Computationally Tractable Algorithms for Finding a Subset of Non-defective Items from a Large Population

Abhay Sharma and Chandra R. Murthy

***Abstract*—In the classical non-adaptive group testing setup, pools of items are tested together, and the main goal of a recovery algorithm is to identify the *complete defective set* given the outcomes of different group tests. In contrast, the main goal of a *non-defective subset recovery* algorithm is to identify a *subset* of non-defective items given the test outcomes. In this paper, we present a suite of computationally efficient and analytically tractable non-defective subset recovery algorithms. By analyzing the probability of error of the algorithms, we obtain bounds on the number of tests required for non-defective subset recovery with arbitrarily small probability of error. Our analysis accounts for the impact of both the additive noise (false positives) and dilution noise (false negatives). By comparing with information theoretic lower bounds, we show that the upper bounds on the number of tests are order-wise tight up to a $\log^2 K$ factor, where K is the number of defective items. We also provide simulation results that compare the relative performance of the different algorithms and reveal insights into their practical utility. The proposed algorithms significantly outperform the straightforward approaches of testing items one-by-one, and of first identifying the defective set and then choosing the non-defective items from the complement set, in terms of the number of measurements required to ensure a given success rate.**

***Index Terms*—Non-adaptive group testing, boolean compressed sensing, non-defective subset recovery, inactive subset identification, linear program analysis, combinatorial matching pursuit, sparse signal models.**

I. Introduction

The general group testing framework [2], [3] considers a large set of N items, in which an unknown subset of K items possess a certain testable property, e.g., the presence of an antigen in a blood sample, presence of a pollutant in an air sample, etc. This subset is referred to as the "defective" subset, and its complement is referred to as the "non-defective" or "healthy" subset. A defining notion of this framework is the *group test*, a test that operates on a *group* of items and provides a binary indication as to whether or not the property of interest is present collectively in the group. A *negative* indication implies that none of the tested items are defective. A *positive* indication implies that at least one of the items is defective. In practice, due to the hardware and test procedure limitations, the group tests are not completely reliable. Using the outcomes of multiple such (noisy) group tests, a basic goal of group testing is to reliably identify the defective set of items with as few tests as possible. The framework of group testing has found applications in diverse engineering fields such as industrial testing [4], DNA sequencing [3], [5], data pattern mining [6]–[8], medical screening [3], multi-access communications [3], [9], data streaming [10], [11], etc.

One of the popular versions of the above theme is non-adaptive group testing (NGT), where different tests are conducted simultaneously, i.e., the tests do not use information provided by the outcome of any other test. NGT is especially useful when the individual tests are time consuming, and hence the testing time associated with adaptive, sequential testing is prohibitive. An important aspect of NGT is how to determine the set of individuals that go into each group test. Two main approaches exist: a combinatorial approach, see e.g., [12]–[14], which considers explicit constructions of test matrices/pools; and a random pooling approach, see e.g., [11], [15], [16], where the items included in the group test are chosen uniformly at random from the population. When the test outcomes are unreliable, the latter is called noisy non-adaptive group testing with random pooling (NNGT-R). It has also been referred to as *boolean* compressed sensing in the recent literature [17], [18].

In this work, in contrast to the defective set identification problem, we study the *healthy/non-defective subset identification* problem, in the NNGT-R framework. There are many applications where the goal is to identify only a small subset of non-defective items. For example, consider the spectrum hole search problem in a cognitive radio (CR) network setup. It is known that the primary user occupancy is sparse in the frequency domain, over a wide band of interest [19], [20]. This is equivalent to having a small subset of defective items embedded in a large set of candidate frequency bins. The secondary users do not need to identify all the frequency bins occupied by the primary users; they only need to discover a small number of unoccupied sub-bands to setup the secondary communications. This, in turn, is a non-defective subset identification problem when the bins to be tested for primary occupancy can be pooled together into group tests [21]. In [22], using information theoretic arguments, it was shown that compared to the conventional approach of identifying the non-defective subset by first identifying the defective set, directly searching for an L-sized non-defective subset offers a reduction in the number of tests, especially when L is small compared to $N - K$. The achievability results in [22] were obtained by analyzing the performance of the

A. Sharma was with the Dept. of ECE, Indian Institute of Science, Bangalore (IISc) Bangalore, India. He is now with the Robert Bosch Centre for Cyber-Physical Systems, IISc Bangalore, India. C. R. Murthy is with the Dept. of ECE, IISc Bangalore, India. Emails: abhay.bits@gmail.com, cmurthy@iisc.ac.in.

This paper was presented in part in [1].

This work was financially supported in part by a research grant from the Ministry of Electronics and Information Technology (MEITy), Govt. of India.

exhaustive search based algorithms which are not practically implementable. In this paper, we develop computationally efficient algorithms for non-defective subset identification in an NNGT-R framework.

We note that the problem of non-defective subset identification is a generalization of the defective set identification problem, in the sense that, when $L = N-K$, the non-defective subset identification problem is identical to that of identifying the K defective items. Hence, by setting $L = N-K$, the algorithms presented in this work can be related to algorithms for finding the defective set. In general, for the NNGT-R framework, three broad approaches have been adopted for defective set recovery [18]. First, the row based approach (also referred to as the "naïve" decoding algorithm) finds the defective set by finding *all* the non-defective items. The survey in [23] lists many variants of this algorithm for finding defective items. More recently, the **CoCo** algorithm was studied in [18], where an interesting connection of the naïve decoding algorithm with the classical coupon-collector problem was established for the noiseless case. The second popular decoding approach, also referred to as the Combinatorial Orthogonal Matching Pursuit (COMP) in the literature, is based on the idea of finding defective items iteratively (or greedily) by matching the column of the test matrix corresponding to a given item with the test outcome vector [3], [18], [24], [25]. For example, in [24], column matching consists of taking set differences between the set of pools where the item is tested and the set of pools with positive outcomes. Another variant of matching is considered in [18], where, for a given column, the ratio of number of times an item is tested in pools with positive and negative outcomes is computed and compared to a threshold. A recent work, [26], investigates the problem of finding zeros in a sparse vector in the compressive sensing framework, and also proposes a greedy algorithm based on correlating the columns of the sensing matrix (i.e., column matching) with the output vector.[1] The connection between defective set identification in group testing and the sparse recovery in compressive sensing was further highlighted in [8], [18], [27], where relaxation based linear programming algorithms have been proposed for defective set identification in group testing. A class of linear programs to solve the defective set identification problem was proposed by letting the boolean variables take real values (between 0 and 1) and setting up inequality or equality constraints to model the outcome of each pool. We refer the interested reader to [3] for an excellent collection of existing results and references on defective set identification.

