Group Testing with Unreliable Elements

Arya Mazumdar and Soheil Mohajer
Department of ECE
University of Minnesota– Twin Cities
Minneapolis, MN 55455
email: {arya,soheil}@umn.edu

***Abstract*—We consider a generalization of the well-known nonadaptive group testing problem. In our generalization, tests or measurements are performed in the presence of a number of unknown but fixed *pretenders*, that will, with certain probability be active (pretend as being defective) during any test. We show some simple extensions of the achievability results of group testing tailored for this case.**

I. Introduction

This note is a companion paper of an invited presentation by the first author in the Allerton conference, and demonstrates the power of simple random choice in the group testing problem.

Combinatorial group testing is an old and well-studied problem. There is a set of n elements among which at most t are *defective*. The smallest number of yes/no tests, that identify the defective items is $\log \sum_{i=0}^{t} \binom{n}{i} \approx t \log n$. The main objective of a search problem is to identify the defective configuration using the number of tests that is as close to this minimum as possible.

In the group testing problem, a *group* of elements are tested together, and if this particular group contains any defective element, the test result is positive. Based on the test results of this kind one *identifies* (with an efficient algorithm) the defective configuration using the smallest possible number of tests. The collection of tests is called a *group testing scheme.* Group testing schemes can be adaptive (see, e.g., [4]), where the design of one test may depend on the results of preceding tests, or non-adaptive, the interest of this paper, where all the tests are designed together. If the number of designed tests is m, then a non-adaptive group testing scheme is equivalent to the design of a binary *test matrix* of size $m \times n$ where the (i, j)th entry is 1 if the ith test includes the jth element; it is 0 otherwise. As the test results, we see the Boolean OR of the columns corresponding to the defective entries. It is known that the matrix must have *disjunct* property (defined later) to be a good group testing matrix. The best known lower bound on the number of required tests in terms of the number of elements n and the maximum number of defective elements t is given by [6], $m = \Omega\left(\frac{t^2}{\log t} \log n\right)$. The existence of non-adaptive group testing schemes with $m = O(t^2 \log n)$ is also known for some time [4], [9], [14].

Generally, constructing groups testing schemes is a difficult problem. As one of the ways of attacking it, it has been suggested to construct schemes that permit a small probability of error (either missing defectives, or allowing false positives). Such schemes were considered under the name of *weakly separated designs* in [11], [12], [15]. With this relaxation it is be possible to reduce the number of tests to be proportional to $t \log n$ [15]. Related notions and constructions are explored in [3], [8], [13].

A. Problem Statement

In this paper we study a generalization of the nonadaptive group testing problem: finding a sparse set of (at most) t defective items in a large population of size n with small number of tests, when there are fixed but unknown *pretenders* in the system. A pretender item is in fact non-defective, but sometimes it happens to pretend like a defective item. We denote the role of a pretender i to be active (pretend like a defective) or passive (act as a healthy item) in the j-th test by X_i^j, and assume X_i^j's are independent across the pretenders and across the tests, and they all distributed as $p = \mathbb{P}[X_i^j = 1]$.

More precisely we have a set of n items indexed by an integer in $[n]$, which includes a set of $|T| \leq t$ defectives. There is another set $S \subset [n]$, with $|S| = \tau$ and $S \cap T = \emptyset$, includes defective items. Each test j corresponds to picking a subset $A_j \subset [n]$ and ask whether A_j includes any defectives. The answer to this question is given by

$$Y_j = \begin{cases} 1 & \text{if } |A_j \cap T| \geq 1, \\ 1 & \text{if } |A_j \cap S| \geq 1 \text{ and } \exists i \in A_j \cap S : X_i^j = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

In detection phase, we have to find $\tilde{T}$, the set of detected defectives, based on the observations:

$$\tilde{T} = g(Y_1, Y_2, \ldots, Y_m),$$

where m is the number of tests.

It is clear that the behavior of the system is random due to the stochastic behavior of the pretenders. Therefore, any desired results would be a stochastic statement. Under the model explained above, we are interested in the following two problems:

(i) **Perfect recovery:** For a given $\delta > 0$, how many measurements do we need to be able to guarantee that

$$\mathbb{P}[\tilde{T} \neq T] < \delta,$$

i.e., to be able to find the set of defective items with probability at least $1 - \delta$?

