BOOLEAN COMPRESSED SENSING: LP RELAXATION FOR GROUP TESTING

Dmitry Malioutov

DRW Holdings
Algorithmic Trading
Chicago, IL, USA

Mikhail Malyutov

Northeastern University
Department of Mathematics
Boston, MA, USA

ABSTRACT

We revisit the well-known problem of boolean group testing which attempts to discover a sparse subset of faulty items in a large set of mostly good items using a small number of pooled (or grouped) tests. This problem originated during the second World War, and has been the subject of active research during the 70's, and 80's. Recently, there has been a resurgence of interest due to the striking parallels between group testing and the now highly popular field of compressed sensing. In fact, boolean group testing is nothing but compressed sensing in a different algebra – with boolean 'AND' and 'OR' operations replacing vector space multiplication and addition. In this paper we review existing solutions for non-adaptive (batch) group testing and propose a linear programming relaxation solution, which has a resemblance to the basis pursuit algorithm for sparse recovery in linear models. We compare its performance to alternative methods for group testing.

Index Terms— group testing, compressed sensing, LP relaxation

1. INTRODUCTION

One incarnation of the group testing problem known to many readers is the job interview puzzle of detecting a few fake coins using the smallest number of weightings with a scale. However, group testing also has a number of serious applications, ranging from blood screening of large groups of subjects (which is in fact the original motivation considered by Dorfman and Rosenblatt [1]), to computational biology, fault discovery in computer networks, and others [2].

In the context of *boolean* group testing, a pooled test combines some subset of the subjects in question and answers whether any of them are faulty (infected). We will focus on the non-adaptive version of the problem where all the tests are administered at the same time, without the ability to use the results of some tests to affect the selection of the other tests.

The main questions of interest in non-adaptive group testing are the characterization of the minimum number of tests that are required to find K faulty subjects out of the total set of N, and the design of optimal pooling matrices A which encode the set of tests: non-zero elements in row i of the matrix encode the subjects participating in test i. A large body of literature characterizes upper and lower bounds on the number of tests in noiseless and noisy group testing, with both combinatorial and random designs of the matrices. In particular, a set of information theoretic bounds for group testing with random designs was established by Malyutov and his coworkers [3, 4, 5]. Very recently these bounds were independently rediscovered [1] after 30 years and extended in [6, 7].

The recent revival of interest in group testing came from the active research area of compressed sensing (CS), where the goal is to infer a sparse high-dimensional vector from a small set of linear measurements. Compressed sensing is set in the context of real vector spaces with additive noise, while group testing studies the same problem in the boolean setting and with Bernoulli noise. These recent works on group testing drew parallels with CS, and bounds on performance were established which also parallel the related bounds in CS. In addition, a few tractable approximation algorithms were proposed for group testing, including a belief propagation solution [7], and matching-pursuit-like solutions [8].

In this paper we describe a simple linear programming (LP) relaxation for both the noiseless and the noisy non-adaptive group testing problem, and compare its performance with alternative solution methods[2]. An initial study of LP for the noiseless boolean model appeared in [9]. In Sections 1 and 2 we set up the notation and describe some existing bounds for group testing. We describe the LP relaxation in Section 3, and present simulation results in Section 4.

2. COMBINATORIAL GROUP TESTING

Suppose we have a boolean vector $\mathbf{x} \in \{0,1\}^N$, with only a small number K of entries which are non-zero. We will also call these non-zero items 'faulty', whereas the items j with $x_j = 0$ are 'normal'. A pooled measurement y_i is obtained by taking the boolean sum (boolean OR) of x_j in some subset $A_i \subset \{1,..,N\}$, i.e. $y_i = \vee_{j \in A_i} x_j$. With a slight abuse of notation, we will represent a collection of M pooled mea-

[1] In [3, 4, 5] K was fixed and $N \to \infty$, while [6, 7] let both $K, N \to \infty$.

