INFORMATION-BASED POOL SIZE CONTROL OF BOOLEAN COMPRESSIVE SENSING FOR ADAPTIVE GROUP TESTING

Yohei Kawaguchi, Tatsuhiko Osa, Shubhranshu Barnwal,
Hisashi Nagano, and Masahito Togami

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji, Tokyo 185-8601, Japan
yohei.kawaguchi.xk@hitachi.com

ABSTRACT

A new method for solving the adaptive-group-testing problem is proposed. To solve the problem that the conventional method for non-adaptive group testing by Boolean compressive sensing needs a larger number of tests when the pool size is not optimized, the proposed method controls the pool size for each test. The control criterion is the expected information gain that can be calculated from the ℓ_0 norm of the estimated solution. Experimental simulation indicates that the proposed method outperforms the conventional method even when the number of defective items is varied and the number of defective items is unknown.

***Index Terms*—** adaptive group testing, compressive sensing, information gain, entropy, sparse signal processing

1. INTRODUCTION

Group testing is the well-known problem that attempts to discover a sparse subset of defective items in a large set of items by using a small number of tests. Each test consists of three processing steps: (1) selecting items for a pool on the basis of a certain method, (2) mixing the selected items into the pool, and (3) observing a single Boolean result by testing the pool. When the proportion of defective items is small, a small number of the tests on the mixed pool is sufficient to detect the defective items; that is, all the items need not be tested directly. Group testing as a subject dates back to the work of Dorfman [1] in 1943, during the Second World War. Dorfman developed this approach in order to test soldiers' blood for syphilis. Group testing has applications such as blood screening, deoxyribonucleic acid (DNA) sequencing, and anomaly detection in computer networks [2].

Traditionally, group testing has been regarded as a combinatorial problem. As for this problem, many researches about the upper and lower bounds on the number of tests required to find all the defective items have been done. A set of information-theoretic bounds for group testing with random mixing was established by Malyutov [3, 4], Atia and Saligrama [5], Sejdinovic and Johnson [6], and Aldridge *et al.* [7]. In addition, several tractable approximation algorithms, such as one based on belief propagation [6] and one based on matching pursuit [8], have been proposed.

In recent years, group testing has drawn interest from the active research area of compressive sensing. Compressive sensing solves a kind of underdetermined linear equation, namely, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is an unknown high-dimensional vector to be estimated, $\boldsymbol{A}$ is a given mixing matrix, and $\boldsymbol{y}$ is a given low-dimensional observed vector. The problem with compressive sensing is similar to that with group testing from the viewpoint that both of them are underdetermined problems such that an unknown high-dimensional vector is decoded from an observed low-dimensional vector. However, while compressive sensing is defined in a real vector space, group testing is defined in a Boolean vector space. To improve the performance of group testing by using compressive sensing, Malioutov and Malyutov [9] proposed a method for converting group testing into compressive sensing through linear-programming relaxation. As for this conversion method, ℓ_1 minimization imposes the sparsity constraint to the solution and solves the uncertainty of the underdetermined problem. It thus outperforms other conventional methods (i.e., the method based on belief propagation [6], the method based on matching pursuit [8], etc.). However, the conventional method is defined in non-adaptive group testing, which has the drawback that it cannot choose the pool for each test based on observation data. In particular, the optimal size of the pool depends on the number of defective items, and the number of defective items is unknown; therefore, in the case that Malioutov's method is applied, a larger number of tests is required when the pool size is not optimized.

To reduce the number of tests of Malioutov's method, a method for adaptive group testing is proposed here. The proposed method controls the pool size for each test. The criterion of the control is the expected information gain that can be calculated from the ℓ_0 norm of the estimated solution. Simulation results indicate that the proposed method outperforms the conventional method even under the condition that the number of defective items is varied and the number of de-

fective items is unknown.