Non-defective subset identification is related to the problem of group testing using list decoding [28]–[31], where the decoder outputs a superset of the true defective set, i.e., a list of items $\mathcal{L}$ (with $|\mathcal{L}| > K$) such that $\mathcal{L}$ contains the defective set. In [30], list decoding has been studied as an intermediate step while decoding the defective set in the conventional group testing setup. A combinatorial approach employing list-disjunct matrices was used to derive bounds on number of tests. These bounds are also applicable for non-defective subset identification, since the complement of the list output by the algorithm can be viewed as a non-defective subset. A very recent work [31] focuses on the $|\mathcal{L}| = o(N)$ case, and remarks that the list decoding viewpoint is more suited when $|\mathcal{L}| = o(N)$ and the non-defective subset identification is more suited when $|\mathcal{L}| = O(N)$.

In this work, we develop novel algorithms for identifying a non-defective subset in an NNGT-R framework. We present error rate analysis for each algorithm and derive non-asymptotic upper bounds on the average error rate. The derivation leads to a theoretical guarantee on the sample complexity, i.e., the number of tests required to identify a subset of non-defective items with arbitrarily small probability of error. We summarize our main contributions as follows:

- We propose a suite of computationally efficient and analytically tractable algorithms for identifying a non-defective subset of given size in a NNGT-R framework: **RoAI** (row based), **CoAl** (column based) and **RoLpAl**, **RoLpAl++**, **CoLpAl** (Linear Program (LP) relaxation based) algorithms.
- We derive bounds on the number of tests that guarantee successful non-defective subset recovery for each algorithm. The derived bounds are a function of the number of defective items, the size of non-defective subset, the population size, and the noise parameters.
  - For our suite of LP based algorithms, we present a novel analysis technique based on characterizing the recovery conditions via the dual variables associated with the LP, which may be of interest in its own right.
- We also derive a lower bound, based on Fano's inequality, characterizing the number of tests required to identify L inactive variables.
  - The upper bounds on the number of tests for different algorithms are within $O(\log^2 K)$ factor of the presented lower bounds.
- Finally, we present numerical simulations to compare the relative performance of the algorithms. The results also illustrate the significant benefit in finding non-defective items directly, compared to using the existing defective set recovery methods or testing items one-by-one, in terms of the number of group tests required.

The rest of the paper is organized as follows. Section II describes the NNGT-R framework and the problem setup. The proposed algorithms and the main analytical results are presented in Section III. Lower bounds on the number of tests are presented in Section IV. In Section V, we discuss the theoretical guarantees obtained and contrast them with available results on defective set recovery. Proofs of the main results are provided in Section VI. Section VII discusses numerical simulation results, and Section VIII presents some concluding remarks. We conclude this section by presenting the notation followed throughout the paper.

Notation: For any positive integer a, $[a] \triangleq \{1, 2, \ldots, a\}$. For any set A, A^c denotes complement operation and $|A|$

[1]Note that directly computing correlations between the column vector for an item and the test outcome vector will not work in case of group testing, as both the vectors are boolean. Furthermore, positive and negative pools have asymmetric roles in the group testing problem.

denotes the cardinality of the set. For any two sets A and B, $A\backslash B = A \cap B^c$. $\{\emptyset\}$ denotes the null set. Scalar random variables (RVs) are represented by capital non-bold alphabets, e.g., $\{Z_1, Z_3, Z_5, Z_8\}$ represent a set of 4 scalar RVs. If the index set is known, we also use the index set as subscript, e.g., Z_S, where $S = \{1, 3, 5, 8\}$. Matrices are denoted using uppercase bold letters and vectors are denoted using an underline. For a given matrix $\mathbf{A}$, $\underline{a}_i^{(r)}$ and $\underline{a}_i$ denote the i^{th} row and column, respectively. For a given index set S, $\mathbf{A}(S, :)$ denotes a sub-matrix of $\mathbf{A}$ where only the rows indexed by set S are considered. Similarly, $\mathbf{A}(:, S)$ or $\mathbf{A}_S$ denotes a sub-matrix of $\mathbf{A}$ that consists only of columns indexed by set S. For a vector $\underline{a}$, $\underline{a}(i)$ denotes its i^{th} component; $\text{supp}(\underline{a}) \triangleq \{j : \underline{a}(j) > 0\}$; $\{\underline{a} = c\}$ denotes the set $\{j : \underline{a}(j) = c\}$ for any c. In the context of a boolean vector, $\underline{a}^c$ denotes the component wise boolean complement of $\underline{a}$. $\underline{1}_n$ and $\underline{0}_n$ denote an all-one and all-zero vector, respectively, of size $n \times 1$. We denote the component wise inequality as $\underline{a} \preccurlyeq \underline{b}$, i.e., it means $\underline{a}(i) \leq \underline{b}(i)\ \forall\ i$. Also, $\underline{a} \circ \underline{b}$ denotes the component-wise product, i.e., $(\underline{a} \circ \underline{b})(i) = \underline{a}(i)\underline{b}(i),\ \forall\ i$. The boolean OR operation is denoted by $\bigvee$. For any $q \in [0, 1]$, $\mathcal{B}(q)$ denotes the Bernoulli distribution with parameter q. $\mathbb{I}_{\mathcal{A}}$ denotes the indicator function and returns 1 if the event $\mathcal{A}$ is true, else returns 0. Note that, $x(n) = O(y(n))$ implies that $\exists\ B > 0$ and $n_0 > 0$, such that $|x(n)| \leq B|y(n)|$ for all $n > n_0$. Further, $x(n) = \Omega(y(n))$ implies that $\exists\ B > 0$ and $n_0 > 0$, such that $|x(n)| \geq B|y(n)|$ for all $n > n_0$. Also, $x(n) = o(y(n))$ implies that for every $\epsilon > 0$, there exists an $n_0 > 0$ such that $|x(n)| \leq \epsilon|y(n)|$ for all $n > n_0$. All logarithms in this paper are to the base e. Also, for any $p \in [0, 1]$, $H_b(p)$ denotes the binary entropy in nats, i.e., $H_b(p) \triangleq -p\log(p) - (1-p)\log(1-p)$.

II. Signal Model

In our setup, we have a population of N items, out of which K are defective. Let $S_d \subset [N]$ denote the defective set, such that $|S_d| = K$. We consider a non-adaptive group testing framework with random pooling [3], [17], [18], [32], where the items to be pooled in a given test are chosen at random from the population. The group tests are defined by a boolean matrix, $\mathbf{X} \in \{0, 1\}^{M \times N}$, that assigns different items to the M group tests (pools). The j^{th} pool tests the items corresponding to the columns with 1 in the j^{th} row of $\mathbf{X}$. We consider an independent and identically distributed (i.i.d.) random Bernoulli measurement matrix [17], where each $X_{ij} \sim \mathcal{B}(p)$ for some $0 < p < 1$. Thus, M randomly generated pools are specified. In the above, p is a design parameter that controls the average group size, i.e., the average number of items being tested in a single group test. In particular, we choose $p = \frac{\alpha}{K}$, and a specific value of α is chosen based on the analysis of different algorithms.

