An Analysis on Non-Adaptive Group Testing based on Sparse Pooling Graphs

Tadashi Wadayama
Department Computer Science and Engineering,
Nagoya Institute of Technology, Nagoya, Japan
Email: wadayama@nitech.ac.jp

***Abstract*—In this paper, an information theoretic analysis on non-adaptive group testing schemes based on sparse pooling graphs is presented. The binary status of the objects to be tested are modeled by i.i.d. Bernoulli random variables with probability p. An (l, r, n)-regular pooling graph is a bipartite graph with left node degree l and right node degree r, where n is the number of left nodes. Two scenarios are considered: a noiseless setting and a noisy one. The main contributions of this paper are direct part theorems that give conditions for the existence of an estimator achieving arbitrary small estimation error probability. The direct part theorems are proved by averaging an upper bound on estimation error probability of the typical set estimator over an (l, r, n)-regular pooling graph ensemble. Numerical results indicate sharp threshold behaviors in the asymptotic regime.**

I. Introduction

The paper by Dorfman [8] introduced the idea of *group testing* and also presented a simple analysis, which indicates advantages of the idea. His main motivation was to devise an *economical way* to detect infected persons within a population by using blood tests. It is assumed that the outcome of a blood test determines if the blood used in the test contains certain target viruses (or bacteria).

Of course, blood tests for every person in the population would clearly distinguish the infected individuals from those who are not infected. Dorfman's idea for reducing the number of tests is the following. We first divide the population into several disjoint groups and then mix the blood of individuals in each group to form *pool*. The test process then consists of two-stages. In the first stage, the pools containing infected blood are determined by blood test of each pool. In the second stage, all the individuals in those groups with positive results are tested. Numerical examples show that the number of tests can be reduced without loss of detection capability [8].

Dorfman's idea triggered the emergence of subsequent theoretical works on group testing and a variety of practical applications, such as the screening of DNA clone libraries and the detection of faulty machines parts [9] [10]. In addition, recent advances in the theory of compressed sensing [6] have stimulated research into the theoretical aspects of group testing.

The group testing scheme due to Dorfman can be classified as *adaptive group testing*, in which the latter part of test design depends on the results of earlier tests. There is also *non-adaptive group testing*, in which the test design is completely determined before conducting any tests. Intuitively, adaptive group testing seems advantageous over non-adaptive group testing because it requires fewer tests. However, there are also advantages to the non-adaptive group testing, since in this design, all the tests can be executed in parallel. Note that adaptive group testing requires sequential tests and thus prevents parallel testing.

In order to develop a non-adaptive group testing scheme with good detection performance, pool design is crucial. In the field of combinatorial group testing, a pooling matrix that defines the set of pools to be tested is constructed by using combinatorial design and combinatorics. The *deterministic counstruction* of a K-disjunct matrix is one of the central themes of combinatorial group testing [9] [10].

Pooling matrices can also be obtained by *random construction*; that is, the $(0, 1)$-elements of a pooling matrix are determined probabilistically. Several reconstruction algorithms have been proposed for such probabilistically constructed pooling matrices. For example, Sejdinovic and Johnson [18], Kanamori et al. [13] recently proposed reconstruction algorithms based on belief propagation. Malioutov and Malyutov [16], Chan et al. [5] studied reconstruction algorithms based on linear programming (LP).

Clarifying the scaling behavior of the number of required test for correct reconstruction has become one of the most important topics in this field. Berger and Levenshtein [2] studied a two-stage group testing scheme and unveiled the scaling law for the number of required tests based on an information theoretic arguments. Mézard and Toninelli [17] provided a novel analysis of two-stage schemes based on theoretical techniques from statistical mechanics. Recently, Atia and Saligrama [1] presented an information theoretic analysis of non-adaptive group testing with and without noise. They presented a direct part theorem that gives a condition for the existence of an estimator achieving arbitrary small estimation error probability and a converse part theorem that gives a condition for the non-existence of good estimators. The arguments in their proof of these theorems are based on the proof of the channel coding theorems for multiple access channels, and they can be applied to both noiseless and noisy observations. For example, in the noiseless case, it was shown that a K-sparse instance of n-objects can be perfectly recovered from the test results if the number of tests is asymptotically $O(K \log n)$.

