On a Hypergraph Approach to Multistage Group Testing Problems

A. G. D'yachkov, I.V. Vorobyev, N.A. Polyanskii and V.Yu. Shchukin
Lomonosov Moscow State University, Moscow, Russia
Email: agd-msu@yandex.ru, vorobyev.i.v@yandex.ru, nikitapolyansky@gmail.com, vpike@mail.ru

***Abstract*—Group testing is a well known search problem that consists in detecting up to s, $s \ll t$, defective elements of the set $[t] = \{1, \ldots, t\}$ by carrying out tests on properly chosen subsets of $[t]$. In classical group testing the goal is to find all defective elements by using the minimal possible number of tests. In this paper we consider multistage group testing. We propose a general idea how to use a hypergraph approach to searching defective elements. For the case $s = 2$ and $t \to \infty$, we design an explicit construction, which makes use of $2\log_2 t(1 + o(1))$ tests in the worst case and consists of 4 stages. For the general case of fixed $s > 2$ and $t \to \infty$, we provide an explicit construction, which uses $(2s-1)\log_2 t(1+o(1))$ tests and consists of $2s-1$ rounds.**

Keywords: Group testing problem, multistage algorithms, hypergraph, construction

I. Introduction

Group testing is a very natural combinatorial problem that consists in detecting up to s *defective elements* of the set of objects $[t] = \{1, \ldots, t\}$ by carrying out tests on properly chosen subsets (pools) of $[t]$. The test outcome is positive if the tested pool contains one or more defective elements; otherwise, it is negative.

There are two general types of algorithms. In *adaptive* group testing, at each step the algorithm decides which group to test by observing the responses of the previous tests. In *non-adaptive* algorithm, all tests are carried out in parallel. There is an intermediate algorithm between these two types, which is called a *multistage* algorithm. For the multistage algorithm all tests are divided into p sequential stages. The tests inside the same stage are performed simultaneously. The tests of the next stages may depend on the responses of the previous. In this context, a non-adaptive group testing algorithm is reffered to as a one stage algorithm.

A. Previous results

We refer the reader to the monographs [1], [2] for a survey on group testing and its applications. In spite of the fact that the problem of estimating the minimum *average* (the set of defective elements is chosen randomly) number of tests has been investigated in many papers (for instance, see [3], [4]), in the given paper we concentrate our attention only on the minimal number of test in the *worst case*.

Dyachkov and Rykov [5] proved that at least

$$\frac{s^2}{2\log_2(e(s+1)/2)} \log_2 t(1 + o(1))$$

tests are needed for non-adaptive group testing algorithm.

If the number of stages is 2, then it was proved that $O(s\log_2 t)$ tests are already sufficient. It was shown by studying random coding bound for disjunctive list-decoding codes [7], [8] and selectors [9]. The recent work [6] has significantly improved the constant factor in the main term of number of tests for two stage group testing procedures. In particular, if $s \to \infty$, then

$$\frac{se}{\log_2 e} \log_2 t(1 + o(1))$$

tests are enough for two stage group testing.

As for adaptive strategies, there exist such ones that attain the information theory lower bound, i.e., the necessary number N of tests for any algorithm satisfies the inequality

$$N \geqslant \left\lceil \log_2 \sum_{i=0}^{s} \binom{t}{i} \right\rceil = s\log_2 t(1 + o(1)),$$

if s is fixed and $t \to \infty$. However, for $s > 1$ the number of stages in well-known optimal strategies is a function of t, and grows to infinity as $t \to \infty$.

