,s}}{p_{n,l,m,p,s}}\right)^W .$$

The proof now follows from the facts that the coefficient r in (15) is $r = \lim_{n\to\infty} p_{n,l,m,p,s}$ for the dominating term and

$$\binom{k}{W} = \frac{k!}{(k-W)!W!} = \frac{k^W}{W!}\ (1+O(1))\,, \quad k\to\infty\,.$$

□

The solution of equation (19) is, of course,

$$N_{as}(0) = \frac{\alpha}{-\ln r}\ln n + \frac{\ln c_*}{-\ln r}\,. \tag{22}$$

This defines the asymptotic expression for N in the case of no lies. In the case of L lies, we need to find the asymptotic expression (as $n\to\infty$) for the solution of (20).

Set $N_1 = N/W$,

$$c_1 = \left(\frac{1-r}{r}\right)\left(\frac{c_*}{W!}\right)^{1/W} \tag{23}$$

and take power $1/W$ of both sides in (20). Then (20) becomes

$$c_1 N_1 n^{\alpha/W} r^{N_1} = 1.$$

This can be written as

$$\frac{c_1}{-\ln r} n^{\alpha/W} = \frac{\exp(N_2)}{N_2}\,,$$

where

$$N_2 = (-\ln r)\,N_1 = \frac{-\ln r}{W} N\,. \tag{24}$$

Applying (18), we obtain

$$N_2 = \ln\left(\frac{c_1}{-\ln r} n^{\alpha/W}\right) + \ln\ln\left(\frac{c_1}{-\ln r} n^{\alpha/W}\right) + c_0(n)$$

where $0 < c_0(n) < 0.4587$ for all $n \ge n_* = \left(e\,(-\ln r)/c_1\right)^{W/\alpha}$.

This and (24) imply that the solution of equation (20) satisfies

$$N_{as}(L) = \frac{\alpha}{-\ln r}\ \ln n\ +\ \frac{W}{-\ln r}\left(\ln\ln\left(c_2\, n^{\alpha/W}\right) + \ln c_2 + c_0(n)\right)\,, \tag{25}$$

where $W = 2L$, $c_2 = c_1/(-\ln r)$, c_1 is defined in (23) and $0 < c_0(n) < 0.4587$ for $n > n_*$.

The derivation above imply the following theorem which can be considered as the main result of this paper.

Theorem 3. *Consider an L-lie group testing problem where $\mathcal{T}$ is either $\mathcal{P}_n^t$ or $\mathcal{P}_n^{\leq t}$, $\mathcal{X} = \mathcal{P}_n^s$, $n \to \infty$, $t/n \to 0$ and $s/n \to \lambda$ as $n \to \infty$ with $0 < \lambda < 1$. Let $N_L(n)$ and $N_0(n)$ be defined by (14) and (15), correspondingly. Assume that the asymptotic (as $n \to \infty$) expression for $N_0(n)$ is $N_0(n) = N_{as}(0) + o(1)$ as $n \to \infty$, where $N_{as}(0)$ is defined in (22) with some constants $c_* > 0$, $\alpha > 0$ and $0 < r < 1$. Then we have $N_L(n) = N_{as}(L) + o(1)$ as $n \to \infty$, where $N_{as}(L)$ is defined in (25).*

Corollary 1. *Consider the L-lie group testing problem as in Theorem 3 and assume that n is large enough. Then the asymptotic upper bound $N_{as}(L)$ satisfies the inequality*

$$N_{as}(L) < \frac{\alpha}{-\ln r}\,\ln n \,+\, \frac{W}{-\ln r}\ln\ln n + \mathrm{const}\,, \tag{26}$$

where

$$\mathrm{const} = \frac{W}{-\ln r}\left(\max\{0, \ln c_2\} + \ln\left(\frac{\alpha c_2}{W}\right) + 0.4587\right)\,,$$

$W = 2L$, $c_2 = c_1/(-\ln r)$ and c_1 is defined in (23).