2. PROBLEM STATEMENT

To state the problem, first, the following notation is fixed. N is the number of items, of which a subset of size K is defective. Defective items are called "positive", and non-defective items are called "negative". $x_n = 1$ indicates that the n-th item is positive, and $x_n = 0$ indicates that the n-th item is negative. For convenience, $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]^T$ is written. T tests, where $T < N$, are then performed. As explained above, in each test, some items are selected from all the items, and they are mixed into the same pool. This selection is defined by a mixing matrix, $\boldsymbol{A}$, which is a $T \times N$ binary matrix. The element of the t-th row and the n-th column of $\boldsymbol{A}$ is given as a_{tn}, where $a_{t,n} = 1$ indicates that the n-th item is mixed into the pool of the t-th test, and $a_{t,n} = 0$ indicates that the n-th item is not mixed into the pool of the t-th test. The observed signal of each test, t, is a single Boolean value, $y_t \in \{0, 1\}$. y_t is obtained by taking the Boolean sum of $\{x_n | a_{tn} = 1\}$. For convenience, $\boldsymbol{y} = [y_1, y_2, \cdots, y_T]^T$ is written. The vector notation

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{1}$$

is used in the following.

The problem of group testing is to estimate unknown vector $\boldsymbol{x}$ from given $\boldsymbol{A}$ and $\boldsymbol{y}$. In addition, the noise of the observation is considered. The noise includes both the false positive and the false negative. The former represents the case that $y_t = 1$ even when the Boolean sum of $\{x_n | a_{tn} = 1\}$ is 0. The latter represents the case that $y_t = 0$ even when the Boolean sum of $\{x_n | a_{tn} = 1\}$ is 1. This observation with noise is represented by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \otimes \boldsymbol{v}, \tag{2}$$

where $\boldsymbol{v}$ is the Boolean vector of errors, and $\otimes$ means the XOR operation.

A number of works have studied the design of $\boldsymbol{A}$ [2]. For example, K-separating and K-disjunct are well-known properties of $\boldsymbol{A}$. When these properties hold, $\boldsymbol{x}$ can be recovered exactly. However, such design is often unsuitable for practical situations because it assumes that the exact number of the positive items (K) before group testing. Moreover, if all T tests cannot be carried out, the performance of the method will not be guaranteed [7]. Therefore, in many works, $\boldsymbol{A}$ is simply designed by the Bernoulli random design, where each element of $\boldsymbol{A}$ is generated independently at random with a probability p corresponding with the size of the pool. That is, a_{tn} is 1 with probability p, and a_{tn} is 0 with probability $1-p$. Bernoulli random design is also used in this study.

One of the problems of non-adaptive group testing is that optimal probability p largely depends on K, although the number of positive items is unknown. The present study thus focuses on adaptive group testing. In adaptive group testing, the mixing vector of the next test after each observation is determined as follows:

$$\boldsymbol{a}_{T+1} = \begin{bmatrix} a_{T+1,1} \\ \vdots \\ a_{T+1,N} \end{bmatrix} = f_T\left(\boldsymbol{A}, \boldsymbol{y}\right), \tag{3}$$

where f_T is a function to determine the next mixing vector after the T-th test. In addition, each test is assumed to have a Bernoulli random design, and control of p is focused on. Accordingly, (3) can be rewritten as

$$a_{T+1,n} \overset{\text{i.i.d}}{\sim} p_{T+1} = g_T\left(\boldsymbol{A}, \boldsymbol{y}\right), \tag{4}$$

where g_T is a function to determine the next Bernoulli probability, p_{T+1}, after the T-th test. In Section 4, a new g_T is proposed.

3. BOOLEAN COMPRESSIVE SENSING FOR GROUP TESTING

3.1. Compressive sensing

Malioutov and Malyutov [9] proposed a conversion of group testing into compressive sensing through a linear-programming relaxation. This conventional method is the basis of our method, which is explained in this section.

Many works on compressive sensing have been reported [10]. In this study, a sparse signal, $\boldsymbol{x} \in \mathbb{R}^N$, is assumed, and it is estimated from M measurements $\boldsymbol{y} \in \mathbb{R}^T$ by using a random measurement matrix $\boldsymbol{A}$, where $M < N$. Compressive sensing, namely, decoding $\boldsymbol{x}$, uses the following ℓ_0 minimization:

$$\min_{\boldsymbol{x}} |\boldsymbol{x}|_0 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{5}$$

However, Eq. (5) is a NP-hard problem, which cannot be solved practically. Candes et al. [10] proved that if certain conditions hold, $\boldsymbol{x}$ can be decoded exactly by the following ℓ_1 minimization:

$$\min_{\boldsymbol{x}} |\boldsymbol{x}|_1 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{6}$$

Since ℓ_1 minimization is a simple linear-programming problem, a number of practicable algorithms can be used to solve it.

