Pool Size Control for Adaptive Group Testing via Boolean Compressed Sensing with Solution Space Reduction

Riho Kawasaki, Kazunori Hayashi and Megumi Kaneko
Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {riho.k, kazunori, meg}@sys.i.kyoto-u.ac.jp

***Abstract*—This paper proposes a new method of adaptive group testing via Boolean compressed sensing. The proposed scheme utilizes solution space reduction at each step using the idea of classical sequential group testing and the pool size for each test is controlled to maximize the expected number of identifiable items by the test. Moreover, for the pool size control, the cardinality of the remaining positive items is directly estimated by using probabilistic zero estimator (PZE), which has been originally proposed for RFID systems. The performance gains of the proposed method against conventional non-adaptive and adaptive group testing methods are demonstrated through computer simulations.**

I. Introduction

Group testing is a method for specifying a small subset of defective (or positive) items in a large set of items efficiently. Each test consists of selecting some items to construct a 'pool', which is a mixture of the selected items, and testing to tell whether at least one positive item is in the pool or not. After some tests, the estimated positive items are obtained by some algorithm using the test results. It is known that, with properly designed pools and algorithm, a smaller number of tests than that of all items is sufficient to detect all positive items. A combinatorial group testing technique was first proposed by Dorfman during WWII for finding out infections from a large number of soldiers by a small number of blood tests [1]. Since then, group testing has been applied to many areas such as blood screening, DNA sequencing [2], and network security [3]. In general, group testing can be classified into non-adaptive and adaptive. The non-adaptive group testing prepares all the pools a priori. Thus, it can complete the test within one round if the tests can be conducted in parallel. On the other hand, the adaptive group testing constructs the pool of each test by using the results of previous tests. Therefore, each test has to be performed one by one, but it requires fewer tests than non-adaptive group testing in general.

In recent years, group testing has attracted interest from the active research area of compressed sensing [4], where the goal is to infer a sparse high-dimensional vector from a small set of linear measurements. Using the idea of compressed sensing, Malioutov and Malyutov have proposed Boolean compressed sensing, which solves non-adaptive group testing through linear programming relaxation [5]. Moreover, Boolean compressed sensing has been applied to adaptive group testing [6]. In [6], the basic idea of the pool construction for each step is to select the pool size so as to maximize the amount of information obtained in the next test. In order to evaluate the amount of information, the number of remaining positive items is estimated from the tentative reconstruction result obtained by Boolean compressed sensing using the available test results at the moment. Since the estimated number might be unreliable especially for the beginning of the testing, the impact of the estimation error is also taken into consideration in [6], however, this causes another new problem of the estimation of the error probability, which can be challenging in general.

In this paper, we propose a new method of adaptive group testing based on Boolean compressed sensing. The proposed method introduces solution space reduction at each test using the idea of classical sequential group testing, where all the items in the pools for which test results are negative (negative pools) can be identified as negative and the item in the positive pools with size one as positive, while those in positive pools with size more than one require further tests. With the solution space reduction, the reconstruction performance by Boolean compressed sensing can be improved because some elements of the unknown sparse vector are determined exactly and the size of the sparse vector is reduced. Moreover, we propose a novel pool size control scheme, where the pool size is selected so as to maximize the expected number of reducible items with the proposed solution space reduction in the next test. Furthermore, for the evaluation of the expected number of reducible items, the cardinality of remaining positive items is directly estimated by using probabilistic zero estimator (PZE), which has been originally proposed for radio frequency identifier (RFID) systems [7], instead of using the number of nonzero elements in the tentative reconstruction results as in [6]. We evaluate the performance of the proposed method and conventional non-adaptive and adaptive group testing method through computer simulations, and compare their performances by the rate of exact recovery at each test iteration. The results indicate that the proposed solution space reduction and novel pool size control scheme contribute to reducing the required number of tests to achieve reliable reconstruction.