Proof. The expression (25) can more conveniently be written as

$$N_{as}(L) = \frac{\alpha}{-\ln r}\,\ln n \,+\, \frac{W}{-\ln r}\left(\ln\left(\frac{\alpha}{W}\ln n + \ln c_2\right) + \ln c_2 + c_0(n)\right)\,. \tag{27}$$

For any y and large enough x (x must satisfy $x > y_+/(e^{y_+}-1)$) we have $\ln(x+y) \leq \ln x + y_+$, where $y_+ = \max\{0, y\}$. Applying this inequality to (27) with $x = \frac{\alpha}{W}\ln n$ and $y = \ln c_2$ we obtain for n large enough

$$N_{as}(L) < \frac{\alpha}{-\ln r}\,\ln n \,+\, \frac{W}{-\ln r}\left(\ln\ln n + (\ln c_2)_+ + \ln\left(\frac{\alpha c_2}{W}\right) + c_0(n)\right)\,,$$

To conclude the proof we apply the inequality $c_0(n) < 0.4587$ which holds for all $n > n_*$. □

4.3 Particular cases

In this section, we consider several particular group testing problems where the test function has the form (1) with $K = 1, 2$ and ∞, $\mathcal{T}$ is either $\mathcal{P}_n^t$ or $\mathcal{P}_n^{\leq t}$, $\mathcal{X} = \mathcal{P}_n^s$, $n \to \infty$, $t/n \to 0$ and $s/n \to \lambda$ as $n \to \infty$, and the value of λ is chosen in an optimal way (to minimize the asymptotic upper bounds). We shall only provide the values of the coefficients α, r and c_*. The asymptotic upper bounds can then be constructed by applying Theorem 3.

Numerical experiments show that if n is sufficiently large, then in all particular cases considered below the asymptotic upper bound (25) with $c_0(n) = c_{\max} = 0.4587$ provides a very good approximation to the non-asymptotic bound (14).

4.3.1 Binary model ($K = 1$)

If $\mathcal{T} = \mathcal{P}_n^t$ (exactly t defective factors) then, see formula (67) in Zhigljavsky (2003), we have

$$\alpha = t + 1, \;\; r = 1 - 2t^t/(t+1)^{t+1}, \;\; c_* = 1/(2(t-1)!)\,.$$

If $\mathcal{T} = \mathcal{P}_n^{\le t}$ ($\le t$ defective factors) then, see formula (70) in Zhigljavsky (2003), we have

$$\alpha = t, \;\; r = 1 - (t-1)^{t-1}/t^t, \;\; c_* = 1/(t-1)!\,.$$

4.3.2 Additive model ($K = \infty$)

If $\mathcal{T} = \mathcal{P}_n^t$ (exactly t defective factors) then, in view of (58) in Zhigljavsky (2003), we have

$$\alpha = t + 1, \;\; r = \frac{1}{2}\,, \;\; c_* = 1/(2(t-1)!)\,.$$

For the additive model, the case $\mathcal{T} = \mathcal{P}_n^{\le t}$ ($\le t$ defective factors) is not very interesting.

4.3.3 Multiaccess channel ($K = 2$)

If $\mathcal{T} = \mathcal{P}_n^t$ (exactly t defective factors) with t fixed, then, see (77) in Zhigljavsky (2003), we have

$$\alpha = t + 1, \;\; r = r_\lambda = 1 - 2\lambda(1-\lambda)^{t-1}(1 + \lambda(t-2))\,, \;\; c_* = 1/(2(t-1)!)\,. \tag{28}$$

The optimal value of λ is $\lambda_t = (t-4+\sqrt{5t^2-12t+8})/(2t^2-2t-4)$; this value minimizes r_λ in (28). If $t \to \infty$ then $\lambda_t = \varphi/t + O(t^2)$ as $t \to \infty$ where $\phi = (\sqrt{5}+1)/2 \simeq 1.618034$ is the golden mean.