3.2. Noise-free case

Equation (1) is similar to constraint equation (6). However, it is not a linear equation in a real vector space but a Boolean equation. It is shown in [9] that (1) can be replaced with a closely related linear formulation: $\mathbf{1} \leq \boldsymbol{A}_{\mathcal{I}}\boldsymbol{x}$, and $\mathbf{0} = \boldsymbol{A}_{\mathcal{J}}\boldsymbol{x}$, where $\mathcal{I} = \{t | y_t = 1\}$ is the set of positive test results, and $\mathcal{J} = \{t | y_t = 0\}$ is the set of negative test results. A linear-programming formulation similar to Eq. (6) is therefore given

as

$$\begin{aligned} &\min_{\boldsymbol{x}} \left\{ \sum_n x_n \right\} \\ &\text{subject to} \quad \mathbf{0} \le \boldsymbol{x} \le \mathbf{1}, \\ &\boldsymbol{A}_{\mathcal{I}}\boldsymbol{x} \ge \mathbf{1}, \quad \boldsymbol{A}_{\mathcal{J}}\boldsymbol{x} = \mathbf{0} \end{aligned} \tag{7}$$

3.3. Noisy case

Because (7) does not model noisy cases, the performance of the method is degraded in noisy cases. One version of [9]'s method thus covers the noisy case by adding slack variables as follows:

$$\begin{aligned} &\min_{\boldsymbol{x},\boldsymbol{\xi}} \left\{ \sum_n x_n + \alpha \sum_t \xi_t \right\} \\ &\text{subject to} \quad \mathbf{0} \le \boldsymbol{x} \le \mathbf{1}, \quad \mathbf{0} \le \boldsymbol{\xi}_{\mathcal{I}} \le \mathbf{1}, \quad \mathbf{0} \le \boldsymbol{\xi}_{\mathcal{J}}, \\ &\boldsymbol{A}_{\mathcal{I}}\boldsymbol{x} + \boldsymbol{\xi}_{\mathcal{I}} \ge \mathbf{1}, \quad \boldsymbol{A}_{\mathcal{J}}\boldsymbol{x} = \boldsymbol{\xi}_{\mathcal{J}}, \end{aligned} \tag{8}$$

where $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_T]$ is the vector composed of the slack variables, and α is the regularization parameter that balances the amount of noise and the sparsity of the solution.

4. PROPOSED METHOD

The proposed method for controlling Bernoulli probability p in adaptive compressive sensing is described as follows. Expected information gain of the next $(T+1)$-th test is introduced as

$$I_{T+1}(p) = q_N I_N + (1 - q_N) I_P, \tag{9}$$

where $I_{T+1}(p)$ is the expected information gain for Bernoulli probability p, q_N is the probability that the result of the $(T+1)$-th test is negative, I_N is the information gain of the negative test, and I_P is the information gain of the positive test. The negative test means that all the items of the pool are negative, so q_N is given by

$$q_N = \frac{\left(\begin{array}{c} N - |\boldsymbol{x}|_0 \\ G \end{array} \right)}{\left(\begin{array}{c} N \\ G \end{array} \right)}, \tag{10}$$

where G is the size of the pool, namely, the number of the non-zero elements of $\boldsymbol{a}_{T+1}$. The negative test gives the information that all the items of the pool are negative, so I_N is the sum of the current entropy of the G items of the pool; therefore, I_N is given by

$$I_N = G\left\{ -r \log r - (1-r)\log(1-r) \right\}, \tag{11}$$

where $r = |\boldsymbol{x}|_0 / N$ is the probability that each item is positive. The positive test gives the information that there is at least one positive item in the pool; therefore, I_P is given as

$$I_P = \left\{ -r^G \log r^G - (1 - r^G)\log(1 - r^G) \right\}, \tag{12}$$

The estimate of $\boldsymbol{x}$, $\hat{\boldsymbol{x}}$, is obtained by using T tests.