[2] The authors of [8] have also independently started to work on an LP relaxation solution (private communication).

surements by an $M \times N$ binary matrix A, where $A_{ij} = 1$ if item j belongs to the subset pooled in test i. We will use the vector notation

$$\mathbf{y} = A \vee \mathbf{x} \tag{1}$$

to represent the operation of obtaining the results of these M tests. In the presence of noise we allow the results of some tests to be corrupted (inverted): we allow both false alarms, i.e. a positive outcome y_i despite all the items x_j tested in A_i are negative, and mis-detections where $y_i = 0$ while at least one x_j in A_i is non-zero. We represent this by

$$\mathbf{y} = (A \vee \mathbf{x}) \otimes \mathbf{n} \tag{2}$$

where $\mathbf{n}$ is the boolean vector of errors, and $\otimes$ is the boolean XOR operation. Group testing studies both the design of the measurement matrix A and the decoding of $\mathbf{x}$ given $\mathbf{y}$. Both deterministic and random designs of A are of interest.

2.1. Group testing theory: exact recovery and bounds

We review some results in combinatorial group testing [10, 2] for exact recovery and recovery with small error probability.

Definition 1 *We call a measurement matrix A K-*separating*, if boolean sums of sets of K columns are all distinct. A is called K-*disjunct*, if the union of any K columns does not contain any other column.*

The K-separating property on A implies that any $\mathbf{x}$ with up-to K non-zeros can be recovered exactly [2]. However, in general, the recovery problem requires searching over all subsets. The K-disjunct property simplifies the search, and a simple algorithm exactly recovers $\mathbf{x}$ [2]: let $\mathcal{J} = \{i \mid y_i = 0\}$. Set $x_j = 0$ for those j for which there exists $i \in \mathcal{J}$ with $A_{ij} = 1$. The remaining entries in x are labeled as 1. We call this algorithm the *baseline algorithm*. It can also be applied to non-disjunct matrices A, albeit loosing the correctness guarantee.

2.2. Required number of tests

For guaranteed recovery in the noiseless setting, a lower bound requiring $K^2/\log(e(K+1)/2)\log(N)$ tests has been established, and a family of combinatorial designs of A was found achieving $(e \log 2)K^2 \log(N)$ tests [10].

By allowing non-zero probability of error in recovery[3] the number of tests can be reduced. Moreover, it turns out that with brute-force decoding the smallest possible number of tests is asymptotically achieved with certain random designs, meeting information theoretic lower bounds. This has been derived for noiseless designs in [5] and generalized to noisy scenarios for arbitrary symmetric models in [11] using the Fano inequality[4]. Now consider an arbitrary analysis procedure F. We define $M^F(K, N, \gamma)$ to be the smallest number of tests such that the analysis procedure F can recover the correct active inputs with probability of error less than γ. A notion of screening capacity is introduced in [3] for recovery procedure F with error probability γ:

$$C^F(K) = \lim_{N\to\infty} (\log N / M^f(K, N, \gamma) \tag{3}$$

It was shown in [3] that simple i.i.d. Bernoulli random matrices achieve the optimal screening capacity under brute-force analysis for $\gamma > 0$ and arbitrary independent noise. The probability p of having 1 in each entry A_{ij} is set to maximize the mutual information between $\mathbf{x}$ and $\mathbf{y}$. For the noiseless case this leads to $p = 2^{-1/K}$, and the required number of tests with brute-force decoding is $M = K \log(N)(1 + o(1))$, see [5]. We now overview existing recovery methods.

2.3. Recovery algorithms

Brute Force (BF) recovery finds $\mathbf{x}$ with the smallest number of non-zero components to satisfy all the tests:

$$\min \|\mathbf{x}\|_0 \quad \text{such that } \mathbf{y} = A \vee \mathbf{x} \tag{4}$$

The complexity is enumerating all the subsets of size up to K.

Separate testing of Inputs (STI) STI is a simple method effective for random designs, and can be traced back to Fisher's ideas on randomization in estimation. STI separately computes the empirical mutual information (EMI) between A_i and y, regarding the influence of other inputs as noise. Then it selects K elements of $\mathbf{x}$ with highest EMI. An analysis and an empirical study of STI was conducted for several models including boolean group testing and linear models in [9], including a preliminary comparison to LP relaxation.