B. Summary of the results

In the given article, we develop some explicit multistage algorithms, in which the number of stages will depend only of s. The necessary notations are introduced in Sect. II. Sect. III presents a general idea how to search defective elements using a hypergraph approach. In Sect. IV, we describe a 4-stage group testing strategy, which detects up to 2 defective elements and uses the asymptotically optimal number of tests $2\log_2 t(1+o(1))$. As far as we know the best result for such a problem was obtained [10] by Damashke et al. in 2013. They provide an exact two stage group testing strategy and use $2.5\log_2 t$ tests. For other constructions for the case of 2 defective elements, we refer to [11], [12]. In Sect. V, we work out a multistage algorithm of detecting up to s defective elements in $2s-1$ rounds. Asymptotically ($t \to \infty$) the algorithm uses $(2s-1)\log_2 t(1 + o(1))$ tests in the worst case. In Sect. VI for certain values of t we present tables of numerical values of the number of tests of the suboptimal 4 stage algorithm discussed in Sect. IV detecting up to 2 defective elements among the set $[t]$.

II. Preliminaries

Throughout the paper we use t, s, p for the number of elements, defective elements, and stages, respectively. Let $\triangleq$

978-1-5090-1805-5/16/$31.00 ©2016 IEEE

denote the equality by definition, $|A|$ – the cardinality of the set A. The binary entropy function $h(x)$ is defined as usual

$$h(x) = -x \log_2(x) - (1-x)\log_2(1-x).$$

A binary $(N \times t)$-matrix with N rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and t columns $\boldsymbol{x}(1), \ldots, \boldsymbol{x}(t)$ (codewords)

$$X = \|x_i(j)\|, \quad x_i(j) = 0, 1, \quad i \in [N], \, j \in [t]$$

is called a *binary code of length N and size t*. The number of 1's in the codeword $x(j)$, i.e., $|\boldsymbol{x}(j)| \triangleq \sum\limits_{i=1}^{N} x_i(j) = wN$, is called the *weight* of $\boldsymbol{x}(j)$, $j \in [t]$ and parameter w, $0 < w < 1$, is the *relative weight*.

One can see that the binary code X can be associated with N tests. A column $\boldsymbol{x}(j)$ corresponds to the j-th sample; a row $\boldsymbol{x}_i$ corresponds to the i-th test. Let $\mathbf{u} \bigvee \mathbf{v}$ denote the disjunctive sum of binary columns $\mathbf{u}, \mathbf{v} \in \{0,1\}^N$. For any subset $\mathcal{S} \subset [t]$ define the binary vector

$$r(X, \mathcal{S}) = \bigvee_{j \in \mathcal{S}} \boldsymbol{x}(j),$$

which later will be called the *outcome vector*.

By $\mathcal{S}_{un}$, $|\mathcal{S}_{un}| \leqslant s$, denote an unknown set of defective elements. Suppose there is a p-stage group testing strategy $\mathfrak{S}$ which finds up to s defective elements. It means that for any $\mathcal{S}_{un} \subset [t]$, $|\mathcal{S}_{un}| \leqslant s$, according to $\mathfrak{S}$:

1) we are given with a code X_1 assigned for the first stage of group testing;
2) we can design a code X_{i+1} for the $i+1$-th stage of group testing, based on the outcome vectors of the previous stages $r(X_1, \mathcal{S}_{un})$, $r(X_2, \mathcal{S}_{un})$, $\ldots$, $r(X_i, \mathcal{S}_{un})$;
3) we can identify all defective elements $\mathcal{S}_{un}$ using $r(X_1, \mathcal{S}_{un})$, $r(X_2, \mathcal{S}_{un})$, $\ldots$, $r(X_p, \mathcal{S}_{un})$.

Let N_i be the number of test used on the i-th stage and

$$N_T(\mathfrak{S}) = \sum_{i=1}^{p} N_i$$

be the maximal total number of tests used for the strategy $\mathfrak{S}$. We define $N_p(t, s)$ to be the minimal worst-case total number of tests needed for group testing for t elements, up to s defective elements, and at most p stages.

III. Hypergraph approach to the search of defective elements

Let us introduce a hypergraph approach to searching defective elements. Suppose a set of vertices V is associated with the set of samples $[t]$, i.e. $V = \{1, 2 \ldots, t\}$.