The main motivation of this work is to provide further information-theoretic analysis of non-adaptive group testing based on sparse pooling graphs. In this paper, we assume that the status (0 or 1) of n-objects is modeled by a Bernoulli random variable with probability p. In other words, we consider the scenario in which the sparsity parameter K scales as $K \simeq pn$ asymptotically. In most conventional analyses, such as [1], K is assumed to be independent of n. Such an assumption is reasonable in order to clarify the dependency of the required number of tests on the sparsity parameter and the number of objects. Although our assumption is different from the conventional one, it is natural from an information-theoretic point of view and is suitable for observing sharp threshold behaviors in the asymptotic regime.

Another new aspect is that the analysis is carried out under the assumption of an (l, r, n)-regular pooling graph ensemble, which is a bipartite graph ensemble with left node degree l and right node degree r, where n is the number of left nodes. This model is suitable for handling a very sparse pooling matrix and is amenable to ensemble analysis. We will present both direct and converse theorems that predict the asymptotic behavior of a group testing scheme with an (l, r, n)-pooling graph. These asymptotic conditions are parameterized by p, l, and r. Therefore, for a given pair (l, r), we can determine the region for p in which we can achieve arbitrarily accurate estimation. Our analysis was inspired by the analysis of Gallager and others [11] [14] [12] of low-density parity-check (LDPC) codes.

II. Preliminaries

In this section, we introduce the two scenarios for group testing that will be discussed in this paper. The first one is the *noiseless system*, where test results can be seen as a function of an input vector. The second one is the *noisy system*, where the test results are disturbed by the addition of noise.

A. Problem setting for noiseless system

The random variable $X \triangleq (X_1, \ldots, X_n)$ represents the status of n-objects. We assume that $X_i (i \in [1, n])$ is an i.i.d. Bernoulli random variable with the probability distribution $Pr(X_i = 0) = 1 - p, Pr(X_i = 1) = p (0 \leq p \leq 1)$. The notation $[a, b]$ represents the set of consecutive integers from a to b. With some slight abuse of notation, the notation $[a, b]$ is also used for representing closed interval over $\mathbb{R}$ when there is no fear of confusion. A realization of X is denoted by $x \triangleq (x_1, \ldots, x_n)$. The test function $OR(z_1, \ldots, z_r) : \{0, 1\}^r \to \{0, 1\}$ is the logical OR (disjunctive) function with r-arguments (r is a positive integer) defined by

$$OR(z_1, \ldots, z_r) \triangleq \begin{cases} 0, & z_1 = z_2 = \cdots = z_r = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

The results of pooling tests which is abbreviated as *test results* are represented by $Y \triangleq (Y_1, \ldots, Y_m)$. A realization of Y is denoted by $y \triangleq (y_1, \ldots, y_m)$.

Let $G \triangleq (V_L, V_R, E)$ be a bipartite graph, called a *pooling graph*, with the following properties. The n-nodes in V_L are called *left nodes* and the other m-nodes in V_R are called *right nodes*. The set E represents the set of edges. For convenience, we assume that the left nodes are labeled from 1 to n. The left node with label $i \in [1, n]$ corresponds to X_i; for simplicity, we will refer to it as left node i. In a similar manner, the right nodes are labeled from 1 to m. In this paper, G is assumed to be an (l, r, n)-regular bipartite graph, which means that any left and right nodes have degrees l and r, recpectively, and that the number of the left nodes is n.