II. Problem Formulation

Suppose N to be the number of all items, with only K of which are positive. We define a Boolean vector $\mathbf{x} =$

$[x_1 x_2 \cdots x_N]^T \in \{0,1\}^N$, where $x_n = 1$ or 0 indicates that the n-th item is positive or negative, respectively. In each test, some items are chosen from the set of items and are mixed to obtain a so-called 'pool'. The pool for the s-th test is defined by a Boolean row vector $\boldsymbol{a}_s = [a_{s,1} a_{s,2} \cdots a_{s,N}] \in \{0,1\}^N$, where $a_{s,n} = 1$ or 0 indicates that the n-th item belongs to the pool for the s-th test or not, respectively, and a $S \times N$ Boolean matrix $\mathbf{A}_S = [\boldsymbol{a}_1^T \boldsymbol{a}_2^T \cdots \boldsymbol{a}_S^T]^T$ defines the pools for S tests. The test result of the s-th test is a single Boolean value $y_s \in \{0,1\}$, where $y_s = 1$ indicates that at least one positive item is included in the pool for the s-th test while $y_s = 0$ indicates that no positive item is in it. Thus a test result y_s is obtained by taking the Boolean sum of $\{x_n | a_{s,n} = 1\}$, which is the set of x_n belonging to the pool for the s-th test. Accordingly, the observation model is represented by

$$y_s = \bigvee_{n=1}^{N} (a_{s,n} \wedge x_n), \tag{1}$$

where $\bigvee_{n=1}^{N} z_n$ denotes the Boolean sum of $z_1, z_2, \ldots, z_N$, and $\wedge$ the Boolean AND operation. For convenience, we define $\mathbf{y}_S = [y_1, y_2, \ldots, y_S]^{\mathrm{T}}$ as the vector of test results obtained by S tests, which is given by

$$\mathbf{y}_S = \mathbf{A}_S \vee \mathbf{x}, \tag{2}$$

where $\mathbf{A}_S \vee \mathbf{x}$ indicates taking Boolean product of corresponding elements of each row of $\mathbf{A}_S$ and $\mathbf{x}$, and then taking Boolean sum of them. The goal of group testing is to estimate the unknown sparse Boolean vector $\mathbf{x}$ from the Boolean matrix $\mathbf{A}_S$ and the test results vector $\mathbf{y}_S$.

For the case of adaptive group testing, the pool for the next test is determined based on all the results of previous tests. Specifically, the row vector $\boldsymbol{a}_{S+1}$, which defines the pool for the $(S+1)$-th test, is determined by using the observed signals y_s $(s = 1, 2, \ldots, S)$ and pooling vectors $\boldsymbol{a}_s$ $(s = 1, 2, \ldots, S)$. In the rest of the paper, $\hat{\mathbf{x}}_S$ indicates the estimate of $\mathbf{x}$ reconstructed from S test results.

III. Conventional Group Testing Method

A. Sequential group testing

Sequential group testing is one of classical adaptive group testing methods, and is based on the idea that all the items in negative pools are identified as negative, while those in positive pools require further test. For instance, assume that we separate all the items into two groups to obtain two pools and test them. If the test results is negative, all the items in the corresponding pool can be identified as negative. If the test result is positive, we divide the items in the corresponding pool into two disjoint sets to have smaller pools and conduct further tests on them. After repeating these operations, an item in a positive pool with size one is identified as positive. The worst-case number of tests required for N items having K positive items is known to be $\lceil \log_2 \binom{N}{K} \rceil \simeq K \log_2(N/K)$, which gives the well-known information theory bound [3].

B. Non-Adaptive group testing via Boolean compressed sensing

Malioutov and Malyutov proposed Boolean compressed sensing, where the non-adaptive group testing is modified into a linear programming problem through relaxing the elements of the vector $\mathbf{x}$ into real numbers [5]. The problem of Boolean compressed sensing for non-adaptive group testing is formulated as

$$\hat{\mathbf{x}}_S = \arg\min_{\mathbf{x}} \sum_n x_n$$
$$\text{such that: } \mathbf{0} \le \mathbf{x} \le \mathbf{1}, \quad \mathbf{A}_S^{(\mathcal{P})} \mathbf{x} \ge \mathbf{1}, \quad \mathbf{A}_S^{(\mathcal{N})} \mathbf{x} = \mathbf{0}, \tag{3}$$

where $\mathbf{A}_S^{(\mathcal{P})}$ and $\mathbf{A}_S^{(\mathcal{N})}$ are the matrices constructed by the rows of $\mathbf{A}_S$ corresponding to positive and negative tests, respectively. It is known that an upper bound on the number of tests required to recover $\mathbf{x}$ with small error probability is $O(K \log(N))$.