If the tests are completely reliable, then the output of the M tests is given by the boolean OR of the columns of $\mathbf{X}$ corresponding to the defective set S_d. However, in practice, the outcome of a group test may be unreliable. Two popular noise models that are considered in the literature on group testing are [17], [18], [24]: (a) An *additive* noise model, where there is a probability, $q \in (0, 0.5)$, that the outcome of a group test containing only non-defective items turns out to be positive (Fig. 1); (b) A *dilution* model, where there is a probability, $u \in (0, 0.5)$, that a given item does not participate in a given group test (see Fig. 1). Let $\underline{d}_i \in \{0, 1\}^M$. Let $\underline{d}_i(j) \sim \mathcal{B}(1-u)$ be chosen independently for all $j = 1, 2, \ldots, M$ and for all $i = 1, 2, \ldots, N$. Let $\mathbf{D}_i \triangleq \text{diag}(\underline{d}_i)$. The output vector $\underline{y} \in \{0, 1\}^M$ can be represented as

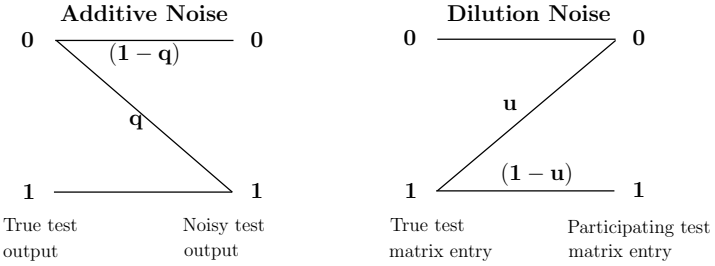


Fig. 1. Impact of different types of noise on the group testing signal model.

$$\underline{y} = \bigvee_{i=1}^{N} \mathbf{D}_i \underline{x}_i \mathbb{I}_{\{i \in S_d\}} \bigvee \underline{w}, \tag{1}$$

where $\underline{x}_i \in \{0, 1\}^M$ is the i^{th} column of $\mathbf{X}$, $\underline{w} \in \{0, 1\}^M$ is the additive noise with the i^{th} component $\underline{w}(i) \sim \mathcal{B}(q)$. Note that, for the noiseless case, $u = 0, q = 0$. Given the test output vector, $\underline{y}$, our goals are as follows:

(a) To find computationally efficient algorithms to find L non-defective items, i.e., an L-sized subset of $[N]\backslash S_d$.
(b) To analyze the performance of the proposed algorithms with the objective of *(i)* finding the number of tests required, and *(ii)* choosing the appropriate design parameters that leads to non-defective subset recovery with high probability of success.

In the literature on defective set recovery in group testing or on sparse vector recovery in compressed sensing, there exist two types of recovery results: (a) *Non-uniform/Per-Instance recovery results*: These derive conditions under which a randomly chosen test matrix leads to non-defective subset recovery with high probability of success for a given *fixed* defective set, and, (b) *Uniform/Universal recovery results*: These derive conditions under which a randomly chosen test matrix leads to a successful non-defective subset recovery with high probability for *all possible* defective sets. It is possible to easily extend non-uniform results to the uniform case using union bounds. Hence, we focus mainly on non-uniform recovery results, and show the extension to the uniform case for one of the proposed algorithms (see Corollary 1).

For later use, we summarize some key facts pertaining to the above signal model in the lemma below. For any $l \in [M]$ and $k \in [N]$, let X_{lk} denote the $(l, k)^{\text{th}}$ entry of the test matrix $\mathbf{X}$ and let $Y_l \triangleq \underline{y}(l)$ denote the l^{th} test output. With u, q and p as defined above, let $\Gamma \triangleq (1-q)\left(1-(1-u)p\right)^K$ and $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$. We claim that,

Lemma 1. *(a)* $\mathbb{P}(Y_l = 0) = \Gamma$.
(b) For any $j \notin S_d$, $\mathbb{P}(Y_l | X_{lj}) = \mathbb{P}(Y_l)$.
(c) For any $i \in S_d$, $\mathbb{P}(Y_l = 0 | X_{li} = 1) = \gamma_0\Gamma$ *and* $\mathbb{P}(Y_l = 0 | X_{li} = 0) = \frac{\Gamma}{1-(1-u)p}$. *Further, using Bayes' rule,* $\mathbb{P}(X_{li} = 1 | Y_l = 0) = p\gamma_0$.
(d) Given Y_l, X_{li} *is independent of* X_{lj} *for any* $i \in S_d$ *and* $j \notin S_d$.

(e) For a given output vector $\underline{y}$, the metric $\mathcal{T}(i,\underline{y})$ computed in (2) (see Section III-B) is a sum of independent RVs.

The proof is provided in Appendix A.

III. Algorithms and Main Results

We now present several algorithms for non-defective/healthy subset recovery. Each algorithm takes the observed noisy test-output vector $\underline{y} \in \{0,1\}^M$ and the test matrix $\mathbf{X} \in \{0,1\}^{M\times N}$ as inputs, and outputs a set of L items, $\hat{S}_L$, that have been declared non-defective. The recovery is successful if the declared set does not contain any defective item, i.e., $\hat{S}_L \cap S_d = \{\emptyset\}$. For each algorithm, we derive upper bounds on the average probability of error, which are further used to obtain sufficient conditions on the number of tests required for successful non-defective subset recovery.

A. Row Based Algorithm

Our first algorithm to find non-defective items makes use of the fact that, in the noiseless case, if the test outcome is negative, then all the items being tested are non-defective.

RoAl (Row based algorithm):

- Compute $\underline{z}(\underline{y}) = \sum_{j\in \text{supp}(\underline{y}^c)} \underline{x}_j^{(r)}$, where $\underline{x}_j^{(r)}$ is the j^{th} row of the test matrix.
- Order entries of $\underline{z}(\underline{y})$ in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

That is, declare the L items that have been tested most number of times in pools with negative outcomes as non-defective items. The above decoding algorithm proceeds by only considering the tests with negative outcomes. Note that, when the test outcomes are noisy, there is a nonzero probability of declaring a defective item as non-defective. In particular, the dilution noise can lead to a test containing defective items in the pool being declared negative, leading to a possible misclassification of the defective items. On the other hand, since the algorithm only considers tests with negative outcomes, additive noise does not lead to misclassification of defective items as non-defective. However, the additive noise does lead to an increased number of tests as the algorithm has to possibly discard many of the pools that contain only non-defective items.

We note that existing row based algorithms for finding the defective set [3], [18], can be obtained as a special case of the above algorithm by setting $L = N - K$, i.e., by looking for all non-defective items. However, the analysis in the past work does not quantify the impact of the parameter L and that is our main goal here. We characterize the number of tests, M, that are required to find L non-defective items with high probability of success using **RoAl** in Theorem 1.