For the right node $j \in [1, m]$, the neighbor set of the node j is defined by $M(j) \triangleq \{i \in [1, n] \mid (i, j) \in E\}$. We are now ready to describe the relationship between X and Y. For a given pooling graph G, $Y_j (j \in [1, m])$ are related to $X_i (i \in [1, n])$ by $Y_j = OR(X_i)_{i \in M(j)}$. The notation $(X_i)_{i \in M(j)}$ represents $(X_{j_1}, \ldots, X_{j_r})$ when $M(j) = \{X_{j_1}, \ldots, X_{j_r}\}$. Namely, a pooling graph G defines a function from X to Y. We will denote this relationship as $Y = F_G(X)$ for short.

The goal of an examiner to infer, as correctly as possible, the realization of a hidden random variable X from the test observation y. Assume that the examiner uses an estimator (i.e., estimation function) $\Phi : \{0, 1\}^m \to \{0, 1\}^n$ for the inference. The estimator gives an estimate of x, $\hat{x} = \Phi(y)$, from the test observation y. The estimator Φ should be chosen so that the *estimation error probability*

$$P_e \triangleq Pr\left(\Phi(F_G(X)) \neq X\right) \tag{2}$$

is as small as possible.

B. Problem setting for noisy system

The setting for the noisy system is almost the same as the setting for the noiseless system, which was described in the previous subsection. The crucial difference between the two is the assumption of observation noises in the noisy system. In this case, the examiner observes a realization of the random variable Z, defined by

$$Z = Y + E = F_G(X) + E, \tag{3}$$

where $E \triangleq (E_1, \ldots, E_m)$ represents the observation noise. We assume that $E_i (i \in [1, m])$ is an i.i.d. Bernoulli random variable with the probability distribution $Pr(E_i = 0) = 1 - q, Pr(E_i = 1) = q (0 \leq q \leq 1)$.

III. Converse Part Analysis

In this section, lower bounds on estimation error probability for the noiseless and noisy systems will be shown. The key to the proofs is Fano's inequality, which ties the estimation error probability to the conditional entropy.

A. Lower bound for noiseless system

Fano's inequality relates the conditional entropy and the estimation error probability, and it has often been used as the main argument in the proof of the converse part of a channel coding theorem [7]. This inequality also plays a crucial role in the following analysis, in which it clarifies the limits of accurate estimation for the noiseless and noisy systems.

Lemma 1 (Fano's inequality): Assume that random variables A, B are given. The cardinality of the domains (alphabets) of A and B are assumed to be finite. For any estimator ϕ for estimating the hidden value of A from the observation of B, the inequality $1 + Pr(A \neq \phi(B)) \log_2 |\mathcal{A}| \geq H(A|B)$ holds. The domain of A is denoted by $\mathcal{A}$. $\square$

We use Fano's inequality for deriving a lower bound on the error probability of an estimation for the noiseless system. Note that this lower bound does not depend on the choice of pooling graph and an estimator. The proof of the theorem resembles the proof of the upper bound on code rate for LDPC codes [11] [4]. Similar argument can be found in [1], [5] as well.

Theorem 1 (Noiseless system): Assume the noiseless system. For any pair of an (l, r, n)-pooling graph and an estimator, the error probability P_e is bounded from below by

$$h(p) - \frac{l}{r} h((1-p)^r) - \frac{1}{n} \leq P_e. \tag{4}$$

(Proof) For any estimator having the error probability P_e, we have

$$\begin{aligned} H(X) &= I(X;Y) + H(X|Y) \\ &\leq I(X;Y) + 1 + P_e \log_2 |\mathcal{X}| \quad (5) \\ &= H(Y) - H(Y|X) + 1 + P_e n \quad (6) \\ &= H(Y) + 1 + P_e n. \quad (7) \end{aligned}$$

The inequality (5) is due to Fano's inequality. Equation (6) holds since $\mathcal{X} = \{0, 1\}^n$. Note that, in the noiseless system, the random variable Y is a function of X, namely $Y = F_G(X)$ and that it implies $H(Y|X) = 0$. The last equality (7) is a consequence of $H(Y|X) = 0$.