(ii) **Majority recovery:** For a given $\delta > 0$ and $\epsilon > 0$, how many measurements do we need to guarantee the total

number of false positives and false negatives does not exceed 2ϵ fraction of the size of defective items:

$$\mathbb{P}[|\tilde{T} \setminus T| + |T \setminus \tilde{T}| \leq 2\epsilon|T|] \geq 1 - \delta.$$

The following notations have been adopted in this paper

n	total number of elements
m	number of tests
A	the $m \times n$ test matrix
t	maximum number of defectives
τ	number of unreliable elements or pre-tenders
p	pretend probability or probability of any pretender being active
$[n]$	set integers $\{1, 2, \ldots, n\}$

Remark 1 (Difference with noisy group testing)**.** *The generalization that we consider above is different from the usual noisy group testing where the output of the test travels through a noisy channel (such as binary symmetric channel or Z-channel) before they are seen [1], [2], [5], [10].*

II. Achievability via random sampling

We have the following result for majority recovery.

Theorem 1. *Suppose there exist at most t defective items among a total n items. There exists an $m \times n$ test-matrix A, such that even in the presence of an unreliable set with size τ and pretend probability p, $1 - \epsilon$ proportion of all defectives item can be identified with probability $1 - o(1)$, $\epsilon > 0$, as long as,*

$$m \geq \frac{3t \ln n}{1 - e^{-\frac{t\epsilon}{t+p\tau}}}. \tag{2}$$

In particular, when $p\tau \gg t$, a sufficient condition is $m \geq \frac{3}{\epsilon}(t + p\tau) \log n$.

We will need the following simple lemma.

Lemma 2. $\sum_{i=0}^{t} \binom{n}{i} \leq \frac{(ne)^t}{t^{t-1}}$.

Proof of Thm. 1: Suppose, $T \subset [n]$ is the set of defectives. The recovery will be successful as long as we return a set T' such that $r \equiv |T \cap T'| \geq (1 - \epsilon)|T|$.

Our object of interest is the probability of error P_e, the probability of existence of a pair T and T', $|T|, |T'| \leq t$ that A fails to distinguish between, where $r < (1 - \epsilon)|T|$.

We assume the matrix A is chosen randomly from the ensemble of all $m \times n$ matrix in the following way. Each entry of A is 1 with probability $q \equiv \frac{1}{t+p\tau}$, and it is zero with the remaining probability. In other words, in each test we include an item with probability q. We will show that the probability of error P_e in this case is $o(1)$ which will imply existence of a matrix A that achieves probability of error $o(1)$ (similar argument appears in random coding, c.f. [7, Sec. 5.5]).

Assume the unknown set of unreliable elements is $S \in [n]$, with $|S| = \tau$. Recall that, in our setting $S \cap (T \cup T') = \emptyset$. A given test will be able to distinguish between T and T' if in that test one of the following two events occur:

- No item from the set S that is active is included and no item from set T is included and at least one item from the set $T' \setminus T$ is included.
- No item from the set S that is active is included and no item from set T' is included and at least one item from the set $T \setminus T'$ is included.

The probability that any one test will be successful to distinguish between T and T' is therefore (here we are taking the sizes of T and T' exactly equal to t and not less than equal to, which is permissible without much loss of generality),

$$\begin{aligned}
&2(1-pq)^{\tau}(1-q)^{t}(1-(1-q)^{t-r}) \\
&= 2\Big(1 - \frac{p}{t+p\tau}\Big)^{\tau}\Big(1 - \frac{1}{t+p\tau}\Big)^{t}\Big(1 - \Big(1 - \frac{1}{t+p\tau}\Big)^{t-r}\Big) \\
&\geq 2 \cdot 3^{-\frac{p\tau}{t+p\tau}} 3^{-\frac{t}{t+p\tau}} \Big(1 - e^{-\frac{t-r}{t+p\tau}}\Big) \\
&\geq \frac{2}{3}\Big(1 - e^{-\frac{\epsilon t}{t+p\tau}}\Big),
\end{aligned}$$

where in the second line we have used inequalities $1-x \leq e^{-x}$ for all x and $1 - x \geq 3^{-x}$ for any $x \leq 0.17$ (which is true for any $t \geq 6$). We have also used the fact that $r < (1 - \epsilon)|T| \leq (1 - \epsilon)t$.

Hence the probability that A fails to distinguish between T and T' is

$$\Big(1 - \frac{2}{3}\Big(1 - e^{-\frac{\epsilon t}{t+p\tau}}\Big)\Big)^{m}.$$

Therefore, for this ensemble, if m is given by Eqn. (2),

$$\begin{aligned}
P_e &\leq \Big(\sum_{i=0}^{t} \binom{n}{i}\Big)^2 \Big(1 - \frac{2}{3}\Big(1 - e^{-\frac{\epsilon t}{t+p\tau}}\Big)\Big)^{m} \\
&\leq \Big(\frac{(ne)^t}{t^{t-1}}\Big)^2 e^{-\frac{2m}{3}\Big(1 - e^{-\frac{\epsilon t}{t+p\tau}}\Big)} \\
&\leq \Big(\frac{e^t}{t^{t-1}}\Big)^2 \to 0,
\end{aligned}$$

as t grows. This proves the theorem. Notice that, for x very small $e^{-x} \sim 1 - x$. Hence for $p\tau \gg t$, we have $m \geq \frac{3}{\epsilon}(t + p\tau) \ln n$. ∎

The above result can be modified slightly to cover the case of perfect recovery, albeit with a price.