B. Column Based Algorithm

The column based algorithm is based on matching the columns of the test matrix with the test outcome vector. A non-defective item does not impact the output and hence the corresponding column in the test matrix should be "uncorrelated" with the output. On the other hand, "most" of the pools that test a defective item should test positive. This forms the basis of distinguishing a defective item from a non-defective one. The specific algorithm is as follows:

CoAl (Column based algorithm): Let $\psi_{cb} \geq 0$ be any constant.

- For each $i = 1,\ldots,N$, compute
$$\mathcal{T}(i,\underline{y}) = \underline{x}_i^T \underline{y}^c - \psi_{cb}(\underline{x}_i^T \underline{y}), \tag{2}$$
where $\underline{x}_i$ is the i^{th} column of $\mathbf{X}$.
- Sort $\mathcal{T}(i,\underline{y})$ in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

We note that, in contrast to the row based algorithm, **CoAl** works with pools of both the negative and positive test outcomes (when the parameter $\psi_{cb} > 0$; its choice is explained below). For both **RoAl** and **CoAl**, by analyzing the probability of error, we can derive the sufficient number of tests required to achieve arbitrarily small error rates. We summarize the main result in the following theorem:

Theorem 1. *(Non-Uniform recovery with **RoAl** and **CoAl**) Let $\Gamma \triangleq (1-q)\left(1-(1-u)p\right)^K$ and $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$. Suppose $K > 1$ and let p be chosen as $\frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$. For **RoAl**, let $\psi_0 \triangleq 0$. For **CoAl**, choose $\psi_0 \triangleq \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$ and set $\psi_{cb} = \psi_0$. Let $c_0 > 0$ be any constant. Then, there exist constants $C_{a1}, C_{a2}, c_0' > 0$ independent of N, L and K, and different for each algorithm, such that, if the number of tests is chosen as*

$$M \geq (1+c_0)\frac{K(1-u)}{(1-q)(1-\gamma_0)^2(1+\psi_0)} \times \left(\frac{C_{a1}\log\left[K\binom{N-K}{L-1}\right]}{(N-K)-(L-1)} + C_{a2}\log K\right), \tag{3}$$

*then, for a given defective set, the algorithms **RoAl** and **CoAl** find L non-defective items with probability exceeding $1 - \exp\left(-c_0\log\left(K\binom{N-K}{L-1}\right)\right) - \exp(-c_0\log K) - 2\exp\left(-c_0' K\log K\right)$.*

The following corollary extends Theorem 1 to uniform recovery of a non-defective subset using **RoAl** and **CoAl**.

Corollary 1. *(Uniform recovery with **RoAl** and **CoAl**) For any positive constant $c_0 > 0$, there exist constants $C_{a1}, C_{a2}' > 0$ independent of N, L and K, and different for each algorithm, such that if the number of tests is chosen as*

$$M \geq (1+c_0)\frac{K(1-u)}{(1-q)(1-\gamma_0)^2(1+\psi_0)} \times \left(\frac{C_{a1}\log\left[K\binom{N-K}{L-1}\binom{N}{K}\right]}{(N-K)-(L-1)} + C_{a2}'\log N\right), \tag{4}$$

*then for **any** defective set, the algorithms **RoAl** and **CoAl** find L non-defective items with probability exceeding $1 - \exp\left(-c_0\log\left(K\binom{N-K}{L-1}\right)\right) - \exp(-c_0\log N) - 2\exp(-c_0 K\log N)$.*

The proof of the above theorem and corollary is presented in Section VI-A. It is tempting to compare the performance of

RoAl and **CoAl** by comparing the sufficient number of tests as presented in (3). However, such comparisons must be done keeping in mind that the bound on the number of tests in (3) is based on an upper bound on the average probability of error. The main objective of these results is to provide a guarantee on successful non-defective subset recovery and highlight the order-wise dependence of the number of tests on the system parameters. For the comparison of the relative performance of the algorithms, we refer the reader to Section VII, where we present numerical results obtained from simulations. From the simulations, we observe that **CoAl** performs better than **RoAl** for most scenarios of interest. This is because, in contrast to **RoAl**, **CoAl** uses the information obtained from pools corresponding to both negative and positive test outcomes.

C. Linear program relaxation based algorithms

In this section, we consider linear program (LP) relaxations to the non-defective subset recovery problem and identify conditions under which such LP relaxations lead to recovery of a non-defective subset with high probability of success. These algorithms are inspired by analogous algorithms studied in the context of defective set recovery in the literature [18], [27]. However, past analysis on the number of tests for the defective set recovery do not carry over to the non-defective subset recovery because the goals of the algorithms are very different. Let $Y_z \triangleq \{l \in [M] : \underline{\mathbf{y}}(l) = 0\}$, i.e., Y_z is the index set of all the pools whose test outcomes are negative, and let $M_z \triangleq |Y_z|$. Similarly, let $Y_p \triangleq \{l \in [M] : \underline{\mathbf{y}}(l) = 1\}$ and $M_p \triangleq |Y_p|$. Define the following linear program, with optimization variables $\underline{z} \in \mathbb{R}^N$ and $\underline{\eta}_z \in \mathbb{R}^{M_z}$:

$$\underset{\underline{z},\underline{\eta}_z}{\text{minimize}} \quad \underline{1}_{M_z}^T \underline{\eta}_z \tag{5}$$

$$\textbf{(LP0)} \quad \text{subject to} \quad \mathbf{X}(Y_z,:)(\underline{1}_N - \underline{z}) - \underline{\eta}_z = \underline{0}_{M_z}, \tag{6}$$
$$\underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N, \quad \underline{\eta}_z \succcurlyeq \underline{0}_{M_z},$$
$$\underline{1}_N^T \underline{z} \leq L.$$

Consider the following algorithm:[2]

RoLpAl (LP relaxation with negative outcome pools only)

- Setup and solve **LP0**. Let $\hat{\underline{z}}$ be the solution of **LP0**.
- Sort $\hat{\underline{z}}$ in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

The above program relaxes the combinatorial problem of choosing L out of N items by allowing the boolean variables to acquire "real" values between 0 and 1 as long as the constraints imposed by negative pools, specified in (6), are met. Intuitively, the variable $\underline{z}$ (or the variable $[\underline{1}_N - \underline{z}]$) can be thought of as the confidence with which an item is being declared as non-defective (or defective). The constraint $\underline{1}_N^T \underline{z} \leq L$ forces the program to assign high values (close to 1) for "approximately" the top L entries only, which are then declared as non-defective.

[2]The other algorithms presented in this sub-section, namely, **RoLpAl++** and **CoLpAl**, have the same structure and differ only in the linear program being solved.