C. Adaptive group testing via Boolean compressed sensing

Boolean compressed sensing described above has been applied for the problem of adaptive group testing in [6], where an information-based pool size control scheme is proposed. In the method, the expected information obtained in the $(S+1)$-th test is introduced as

$$I_{S+1}(G) = q_{\mathrm{N}} I_{\mathrm{N}} + (1 - q_{\mathrm{N}}) I_{\mathrm{P}}, \tag{4}$$

and the pool size G is set to maximize this value. Here, q_{N} is the probability that the result of the $(S+1)$-th test is negative, I_{N} is the information obtained from the negative test, and I_{P} is the information from the positive test. Specifically, they are represented as

$$q_{\mathrm{N}} = \frac{\binom{N-K}{G}}{\binom{N}{G}}, \tag{5}$$

$$I_{\mathrm{N}} = G\{-r \log r - (1-r)\log(1-r)\}, \tag{6}$$

$$I_{\mathrm{P}} = -(1-r)^G \log(1-r)^G - \{1-(1-r)^G\} \log\{1-(1-r)^G\}, \tag{7}$$

where $r = K/N$ is assumed to be the probability that each item is positive.

In order to evaluate (4), the number of positive items K has to be estimated, since it is unknown in general. In [6], K is estimated by using the tentative reconstruction result of $\hat{\mathbf{x}}_S$, namely, $\hat{K} = ||\hat{\mathbf{x}}_S||_0$. However, $\hat{\mathbf{x}}_S$ may include some estimation errors, especially when the result is based on a small number of tests. Therefore, the objective function (4)

was revised to cope with such estimation errors as

$$\bar{I}_{S+1}(G) = \sum_{\{K'|K'=||\hat{\mathbf{x}}_S||_0-a+b\}} I_{S+1}(G) \times \begin{pmatrix} ||\hat{\mathbf{x}}_S||_0 \\ a \end{pmatrix} \epsilon^a (1-\epsilon)^{||\hat{\mathbf{x}}_S||_0-a} \times \begin{pmatrix} N-||\hat{\mathbf{x}}_S||_0 \\ b \end{pmatrix} \epsilon^b (1-\epsilon)^{N-||\hat{\mathbf{x}}_S||_0-b}, \quad (8)$$

where ϵ is the estimation error probability of each element of $\hat{\mathbf{x}}_S$, a and b are the number of false positives and false negatives. Using G, which maximizes $\bar{I}_{S+1}(G)$, each element of $\boldsymbol{a}_{S+1}$ is generated independently as a realization of a Bernoulli random variable with probability $p = G/N$. However, the revised objective function requires the information of the estimation error probability ϵ, which would be difficult to obtain. Moreover, probabilities of false positive and false negative are supposed to be the same, but this will not be true in general.

IV. Proposed Method

The proposed method has three key advantages. We explain each of them in the following sections.

A. Solution space reduction

If test results are not deteriorated by noise, by using the idea of classical sequential group testing, we can identify all the items in the negative pools as negative and the item in the positive pools with size one as positive. Using these properties, the elements of the unknown vector which correspond to the identified items can be determined, and the size of the unknown vector composed by "candidate items", namely, items to be further tested, can be reduced. The improvement of the performance of reconstruction by Boolean compressed sensing could be expected with this manipulation, because the required number of tests to achieve reliable reconstruction is monotonically increasing for N from the upper bound, $O(K\log(N))$.

B. Pool size control

For efficient reconstruction of the unknown vector, the pool size of the next test has to be appropriately selected based on the information of the previous test results. In the proposed scheme, making much account of effective solution space reduction, we employ the expected number of reducible items with the solution space reduction in the next test as the criterion for the pool size control. Thus, we select G in order to maximize the expected number of reducible items, which can be calculated as

$$J_{S+1}(G) = q'_{\mathrm{N}} G, \quad (9)$$

where q'_{N} is the probability that the $(S+1)$-th test is negative, which is given by

$$q'_{\mathrm{N}} = \frac{\begin{pmatrix} N_{S+1}-K \\ G \end{pmatrix}}{\begin{pmatrix} N_{S+1} \\ G \end{pmatrix}}. \quad (10)$$

Note that N_{S+1} is the number of candidate items for the $(S+1)$-th test. The pool size G for the $(S+1)$-th test is determined to maximize $J_{S+1}(G)$, and each element of the row vector $\boldsymbol{a}_{S+1}$ defining the pool for the $(S+1)$-th test is generated by selecting G elements randomly out of N_{S+1} candidate elements by setting 1 for them, and 0 for the other elements.

