Randomized Recovery for Boolean Compressed Sensing

Mitra Fatemi and Martin Vetterli
Laboratory of Audiovisual Communication
École Polytechnique Fédéral de Lausanne (EPFL)
Email: {mitra.fatemi, martin.vetterli}@epfl.ch

***Abstract*—We consider the problem of boolean compressed sensing, which is also known as group testing. The goal is to recover a small number of defective items in a large set from a few collective binary tests. This problem can be formulated as a binary linear program, which is NP hard in general. To overcome the computational burden, it was recently proposed to relax the binary constraint on the variables, and apply a rounding to the solution of the relaxed linear program. In this paper, we introduce a randomized algorithm to replace the rounding procedure. We show that the proposed algorithm considerably improves the success rate with only a slight increase in computational cost.**

***Index Terms*—Boolean compressed sensing, Group testing, Linear programming, Randomized algorithm.**

I. Introduction

The group testing problem is about distinguishing a small number of defective items among a large group via a few random collective measurements. This problem was first studied by Dorfman [1] for the blood screening of large groups and then found applications in many other fields such as computational biology (e.g. DNA library screening), multiple access control protocols and data streams. In non-adaptive group testing, the structure of tests does not change based on the previous test outcomes. This allows the parallel implementation of different tests. In this paper we only study non-adaptive group testing.

Let $\mathbf{x} \in \mathbb{R}^n$ denote a binary vector with entries as indicators of state of the n items involved in measurements, i.e. it contains 1s exactly in the places corresponding to the defective items. Introduce $\mathbf{y}$ as the binary outcome of the m measurements and let γ_j represent the group of items contributing to the jth test. Then $[\mathbf{y}]_j = \vee_{i\in\gamma_j}[\mathbf{x}]_i$ which means that a test outcome is positive it it contains at least one defective item.

We assume that the number k of defective items is very small compared to the total number of participating items. In this case, we call $\mathbf{x}$ a k-sparse vector and we represent it by $\|\mathbf{x}\|_{\ell_0} = k \ll n$. The goal of the group testing problem is to efficiently recover the small subset of defective items from the test outcomes while reducing the total number of tests (measurement). We can formulate this problem as a boolean linear matrix equation

$$\mathbf{y} = \mathbf{\Gamma} \vee \mathbf{x}. \tag{1}$$

In the above equation, we use $\vee$ to remind that the summation is replaced by the logical OR operation. Here, $\mathbf{\Gamma}$ represents a binary matrix in $\mathbb{R}^{m\times n}$ with 1s in the jth row located by the indices in γ_j.

In addition to the noiseless scenario, we may also consider the noisy-variant of the group testing problem, in which the test outcomes may differ from the true results. We model the noisy measurements as

$$\mathbf{y} = \mathbf{\Gamma} \vee \mathbf{x} \oplus \mathbf{n}, \tag{2}$$

where $\oplus$ denotes XOR operation and $\mathbf{n}$ represents the noise vector with *i.i.d.* Bernoulli distributed entries. In this case, the estimation of defective items is more challenging and requires more measurements.

The formulations of the group testing problem in equations (1) and (2) are very similar to the well-known problem of compressed sensing (CS) [2]–[4], where the goal is to estimate a large-size sparse vector from a small number of linear measurements. The major differences are that the latter involves operations in the field of real numbers with Gaussian noise while the former involves boolean operations and Bernoulli noise. Therefore, the group testing problem is sometimes referred to as boolean compressed sensing [5], [6]. Moreover, a number of solutions to this problem have parallels in CS; for example, the combinatorial basis pursuit (CBP) and orthogonal matching pursuit (COMP) algorithms in [7].