For the purpose of analysis, we first derive sufficient conditions for correct non-defective subset recovery with **RoLpAl** in terms of the dual variables of **LP0**. We then derive the number of tests required to satisfy these sufficiency conditions with high probability. The following theorem summarizes the performance of the above algorithm:

Theorem 2. *(Non-Uniform recovery with **RoLpAl**) Let $K > 1$ and let p be chosen as $\frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$. If the number of tests is chosen as in (3) with $\psi_0 = 0$, then for a given defective set there exist constants $C_{a1}, C_{a2} > 0$ independent of N, L and K, such that **RoLpAl** finds L non-defective items with probability exceeding $1 - \exp\left(-c_0 \log\left(K\binom{N-K}{L-1}\right)\right) - \exp(-c_0 \log K) - 2\exp\left(-c_0' K \log K\right)$.*

The proof of the above theorem is presented in Section VI-B. Note that **LP0** operates only on the set of pools with negative outcomes and is, thus, sensitive to the dilution noise which can lead to a misclassification of a defective item as non-defective. To combat this, we can leverage the information available from the pools with positive outcomes also, by incorporating constraints for variables involved in these tests. Consider the following linear program with optimization variables $\underline{z} \in \mathbb{R}^N$ and $\underline{\eta}_z \in \mathbb{R}^{M_z}$:

$$\underset{\underline{z},\underline{\eta}_z}{\text{minimize}} \quad \underline{1}_{M_z}^T \underline{\eta}_z \tag{7}$$

$$\textbf{(LP1)} \quad \text{subject to} \quad \mathbf{X}(Y_z,:)(\underline{1}_N - \underline{z}) - \underline{\eta}_z = \underline{0}_{M_z}$$
$$\mathbf{X}(Y_p,:)(\underline{1}_N - \underline{z}) \succcurlyeq (1-\epsilon_0)\underline{1}_{M_p} \tag{8}$$
$$\underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N, \quad \underline{\eta}_z \succcurlyeq \underline{0}_{M_z}$$
$$\underline{1}_N^T \underline{z} \leq L.$$

In the above, $0 < \epsilon_0 \ll 1$ is a small positive constant. Note that (8) attempts to model, in terms of real variables, a boolean statement that at least one of the items tested in tests with positive outcomes is a defective item. We refer to the algorithm based on **LP1** as **RoLpAl++**. We expect **RoLpAl++** to outperform **RoLpAl**, as the constraint (8) can provide further differentiation between items that are indistinguishable just on the basis of negative pools. Note that, due to the constraint $\underline{1}_N^T \underline{z} \leq L$, the entries of $\hat{\underline{z}}$ in $[N] \backslash \hat{S}_L$ are generally assigned small values. Hence, when L is small, for many of the positive pools, the constraint (8) may not be active. Thus, we expect **RoLpAl++** to perform better than **RoLpAl** as the value of L increases; this will be confirmed via simulation results in Section VII. Due to the difficulty in obtaining estimates for the dual variables associated with the constraints (8), it is hard to derive theoretical guarantees for **RoLpAl++**.

Motivated by the connection between **RoAl** and **RoLpAl**, as revealed in the proof of Theorem 2 (see Section VI-B), we now propose another LP based non-defective subset recovery algorithm that incorporates both positive and negative pools, which, in contrast to **RoLpAl++**, turns out to be analytically tractable. By incorporating (8) in an unconstrained form and by using the *same* weights for all the associated Lagrangian

multipliers in the optimization function, we get

$$\begin{aligned} \underset{\underline{z}}{\text{minimize}} \quad & \underline{1}_{M_z}^T \mathbf{X}(Y_z,:)(\underline{1}_N - \underline{z}) \\ & - \psi_{lp}\left[\underline{1}_{M_p}^T \mathbf{X}(Y_p,:)(\underline{1}_N - \underline{z})\right] \quad (9) \\ \textbf{(LP2)} \quad \text{subject to} \quad & \underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N, \\ & \underline{1}_N^T \underline{z} \leq L, \end{aligned}$$

where $\psi_{lp} > 0$ is a positive constant that assigns appropriate weight to the two different type of cumulative errors. Note that, compared to **LP1**, we have also eliminated the equality constraints in the above program. The intuition is that, by using (8) in an unconstrained form, i.e., by maximizing $\sum_{j\in Y_z} \mathbf{X}(j,:)(\underline{1}_N - \underline{z})$, the program will tend to assign higher values to $(1-\hat{\underline{z}}(i))$ (and hence lower values to $\hat{\underline{z}}(i)$) for $i \in S_d$ since for random test matrices with i.i.d. entries, the defective items are likely to be tested more number of times in the pools with positive outcomes. Also, in contrast to **LP1** where different weight is given to each positive pool via the value of the associated dual variable, **LP2** gives the same weight to each positive pool, but it adjusts the overall weight of positive pools using the constant ψ_{lp}. We refer to the algorithm based on **LP2** as **CoLpAl**. The theoretical analysis for **CoLpAl** follows on similar lines as **RoLpAl** and we summarize the main result in the following theorem:

Theorem 3. *(Non-Uniform recovery with **CoLpAl**) Let* $\Gamma \triangleq (1-q)\left(1-(1-u)p\right)^K$ *and* $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$. *Let* $K > 1$ *and let* p *be chosen as* $\frac{\alpha}{K}$ *with* $\alpha = \frac{1}{(1-u)}$. *Let* $\psi_0 \triangleq \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$ *and set* $\psi_{lp} = \psi_0$. *Then, for any positive constant* c_0, *there exist constants* $C_{a1}, C_{a2}, c_0' > 0$ *independent of* N, L *and* K, *such that, if the number of tests is chosen as in (3), then for a given defective set* **CoLpAl** *finds* L *non-defective items with probability exceeding* $1 - \exp\left(-c_0 \log\left(K\binom{N-K}{L-1}\right)\right) - \exp(-c_0 \log K) - 2\exp\left(-c_0' K \log K\right)$.

The proof of the above theorem is presented in Section VI-C.

IV. Necessary Number of Observations

In this section, we derive information theoretic lower bounds on the number of tests required to identify L non-defective items. We consider the general sparse signal model employed in [17], [32] in the context of support recovery problem. It is a generalization of the signal models employed in some of the popular non-adaptive measurement system signal models such as compressed sensing and non-adaptive group testing. Thus, the lower bounds obtained here are more general, and can be applied in a variety of practical scenarios. We start by briefly describing the signal model before presenting the main result of this section. For more details on the signal model, we refer the reader to [17], [22], [32].