Since we have assumed that $X = (X_1, \ldots, X_n)$ is an n-tuple of i.i.d. Bernoulli random variables, the entropy of X is given by $H(X) = nh(p)$, where $h(p)$ is the binary entropy function defined by $h(p) \stackrel{\triangle}{=} -p \log_2 p - (1-p) \log_2 (1-p)$. We thus have $nh(p) \leq H(Y) + 1 + P_e n$.

Next, we need to evaluate $H(Y) = H(Y_1, \ldots, Y_m)$. It should be noted that the random variables $Y_1, Y_2, \ldots, Y_n$ are binary random variables, and they are correlated in general. A simple upper bound on $H(Y)$ can be obtained as $H(Y_1, Y_2, \ldots, Y_m) \leq \sum_{i=1}^m H(Y_i)$. This is simply due to the chain rule and a property of the conditional probabilities (i.e., conditioning reduces entropy [7]). From our assumptions that $Y_j = OR(X_i)_{i \in M(j)} (j \in [1, m])$ and that $|M(j)| = r (j \in [1, m])$, we have $H(Y_j) = h((1-p)^r)$ because $Pr[Y_j = 0] = (1-p)^r$. Combining these results, we obtain an inequality $nh(p) \leq mh((1-p)^r) + 1 + P_e n$. From this inequality, we immediately obtain the claim of the theorem. $\square$

B. Lower bound for noisy system

Let us recall the problem setup for the noisy system. The random variable $Z \stackrel{\triangle}{=} (Z_1, \ldots, Z_m)$, representing a noisy observation, is defined by $Z = Y + E = F_G(X) + E$. As in the case of the noiseless system, a lower bound on the error probability for the noisy system can be derived based on Fano's inequality.

Theorem 2 (Noisy system): Assume the noisy system described above. For any pair of an (l, r, n)-pooling graph and an estimator, the error probability P_e is bounded from below by

$$h(p) + \frac{l}{r} h(q) - \frac{l}{r} h((1-p)^r (1-q) + (1-(1-p)^r) q) - \frac{1}{n} \leq P_e. \tag{8}$$

IV. Direct Part Analysis

In the previous section, we discussed the limitation of accurate estimation of any estimator, i.e., a lower bound on error probability. This result is similar to the converse part of a coding theorem. In this section, we shall discuss the direct part; i.e., the existence a sequence of estimators achieving arbitrary small error probability. As in the case of coding theorems, we rely on the standard *bin coding* argument [7] to prove the main theorems. In order to apply such an information theoretic argument, we here introduce *typical set estimator*.

A. Pooling graph ensemble

In the following analysis, we will take ensemble average of the error probability of the typical set estimator over an ensemble of pooling graphs. The pooling graph ensemble introduced below resembles the bipartite graph ensemble for regular LDPC codes. The following definition gives the details of the pooling graph ensemble [14].

Definition 1 (Pooling graph ensemble): Let $G_{l,r,n}$ be the set of all (l, r, n)-regular bipartite graphs with n-left and $m = (l/r)n$-right nodes. The cardinality of $G_{l,r,n}$ is $(nl)!$. Assume that equal probability $P(G) = 1/(nl)!$ is assigned for each graph $G \in G_{l,r,n}$. The ensemble of graphs defined based on the pair $(G_{l,r,n}, P)$ is called the (l, r, n)-pooling graph ensemble. $\square$

In order to prove the direct part theorems, we need to evaluate the expectation of the number of typical sequences x satisfying $y = F_G(x)$ over the (l, r, n)-pooling graph ensemble. The next lemma is required for deriving the main theorems.