It is recently proposed to use relaxed linear programming (LP) to solve the group testing problem [6]. The

978-1-4799-0446-4/13/$31.00 ©2013 IEEE

LP algorithm of [6] bypasses the binary constraints and solves a linear problem. Then, the outcome undergoes rounding to recover a binary vector. Unfortunately, the final result is often less sparse than the original vector. In this paper, we replace the rounding procedure with a random assignment of 1's to the most likely defective entries; the probabilities of the random assignments are determined by the solution of the linear program. In this paper, we only consider the noiseless measurement scenario. The more involved case of noisy measurements will be addressed in future work.

The paper is organized as follows. In Section II, we review the bounds on the number of measurements that guarantee exact signal reconstruction in CS and group testing. In Section III, we review the LP algorithm for the noiseless and noisy measurements. We present the randomized algorithm in Section IV, accompanied with an analysis of the algorithm. We also highlight the link between the group testing problem and the classical problem of set covering. The performance comparison of the ordinary and randomized algorithms is presented in Section V. Finally, we conclude the paper in Section VI.

II. CS and Group Testing: Recovery Bounds

In CS, the goal is to recover a sparse vector $\|\mathbf{x}^*\|_{\ell_0} \leq k$ from the smallest possible number of linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}^*$. Combinatorial solutions to this problem solve the equation

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_{\ell_0} \quad s.t. \;\; \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{3}$$

Another possible solution for CS can be obtained by substituting the non-convex ℓ_0 norm in (3) with the convex ℓ_1 norm. This results to the *Basis Pursuit* algorithm that uses linear programming to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_{\ell_1} \quad s.t. \;\; \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{4}$$

It is shown that if the sensing matrix $\mathbf{A}$ has random *i.i.d* Gaussian entries with $m = O\Big(k\log(n/k)\Big)$ measurements, both the equations in (3) and (4) recover the solution exactly, i.e. $\mathbf{x} = \mathbf{x}^*$. The reason is that these matrices satisfy the so-called *Restricted Isometry Property* (RIP), which ensures that different k-sparse vectors are mapped to different measurements.

In group testing, there are two types of matrices that ensure the identifiablity of k-sparse vectors.

Definition 1. *A binary matrix $\mathbf{\Gamma}$ is k-separating if the boolean sums of sets of k columns are all distinct.*

Matrices with k-separating property ensure that different k-sparse vectors produce distinct measurements and therefore, they guarantee recovery of a unique k-sparse solution. A stronger notion is k-disjunct property.

Definition 2. *A binary matrix is called k-disjunct if the boolean sum of any k columns does not contain any other column.*

Matrices that satisfy this property are desirable not only because they ensure identifiability but they also lead to efficient decoding. Combinatorial constructions of k-disjunct matrices were extensively developed in [8], [9].

A different approach to the group testing problem is based on probabilistic methods. In [8], [10], the authors establish upper and lower bonds on the number of rows m for a matrix to be k-disjunct. They show that in the noiseless scenario, m should scale as $O\Big(\frac{k^2\log n}{\log k}\Big)$ for exact recovery with worst-case input.

The authors of [5] have recently studied the noisy counterpart of group testing problem in equation (2). They show that the number of measurements must scale as $O\Big(\frac{k^2\log n}{(1-q)\log k}\Big)$ for a worst-case error criterion, where the noise distribution is Bernoulli(q).

III. LP relaxation

An LP relaxation of the group testing problem is proposed in [6] that parallels the LP relaxation of basis pursuit in CS. Let $\mathcal{I}$ and $\mathcal{J}$ denote the index of positive and negative tests, respectively;i.e. $\mathcal{I} = \{i|\ [\mathbf{y}]_i = 1\}$, $\mathcal{J} = \{1,...,m\}\backslash\mathcal{I}$. Also, let $\mathbf{\Gamma}_\mathcal{I}$ denote the rows of $\mathbf{\Gamma}$ indexed by $\mathcal{I}$. The following equation gives a boolean linear programming formulation of the group testing problem.

$$\begin{aligned} &\min \|\mathbf{x}\|_{\ell_0} \\ &\text{s.t.} \;\; \mathbf{x}\in\{0,1\}^n, \;\; \mathbf{\Gamma}_\mathcal{I}\mathbf{x} \geq \mathbf{y}_\mathcal{I}, \;\; \mathbf{\Gamma}_\mathcal{J}\mathbf{x} = 0. \end{aligned} \tag{5}$$

