LETTER

Density of Pooling Matrices vs. Sparsity of Signals for Group Testing Problems*

Jin-Taek SEONG[†a)], ***Member***

SUMMARY In this paper, we consider a group testing (GT) problem. We derive a lower bound on the probability of error for successful decoding of defected binary signals. To this end, we exploit Fano's inequality theorem in the information theory. We show that the probability of error is bounded as an entropy function, a density of a pooling matrix and a sparsity of a binary signal. We evaluate that for decoding of highly sparse signals, the pooling matrix is required to be dense. Conversely, if dense signals are needed to decode, the sparse pooling matrix should be designed to achieve the small probability of error.

key words: *group testing, lower bound, pooling matrix, sparsity*

1. Introduction

Group Testing (GT) was first introduced in 1943 by Dorfman [1], which is a class of the combinatorial problems. Since then, a lot of algorithms for solving this combinatorial problem were developed. And recently, GT has extended to probabilistic approaches. In 2006, Compressed Sensing (CS) introduced by Donoho [2] is a linear inverse problem which dealt with sparse signals. CS is a variant of GT problems [3]. The advent of CS has led to the rediscovery of the traditional GT. So far, a number of theoretic results for GT problems have been presented over the past decade [4]–[7].

First, GT was found in a report written by Dorfman [1]. During World War II, the government of the United State embarked a project to find out all syphilitic men out of a number of soldiers. At the time, individual syphilis testing was expensive and inefficient as well as long took to inspect all soldiers individually. Suppose that the number of syphilitic men is very small out of all soldiers. In fact, it makes sense. Since most soldiers were not infected, the result of the sample test combined with two or three soldier's blood samples would be negative. Grouping blood samples from a few soldiers not only reduces the number of tests, but also enables fast testing for efficiency. This is the background in which GT has emerged at first. A number of remarkable results for GT problems have been presented and improved by the basic idea of this GT.

GT is briefly described as follows. The core question of GT problems is that when K samples out of N samples have defects, how many measurements are needed to find all defective samples. This is like a question studied in CS. The difference between both of them is an operation method. In other words, while CS deals with linear systems in real or complex numbers, but GT is about logical systems, e.g., AND and OR. Except for the logical or real-valued operation, both systems are much similar.

Let us consider a syphilis testing as follows. Blood samples from several soldiers are collected and measured whether the corresponding result is positive or negative. And then, a subset of blood samples from soldiers mixed for a syphilis testing is called a pool. If the result of this pool is positive, it is observed that at least one of the soldiers is infected with syphilis. On the other hand, if the result is negative, it means that all the soldiers in this pool were not infected with syphilis. Such blood test is called the GT [3]. In other words, GT is a class of logical systems with operation of AND and OR.

In this paper, we aim to analyze performance of GT problems with respect to density of pooling matrices and sparsity of binary signals. The density of the pooling matrix determines the size of pools participating in a single group testing. This is a constraint condition to design a pool. The performance of GT problems depends on how much sparse of pooling matrices with different sparsity of binary signals. The remain part of this paper will provide a low bound on the number of tests using Fano's inequality. And we show how much relationship between density of the pooling matrix and sparsity of the binary signal for a necessary condition on successful decoding of all defective samples.

2. Group Testing

2.1 Problem Statement

In this section, we will more clearly define the GT problem with mathematical expressions. Let $\mathbf{x} \in \{0,1\}^N$ be the binary signal with size of N. If the ith entry of the binary signal $\mathbf{x}$ is defective, it is represented as $x_i = 1$. Otherwise, $x_i = 0$. In this paper, each element of the binary signal $\mathbf{x}$ is identically independent distributed (i.i.d.) as the following Bernourlli distribution,

$$\Pr\{x_i = \theta\} = \begin{cases} 1-\delta & \text{if } \theta = 0, \\ \delta & \text{if } \theta = 1, \end{cases} \tag{1}$$

where $\delta := K/N$ is the sparsity ratio of the binary signal $\mathbf{x}$

Manuscript received September 17, 2018.
Manuscript revised December 18, 2018.
Manuscript publicized February 4, 2019.

[†]The author is with the Department of Convergence Software, Mokpo National University, Republic of Korea.