Using the ℓ_0 norm of $\hat{\boldsymbol{x}}$, $|\hat{\boldsymbol{x}}|_0$ makes it possible to optimize p by maximizing $I_{T+1}(p)$ of (9). However, $\hat{\boldsymbol{x}}$ may include an estimation error because $\hat{\boldsymbol{x}}$ is only a temporary result based on a small number of tests. The control of p is degraded by the estimation error; therefore, the objective function (9) is revised in consideration of the estimation error as follows:

$$\begin{aligned} \bar{I}_{T+1}(p) \quad = \quad & \sum_{\left\{K' | K' = |\hat{\boldsymbol{x}}|_0 - a + b\right\}} I_{T+1}(p) \\ & \times \left(\begin{array}{c} K' \\ a \end{array} \right) \epsilon^a (1-\epsilon)^{K'-a} \\ & \times \left(\begin{array}{c} N - K' \\ b \end{array} \right) \epsilon^b (1-\epsilon)^{N-K'-b}, \end{aligned} \tag{13}$$

where ϵ is the probability of the estimation error, a is the number of the false-positive items, and b is the number of the false-negative items. p can be optimized by maximizing (13).

The convergence of the above-described adaptation of the case of no noise is discussed as follows. $\hat{\boldsymbol{x}}_T$ is defined as the estimates of $\boldsymbol{x}$ by using T tests. $\hat{\boldsymbol{x}}_{T+1}$ is defined as the estimates of $\boldsymbol{x}$ by using $(T+1)$ tests. $\hat{\boldsymbol{x}}_T$ is given by Eq. (7). When the $(T+1)$-th test is positive, $\hat{\boldsymbol{x}}_{T+1}$ is given by

$$\begin{aligned} \hat{\boldsymbol{x}}_{T+1} = \arg\min_{\boldsymbol{x}} &\left\{ \sum_n x_n \right\} \\ &\text{subject to} \quad \mathbf{0} \le \boldsymbol{x} \le \mathbf{1}, \\ &\boldsymbol{A}_{\mathcal{I}}\boldsymbol{x} \ge \mathbf{1}, \quad \boldsymbol{A}_{\mathcal{J}}\boldsymbol{x} = \mathbf{0}, \\ &{\boldsymbol{a}_{\boldsymbol{T+1}}}^T \boldsymbol{x} \ge 1. \end{aligned} \tag{14}$$

Because (14) is (7) with an additional constraint, i.e., ${\boldsymbol{a}_{T+1}}^T \boldsymbol{x} \ge 1$, $|\hat{\boldsymbol{x}}_T|_0 \le |\hat{\boldsymbol{x}}_{T+1}|_0$. By the additional constraint, $|\hat{\boldsymbol{x}}_{T+1}|_0$ may increase from $|\hat{\boldsymbol{x}}_T|_0$ by no more than one. When the $(T+1)$-th test is negative, $\hat{\boldsymbol{x}}_{T+1}$ is given by

$$\begin{aligned} \hat{\boldsymbol{x}}_{T+1} = \arg\min_{\boldsymbol{x}} &\left\{ \sum_n x_n \right\} \\ &\text{subject to} \quad \mathbf{0} \le \boldsymbol{x} \le \mathbf{1}, \\ &\boldsymbol{A}_{\mathcal{I}}\boldsymbol{x} \ge \mathbf{1}, \quad \boldsymbol{A}_{\mathcal{J}}\boldsymbol{x} = \mathbf{0}, \\ &{\boldsymbol{a}_{T+1}}^T \boldsymbol{x} = 0. \end{aligned} \tag{15}$$