We remind that for a boolean vector $\mathbf{x}$, $\|\mathbf{x}\|_{\ell_0} = \|\mathbf{x}\|_{\ell_1}$. By relaxing the binary constraint on $\mathbf{x}$, we obtain a tractable linear program

$$\begin{aligned} &\min \|\mathbf{x}\|_{\ell_1} \\ &\text{s.t.} \;\; 0 \leq \mathbf{x} \leq 1, \;\; \mathbf{\Gamma}_\mathcal{I}\mathbf{x} \geq \mathbf{y}_\mathcal{I}, \;\; \mathbf{\Gamma}_\mathcal{J}\mathbf{x} = 0. \end{aligned} \tag{6}$$

In the case of non-integral $[\mathbf{x}]_i$, we set them to 1.

We can use a vector $\boldsymbol{\xi} \in \mathbb{R}^m$ to get an LP relaxation of group testing in the noisy scenario:

$$\begin{aligned} &\min \|\mathbf{x}\|_{\ell_1} + \alpha\|\boldsymbol{\xi}\|_{\ell_1} \\ &\text{s.t.} \;\; 0 \leq \mathbf{x} \leq 1, \;\; 0 \leq \boldsymbol{\xi}, \;\; \boldsymbol{\xi}_\mathcal{I} \leq 1, \\ &\mathbf{\Gamma}_\mathcal{I}\mathbf{x} + \boldsymbol{\xi}_\mathcal{I} \geq \mathbf{y}_\mathcal{I}, \;\; \mathbf{\Gamma}_\mathcal{J}\mathbf{x} = \boldsymbol{\xi}_\mathcal{J}. \end{aligned} \tag{7}$$

Algorithm 1 RLP for noiseless measurements

Input: $\mathbf{y}$, $\mathbf{\Gamma}$, ϵ.

Output: $\hat{\mathbf{x}} \in \{0,1\}^n$ such that $\mathbf{y} = \mathbf{\Gamma} \vee \hat{\mathbf{x}}$ (with prob. $\geq 1-\epsilon$).

Initialization:

1. $\hat{\mathbf{x}} = \mathbf{0}$, $\mathcal{I} = \{i|\ [\mathbf{y}]_i = 1\},\ \mathcal{J} = \{i|\ [\mathbf{y}]_i = 0\}$;
2. Set $\mathbf{x}_p$ as the minimizer of (6);

for $\ell := 1$ **to** $\left\lceil \log \frac{|\mathcal{I}|}{\epsilon} \right\rceil$ **do**

1. Generate a vector $\mathbf{x}_\ell$ according to the distribution $[\mathbf{x}_\ell]_i \sim \text{Bernoulli}([\mathbf{x}_p]_i),\ i = 1, \ldots, n$;
2. $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} \vee \mathbf{x}_\ell$;

if $(\mathbf{\Gamma} \vee \hat{\mathbf{x}} = \mathbf{y})$ **then**
Stop;
end if
end for

IV. Randomized Recovery

LP algorithm in equation (6) provides the optimal solution $\mathbf{x} = \mathbf{x}^*$ if $\mathbf{x}^*$ is k-sparse and the matrix $\mathbf{\Gamma}$ is k-disjunct [6]. Otherwise, it may yield a non-integral solution with minimum ℓ_1 norm and a large number of non-zero entries. Rounding the non-integral entries to 1 gives a solution with a large number of defective items.

In this section, we propose a randomized LP algorithm (RLP) based on the LP relaxation in (6). The new algorithm provides arbitrary small measurement error probability and sparser solutions compared to the LP algorithm described above.

The new recovery method is summarized in Algorithm 1 for the case of noiseless measurements. Next, we prove that this algorithm recovers a boolean vector that satisfies the test outcomes with a probability not less than $1-\epsilon$.