First stage: Let X_1 be the code corresponding to the first stage of group testing. For the outcome vector $r = r(X_1, \mathcal{S}_{un})$ let $E(r, s)$ be the set of subsets of $\mathcal{S} \subset V$ of size at most s such that $r(X, \mathcal{S}) = r(X, \mathcal{S}_{un})$. So, the pair $(V, E(r, s))$ forms the hypergraph H. We will call two vertices *adjacent* if they are included in some hyperedge of H. Suppose there exist a *good* vertex colouring of H in k colours, i.e., assignment of colours to vertices of H such that no two adjacent vertices share the same colour. By $V_i \subset V$, $1 \leqslant i \leqslant k$, denote vertices corresponding to the i-th colour. One can see that all these sets are pairwise disjoint.

Second stage:

Now we can perform k tests to check which of monochromatic sets V_i contain a defective element. Here we find the cardinality of set $\mathcal{S}_{un}$ and $|\mathcal{S}_{un}|$ sets $\{V_{i_1}, \ldots, V_{i_{|\mathcal{S}_{un}|}}\}$, each of which contains exactly one defective element.

Third stage:

Carrying out $\lceil \log_2 |V_{i_1}| \rceil$ tests we can find a vertex v, corresponding to the defective element, in the suspicious set V_{i_1}. Observe that actually by performing $\sum\limits_{j=1}^{\mathcal{S}_{un}} \lceil \log_2 |V_{i_j}| \rceil$ tests we could identify all defective elements $\mathcal{S}_{un}$ on this stage.

Fourth stage:

Consider all hyperedges $e \in E(r, s)$, such that e contains the found vertex v and consists of vertices of $v \cup V_{i_2} \cup \ldots \cup V_{i_{|\mathcal{S}_{un}|}}$. At this stage we know that the unknown set of defective elements coincides with one of this hyperedges. To check if the hyperedge e is the set of defective elements we need to test the set $[t] \backslash e$. Hence, the number of test at fourth stage is equal to the degree of the vertex v.

IV. Optimal search of 2 defective elements

Now we consider a specific construction of 4-stage group testing. Then we upper bound number of tests N_i at each stage.

First stage:

Let $C = \{0, 1, \ldots q-1\}^{\hat{N}}$ be the q-ary code, consisting of all q-ary words of length $\hat{N}$ and having size $t = q^{\hat{N}}$. Let D be the set of all binary words with length N' such that the weight of each codeword is fixed and equals wN', $0 < w < 1$, and the size of D is at least q, i.e., $q \leqslant \binom{N'}{wN'}$. On the first stage we use the concatenated binary code X_1 of length $N_1 = \hat{N} \cdot N'$ and size $t = q^{\hat{N}}$, where the inner code is D, and the outer code is C. We will say X_1 consists of $\hat{N}$ layers. Observe that we can split up the outcome vector $r(X_1, \mathcal{S}_{un})$ into $\hat{N}$ subvectors of lengths N'. So let $r_j(X_1, \mathcal{S}_{un})$ correspond to $r(X_1, \mathcal{S}_{un})$ restricted to the j-th layer. Let w_j, $j \in [\hat{N}]$, be the relative weight of $r_j(X_1, \mathcal{S}_{un})$, i.e., $|r_j(X_1, \mathcal{S}_{un})| = w_j N'$ is the weight of the j-th subvector of $r(X_1, \mathcal{S}_{un})$.

If $w_j = w$ for all $j \in [\hat{N}]$, then we can say that $\mathcal{S}_{un}$ consists of 1 element and easily find it.

If there are at least two defective elements, then suppose for simplicity that $\mathcal{S}_{un} = \{1, 2\}$. The two corresponding codewords of C are c_1 and c_2. There exists a coordinate $i, 1 \leqslant i \leqslant \hat{N}$, in which they differs, i.e., $c_1(i) \neq c_2(i)$. Notice that the relative weight w_i is bigger than w. For any $i \in [\hat{N}]$ such that $w_i > w$, we can colour all vertices V in q colours, where the colour of j-th vertex is determined by the corresponding q-nary symbol $c_i(j)$ of code C. One can check that such a colouring is a good vertex colouring.