Theorem 3. *There exists an $m \times n$ test-matrix A that, even in the presence of τ pretender items with pretend probability p, can identify the exact set of defective items with an overwhelming probability, provided that*

$$m \geq 3t(t + p\tau) \ln n. \tag{3}$$

Proof: As before, let the test matrix A be an $m \times n$ binary matrix with i.i.d. entries, where each entry is 1 with probability $q \equiv \frac{1}{t+p\tau}$.

Let S be the set of pretenders, T be the true set of defectives, and T' be an alternative set of size T. Define $r \equiv |T \cap T'|$. A test (a single row of the measurement matrix) can distinguish between T and T' iff $Y = 0$ for one and $Y = 1$ for the other one. That happens only if

- Every elements for S in the pool is negative, and pool does not include any item from T, but at least one item from $T' \setminus T$; or
- Every elements for S in the pool is negative, and pool does not include any item from T', but at least one item from $T \setminus T'$;

Therefore, the probability for a test to be able to distinguish between T and T' is given by

$$\begin{aligned}
&2(1-pq)^{\tau}(1-q)^{\tau}(1-(1-q)^{t-r}) \\
&= 2\Big(1-\frac{p}{t+p\tau}\Big)^{\tau}\Big(1-\frac{1}{t+p\tau}\Big)^{t}\Big(1-\Big(1-\frac{1}{t+p\tau}\Big)^{t-r}\Big) \\
&\geq 2\cdot 3^{-\frac{p\tau}{t+p\tau}}3^{-\frac{t}{t+p\tau}}\Big(1-\Big(1-\frac{1}{t+p\tau}\Big)^{t-(t-1)}\Big) \\
&\geq \frac{2}{3(t+p\tau)},
\end{aligned}$$

where we have used $r \leq t-1$. Assuming the measurement matrix has m independent rows, the probability of failure in distinguishing between T and T' will be bounded by

$$\Big(1-\frac{2}{3(t+p\tau)}\Big)\Big)^{m} \leq e^{-\frac{2m}{3(t+p\tau)}}.$$

Taking union bound over all choices of T, T' we have the probability of error bounded by,

$$\Big(\frac{(ne)^t}{t^{t-1}}\Big)^2 e^{-\frac{2m}{3(t+p\tau)}} \leq \Big(\frac{(e)^t}{t^{t-1}}\Big)^2 e^{-\frac{2m}{3(t+p\tau)}+2t\ln n} \to 0,$$

where we have substituted the value of m from Eqn. (3). ■

REFERENCES

[1] G. K. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *Information Theory, IEEE Transactions on*, 58(3):1880–1901, 2012.

[2] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi. Grotesque: Noisy group testing (quick and efficient). *arXiv preprint arXiv:1307.2811*, 2013.

[3] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1837–1841. IEEE, 2012.

[4] D. Z. Du and F. Hwang. *Combinatorial group testing and its applications, 2nd Ed.* World Scientific, 2000.

[5] A. Dyachkov, V. Rykov, and A. Rashad. Superimposed distance codes. *problems of control and information theory-problemy upravleniya i teorii informatsii*, 18(4):237–250, 1989.

[6] A. G. D'yachkov and V. V. Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.

[7] R. G. Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[8] A. C. Gilbert, B. Hemenway, A. Rudra, M. J. Strauss, and M. Wootters. Recovering simple signals. In *Information Theory and Applications Workshop (ITA), 2012*, pages 382–391. IEEE, 2012.

[9] F. Hwang and V. Sós. Non-adaptive hypergeometric group testing. *Studia Sci. Math. Hungar*, 22:257–263, 1987.

[10] E. Knill, W. J. Bruno, and D. C. Torney. Non-adaptive group testing in the presence of errors. *Discrete applied mathematics*, 88(1):261–290, 1998.

[11] M. Malyutov. Mathematical models and results in the theory of screening experiments. *Problems in Cybernatics 35, Moscow*, pages 5–69, 1977.

[12] M. B. Malyutov. The separating property of random matrices. *Mathematical Notes*, 23(1):84–91, 1978.

[13] A. Mazumdar. On almost disjunct matrices for group testing. In *Algorithms and Computation*, pages 649–658. Springer, 2012.

[14] E. Porat and A. Rothschild. Explicit non-adaptive combinatorial group testing schemes. In *Automata, Languages and Programming*, pages 748–759. Springer, 2008.

[15] A. Zhigljavsky. Probabilistic existence theorems in group testing. *Journal of statistical planning and inference*, 115(1):1–43, 2003.