Loopy belief propagation decoder (LBP) The decoder proposed in [7] invokes the max-product form of the loopy belief propagation algorithm, a popular but enigmatic approximate inference approach for graphical models. LBP has been used with success for many combinatorial optimization problems, e.g. [12]. Given a probabilistic model for noisy group testing represented as a factor graph, [7] derive the corresponding LBP message update equations. LBP is an iterative algorithm, and may or may not converge, and when it converges it may not give the optimal answer. So far no guarantees on BP performance for group testing are known.

Combinatorial basis pursuit (CBP) and matching pursuit (CMP) These algorithms proposed in [8] loosely mimic the corresponding algorithms from compressed sensing. The CBP algorithm invokes the baseline decoder that we described earlier in Section 2.1: it sets $x_j = 0$ if there exists i such that both $A_{ij} = 1$ and $y_i = 0$. All the other x_j are set to 1. This is guaranteed to be correct if the matrix A is disjunct. CMP proceeds columnwise and sets x_j to 1 if $A_j \leq \mathbf{y}$. Note that in the noiseless setting these algorithms are almost

[3] Assuming that K-subsets of defective items are uniformly distributed.

[4] For asymmetric models, more involved ideas of capacity bounds in Multiple Access Communication models are used for sharp lower bounds [5, 9].

equivalent except when A has all-zero columns. The analysis in [8] establishes that an upper bound on the number of tests required to recover $\mathbf{x}$ with error-probability at most $N^{-\delta}$ is $2e(1+\delta)K\log(N)$ for CBP and $e(1+\delta)K\log(N)$ for CMP. They also analyze the performance of the modified CBP and CMP algorithms in the noisy setting, and show that they noisy CMP is also $O(K\log(N)$.

2.4. Parallels with compressed sensing

Many authors have commented on the parallels between group testing and compressed sensing, where one has a sparse signal $\mathbf{x} \in \mathbb{R}^N$, and tries to find it from $M \ll N$ measurements using a random measurement matrix A. The combinatorial problem is

$$\min \|\mathbf{x}\|_0 \quad \text{such that } \mathbf{y} = A\mathbf{x} \tag{5}$$

If we replace the measurement model with the boolean one, we obtain the group testing problem in (4). The *basis pursuit* algorithm for compressed sensing alluded to earlier solves the ℓ_1 relaxation of this subset selection problem:

$$\min \|\mathbf{x}\|_1 \quad \text{such that } \mathbf{y} = A\mathbf{x} \tag{6}$$

The minimum number of measurements to recover $\mathbf{x}$ exactly have been established for various constructs of A, e.g. $O(K\log(N/K))$ tests are required for i.i.d. Gaussian measurement vectors. Next, we generalize this LP relaxation to group testing for both noiseless and noisy settings.

3. LP RELAXATION FOR GROUP TESTING

We now describe a linear programming relaxation solution which parallels the basis pursuit algorithm for compressed sensing. The challenge is of course that $\mathbf{y} = A \vee \mathbf{x}$ is not linear. However we note that we can replace it with a closely related linear formulation: $A_{\mathcal{I}}\mathbf{x} \geq \mathbf{y}_{\mathcal{I}}$, and $A_{\mathcal{J}}\mathbf{x} = \mathbf{0}$ where $\mathcal{I} = \{i \mid y_i = 1\}$ are the positive tests, and $\mathcal{J} = \{i \mid y_i = 0\}$ are the negative tests. This gives us an equivalent binary linear programming formulation:

$$\begin{aligned} \min \sum_j x_j \quad \text{such that:} \quad & \mathbf{x} \in \{0,1\} \\ & A_{\mathcal{I}}\mathbf{x} \geq \mathbf{y}_{\mathcal{I}}, \quad A_{\mathcal{J}}\mathbf{x} = \mathbf{0}. \end{aligned} \tag{7}$$

This problem is equivalent to the group testing problem, and it is also NP-hard in general. By relaxing the boolean constraint on $\mathbf{x}$ we obtain a tractable LP relaxation:

$$\begin{aligned} \min \sum_j x_j \quad \text{such that:} \quad & 0 \leq \mathbf{x} \leq 1 \\ & A_{\mathcal{I}}\mathbf{x} \geq \mathbf{y}_{\mathcal{I}}, \quad A_{\mathcal{J}}\mathbf{x} = \mathbf{0}. \end{aligned} \tag{8}$$

In case of non-integral x_j in the solution we set them to 1.