Lemma 2: Assume that $s \in [0, m]$ and $w \in [0, n]$ are given. Let $y_s \in \{0, 1\}^m$ be a binary m-tuple with weight s and $x_w \in \{0, 1\}^n$ be a binary n-tuple with weight w. The probability of the event $y_s = F_G(x_w)$ is given by

$$\mathsf{E}\left[\mathbb{I}[y_s = F_G(x_w)]\right] = \frac{1}{\binom{nl}{wl}} \mathsf{Coeff}[((1+z)^r - 1)^s, z^{lw}], \tag{9}$$

where $\mathsf{Coeff}[g(z), z^i]$ represents the coefficient of z^i in the polynomial $g(z)$. The function $\mathbb{I}[cond]$ is the indicator function taking value 1 if $cond$ is true; otherwise it gives value 0.

The combinatorial argument presented in the proof of Lemma 2 is closely related to the derivation of an average input-output weight distribution of LDPC codes over a regular bipartite graph ensemble due to Hsu and Anastasopoulos [12].

B. Analysis on error probability for noiseless system

In this subsection, we define the typical set estimator for the noiseless system and analyze its error performance. Before describing the typical set estimator, we define the typical set [7] as follows.

Definition 2 (Typical set): Assume that an i.i.d. random variables $A_i (i \in [1, n])$, a positive constant ϵ and a positive integer n are given. The typical set $T_{n,\epsilon}$ is defined by

$$T_{n,\epsilon} \stackrel{\triangle}{=} \{(a_1, \ldots, a_n) \in \mathcal{A}^n \mid |\mathcal{H} - \kappa(a_1, \ldots, a_n)| \le \epsilon\}, \tag{10}$$

where $\mathcal{A}$ is the finite alphabet of A_i and $\mathcal{H} = H(A_i)$ holds for $i \in [1, n]$. The function κ is defined by $\kappa(a_1, \ldots, a_n) \stackrel{\triangle}{=} (-1/n) \log_2 Pr(a_1, \ldots, a_n)$. □

The typical set estimator defined below is almost the same as the typical set decoder assumed in the proof of several coding theorems, such as in [15]. It is exploited in order to simplify the proof, and it is, in general, computationally infeasible. Despite its computational complexity, the performance of the typical set estimator can be used as a benchmark for other estimation algorithms.

Definition 3 (Typical set estimator): Assume the noiseless system. Suppose that an (l, r, n)-pooling graph $G \in G_{l,r,n}$ and a positive real value ϵ are given. The typical set estimator $\Phi : \{0, 1\}^m \to \{0, 1\}^n \cup \{E\}$ is defined by

$$\Phi(y) \stackrel{\triangle}{=} \begin{cases} x \in D(y), & \text{if } |D(y)| = 1, \\ E, & otherwise, \end{cases} \tag{11}$$

where $D(y)(y \in \{0, 1\}^m)$ is the decision set defined by $D(y) \stackrel{\triangle}{=} \{x \in T_{n,\epsilon} \mid y = F_G(x)\}$. The symbol E represents failure of estimation. □

The typical set estimator Φ depends on the *bins* defined on the typical set $T_{n,\epsilon}$. A bin $D(y)$ consists of the inverse image of y in the typical set. For an observed vector y, if the cardinality of the bin $D(y)$ is 1, the estimator declare that $x \in D(y)$ has occurred. The failure of estimation occurs when the cardinality of $D(y)$ is greater than 1. For evaluating the error probability of the typical set estimator, an analysis for this event is indispensable and it will be the main topic of the following analysis.

The next lemma proves the existence of a pair (G, Ψ) achieving a given upper bound on the error probability, which can be regarded as a counter part of the direct part of a coding theorem.

Lemma 3: Assume the noiseless system. If $\gamma > 0$ satisfying

$$-(l-1)h(p) + \max_{\sigma \in [0, l/r]} \left[\log_2 \inf_{z>0} \frac{((1+z)^r - 1)^\sigma}{z^{lp}}\right] + \gamma < 0 \tag{12}$$

exists, then there exists a pair $(G \in G_{l,n,r}, \Phi)$ with the error probability smaller than γ. □

(Proof) The proof is based on the bin-coding argument. Assume that a positive real number ϵ is given (later we will see that ϵ is determined according to γ but, for now we consider that ϵ is given). Note that there are two events that the typical set estimator fails to correctly estimate . By Event I, we denote the event in which a realization of X, x, is not a typical sequence. Event II corresponds to the case in which a realization x is a typical sequence, but $|D(F_G(x))| > 1$ holds.