C. Cardinality estimation

Since the proposed pool size control algorithm requires the number of positive items K, which is unknown in general, we run an algorithm to directly estimate K from the test results at each step in parallel with the adaptive group testing. Specifically, we apply the probabilistic zero estimator (PZE), which has been originally proposed for RFID systems [7], instead of using the number of nonzero elements in the tentative reconstruction results as in [6].

The probability p_S that each item is included in the pool for the S-th test is represented by $p_S = G/N_S$, where G is the number of items in the pool. If the number of positive items is K, the probability q_S that the pool is negative equals to the probability that all the K positive items are not in the pool, represented as

$$q_S = (1-p_S)^K. \quad (11)$$

Therefore, the estimated value of K, $\hat{K}$, is calculated by

$$\hat{K} = \frac{\log Q}{\log(1-p_S)}, \quad (12)$$

where Q is the rate that the test results are negative. When we estimate K using m test results, $Q = ||\bar{\mathbf{y}}||_0/m$, where $\bar{\mathbf{y}}$ is the vector whose elements are the m Boolean values of the tests. $p_S = \min(1, 1.59/K)$ is known as the optimum value of p_S for PZE. However, we cannot use this value, because the pool size control for the group testing is prioritized over the estimation of K, i. e., p_S is determined based on G that maximizes (9). In the proposed method, K is estimated by (12) every m tests, and the estimated value of K to be used for pool size control is calculated by taking an average of with all previously estimated values.

The process of the proposed method is shown in Figure 1. Given the initial value $N_1 = N$ and $p_1 = \min(1, 1.59/K_0)$, where K_0 is initially set to a random value[1], a trial of the proposed group testing is started. First, we construct a pooling vector $\boldsymbol{a}_S$ and get a test result y_S, then N_{S+1} is updated at

[1] K_0 has been set to 5 in our simulations for all true values $K = 2, 6, 10$. In all cases, the proposed method can outperform conventional methods and hence is not sensitive to the setting of K_0.

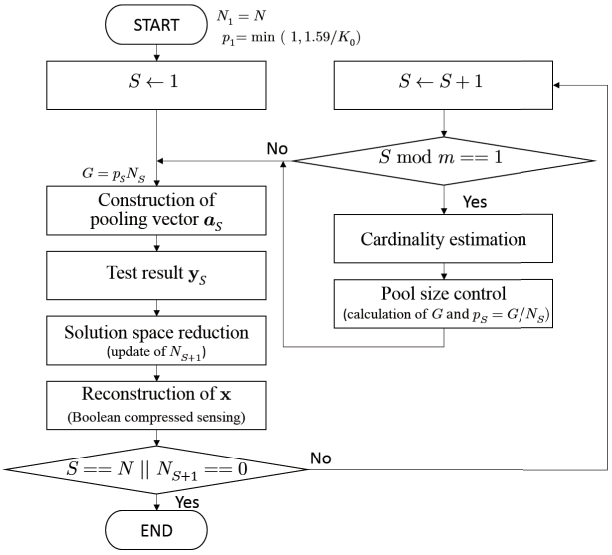

Fig. 1. Flowchart of the proposed method

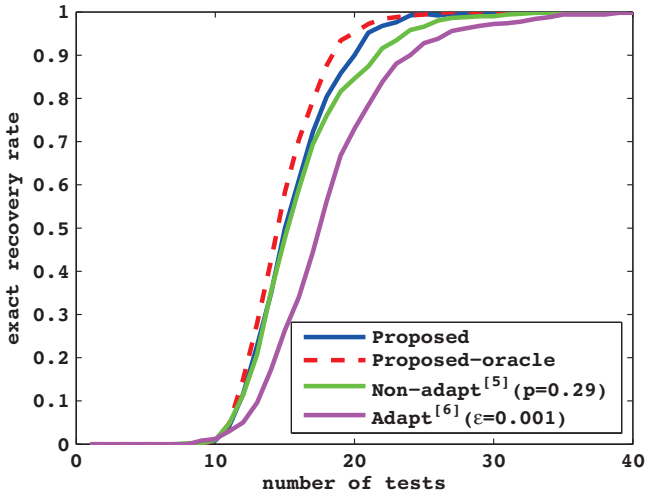

Fig. 2. The rate of the exact recovery versus the number of tests ($N = 150, K = 2$)