3.1. LP relaxation in the noisy setting

By adding slack variables $\boldsymbol{\xi} = \{\xi_1, .., \xi_M\}$ we obtain an LP relaxation for the noisy version of group testing:

$$\begin{aligned} \min \sum_j x_j + \alpha \sum_i \xi_i \\ \text{such that:} \quad 0 \leq \mathbf{x} \leq 1, \;\; 0 \leq \boldsymbol{\xi}, \;\; \boldsymbol{\xi}_{\mathcal{I}} \leq 1 \\ A_{\mathcal{I}}\mathbf{x} + \boldsymbol{\xi}_{\mathcal{I}} \geq \mathbf{y}_{\mathcal{I}}, \quad A_{\mathcal{J}}\mathbf{x} = \mathbf{0} + \boldsymbol{\xi}_{\mathcal{J}}. \end{aligned} \tag{9}$$

Note that the sum $A_{\mathcal{J}}\mathbf{x}$ may exceed 1 if multiple entries in $\mathbf{x}$ are active for a given row, hence we *do not* impose the $\boldsymbol{\xi} \leq 1$ constraint for $\mathcal{J}$. The regularization parameter α balances the amount of noise versus the sparsity of the solution. By using parametric linear programming it is possible to trace the whole solution path as a function of α, which allows one to pick a solution with the desired number of non-zero elements.

3.2. Remarks on LP relaxation.

We now establish some guarantees for noiseless LP in (8). If the solution $\hat{\mathbf{x}}$ to LP is integral – then it is also the optimal solution $\mathbf{x}^*$ to the combinatorial group testing problem. Furthermore, we show that LP relaxation is a strict improvement (in probability of exact recovery) over the baseline algorithm:

Lemma 1 *If the matrix A is K-disjunct and $\mathbf{x}^*$ is K-sparse then LP relaxation provides the optimal solution, i.e. $\hat{\boldsymbol{x}} = \mathbf{x}^*$.*

Proof. First, $\mathbf{x}^*$ is a feasible solution to LP. Consider rows of A corresponding to positive tests, and columns of A which are not eliminated via the zero-rows. There are exactly K such columns, and the K-disjunct property implies that the matrix is full-rank (in Euclidean sense). For each column A_j there is at least one non-zero entry i which does not appear in any other column. Since $y_i = 1$, and $A_{ij} = 1$, then $x_j \geq 1$. Now $x_j^* = 1$ for all $j \in \mathcal{I}$, so it is the unique optimal solution. $\diamond$

By similar logic, if $\mathbf{x}^*$ is the unique optimal solution to (4) and the baseline algorithm recovers $\mathbf{x}^*$, then so does the LP relaxation. Hence, the sample-complexity of LP in the noiseless case allowing small error probability is upper bounded by that of CBP, and is also $O(K\log(N)$.

4. EXPERIMENTS

We now present simulation comparing the baseline algorithm (we label it CBP), CMP, STI, LBP, and our LP relaxation. First we describe some improvements we considered for LBP.