$|\hat{\boldsymbol{x}}_{T+1}|_0$ does not increase from $|\hat{\boldsymbol{x}}_T|_0$ because $|\hat{\boldsymbol{x}}_T|_0$ has been already minimized at the time of the T-th test. From the above, $|\hat{\boldsymbol{x}}_T|_0 \le |\hat{\boldsymbol{x}}_{T+1}|_0$, and $|\hat{\boldsymbol{x}}_T|_0$ weakly monotonically increases as T increases. In addition, it is obvious that $|\hat{\boldsymbol{x}}|_0 \le |\boldsymbol{x}|_0$ because $|\hat{\boldsymbol{x}}|_0$ is minimized under the constraints that also holds for $\boldsymbol{x}$, so $|\boldsymbol{x}|_0$ is an upperbound of $|\hat{\boldsymbol{x}}_T|_0$. Therefore, $|\hat{\boldsymbol{x}}_T|_0$ moves to $|\boldsymbol{x}|_0$, and the adaptation of the pool size converges as T increases.

5. EXPERIMENTAL RESULTS

The performance of the proposed method was evaluated by simulation. In particular, the averaged probability of correct

estimation was computed over 100 trials as a function of T, for $N = 150$. N items were generated independently for each trial. In this simulation, $\hat{\boldsymbol{x}} = \boldsymbol{x}$ was considered to be the correct case. The proposed method was compared with the non-adaptive conventional method [9] . To evaluate the robustness against the difference in the number of positive items, K, the simulation was conducted for two cases: $K = 2$ and $K = 6$. Moreover, the optimal p for $K = 2$, i.e., $p = 0.31$, that for $K = 4$, i.e., $p = 0.2$, and that for $K = 6$, i.e., $p = 0.14$, were calculated by simulation. As for the non-adaptive conventional method, these two fixed optimal values of p were used. As for the proposed method, $\boldsymbol{a}_T$ of the T-th row vector of $\boldsymbol{A}$ was computed by random design of Bernoulli probability p_T. The original version of the information gain, (9), and the revised version, (13), were then compared.

First, the performance of the proposed method in the case of no noise was computed. Figure 1 shows the probability of exact recovery in the case of $K = 2$, and Figure 2 shows that in the case of $K = 6$. NON-ADAPT means the non-adaptive conventional method [9], ADAPT means the proposed method maximizing (9), and REVISED-ADAPT means the proposed method maximizing (13). In both cases, the proposed method, namely, "REVISED-ADAPT", is better than the non-adaptive method in the worst cases, and the performance of the proposed method is near the level of that using the optimal pool size. These results indicate that the proposed method can effectively control pool size. Figure 2 shows that the performance of the ADAPT is low. This result indicates that ADAPT is degraded by the estimation error.

The performance of the proposed method in the noisy case was simulated next. In the simulation, noise with i.i.d 5% probability of flipping each bit of $\boldsymbol{y}$ was added. Figure 3 shows the probability of exact recovery in the case of $K = 2$, and Figure 4 shows that in the case of $K = 6$. NON-ADAPT means the non-adaptive conventional method [9], ADAPT means the proposed method maximizing (9), and REVISED-ADAPT means the proposed method maximizing (13). According to these results, the proposed method (REVISED-ADAPT) is better than the non-adaptive method in the worst case, and the performance of the proposed method is near the level of that using the optimal pool size. These results indicate that the proposed method can effectively control the pool size even under noisy conditions.

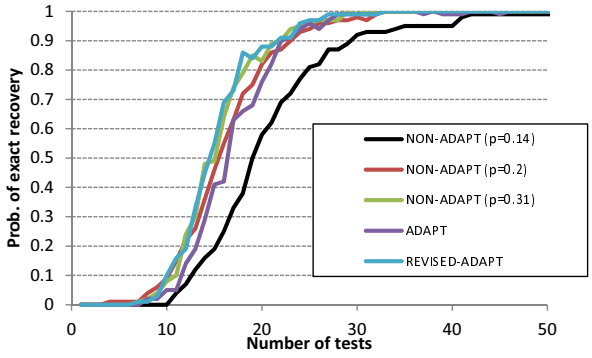

Fig. 1. Probability of exact recovery in noiseless case as a function of number of tests, T. $N = 150$, $K = 2$.