*This paper was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (NRF-2017R1C1B5075823).

a) E-mail: jtseong@mokpo.ac.kr

DOI: 10.1587/transinf.2018EDL8200

and θ is a dummy variable with binary field, 0 or 1. The sparsity ratio of the binary signal $\mathbf{x}$ is assumed as $0 < \delta < 0.5$. This can be called as a defective rate.

Next let $\mathbf{A} \in \{0,1\}^{M\times N}$ be the pooling matrix with M rows and N columns. If the ith entry of the binary signal $\mathbf{x}$ is pooled in the jth test, the corresponding element of the pooling matrix is expressed as $A_{ji} = 1$. Otherwise, $A_{ji} = 0$. In other words, the pooling matrix has a role of collection of entries of the binary signal $\mathbf{x}$ participating in a single group test. Each element of the pooling matrix $\mathbf{A}$ with i.i.d. is defined as follows,

$$\Pr\{A_{ji} = \theta\} = \begin{cases} 1-\gamma & \text{if } \theta = 0, \\ \gamma & \text{if } \theta = 1, \end{cases} \tag{2}$$

where γ is the density ratio of the pooling matrix. If the density ratio is large, it means that the probability that the element of the pooling matrix has 1 is high. That is, most of the elements of the vector $\mathbf{x}$ is pooled in each test. Note that to collect many elements of the binary signal $\mathbf{x}$ is costly. In order to perform efficient testing, we need to be small density of the pooling matrix.

Using the binary signal $\mathbf{x}$ and the pooling matrix $\mathbf{A}$, the mathematical expression of the GT problem follows as

$$\mathbf{y} = \mathbf{A} \odot \mathbf{x} \tag{3}$$

where $\mathbf{y}$ is the testing signal and the symbol $\odot$ denote the *element-wise* logical operation. The following example more clearly describes the mathematical expression in the GT problem,

$$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{4}$$

where the first element of the left hand side in (4) is 1 because $\begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \odot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ = (0 AND 0) OR (1 AND 0) OR (1 AND 1) = 1. Other elements are obtained by using the same logical operation. The result for a pool participating one or more defective samples as shown in the example above is positive. Conversely, if all the elements participating in the pool are all negative, the result is negative.

Next, we describe a decoding scheme to find defective samples. We assume that there is a decoder to estimate all defective samples. This decoder finds a binary signal $\mathbf{x}$ using the given information of the pooling matrix $\mathbf{A}$ and the testing signal $\mathbf{y}$. In this paper, we assume that the decoder uses an estimated function Ψ which determines a binary signal $\mathbf{x}$. Using this function Ψ, we obtain an estimated signal $\hat{\mathbf{x}}$ of the binary signal $\mathbf{x}$ from $\hat{\mathbf{x}} = \Psi(\mathbf{A}, \mathbf{y})$. The probability of error P_e for this decoder is defined as follows,

$$P_e = \Pr\{\Psi(\mathbf{A}, \mathbf{y}) \neq \mathbf{x}\} \tag{5}$$

The above decoder is aimed at minimizing the probability of error. Assume the decoder is used so that the probability of error is as small as possible.

2.2 Related to Bounds for Performance Evaluation

In this section, we discuss bounds on performance of GT problems known so far. In [3], the answer to the critical questions of GT problems was reported by analyzing a lower bound on the performance of the probability of error. If we use an algorithm of the GT problem with that the probability of error is 0, the minimum number needed to find K defect samples out of N samples is

$$M \geq \log_2 \binom{N}{K}. \tag{6}$$

This is bounded by the fact that the sample space is split into two disjoint subsets for each test which corresponds to one of two feasible sets. The theoretic lower bound as called the information bound in even small GT problems is unachievable [4]. Recently, in [7], the authors showed that the individual testing scheme for more than 0.3471 of the sparsity ratio is optimal. It means that it is impossible for the GT problems to decode the range out of highly dense signals.

The upper bound on the probability of success is obtained with the number of tests. One algorithm of the GT problem satisfies the following probability of success P_s with respect to the number of tests [4],

$$P_s \leq M \log_2 \binom{N}{K}^{-1}, \tag{7}$$

This is the same of the upper bound for the adaptive group testing schemes.

3. A New Lower Bound

3.1 Derivation of Lower Bound

In this section, we aim at deriving a new lower bound on the probability of error. In the field of the information theory, Fano's inequality is used for a converse proof between the probability of error and conditional entropy [8]. For the GT schemes, we exploit Fano's inequality to derive the lower bound on the probability of error. In this work, we assume that there is a noiseless system and the decoding error for an arbitrary decoder in (5) is ignored.