We therefore have

$$P_e = Pr[X \neq \Phi(F_G(X))] = P_I + P_{II}(G), \tag{13}$$

where $P_I, P_{II}(G)$ are probabilities corresponding to Event I and II, respectively. Note that the probability P_I depends only on the parameters n and ϵ.

We first consider the probability $P_{II}(G)$, for which the upper bound is as follows:

$$\begin{aligned} &P_{II}(G) \\ &= \sum_{x \in T_{n,\epsilon}} Pr(x) \mathbb{I}[\exists x' \in T_{n,\epsilon}, x' \in D(F_G(x)), x' \neq x] \\ &\le \sum_{x \in T_{n,\epsilon}} Pr(x) \sum_{x' \in T_{n,\epsilon}, x \neq x'} \mathbb{I}[F_G(x) = F_G(x')]. \end{aligned} \tag{14}$$

By taking the expectation of (14) over the (l, r, n)-pooling graph ensemble, we obtain

$$\begin{aligned} &\mathsf{E}[P_{II}(G)] \\ &\le \sum_{x \in T_{n,\epsilon}} Pr(x) \sum_{x' \in T_{n,\epsilon}, x \neq x'} \mathsf{E}[\mathbb{I}[F_G(x) = F_G(x')]] \\ &\le |T_{n,\epsilon}| \max_{s \in [0,m]} \max_{w \in [w_{min}, w_{max}]} \mathsf{E}[\mathbb{I}[y_s = F_G(x_w)]], \end{aligned} \tag{15}$$

where w_{min} and w_{max} are defined by

$$w_{max} = \max_{x \in T_{n,\epsilon}} wt(x), \quad w_{min} = \min_{x \in T_{n,\epsilon}} wt(x), \tag{16}$$

where $wt(x)$ represents the Hamming weight of x. The vector y_s is an arbitrary binary m-tuple with weight s and x_w is an arbitrary binary n-tuple with weight w.

Applying the upper bound on the size of the typical set and Lemma 2 to (15), we have

$$\mathsf{E}[P_{II}(G)] \le 2^{n(h(p)+\epsilon)} \max_{s \in [0,m]} \max_{w \in [w_{min}, w_{max}]} T(n, l, r, w, s) \tag{17}$$

where $T(n, l, r, w, s) \stackrel{\triangle}{=} \frac{1}{\binom{nl}{wl}} \mathsf{Coeff}[((1 + z)^r - 1)^s, z^{lw}]$. By letting $\omega \stackrel{\triangle}{=} w/n$ and $\sigma \stackrel{\triangle}{=} s/n$, the above inequality (17) can be rewritten as $\mathsf{E}[P_{II}(G)] \le 2^{n(h(p)+\epsilon+Q)}$ where

$$Q \stackrel{\triangle}{=} \frac{1}{n} \log_2 \left[\max_{s \in [0,m]} \max_{w \in [w_{min}, w_{max}]} T(n, l, r, w, s)\right].$$