If $\mathcal{T} = \mathcal{P}_n^{\le t}$ ($\le t$ defective factors) then, see Theorem 5.7 in Zhigljavsky (2003), for $t = 2$ and 3 the values of α, r and c_* are as in (28) and for $t \ge 4$

$$\alpha = t, \;\; r = r_\lambda = 1 - \lambda(1-\lambda)^{t-2}(1 + \lambda(t-2)\,, \;\; c_* = 1/(t-1)!\,. \tag{29}$$

The optimal value of λ (minimizing r_λ in (29)) is $\lambda_{\le t} = (t-3+\sqrt{5t^2-14t+9})/(2t^2-4t)$. Similarly to the previous case, $\lambda_{\le t} = \varphi/t + O(t^2)$ as $t \to \infty$.

References

[1] Ahlswede R. and Wegener I. (1987). *Search Problems,* Wiley and Sons, N.Y.

[2] Alon, N., Spencer J. and Erdős P. (1992). *The Probabilistic Method*, Wiley and Sons, N.Y.

[3] Barry D.A., Parlange J.-Y., Li L., Prommer H., Cunnigham C.J. and Stagnitti F. (2000) Analytical approximations for real values of the Lambert W-function, Mathematics and Computers in Simulation, 53, 95-103.

[4] Chen, H.-B. and Hwang, F. K. (2008) A survey on nonadaptive group testing algorithms through the angle of decoding, Journal of Combinatorial Optimization, 15, 49-59.

[5] Corless R.M., Gonnet G.H., Hare D.E.G., Jeffrey D.J. , and Knuth D.E. (1996) On the Lambert W-function, Advances in Computational Mathematics, 5, 329-359.

[6] Deppe, C. (2006) Coding with Feedback and Searching with Lies, In: *Entropy, Search, Complexity.* Bolyai Society Mathematical Studies, 16. (eds I. Csisz'ar, G.O.H. Katona, G. Tardos), Springer, Berlin, 27-70.

[7] Dorfman, R. (1943). The detection of defective numbers of large population, *Ann. Math. Statist.* 14, 436–440.

[8] Du, D.Z. and Hwang, F.K. (2000) *Combinatorial Group Testing,* 2nd edition, World Scientific, Singapore.

[9] Du, D.Z. and Hwang, F.K. (2006) *Pooling Designs and Nonadaptive Group Testing,* World Scientific, Singapore.

[10] Dyachkov A.G. and Rykov V.V. (1983) A survey of superimposed code theory, *Problems Control Inform. Thy.* 12, 229–242.

[11] Erdős, P. and Rényi A. (1963) On two problems of information theory, *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 8A, 229–243.

[12] Ghosh, S. and Avila, D. (1985). Some new factor screening algorithms using the search linear model, *J. Statist. Planning and Inference* 11, 259–266.

[13] Katona, G.O.H. (1966) On separating systems of a finite set, *J. Combinatorial Theory*, 1, 174–194.

[14] Katona, G.O.H. (2002) Search with small sets in presence of a liar, *J. Statist. Plann. Inference*, 100, 319–336.

[15] Katona, G. and Srivastava J.N. (1983). Minimal 2-coverings of a finite affine space of GF(2), *J. Statist. Planning and Inference* 8, 375–388.

[16] Macula, A.J. (1997) Error-correcting nonadaptive group testing with d(e)-disjunct matrices *Discrete Applied Mathematics*, 80, 217–222.

[17] O'Geran, J.H., Wynn, H.P. and Zhigljavsky, A.A. (1991), Search, *Acta Applicandae Mathematicae*, 25, 241–276.

[18] Patel, M.S. (ed.) (1987), Experiments in factor screening, *Commun. Stat. – Th. and Meth.*, 16, No. 10.

[19] Porat, E., Rothschild, A. (2008), Explicit Non-adaptive Combinatorial Group Testing Schemes. *Proceedings of the International Colloquium on Automata, Languages and Programming,* vol. 1, 748-759.

[20] Rényi, A. (1965). On theory of random search, *Bull. Amer. Math. Soc.* 71, 809–828.

[21] Zhigljavsky, A. (2003). Probabilistic existence theorems in group testing, *J. of Statistical Planning and Inference*, 115, 1–43.

[22] Zhigljavsky, A. and Zabalkanskaya, L. (1996). Existence theorems for some group testing strategies, *J. of Statistical Planning and Inference*, 55, 151–173.