the solution space reduction step and the unknown vector $\mathbf{x}$ is estimated using Boolean compressed sensing. Every m tests, K is re-estimated at the cardinality estimation step, and the pool size G is calculated at the pool size control step. After repeating these operations, if all N tests have been conducted or no candidate items remain, the algorithm is finished.

V. Numerical Experiments

The performance of the proposed adaptive group testing scheme has been evaluated through computer simulations. We compare the performance of the proposed method and of conventional methods in terms of the rate of exact recovery versus the number of tests. Specifically, suppose that the total number of items is $N = 150$, the exact recovery rate is defined as the rate that the unknown vector $\mathbf{x}$ is recovered correctly averaged over 500 trials at each number of tests $S = 1, 2, \ldots, 150$. N items are randomly generated for each trial so that K of them are positive. To evaluate the impact on the performance of the number of positive items K, we have conducted simulations for three cases: $K = 2$, 6 and 10. The simulation results for each case are shown in Figures 2, 3 and 4, respectively. In the proposed method (*Proposed*), we estimate K every $m = 10$ tests using PZE, and update the pool size G by maximizing (9). *Proposed-oracle* is the result of the proposed method assuming the perfect knowledge of K, that is, there is no need to estimate K and it updates the pool size G by maximizing (9) at each test. *Non-adapt* stands for the non-adaptive group testing via Boolean compressed sensing, where the measurement matrix $\mathbf{A}_S$ is an i.i.d. Bernoulli random matrix with probability p_o of having 1 in each entry $a_{i,j}$. The probability p_o with the best performance is selected for each K, namely $p_o = 0.29$, 0.14 and 0.09 for $K = 2$, 6, and 10, respectively. *Adapt* means the adaptive group testing shown in III-C selecting the pool size G by maximizing (8). The probability of the estimation error ϵ were obtained via simulations, and we use $\epsilon = 0.001$, 0.03 and 0.04 for $K = 2$, 6, and 10, respectively. Note that *Proposed-oracle* and *Non-adapt* are assumed to have the perfect information of K while *Proposed* and *Adapt* do not. From the figures, we can see that the proposed method outperforms the conventional methods in all cases. Especially, the performance gain against the conventional methods is greater for larger number of positive items, while the performance gap with the case of *Proposed-oracle* is also larger. This indicates that more accurate estimation of the cardinality of remaining positive items could improve the performance of the proposed method especially for larger K.

We have finally evaluated the number of tests required to achieve 99% exact reconstruction of $\mathbf{x}$ versus the number of tests used in each cardinality estimation, namely m, in Figure 5. In case of $K = 2$, the results are almost the same for all m, while for larger K such as $K = 6$ and $K = 10$, the case with $m = 10$ has shown the best performance.

VI. Conclusion

We have proposed a novel method of adaptive group testing via Boolean compressed sensing, which adopts solution space reduction at each test from the idea of classical sequential group testing. In the proposed method, the pool size is controlled to maximize the expected number of reducible items with the proposed solution space reduction in the next test. Moreover, for the evaluation of the expected number of reducible items, the cardinality of remaining positive items is directly estimated by using PZE, which has been originally proposed for RFID systems. We evaluate the performance of the proposed method and conventional non-adaptive and adaptive group testing methods through computer simulations and the validity of the proposed method is confirmed.