Lemma 1 (Fano's inequality [8]): For arbitrary estimator function Ψ such that $\mathbf{x} \to (\mathbf{A}, \mathbf{y}) \to \hat{\mathbf{x}}$, we have the following inequality,

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(\mathbf{x}|\hat{\mathbf{x}}) \geq H(\mathbf{x}|(\mathbf{A}, \mathbf{y})) \tag{8}$$

where the probability P_e is the probability of error for the estimator function Ψ and the cardinality of the random variable $\mathbf{x}$ is denoted by $|\mathcal{X}|$.

We use Fano's inequality to derive the lower bound on the probability of error for any lossless systems. Note that our new lower bound is independent of the estimator function Ψ and the pooling matrix $\mathbf{A}$.

Theorem 2: For any pooling matrix in (2) and any decoder in (5), if $0 < \delta \leq 1/2$, the probability of error is bounded by

$$P_e \geq H_b(\delta) - \frac{M}{N} H_b(\epsilon) - \frac{1}{N} \tag{9}$$

where the function $H_b(\cdot)$ is the binary entropy, ϵ is the probability that the element of $\mathbf{y}$ is 0.

Proof: The probability of error using an arbitrary decoder can be obtained as follows,

$$\begin{aligned} H(\mathbf{x}) &= I(\mathbf{x};\mathbf{y}) + H(\mathbf{x}|\mathbf{y}) \\ &\overset{(a)}{\leq} I(\mathbf{x};\mathbf{y}) + H(P_e) + P_e \log_2 (|\mathcal{X}| - 1) \\ &\overset{(b)}{\leq} I(\mathbf{x};\mathbf{y}) + 1 + P_e \log_2 |\mathcal{X}| \\ &\overset{(c)}{=} H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) + 1 + NP_e \\ &\overset{(d)}{=} H(\mathbf{y}) + 1 + NP_e \end{aligned} \tag{10}$$

where $I(\cdot)$ denotes the mutual information and the first line of (10) comes from the definition of the mutual information: $I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$. And inequality (a) is from Fano's inequality. Inquality (b) is due to the fact that an error event is a binary variable, i.e., correct or wrong. Therefore, $H(P_e) \leq 1$. Equality (c) uses the cardinality of the set: $|\mathcal{X}| = 2^N$ and $\mathcal{X} \in \{0,1\}^N$. In the noiseless case, since $\mathbf{y}$ is the function of $\mathbf{x}$, i.e., $\mathbf{y} = \mathbf{A} \odot \mathbf{x}$, if we know $\mathbf{x}$, there is no randomness in $\mathbf{y}$. Hence, we lead to the following conditional entropy as $H(\mathbf{y}|\mathbf{x}) = 0$. Equality (d) of (10) holds as follows:

$$NH_b(\delta) \leq H(\mathbf{y}) + 1 + NP_e \tag{11}$$

Next we aim to find out the entropy of $\mathbf{y}$: $H(\mathbf{y}) = H(y_1, y_2, y_3, \cdots y_M)$. We know that each element of $\mathbf{y}$ is binary. Using chain rule of conditional entropy, we obtain the entropy of $\mathbf{y}$ as follows,

$$\begin{aligned} H(\mathbf{y}) &= H(y_1) + H(y_2|y_1) + H(y_3|y_2, y_1) \\ &\qquad + \cdots + H(y_M|y_{M-1}, \cdots, y_2, y_1) \\ &\overset{(a)}{\leq} H(y_1) + H(y_2) + \cdots + H(y_M) \\ &\overset{(b)}{=} MH(y_1) \end{aligned} \tag{12}$$

where the first equality of (12) is from the chain rule of the conditional entropy and inequality (a) derives from the fact that the conditional entropy is greater than or equal to the entropy. Inequality (b) exploits the following fact that the random variables $y_1, y_2, \cdots, y_M$ are independent with each other. The binary entropy $H(y_1)$ can be obtained from the following definition,

$$H_b(\epsilon) = -\epsilon \log_2 \epsilon - (1-\epsilon) \log_2 (1-\epsilon) \tag{13}$$