Let $X_{[N]} = \left[X_1, X_2, \ldots, X_N\right]$ denote a set of N independent and identically distributed input random variables (or *items*). Let each X_j belong to a finite alphabet denoted by $\mathcal{X}$ and be distributed as $\Pr\{X_j = x\} = Q(x), x \in \mathcal{X}$, $j = 1, 2, \ldots, N$. For a group of input variables, e.g., $X_{[N]}$, $Q(X_{[N]}) = \prod_{j\in[N]} Q(X_j)$ denotes the known joint distribution for all the input variables. We consider a sparse signal model where only a subset of the input variables are *active* (or *defective*), in the sense that only a subset of the input variables contribute to the output. Let $S \subset [N]$ denote the set of active input variables, with $|S| = K$. We assume that K is known. Let $S^c \triangleq [N]\backslash S$ denote the set of variables that are *inactive* (or *non-defective*). Let the output belong to a finite alphabet denoted by $\mathcal{Y}$. We assume that Y is generated according to a known conditional distribution $P(Y|X_{[N]})$. Then, in our observation model, we assume that given the active set, S, the output signal, Y, is independent of the other input variables. That is, for every $Y \in \mathcal{Y}$,

$$P(Y|X_{[N]}) = P(Y|X_S). \quad (10)$$

We observe the outputs corresponding to M independent realizations of the input variables, and denote the inputs and the corresponding observations by $\{\mathbf{X}, \underline{\mathbf{y}}\}$. Here, $\mathbf{X}$ is an $M \times N$ matrix, with its i^{th} row representing the i^{th} realization of the input variables, and $\underline{\mathbf{y}}$ is an $M \times 1$ vector, with its i^{th} component representing the i^{th} observed output. Note that, the independence assumption across the input variables and across different observations implies that entries of $\mathbf{X}$ are i.i.d. Let $L \leq N - K$. We wish to find *a set* of L inactive variables, i.e., an index set $S_H \subset S^c$ such that $|S_H| = L$, given the observation set, $\{\mathbf{X}, \underline{\mathbf{y}}\}$. In this setting, our goal is to derive an information theoretic lower bound on the number of observations (measurements/group tests) required to find a set of L inactive variables with the probability of error exponentially decreasing with the number of observations. Here, an error event occurs if the chosen inactive set contains one or more active variables (or *items*).

We now relate this model to the group testing signal model described in (1) in Section II. Note that, $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$. Each item in the group testing framework corresponds to one of the N input covariates. The i^{th} *row* of the test matrix, which specifies the i^{th} random pool, corresponds to the i^{th} realization of the input covariates. From (1), given the defective set S_d, the i^{th} test outcome $\underline{\mathbf{y}}(i)$ is independent of values of input variables from the set $[N]\backslash S_d$. That is, with regards to test outcome, it is *irrelevant* whether the items from the set $[N]\backslash S_d$ are included in the test or not. Thus, S_d corresponds to the active set S. The probability distribution functions $P(\underline{\mathbf{y}}|\mathbf{X}_{S_d})$ for any S_d, are fully determined from (1) and the statistical models for the dilution and additive noise. Thus, it is easy to see that the group testing framework is a special case of the general sparse model that we have considered, and, the number of group tests correspond directly to the number of observations in the context of sparse models.

We now derive lower bound on the number of observations required to find a set of L inactive variables, in the sense that if the number of observations is lower than the bound, the probability of error will be bounded strictly away from zero, regardless of the decoding algorithm used. Here, we need to lower bound the probability of error in choosing a set of L inactive variables. To this end, we employ an adaptation of Fano's inequality [33], [34].

Let $\omega \in \mathcal{I}^d \triangleq \left\{1, 2, \ldots, \binom{N}{K}\right\}$ denote the index of the defective set such that $S_\omega \subset [N]$ and $|S_\omega| = K$. For each $\omega \in \mathcal{I}^d$ let us associate a collection of sets, $\mathcal{S}_\omega^h \triangleq \left\{S_{\alpha_1}, S_{\alpha_2}, \ldots, S_{\alpha_{\binom{N-K}{L}}}\right\}$, such that $|S_{\alpha_i}| = L$ and

$S_{\alpha_i} \cap S_\omega = \{0\}$, $i \in \mathcal{I}_\omega^h \triangleq \left\{1, 2, \ldots, \binom{N-K}{L}\right\}$. That is, $\mathcal{S}_\omega^h$ is the collection of all L-sized subsets of all-inactive variables when S_ω represents the active set. Also, let $\mathcal{S}^H$ denote the set of all L-sized subsets of $[N]$. Note that $|\mathcal{S}^H| = \binom{N}{L}$. Given the observation vector, $\underline{\mathbf{y}} \in \mathcal{Y}^M$, let $\phi : \mathcal{Y}^M \times \mathcal{X}^{M\times N} \to \mathcal{S}^H$ denote a decoding function, such that $\hat{S} = \phi(\underline{\mathbf{y}}, \mathbf{X})$ is the decoded set of L inactive variables. Given an active set ω and an observation vector $\underline{\mathbf{y}}$, an error occurs if $\hat{S} \notin \mathcal{S}_\omega^h$. Define,

$$P_e = \Pr(\hat{S} \notin \mathcal{S}_\omega^h). \tag{11}$$

Define a binary error random variable $E \triangleq \mathbb{I}_{\left\{\hat{S} \notin \mathcal{S}_\omega^h\right\}}$. Note that $P_e = \Pr(E = 1)$. Finally, we define a mutual information term that will be used in the result below. Let S be a given active set. For any $1 \le j \le K$, let $S^{(j)}$ and $S^{(K-j)}$ represent a partition of S such that $S^{(j)} \cup S^{(K-j)} = S$, $S^{(j)} \cap S^{(K-j)} = \{\emptyset\}$ and $|S^{(j)}| = j$. Define $I^{(j)} \triangleq I(Y, X_{S^{(K-j)}}; X_{S^{(j)}}) = I(Y; X_{S^{(j)}} | X_{S^{(K-j)}})$ as the mutual information between $\{Y, X_{S^{(K-j)}}\}$ and $X_{S^{(j)}}$ [33], [34]. Mathematically,

$$I^{(j)} = \sum_{Y\in\mathcal{Y}} \sum_{X_{S^{(K-j)}} \in \mathcal{X}^{K-j}} \sum_{X_{S^{(j)}} \in \mathcal{X}^j} P(Y, X_{S^{(K-j)}} | X_{S^{(j)}}) \times Q(X_{S^{(j)}}) \log \frac{P(Y, X_{S^{(K-j)}} | X_{S^{(j)}})}{P(Y, X_{S^{(K-j)}})}. \tag{12}$$

Using the independence assumptions in the signal model, by the symmetry of construction of the test matrix $\mathbf{X}$, for a given j, $I^{(j)}$ is independent of the specific choice of S, and of the specific partitions of S.

We state a necessary condition on the number of observations in the following theorem.

Theorem 4. *Let N, M, L and K be as defined before. Let $I^{(j)}$ and P_e be as defined in (12) and (11), respectively. A necessary condition on the number of observations M required to find L inactive variables with asymptotically vanishing probability of error, i.e., $\lim_{N\to\infty} P_e = 0$, is given by*

$$M \ge \max_{1\le j\le K} \frac{\Gamma_l(L, N, K, j)}{I^{(j)}}(1-\eta), \quad \textit{where} \quad \Gamma_l(L, N, K, j) \triangleq \log\left[\frac{\binom{N-K+j}{j}}{\binom{N-K+j-L}{j}}\right], \tag{13}$$

for some $\eta > 0$.

The proof is provided in Section VI-D.