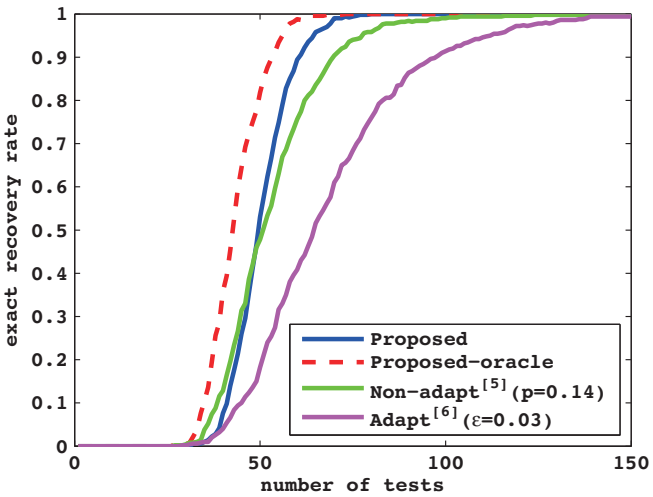

Fig. 3. The rate of the exact recovery versus the number of tests $(N = 150, K = 6)$

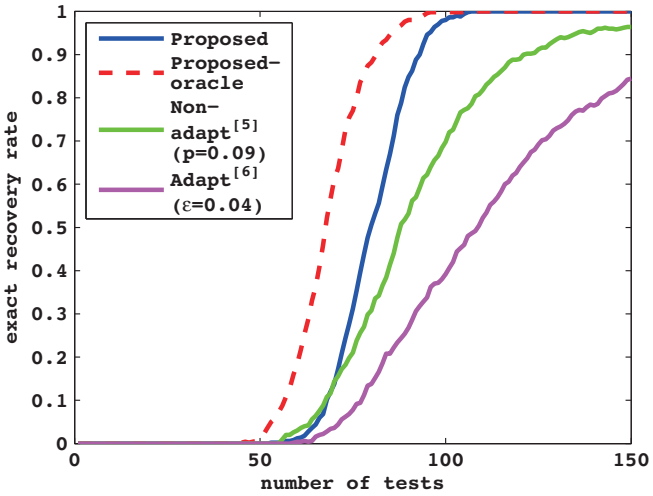

Fig. 4. The rate of the exact recovery versus the number of tests $(N = 150, K = 10)$

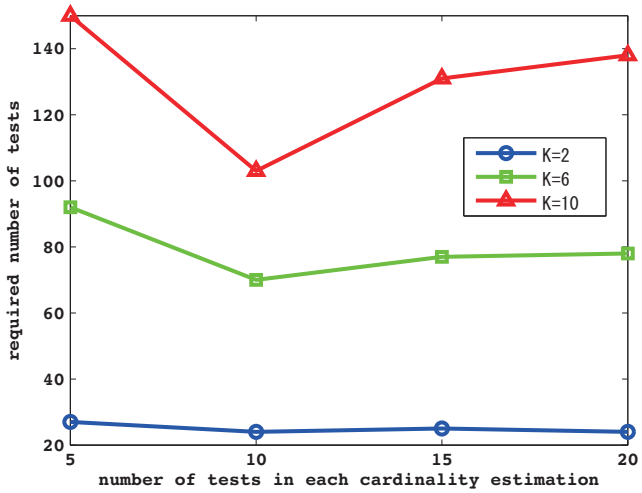

Fig. 5. The number of tests required for reliable reconstruction versus number of tests used in each cardinality estimation m

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers 26820143, 15K06064, and 15H2252.

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, no. 6, pp. 436-440, Dec. 1943.
[2] D. Du and F. K. Hwang, *Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing* Series on Applied Mathematics Vol. 18, World Scientific Publishing, Jun. 2006.
[3] M. T. Thai, *Group Testing Theory in Network Security*, Springer New York, 2012.
[4] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, Vol. 25, Issue 2, pp. 21-30, Mar. 2008.
[5] D. Malioutov and M. Malyutov, "Boolean compressed sensing: LP relaxation for group testing," in *Proc. ICASSP*, Mar. 2012.
[6] Y. Kawaguchi *et al.*, "Information-based pool size control of Boolean compressive sensing for adaptive group testing," in *Proc. EUSIPCO*, Sept. 2014.
[7] M. Kodialam and T. Nandagopal, "Fast and Reliable Estimation Schemes in RFID Systems,"in *Proc. MobiCom*, pp.322-333, Sept. 2006.
[8] W. H. Kautz and R. C. Singleton, "Nonrandom Binary Superimposed Codes," *Information Theory, IEEE Trans*. Vol. 10, Issue 4, pp. 363-377, Oct. 1964.
[9] A. G. Dyachkov and V. V. Rykov, "A Survey of Superimposed code theory," Problems of Control and Information Theory, Vol. 12, Issue 4, pp. 1-13, 1983.