Second stage:

We perform q tests to find which coloured group contain 1 defective element.

Third stage:

Let us upper bound the size $\hat{t}$ of one of such suspicious group:

$$\hat{t} \leqslant \binom{w_1 N'}{wN'} \cdot \ldots \cdot \binom{w_{\hat{N}} N'}{wN'}.$$

In order to find one defective element in the group we may perform $\lceil \log_2 \hat{t} \rceil$ tests.

Fourth stage:

On the final step, we have to bound the degree of the found vertex $v \in V$ in the graph. The degree $\deg(v)$ is bounded as

$$\deg(v) \leqslant \binom{wN'}{(2w - w_1)N'} \cdot \ldots \cdot \binom{wN'}{(2w - w_{\hat{N}})N'}.$$

We know that the second defective element corresponds to one of the adjacent to v vertices. Therefore, to identify it we have to make $\lceil \log_2 \deg(v) \rceil$ tests.

Letting $\hat{N}$ tends to infinity we obtain the following bound:

$$\frac{N_T}{\log_2 t} \leqslant \frac{\hat{N} \cdot N' + \max\limits_{w_i} \left(\log_2 \hat{t} + \log_2 \deg(v) \right)}{(1 + o(1)) \hat{N} \log_2 \binom{N'}{wN'}}.$$

It is easy to see that in the worst case all values of w_i are the same, hence

$$\frac{N_T}{\log_2 t} \leqslant \frac{\hat{N} \cdot N' + \max\limits_{w'} \log_2 \left(\binom{w'N'}{wN'} \binom{wN'}{(2w - w')N'} \right)}{(1 + o(1)) \hat{N} \log_2 \binom{N'}{wN'}}. \quad (1)$$

By choosing the optimal parameter w, $wN' \in \mathcal{Z}$, we can minimize the number of tests for fixed value of q.

If $q \to \infty$, then we can rewrite (1) as follows

$$\frac{N_T}{\log_2 t} \leqslant \sup_{w \leqslant w' \leqslant \min(1, 2w)} f(w, w')(1 + o(1)),$$

where

$$f(w, w') = \frac{1 + w' \cdot h\left(\frac{w}{w'}\right) + w \cdot h\left(\frac{2w - w'}{w}\right)}{h(w)}.$$

Finally, we obtain the following bound

$$\frac{N_T}{\log_2 t} \leqslant \inf_{0 < w < 1} \sup_{w \leqslant w' \leqslant \min(1, 2w)} f(w, w'). \quad (2)$$

Let us find extreme value on y of

$$g(x, y) = y \cdot h(x/y) + x \cdot h((2x - y)/x).$$

$$\frac{dg(x, y)}{dy} = h(x/y) - \frac{x}{y} h'(x/y) - h'((2x - y)/x) =$$
$$= \log_2 y - 2 \log_2(y - x) + \log_2(2x - y).$$

This implies

$$(y - x)^2 - 2xy + y^2 = 0.$$

Hence, if we take $w = 1/(2 + \sqrt{2})$, then the supremum in (2) is attained at $w' = 1/2$, and achievable number of tests by 4-stage group testing procedure is $2 \log_2 t(1 + o(1))$.

Observe that for fixed q we can obtain only finite amount of rational values for parameter w, we could not provide an explicit construction of searching procedure with $2 \log_2 t(1 + o(1))$ tests. But if $q \to \infty$, then the minimal number of test N_T tends to $2 \log_2 t(1 + o(1))$.