Let $0 \leq \mathbf{x}_p \leq 1$ and J_{LP} represent the fractional minimizer and the corresponding minimum value of equation (6), respectively. Let $\hat{\mathbf{x}}$ indicate the output of Algorithm 1 after 1 iteration. Then, $P([\hat{\mathbf{x}}]_i = 1) = [\mathbf{x}_p]_i,\ P([\hat{\mathbf{x}}]_i = 0) = 1 - [\mathbf{x}_p]_i$ and

$$\mathbb{E}(\|\hat{\mathbf{x}}\|_{\ell_0}) = \sum_{i=1}^{n} 1\ .\ [\mathbf{x}_p]_i = J_{LP} \leq J, \tag{8}$$

where $\mathbb{E}$ is the expected value and J is the optimal value of the boolean group testing problem in (5). Equation (8) shows that the expected number of defective items in $\hat{\mathbf{x}}$ is J_{LP}. Therefore, the average number of defective items of the output in Algorithm 1 after $c = \left\lceil \log \frac{|\mathcal{I}|}{\epsilon} \right\rceil$ iterations is not larger than $c\, J_{LP}$.

Lemma 1. *The output vector $\hat{\mathbf{x}}$ of Algorithm 1 after $c = \left\lceil \log \frac{|\mathcal{I}|}{\epsilon} \right\rceil$ iterations coincides with the measurements $\mathbf{y}$ with a probability greater than $1-\epsilon$, i.e.*

$$P(\mathbf{y} = \mathbf{\Gamma} \vee \hat{\mathbf{x}}) > 1 - \epsilon. \tag{9}$$

Proof. It is not hard to verify that the vector of test outcomes $\mathbf{y}$ contains the measurements corresponding to the output of Algorithm 1, i.e. $\mathbf{y} \geq \mathbf{\Gamma} \vee \hat{\mathbf{x}}$. Therefore, we only need to calculate the probability of $\mathbf{\Gamma}_\mathcal{I} \vee \hat{\mathbf{x}} = \mathbf{y}_\mathcal{I}$ or equivalently, $\mathbf{\Gamma}_\mathcal{I} \hat{\mathbf{x}} \geq \mathbf{y}_\mathcal{I}$.

Let $j \in \mathcal{I}$ and $|\gamma_j|$ denote the number of items contributing to the jth positive test. If $\hat{\mathbf{x}}$ denotes the output of Algorithm 1 after 1 iteration, we have

$$P(\mathbf{\Gamma}_j \hat{\mathbf{x}} < [\mathbf{y}]_j) \quad = \quad P(\mathbf{\Gamma}_j \hat{\mathbf{x}} = 0) = \prod_{i \in \gamma_j} (1 - [\mathbf{x}_p]_i).$$

Note that $\log(1-x)$ is a concave function which by using Jensen's inequality yields

$$\sum_{i \in \gamma_j} \frac{1}{|\gamma_j|} \log(1 - [\mathbf{x}_p]_i) \leq \log(1 - \frac{\sum_{i \in \gamma_j} [\mathbf{x}_p]_i}{|\gamma_j|})$$
$$\Longrightarrow \prod_{i \in \gamma_j} (1 - [\mathbf{x}_p]_i) \overset{(a)}{\leq} (1 - \frac{1}{|\gamma_j|})^{|\gamma_j|} \leq \frac{1}{e},$$

where (a) is the result of the fact that $\sum_{i \in \gamma_j} [\mathbf{x}_p]_i = \sum_{i=1}^{n} \mathbf{\Gamma}_{ji} [\mathbf{x}_p]_i \geq 1$. Therefore, after c iterations, we have

$$P(\mathbf{\Gamma}_j \hat{\mathbf{x}} = 0) \leq (\frac{1}{e})^c.$$

Finally, from the union bound, we get

$$P(\exists j \in \mathcal{I} :\ \mathbf{\Gamma}_j \hat{\mathbf{x}} = 0) \quad \leq \quad |\mathcal{I}| (\frac{1}{e})^{\left\lceil \log \frac{|\mathcal{I}|}{\epsilon} \right\rceil} \leq \epsilon,$$

which proves that $P(\mathbf{y} = \mathbf{\Gamma} \vee \hat{\mathbf{x}}) > 1 - \epsilon$. □