Modified LBP algorithm A typical implementation of LBP computes the beliefs (approximate max-marginals) $\tilde{p}(x_i)$, and sets $x_i = 1$ when $\tilde{p}(x_i = 1) > \tilde{p}(x_i = 0)$. In addition to plain LBP, we also used a heuristic of selecting only K items with the top K beliefs $\tilde{p}(x_i = 1)$. We call this version of LBP oracle-LBP, as it has the knowledge of K. To

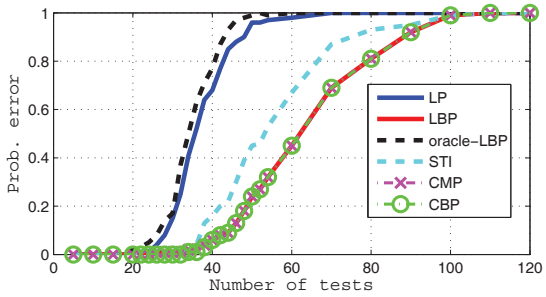

Fig. 1. Probability of exact recovery in the noiseless case as a function of number of tests M: LP, BP, Oracle-BP, STI, and simple-decoder, $N = 150$, $K = 4$.

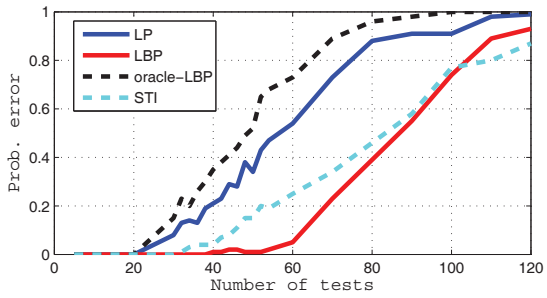

Fig. 2. Probability of exact recovery in the noisy case vs. M: LP, BP, Oracle-BP, and simple-decoder, $N = 150$, $K = 4$.

improve convergence we also used message damping (it does not affect the fixed points of LBP).

In our first experiment we computed the probability of error over 100 trials as a function of M, for $N = 150$, $K = 4$, with no noise. The plot appears in Figure 1. LP-relaxation and oracle-LBP have the best performance, followed by STI. Note that unlike oracle-LBP and STI, LP-relaxation does not assume the knowledge of K.

For our second experiment we add noise, with i.i.d. 5% probability of flipping each bit of $\mathbf{y}$. We compare LP, LBP, oracle-LBP, and STI. As expected the addition of noise increases the number of tests needed for recovery, however, we see that oracle-BP and LP still have the best performance. Note that we simply set $\alpha = 1$ in (9), and an adaptive choice of α would improve the performance.

5. CONCLUSION

We presented a simple yet effective linear programming relaxation for the group testing problem. We compared its performance on noiseless and noisy settings to other existing approximate solutions. For future work we are interested to characterize tight bounds on the number of tests required for exact recovery via LP relaxation. We are also interested to study LP in adaptive group testing, perhaps by drawing parallels to the sequential compressed sensing formulation [13].

6. REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, no. 4, 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 *Proceedings of 1975 IEEE-USSR Workshop on Inform. Theory*, 1976, pp. 144–147.

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

[5] M. B. Malyutov, "Maximal rates of screening designs," *Probability and its Applic.*, vol. 24, pp. 655–657, 1979.

[6] G. Atia and V. Saligrama, "Noisy group testing: An inform. theoretic approach," in *IEEE Allerton Conf.*, 2009.

[7] D. Sejdinovic and O. Johnson, "Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction," in *IEEE Allerton Conf.*, Sep. 2010.

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

[9] M.B. Malyutov, "Recovery of sparse active inputs in general systems: A review," in *Int. Conf. on Comp. Tech. in Electrical and Electronics Eng., Region 8, SIBIRCON*, 2010, pp. 15 – 22.

[10] A.G. Dyachkov and V. V. Rykov, "A survey of superimposed code theory," *Problems of Control and Information theory*, vol. 12, no. 4, pp. 229–242, 1983.

[11] M.B. Malyutov and P.S. Mateev, "Screening designs for non-symmetric response function," *Mat. Zametki*, vol. 27, pp. 109–127, 1980.

[12] S. Sanghavi, D. M. Malioutov, and A. S. Willsky, "Belief propagation and LP relaxation for weighted matching in general graphs," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2203 – 2212, April 2011.

[13] D. M Malioutov, S. Sanghavi, and A. S. Willsky, "Sequential compressed sensing," *IEEE Special Topics in Signal Proc.*, vol. 4, no. 2, pp. 435–444, 2010.