V. Search of s defective elements

Here we will use combination of the first two stages of the previous algorithm. Suppose the number of defective elements is at most s. In fact, we don't use this fact in our algorithm. Let $C = \{0, 1, \ldots q - 1\}^{\hat{N}}$, $|C| = q^{\hat{N}}$, be the set of all q-ary words of length $\hat{N}$. Let D be the set of all binary words of length N' such that the weight of each codeword is fixed and equals $N'/2$, and the size of $|D|$ is at least q. On the first stage we use the concatenated binary code X of length $\hat{N} \cdot N'$ and size $q^{\hat{N}}$, where the inner code is D, and the outer code is C. Notice that if the number of defective elements is one, then we are assumed to identify defective element basing on the outcome vector $r_1(X, \mathcal{S}_{un})$. If this number is at least two than there exists a coordinate i in which the corresponding q-ary words differs. It means that the outcome vector restricted on the i-th coordinate has the relative weight bigger than w. Split up all vertices V in q groups according to q-ary symbol in the i-th coordinate. On the next stage we perform q tests and find which groups contain at least one defective element. Then we will deal with each such group separately. If we could not divide a group into subgroups, then we easily find the unique defective element in this group. In the worst case scenario, we have to perform $2s - 1$ group testing stages, and the total number N_T of tests is upper bounded by the sum of number of tests, which served for separating defective elements into disjoint groups, and number of tests, which used for finding 1 defective element among different groups. Thus, we have

$$N_T \leqslant (s - 1)\hat{N} \cdot N' + s\hat{N} \cdot N' + q(s - 1).$$

In asymptotic regime, the total number of tests

$$N_T \leqslant (2s - 1) \log_2 t(1 + o(1)).$$

VI. Tables of numerical values

In this section we apply our 4 stage procedure from Sect. IV to specific values of t. Let us calculate the numbers of tests at each stage more properly. Recall that the number of tests at the first stage N_1 is equal to $\hat{N} \cdot N'$. In the case $|S_{un}| = 1$ we can find the defective element based only on the outcome of the first stage of group testing.

Let $W = wN'$ and $W_i = w_i N'$. If our colouring is determined by symbols from i-th layer of the code X_1, then the actual number of suspicious sets equals $\binom{W_i}{W}$. Since we know the exact number of defective elements it is sufficient to use $\binom{W_i}{W} - 1$ tests. Also note that we need to determine only one set with a defective element, therefore we can make $\binom{W_i}{W} - 2$ tests at the second stage.

The total number of elements in all suspicious groups is equal to

$$\binom{W_1}{W} \cdot \ldots \cdot \binom{W_{\hat{N}}}{W}.$$

One can see that the numbers of elements of each colour are the same. Hence the cardinality $\hat{t}$ of one suspicious set is equal to

$$\hat{t} = \binom{W_1}{W} \cdot \ldots \cdot \binom{W_{\hat{N}}}{W} \Big/ \binom{W_i}{W}$$

So, at the third stage we need to perform $\lceil \log_2 \hat{t} \rceil$ tests. Before the last stage we have already known one of the defective elements. At each layer $j \neq i$ we have $\binom{W}{2W-W_j}$ ways to choose q-nary coordinate of the second defective element, but at the i-th layer we have only 2 suspicious coordinates left in the worst case. Therefore, the number of tests at the fourth stage is at most

$$\left\lceil \log_2 \left(2 \frac{\binom{W}{2W-W_1} \cdot \ldots \cdot \binom{W}{2W-W_{\hat{N}}}}{\binom{W}{2W-W_i}} \right) \right\rceil.$$

We provide three tables with optimal values of tests for small $t \leqslant 1000$, for $t = 10^k$, $3 \leqslant k \leqslant 18$, and for some values of t, for which we have a small ratio of the number of tests to $\log_2 t$.