The probability ϵ can be obtained from the definitions (1) and (2) as follows,

$$\epsilon = (1 - \delta\gamma)^N \tag{14}$$

Using equations from (11) to (14), we finally derive the lower bound on the probability of error as follow,

$$P_e \geq H_b(\delta) - \frac{M}{N} H_b(\epsilon) - \frac{1}{N} \tag{15}$$

This is the end of the proof for Theorem 2. □

Equation (15) means that the probability of error is larger than or equal to the right hand side of (15) even if we use any decoder for finding all defective samples. In order for the probability of error P_e to converge to zero, the right hand side of (15) should be negative. This is a necessary condition for error free. Therefore, This result in the following necessary condition to solve the GT problems defined by (3).

$$\begin{aligned} M &> \frac{NH_b(\delta) - 1}{H_b(\epsilon)} \\ &\geq NH_b(\delta) - 1 \end{aligned} \tag{16}$$

where the second line for inequality derives from $H_b(y_1) \leq 1$. From (16), we see two interesting results. First, the number of tests for successfully decoding all defective samples in GT problems satisfies the following condition: $M > NH_b(\delta)$. Second, when the following condition $H_b(\epsilon) = 1$ satisfies, the necessary condition for successfully decoding without the probability of error holds. In other words, we can rewrite this condition for $H_b(\epsilon) = 1$,

$$\delta\gamma = 1 - \left(\frac{1}{2}\right)^{\frac{1}{N}} \tag{17}$$

From (17), we show that both the sparsity ratio δ and the density ratio γ are inverse with each other. Therefore, for completely estimating of densely defective samples we need a very sparse pooling matrix. However, if there are sparsely defective samples for decoding, a more denser pooling matrix is required to successfully decode their unknown samples.

3.2 Numerical Evaluation

In this section, we show numerical results for the lower bound. The lower bound on the probability of error has been derived from in Theorem 2. The lower bound is a function of the sparsity ratio δ, the density ratio γ and the number of the total samples N. In our evaluation, we use the number of samples $N = 1000$ for Fig. 1 to 3.

In Fig. 1, we evaluate the number of tests M for successful decoding of the GT problem in (16). For example, if we decode the sparsity ratio $\delta = 0.05$ of the binary signal, i.e., averaged 50 defective samples out of total 1000 samples, the number of tests M is at least 287 which is obtained in 0.013 density ratio of the pooling matrix. This density ratio means that the number of samples in a pool is included averaged 13 different samples out of 1000 samples. As shown in Fig. 1, the curves of the number of tests are convex for different sparsity ratios δ. The interesting point in Fig. 1 is that as the sparsity ratio δ increases, the optimal band of the density ratio γ becomes narrow. This observation suggests that for successful decoding with high

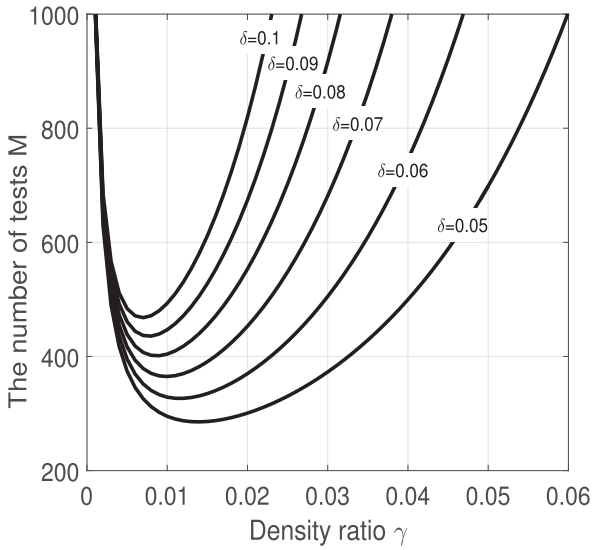

Fig. 1 The number of tests M for successful decoding of the GT scheme with various sparsity ratio δ and density ratio γ and $N = 1000$.