Given a specific application, we can bound $I^{(j)}$ for each $j = 1, 2, \ldots, K$, and obtain a characterization on the necessary number of observations/group tests. To compute the lower bounds on the number of tests, we need to upper bound the mutual information term, $I^{(j)}$, for the group testing signal model given in (1). Using the bounds on $I^{(j)}$ [35], with[3] $p = \frac{1}{K}$ and $u \le 0.5$, we summarize the order-accurate lower bounds on the number of tests to find a set of L non-defective items in Table I. A brief sketch of the derivation of these results is provided in Appendix E.

V. Discussion on the Theoretical Guarantees

We now present some interesting insights by analyzing the number of tests required for correct non-defective subset identification by the proposed recovery algorithms. We note that the expression in (3) adapted for different algorithms differs only on account of the constants involved. This allows us to present a unified analysis for all the algorithms.

(a) Asymptotic analysis of M as $N \to \infty$: We consider the parameter regimes where $K, L \to \infty$ as $N \to \infty$. We note that, under these regimes, when the conditions specified in the theorems are satisfied, the probability of decoding error can be made arbitrarily close to zero. In particular, we consider the regime where $\frac{K}{N} \to \beta_0$, $\frac{L}{N} \to \alpha_0$, as $N \to \infty$, where $0 \le \beta_0 < \alpha_0 < 1$, $\alpha_0 + \beta_0 < 1$. Define $\zeta \triangleq \frac{L-1}{N-K}$, and note that $\zeta \to \zeta_0 \triangleq \frac{\alpha_0}{1-\beta_0}$ as $N \to \infty$. Also, note that $\gamma_0 \to u$ as $N \to \infty$. Using Stirling's formula, it can be shown that $\lim_{N\to\infty} \frac{\log\binom{N-K}{L-1}}{(N-K)-(L-1)} \le \frac{H_b(\zeta_0)}{1-\zeta_0}$ (see [22]), where $H_b(\cdot)$ is the binary entropy function. Further, let $g(\zeta) \triangleq \frac{H_b(\zeta)}{1-\zeta}$. Now, since $g(\zeta_0)$ is a constant, the sufficient number of tests M for the proposed algorithms depends on K as $M \ge C_0 \frac{K}{(1-u)(1-q)} \left(C_{a1} g(\zeta_0) + C_{a2} \log K + o(1)\right)$. Here, C_0, C_{a1} and C_{a2} are constants independent of N, K, L, u and q.

We compare the above with the sufficient number of test required for the defective set recovery algorithms. When K grows sub-linearly with N (i.e., $\beta_0 = 0$), the sufficient number of tests for the proposed decoding algorithms is $O(K \log K)$, which is better than the sufficient number of tests for finding the defective set, which scales as $O(K \log N)$ [18], [24]. On the other hand, when K grows linearly with N (i.e., $\beta_0 > 0$), the performance of the proposed algorithms is order-wise equivalent to defective set recovery algorithms.

We also compare the uniform recovery results. The sufficient number of tests for uniform recovery as given in Corollary 1 for the algorithm **RoAl** and **CoAl** is $M = O(K \log N)$, which is significantly better than the defective set recovery algorithms, where the sufficient number of tests scale as $O(K^2 \log(\frac{N}{K}))$ [24].

(b) Variation of M with L: Let ζ and $g(\zeta)$ be as defined above. We note that the parameter L impacts M *only* via the function $g(\zeta)$. Lemma 3 in Appendix F shows that for small values (or even moderately high values) of ζ, $g(\zeta)$ is upper bounded by an affine function in ζ. This, in turn, shows that the sufficient number of tests is also approximately affine in L; this is also confirmed via simulation results in Section VII.

(c) Comparison with the information theoretic lower bounds: We compare with the lower bounds on the number of tests for non-defective subset recovery, as tabulated in Table I. For the noiseless case, i.e., $u = 0, q = 0$, the sufficient number of tests are within $O(\log^2 K)$ factor of the lower bound. For the additive noise only case, M depends on q via the multiplicative term $\frac{1}{1-q}$. In contrast, the lower bounds indicate that the number of tests is insensitive to additive noise, when q is close to 0 (in particular, when $q < 1/K$). For the dilution noise case, M depends on u via a multiplicative factor $\frac{1}{(1-u)}$, which is the same as in the lower bound. We have also compared the number of tests obtained via simulations with an exact computation

[3] In general, $p = \frac{\alpha}{K}$, with α depending upon u and q, is useful for bounding the mutual information terms $I^{(j)}$ [17], [35].

of the lower bounds, and, interestingly, the algorithms appear to fall within $O(\log K)$ factor of the lower bounds (see Figure 4, Section VII).

Finally, we also compare our uniform recovery results (from Corollary 1) with the lower bound presented in [30, Theorem 5]. For the noiseless case, and in the parameter regime of interest, i.e., $\frac{K}{N} \to \beta_0$, $\frac{L}{N} \to \alpha_0$, as $N \to \infty$, the sufficient number of tests are within $O(\log N)$ of the lower bound presented in [30, Theorem 5].

(d) Defective set recovery via non-defective subset recovery: It is interesting to note that by substituting $L = N - K$ in (3), we get $M = O\left(\frac{K \log(N-K)}{(1-u)(1-q)}\right)$, which is order-wise the same as the number of tests required for defective set identification derived in the existing literature [18], [24].

(e) Robustness under uncertainty in the knowledge of K: The theoretical guarantees presented in the above theorems hold provided the design parameter p is chosen as $O(\frac{1}{(1-u)K})$. This requires the knowledge of u and K. Note that the implementation of the recovery algorithms do not require us to know the values of K or u. These system model parameters are only required to choose the value of p for constructing the test matrix. If u and K are unknown, similar guarantees can be derived, with a penalty on the number of tests. For example, choosing p as $O(1/K)$, i.e., independent of u, results in a $\frac{1}{1-u}$ times increase in the number of tests. The impact of using an imperfect value of K can also be quantified. Let $\hat{K}$ be the value used to design the test matrix and let $\Delta_k > 0$ be such that $\hat{K} = \Delta_k K$. That is, Δ_k parametrizes the estimation error in K. The number of measurements M depend on p as $M \propto \frac{1}{p\Gamma}$.[4] Using the fact that $(1-\alpha/n)^n \approx \exp(-\alpha)$ for large n to simplify Γ, with $p = O(\frac{1}{\Delta_k K})$, the number of tests increases approximately by a factor of $f_M(\Delta_k) \triangleq \Delta_k \exp\left((1-u)\left(\frac{1}{\Delta_k} - 1\right)\right)$ compared to the case with perfect knowledge of K, i.e., with $p = O(1/K)$. Thus, the proposed algorithms are robust to uncertainty in the knowledge of K. For example, with $u = 0.05$, $f_M(1.5) = 1.09$, i.e., a 50% error in the estimation of K leads to an increase in the number of tests by a factor of 1.09. Furthermore, the asymmetric nature of $f_M(\Delta_k)$ (e.g., $f_M(1.5) = 1.09$ and $f_M(0.5) = 1.3$) suggests that the algorithms are more robust when $\Delta_k > 1$ as compared to the case when $\Delta_k < 1$. We corroborate this behavior via numerical simulations, see Table III.