For evaluating the coefficient of the generating function, a theorem by Burshtein and Miller [3] is exploited and Q can be expressed as

$$\begin{aligned} Q &= -h(p) + \max_{\sigma \in [0, l/r]} \left[\log_2 \inf_{z>0} \frac{((1+z)^r - 1)^\sigma}{z^{lp}}\right] \\ &+ \delta(n) + \xi(\epsilon), \end{aligned} \tag{18}$$

where $\xi(\epsilon)$ is a function of ϵ such that $\xi(\epsilon) \to 0$ as $\epsilon \to 0$ and the function $\delta(n)$ is a function such that $\delta(n) \to 0$ as $n \to \infty$. Assume that a positive real number γ is given and

$$-(l-1)h(p) + \max_{\sigma \in [0, l/r]} \left[\log_2 \inf_{z>0} \frac{((1+z)^r - 1)^\sigma}{z^{lp}}\right] + \gamma < 0 \tag{19}$$

holds. For sufficiently large n and sufficiently small ϵ, there exists a pair (n, ϵ) satisfying $\epsilon+\delta(n)+\xi(\epsilon) < \gamma$ and the following two conditions. The first condition is that $\mathsf{E}[P_{II}(G)] < \frac{\gamma}{2}$. Note that, due to the assumption (19), the exponential growth rate of the upper bound on $\mathsf{E}[P_{II}(G)]$ is negative, and thus the upper bound on $\mathsf{E}[P_{II}(G)]$ can be arbitrary small as $n \to \infty$. The second condition is that $P_I < \gamma/2$ which is guaranteed by the asymptotic equipartition property (AEP) for the typical set [7]. As a result, we have $\mathsf{E}[P_E] = P_I + \mathsf{E}[P_{II}(G)] < \gamma$, and this implies the existence of a pair $(G \in G_{l,n,r}, \Phi)$ for which the error probability is smaller than γ. □

From this lemma, we can immediately derive the following direct part result.

Theorem 3 (Achievability for noiseless system): Assume the noiseless system. If $\gamma > 0$ satisfying

$$-(l-1)h(p) - lp\log_2(2^{1/r} - 1) + \gamma < 0$$

exists, then there exists a pair $(G \in G_{l,n,r}, \Phi)$ with the error probability smaller than γ. □

From Theorem 1 and Theorem 3, it is natural to conjecture the existence of the threshold value $p^*(l,r)$ partitioning the range of p into two regions. Namely, if $p < p^*(l,r)$, arbitrary accurate estimation is possible. Otherwise, i.e., $p > p^*(l,r)$, no estimator achieving arbitrary small error probability exists in the asymptotic limit $n \to \infty$. An upper bound on the threshold can be obtained from Theorem 1. The upper bound $p_U^*(l,r)$ is given by $p_U^*(l,r) \stackrel{\triangle}{=} \inf\{p \mid p \text{ satisfies } h(p) - (l/r)h((1-p)^r > 0\}$. On the other hand, a lower bound on the threshold is defined by $p_L^*(l,r) \stackrel{\triangle}{=} \sup\left\{p \mid p \text{ satisfies } -(l-1)h(p) - lp\log_2(2^{1/r}-1) < 0\right\}$, which is a direct consequence of Theorem 3. Table I presents the values of the lower and upper bounds on the threshold for the two cases $l/r = 1/2$ and $l/r = 1/4$.

TABLE I
THRESHOLD BOUNDS FOR NOISELESS SYSTEMS

(l,r)	$P_L^*(l,r)$	$P_U^*(l,r)$
$(2,4)$	0.092763	0.097350
$(3,6)$	0.110022	0.110023
$(4,8)$	0.104629	0.105999
$(5,10)$	0.096091	0.099480
$(6,12)$	0.087848	0.093027
$(2,8)$	0.022022	0.026824
$(3,12)$	0.038651	0.039535
$(4,16)$	0.041685	0.041687
$(5,20)$	0.040693	0.040978
$(6,24)$	0.038556	0.039427

C. Analysis on error probability for noisy system

As in the case of the noiseless system, we can derive a direct part theorem for the noisy system shown below.