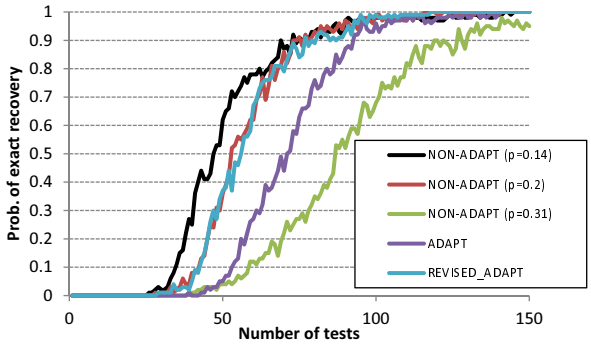

Fig. 2. Probability of exact recovery in noiseless case as a function of number of tests, T. $N = 150$, $K = 6$.

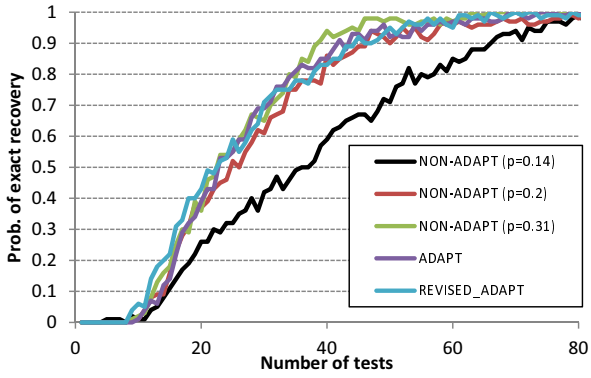

Fig. 3. Probability of exact recovery in noisy case as a function of number of tests, T. $N = 150$, $K = 2$, and 5% noise was added.

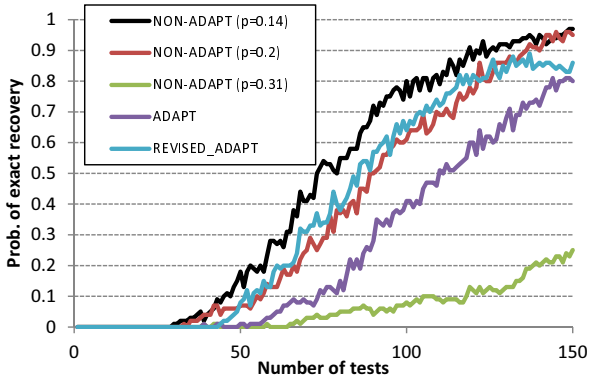

Fig. 4. Probability of exact recovery in noisy case as a function of number of tests T. $N = 150$, $K = 6$, and 5% noise was added.

6. CONCLUSION

A new method for solving the adaptive group-testing problem is proposed. The proposed method controls pool size adaptively by using information gain calculated from the ℓ_0 norm of the estimated solution. Moreover, to improve the robustness of group testing against estimation error, smoothing of the information gain in consideration of the estimation error is applied. An experimental simulation showed that the proposed method outperforms the conventional method even when the number of defective items is varied and the number of defective items is unknown.

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, no. 6, pp. 436–440, 1943.

[2] D.Z. Du and F.K. Hwang, *Pooling designs and nonadaptive group testing: Important tools for DNA sequencing*, World Scientific, 2006.

[3] M.B. Malyutov, "On planning of screening experiments," in *1975 IEEE-USSR Workshop on Information Theory*, 1976, pp. 144–147.

[4] M.B. Malyutov, "The separating property of random matrices," *Matematicheskie Zametki*, vol. 23, pp. 155–167, 1978.

[5] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1880–1901, 2012.

[6] D. Sejdinovic and O. Johnson, "Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction," in *Allerton conference on communication, control and computing*, 2010.

[7] M. Aldridge and L. Baldassini and O. Johnson, "Group testing algorithms: Bounds and simulations," in *arXiv:1306.6438*, 2013.

[8] C.L. Chan, P.K. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *arXiv:1107.4540*, 2011.

[9] D. Malioutov and M. Malyutov, "Boolean compressed sensing: LP relaxation for group testing," in *ICASSP*, 2012.

[10] E. Candes and M.B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.