The LP algorithm in [6] can be regarded as a special case of our algorithm with infinite iterations so that every $[\hat{\mathbf{x}}]_i$ that has a probability $[\mathbf{x}_p]_i$ larger than 0 is set to 1. Therefore, it generates a less sparse binary vector compared to the output of Algorithm 1. We remind that when $\mathbf{y}$ corresponds to a k-sparse vector $\mathbf{x}^*$ and $\mathbf{\Gamma}$ is k-disjunct, equation (6) has a binary solution and therefore, both algorithms recover the optimal solution $\mathbf{x} = \mathbf{x}^*$.

A. Link to the set covering problem

A related problem to the boolean CS is the classical set covering problem (SCP). Given a set of elements $U = \{1, \ldots, m\}$ (called a universe) and n sets whose union comprises the universe, SCP is to identify the smallest number of sets whose union still contains all elements of the universe. The boolean CS problem can be modeled as SCP by considering $\mathbf{y}_\mathcal{I}$ as the universe and columns of $\mathbf{\Gamma}_\mathcal{I}$ as different sets. In this regard, the RLP method of this paper is one of the solutions to the SCP [11].

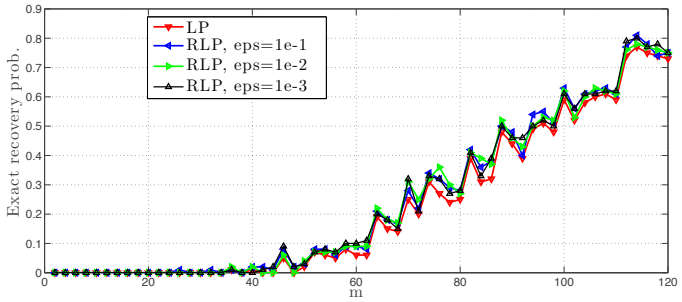

Fig. 1. Probability of exact signal reconstruction in LP and RLP algorithms for $n = 150, k = 4$ and noiseless measurements. Averages over 100 trials.

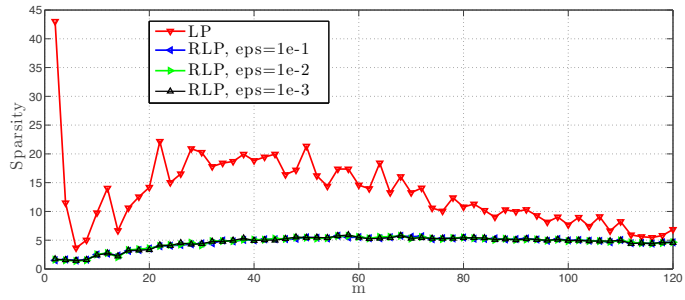

Fig. 2. Sparsity of the recovered signals in LP and RLP algorithms for $n = 150, k = 4$ and noiseless measurements. Averages over 100 trials.

V. Simulation Results

In [6], the authors compare the performance of LP algorithm with a number of algorithms such as CBP, COMP and loopy belief propagation (LBP) [12]. This comparison shows that LP outperforms the other algorithms in terms of the probability of exact recovery. We now present experiment results comparing our randomized algorithm with LP. For better comparison, we follow the experiment setup in [6].

In our first experiment, we computed the probability of exact signal recovery and the sparsity of the recovered signals in 100 trials as functions of m, for $n = 150$ and $k = 4$ without any noise. For each value of m, we generated a boolean sensing matrix with 50% of its entries set to 1. We computed the results of RLP for three different error probabilities $\epsilon = 0.1, 0.01$ and 0.001. The results appear in Figures 1 and 2. These plots show that our RLP algorithm outperforms LP in terms of both the sparsity and the exact reconstruction probability.