Theorem 4 (Achievability for noisy system): Assume the noisy system. If $\gamma > 0$ satisfying $-(l-1)h(p) + \frac{l}{r}h(q) - lp\log_2(2^{1/r}-1) + \gamma < 0$ exists, then there exists a pair $(G \in G_{l,n,r}, \Phi)$ with the error probability smaller than γ. □

V. Conclusion

The analysis presented in this paper was inspired by the theoretical works on LDPC codes [15] [4]. From numerical evaluation, it was shown that the gap between the upper bound $p_U^*(l,r)$ and the lower bound $p_L^*(l,r)$ is usually quite small. This suggests the existence of a sharp threshold, which is similar to the Shannon limit for a channel coding problem. The details of the proofs and discussion on related topics can be found in the full paper version arXiv:1301.7519.

Acknowledgement

The author would like to thank the anonymous reviewers for their constructive comments. This work was supported by JSPS Grant-in-Aid for Scientific Research (B) Grant Number 25289114.

References

[1] G.Atia and V.Saligrama, "Boolean compressed sensing and noisy group testing," IEEE Trans. on Information Theory, vol.58 (3), pp. 1880 – 1901, 2012.

[2] T.Berger and V.I.Levenshtein, "Asymptotic efficiency of two-stage disjunctive testing, " IEEE Trans. on Information Theory, vol.48 (7), pp. 1741 – 1749, 2002.

[3] D.Burshtein and G.Miller, "Asymptotic enumeration methods for analyzing LDPC codes, " IEEE Trans. on Information Theory, vol.50 (6), pp. 1115 – 1131, 2004.

[4] D. Burshtein, M. Krivelevich, S. Litsyn and G. Miller, "Upper bounds on the rate of LDPC codes," IEEE Transactions on Information Theory, vol.48 (9), pp. 2437-2449, 2002.

[5] C.L.Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: explicit bounds and novel algorithms," arXiv:1202.0206v4 [cs.IT], 2012.

[6] E.Candes, J. Romberg and T.Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," IEEE Trans. on Information Theory, vol.52 (2), pp. 489 – 509, 2006.

[7] T.M.Cover and J.A.Thomas, "Elements of Information Theory", 2nd ed. Wiley-Interscience 2006.

[8] R.Dorfman, "The detection of defective members of large populations," Ann. ath. Stat., vol. 14, pp.436–440, 1943.

[9] D.-Z Du and F.K.Hwang, "Combinatorial Group Testing and Its Applications," 2nd ed. World Scientific Publishing Company, 2000.

[10] D.-Z Du and F.K.Hwang, "Pooling Designes and Non-Adaptive Group Testing: Important Tools for DNA Sequencing, " Singapore: World Scientific, 2006.

[11] R.G.Gallager, "Low Density Parity Check Codes". Cambridge, MA:MIT Press 1963.

[12] C.H.Hsu and A. Anastasopoulos, "Capacity-achieving codes with bounded graphical complexity and maximum likelihood decoding," IEEE Trans. Inform. Theory, pp.992-1006, vol.56, no. 3, Mar. 2010.

[13] T. Kanamori, H. Uehara, M. Jimbo, "Pooling design and bias correction in DNA library screening, " Journal of Statistical Theory and Practice, vol. 6, issue 1, pp. 220-238, 2012.

[14] S.Litsyn and V. Shevelev, "On ensembles of low-density parity-check codes: asymptotic distance distributions," IEEE Trans. Inform. Theory, vol. IT-48, pp.887–908, Apr. 2002.

[15] D.J.C.MacKay, "Good error-correcting codes based on very sparse matrices," IEEE Trans. Inform. Theory, vol. IT-45 (2), pp.399-431, 1999.

[16] D.Malioutov and M. Malyutov, "Boolean compressed sensing: LP relaxation for group testing, " IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[17] M.Mézard and C. Toninelli, "Group teasing with random pools: optimal two-stage algorithms, " IEEE Trans. Inform. Theory, vol. IT-57 (3), pp.1736–1745, 2011.

[18] D. Sejdinovic and O.Johnson, "Note on noisy group testing: asymptotic bounds and belief propagation reconstruction," Forty-Eight Annual Allerton Conference on Communication, Control, and Computing, pp. 998-1003, 2010.