TABLE I
NUMBER OF TESTS FOR $t \leqslant 1000$

t	tests	t	tests	t	tests
8-9	8	29-36	14	126-256	20
10-16	10	37-64	15	257-441	22
17-27	12	65-81	16	442-784	24
28	13	82-125	18	785-1000	25

In Table II and Table III we also present the information theory bound $\underline{N}$, which is the minimum integer such that

$$2^{\underline{N}} \geqslant 1 + \binom{t}{1} + \binom{t}{2}.$$

TABLE II
NUMBER OF TESTS FOR $t = 10^k$

$t = q^{N_1}$	tests	N	tests / $\log_2 t$
10^3	26	19	2.609
10^4	33	26	2.483
10^5	41	33	2.468
10^6	48	39	2.408
10^7	56	46	2.408
10^8	64	53	2.408
10^9	71	59	2.375
10^{10}	79	66	2.378
10^{11}	86	73	2.354
10^{12}	94	79	2.358
10^{13}	102	86	2.362
10^{14}	109	93	2.344
10^{15}	117	99	2.348
10^{16}	124	106	2.333
10^{17}	132	112	2.337
10^{18}	139	119	2.325

TABLE III
NUMBER OF TESTS FOR t WITH SMALL RATIO TESTS / $\log_2 t$

$q^{N_1} = t$	tests	N	tests / $\log_2 t$
$28^2 = 784$	24	19	2.496
$15^3 = 3375$	29	23	2.474
$21^3 = 9261$	32	26	2.428
$28^3 = 21952$	35	28	2.427
$15^4 = 50625$	37	31	2.368
$21^4 = 194481$	41	35	2.334
$21^5 = 4084101$	51	43	2.322
$15^6 = 11390625$	54	46	2.304
$21^6 = 85766121$	60	52	2.277
$21^9 = 794280046581$	89	79	2.251
$21^{11} \approx 3.5 \cdot 10^{14}$	108	96	2.235

REFERENCES

[1] *Du D.Z.*, *Hwang F.K.*, Combinatorial Group Testing and Its Applications, 2nd ed., *Series on Applied Mathematics*, vol. 12, 2000.
[2] *Cicalese F.* Fault-Tolerant Search Algorithms. *Monographs in Theoretical Computer ScienceAn EATCS Series*, Springer-Verlag, vol. 15, 2013.
[3] *Damaschke P.*, *Sheikh Muhammad A.*, *Triesch E.*, Two new perspectives on multi-stage group testing, *Algorithmica*, vol. 67, no. 3, pp. 324-354, 2013.
[4] *Mzard M.*, *Toninelli, C.*, Group testing with random pools: Optimal two-stage algorithms, *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1736-1745, (2011).
[5] *D'yachkov A.G.*, *Rykov V.V.*, Bounds on the Length of Disjunctive Codes, *// Problems of Information Transmission*, vol. 18. no 3. pp. 166-171, 1982.
[6] *D'yachkov A.G.*, *Vorobyev I.V.*, *Polyanskii N.A.*, *Shchukin V.Yu.*, Bounds on the Rate of Disjunctive Codes, *Problems of Information Transmission*, vol. 50, no. 1, pp. 27-56, 2014.
[7] *Rashad A.M.*, Random Coding Bounds on the Rate for List-Decoding Superimposed Codes. *Problems of Control and Inform. Theory.*, vol. 19, no 2, pp. 141-149, 1990.
[8] *D'yachkov A.G.*, Lectures on Designing Screening Experiments, *Lecture Note Series 10*, Combinatorial and Computational Mathematics Center, Pohang University of Science and Technology (POSTECH), Korea Republic, Feb. 2003, (survey, 112 pages).
[9] *De Bonis A.*, *Gasieniec L.*, *Vaccaro U.*, Optimal two-stage algorithms for group testing problems, *SIAM J. Comp.*, vol. 34, no. 5 pp. 1253-1270, 2005.
[10] *Damaschke P.*, *Sheikh Muhammad A.*, *Wiener G.* Strict group testing and the set basis problem. *Journal of Combinatorial Theory, Series A*, vol. 126, pp. 70-91, August 2014.
[11] *Macula A.J.*, *Reuter G.R.*, Simplified searching for two defects, *Journal of statistical planning and inference*, vol. 66, no. 1, pp 77-82, 1998.
[12] *Deppe C.*, *Lebedev V.S.*, Group testing problem with two defects, *Problems of Information Transmission*, vol. 49, no. 4, pp. 375-381, 2013.