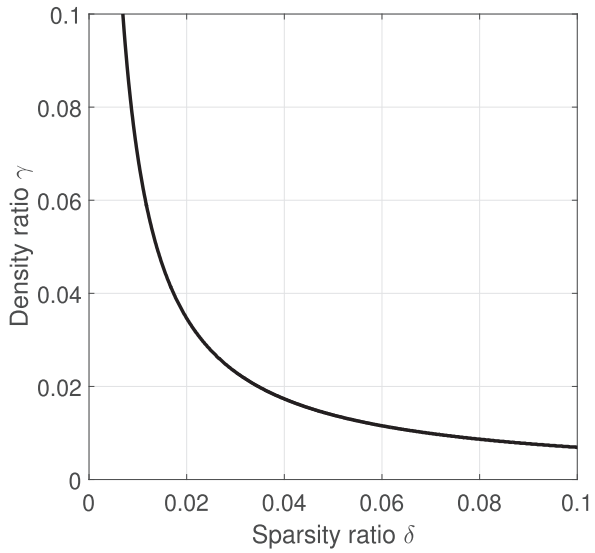

Fig. 2 The relationship between the sparsity ratio δ and the density ratio γ for achieving the number of tests M with $N = 1000$.

sparsity ratios the pooling matrix should be designed in a narrow range of the density ratio γ without different classes of GT problems.

For the evaluation of the relationship between the sparsity ratio δ and the density ratio γ we show Fig. 2 for archiving the number of tests with $N = 1000$. This result provides an important core for the design of the pooling matrix. In other words, the density of the pooling matrix should be designed by the sparsity of the binary signals. If unknown binary signals are very sparse, we have to use dense pooling matrices. Conversely, sparse pooling matrices are used to decode dense binary signals. The core question of the GT problems is how many tests are required for successful decoding. This answer is shown in Fig. 3.

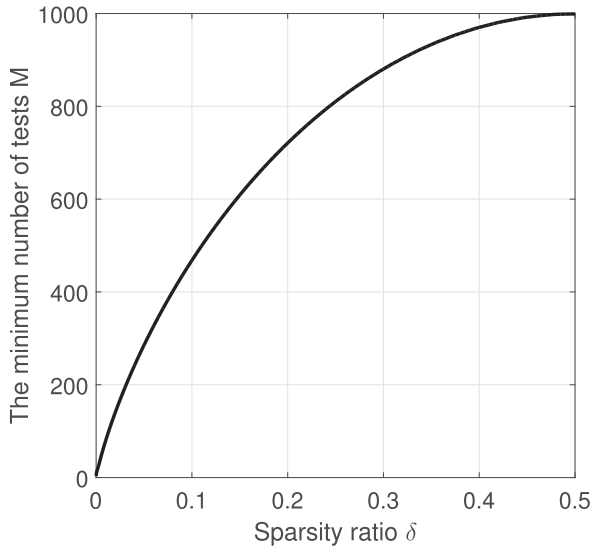

Fig. 3 The number of tests M for successful decoding of the GT scheme with different sparsity ratio δ with $N = 1000$.

4. Conclusion

In this paper, we considered the GT problem. We derived the lower bound on the probability of error. To this end, we used Fano's inequality exploited in the information theory. We showed that the probability of error is expressed as the entropy function of the density ratio of the pooling matrix and sparsity ratio of the binary vector. We showed that for decoding of highly sparse signals, the pooling matrix is required to be dense. Conversely, if dense signals are needed to decode, the sparse pooling matrix should be designed to achieve the small probability of error.

References

[1] R. Dorfman, "The detection of defective members of large populations," The Annals of Mathematical Statistics, vol.14, no.4, pp.436–440, Dec. 1943.
[2] D.L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, April 2006.
[3] D.-Z. Du and F.-K. Hwang, Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing, World Scientific, 2006.
[4] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," IEEE Int. Sym. Inf. Theory (ISIT), pp.2676–2680, Oct. 2013.
[5] M. Aldridge, L. Baldassini, and O. Johnson, "Group tesing algorithm: Bounds and simulations," IEEE Trans. Inf. Theory, vol.60, no.6, pp.3671–3687, June 2014.
[6] T. Wadayama, "Nonadaptive group testing based on sparse pooling graphs," IEEE Trans. Inf. Theory, vol.63, no.3, pp.1525–1534, March 2017.
[7] A. Agarwal, S. Jaggi, and A. Mazumdar, "Novel impossibility results for group-testing," IEEE Int. Sym. Inf. Theory (ISIT), pp.2579–2583, June 2018.
[8] T.M. Cover and J.A. Thomas, Elements of Information Theory, 2nd Edition, Wiley, 2006.