In the next experiment, we examine the computational complexity of RLP. For this purpose, we run RLP until the recovered signal $\hat{\mathbf{x}}$ produces the same measurement vector $\mathbf{y}$. The average number of iterations is depicted in Figure 3 as a function of m, for the same setup of the previous experiment. This plot shows that the RLP algorithm requires a small number of iterations. Note that, each iteration consists of generating a random Bernoulli vector $\mathbf{x}_l$ and the boolean operations involved in $\mathbf{\Gamma} \vee \hat{\mathbf{x}}$ and $\hat{\mathbf{x}} \vee \mathbf{x}_l$. These results show that RLP achieves a considerable performance improvement over LP by slightly increasing the computational complexity.

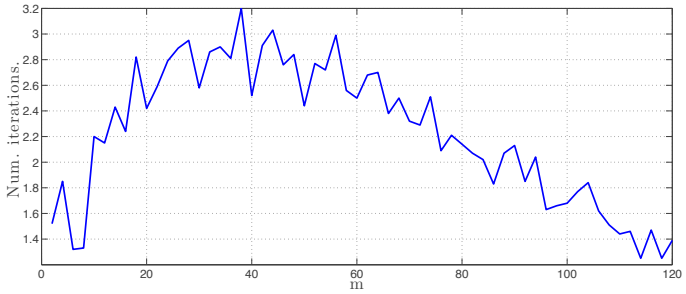

Fig. 3. Computational complexity of RLP: the average number of iterations required for generating a vector $\hat{\mathbf{x}}$ that coincides with measurements $\mathbf{y}$. Averages over 100 trials for $n = 150$ and $k = 4$.

VI. Conclusion

We considered the problem of boolean CS, where the unknown variables $[\mathbf{x}]_i$ are constrained to be in $\{0, 1\}$. Although the measurement process is linear with respect to x_i, due to the binary constraints, the linear program is NP hard. We applied the relaxation $[\mathbf{x}]_i \in [0, 1]$ in the linear program and obtained fractional solutions. To map the fractional values onto binary values, instead of the common rounding techniques, we considered a randomized approach; i.e., each value is randomly mapped to 0 or 1, with a probability determined by the fractional value. The simulation results indicate that the randomized algorithm considerably outperforms the common methods with a slight increase in computational cost.

References

[1] R. Dorfman, "The detection of defective members of large populations," *Annals of Math. Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

[2] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Info Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[5] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, Mar. 2012.
[6] D. Malioutov and M. Malyutov, "Boolean compressed sensing: Lp relaxation for group testing," in *IEEE ICASSP Conf.*, Mar. 2012, pp. 3305–3308.
[7] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *IEEE Allerton Conf.*, Sep. 2011, pp. 1832–1839.
[8] A. G. Dyachkov and V. V. Rykov, "A survey of superimposed code theory," *Problem of Control and Inf. Theory*, vol. 12, no. 4, pp. 1–13, 1983.
[9] P. Erdos, P. Frankl, and Z. Furedi, "Family of finite sets in which no set is covered by the union of n others," *Israel J. Math.*, vol. 51, no. 1-2, pp. 79–89, Dec. 1985.
[10] A. G. Dyachkov, V. V. Rykov, and M. Rashad, "Bounds of the length of disjunct codes," *Problem of Control and Inf. Theory*, vol. 11, pp. 7–13, 1982.
[11] R. Raz and S. Safra, "A sub-constant error-probability low-degree test, and a sub-constant error-probability pcp characterization of np," in *Proc. 29th ACM Symp. on Theory of Comput.*, 1997, pp. 475–484.
[12] D. Sejdinovic and O. Johnson, "Note on the noisy group testing: Asymptotic bounds and belief propagartion reconstruction," in *IEEE Allerton Conf.*, Sep. 2010, pp. 998–1003.