(f) Operational complexity: The execution of **RoAl** and **CoAl** requires $O(MN)$ operations, where M is the number of tests. The complexity of the LP based algorithms **RoLpAl**, **RoLpAl++** and **CoLpAl** are implementation dependent, but are, in general, much higher than **RoAl** and **CoAl**. For example, an interior-point method based implementation will require $O(N^2(M+N)^{3/2})$ operations [36]. Although this is higher than that of **RoAl** and **CoAl**, it is still attractive in comparison to the brute force search based maximum likelihood methods, due to its polynomial-time complexity.

[4]This follows from the proof for the Theorem 1 in section VI-A; see e.g., (25), (28).

TABLE I

FINDING A SUBSET OF L NON-DEFECTIVE ITEMS: ORDER RESULTS FOR NECESSARY NUMBER OF GROUP TESTS WHICH HOLD ASYMPTOTICALLY AS $N \to \infty$, $\frac{K}{N} \to \beta_0$, $\frac{L}{N} \to \alpha_0$ AND $\alpha_0 + \beta_0 < 1$ (SEE THEOREM 4 AND APPENDIX E).

No Noise ($u = 0, q = 0$)	$\Omega\left(\frac{K}{\log K} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$
Dilution Noise ($u > 0, q = 0$)	$\Omega\left(\frac{K}{(1-u)\log K} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$
Additive Noise ($u = 0, q > 0$)	$\Omega\left(\frac{K}{\min\left\{\log \frac{1}{q}, \log K\right\}} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$

TABLE II

FINDING A SUBSET OF L NON-DEFECTIVE ITEMS: RESULTS FOR SUFFICIENT NUMBER OF GROUP TESTS FOR THE PROPOSED ALGORITHMS WHICH HOLD ASYMPTOTICALLY AS $(N, K, L) \to \infty$, $\frac{K}{N} \to \beta_0$, $\frac{L}{N} \to \alpha_0$ WITH $0 < \beta_0 \leq \alpha_0 < 1$ AND $\alpha_0 + \beta_0 < 1$. DEFINE $g(\zeta_0) \triangleq \frac{H_b(\zeta_0)}{1-\zeta_0}$, WHERE $\zeta_0 \triangleq \frac{\alpha_0}{1-\beta_0}$. THE CONSTANTS $C_0, C_{a1}, C_{a2} > 0$ MAY DIFFER FOR DIFFERENT ALGORITHMS AND ARE INDEPENDENT OF N, L, K, u AND q.

No Noise	$C_0 K \left[C_{a1} g(\zeta_0) + C_{a2} \log K + o(1)\right]$
Dilution Noise ($u > 0, q = 0$)	$\frac{C_0 K}{1-u} \left[C_{a1} g(\zeta_0) + C_{a2} \log K + o(1)\right]$
Additive Noise ($u = 0, q > 0$)	$\frac{C_0 K}{1-q} \left[C_{a1} g(\zeta_0) + C_{a2} \log K + o(1)\right]$

VI. PROOFS OF THE MAIN RESULTS

We begin by defining some quantities and terminology that is common to all the proofs. In the following, we denote the defective set by S_d, such that $S_d \subset [N]$ and $|S_d| = K$. We assume that S_d is fixed (but arbitrary) and unknown. We denote the set of L non-defective items output by the decoding algorithm by $\hat{S}_L$. For a given defective set S_d, $\mathcal{E} \triangleq \left\{\hat{S}_L \cap S_d \neq \{\emptyset\}\right\}$ denotes the error event, i.e., the event that a given decoding algorithm outputs an incorrect non-defective subset, and $\Pr(\mathcal{E})$ denotes its probability. Also, let $\mathcal{E}(\underline{y})$ denote the error event for a given output $\underline{y}$, i.e., $\Pr(\mathcal{E}(\underline{y})) = \Pr(\mathcal{E}|\underline{y})$. Define $N_0 \triangleq (N-K) - (L-1)$. We further let $S_z \subset [N] \backslash S_d$ denote any set of non-defective items such that $|S_z| = N_0$. Also, we let $\mathcal{S}_z$ denote all such sets possible. Note that $|\mathcal{S}_z| = \binom{N-K}{L-1}$. Finally, recall from Lemma 1 (Section II), $\Gamma \triangleq (1-q)\left(1-(1-u)p\right)^K$ and $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$.

A. Proof of Theorem 1 and Corollary 1

The proof involves upper bounding the probability of non-defective subset recovery error of the decoding algorithms, **RoAl** and **CoAl**, and identifying the parameter regimes where they can be made sufficiently small.

For **CoAl**, recall that for a given output $\underline{y}$ we compute the metric $\mathcal{T}(i, \underline{y}) \triangleq \underline{x}_i^T \underline{y}^c - (\psi_{cb}) \underline{x}_i^T \underline{y}$ for each item i and output the set of items with the L largest metrics as the non-defective set. Clearly, for any item $i \in S_d$, if $i \in \hat{S}_L$, then there exists a set S_z of non-defective items such that for all items $j \in S_z$, $\mathcal{T}(j, \underline{y}) \leq \mathcal{T}(i, \underline{y})$. Thus, for **CoAl**,

$$\mathcal{E}(\underline{y}) \subset \bigcup_{i \in S_d} \{i \in \hat{S}_L\} \subset \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \left[\bigcap_{j \in S_z} \{\mathcal{T}(j, \underline{y}) \leq \mathcal{T}(i, \underline{y})\}\right]. \tag{14}$$

The algorithm **RoAl** succeeds when there exists a set of at least L non-defective items that have been tested more number of times than any of the defective items, in the tests with

negative outcomes. The number of times an item i is tested in tests with negative outcomes is given by $\underline{z}(i,\underline{y}) = \underline{x}_i^T \underline{y}^c$, which is computed by **RoAl**. Hence, for any item $i \in \hat{S}_d$, if $i \in \hat{S}_L$, then there exists a set S_z of non-defective items such that for all items $j \in S_z$, $\underline{z}(j,\underline{y}) \leq \underline{z}(i,\underline{y})$. Thus, (14) applies for **RoAl** also, except with $\mathcal{T}$ replaced with $\underline{z}$. Also, note that $\underline{z}(i,\underline{y}) = \mathcal{T}(i,\underline{y})|_{\psi_{cb}=0}$. This allows us to unify the subsequent steps in the proof for the two algorithms. We first work with the quantity $\mathcal{T}(i,\underline{y})$ and later specialize the results for each algorithm.

The overall intuition for the proof is as follows: For a given output and any i, since $\mathcal{T}(i,\underline{y})$ is a sum of independent random variables, it will tend to concentrate around its mean value. For any $i \in S_d$ and $j \notin S_d$, we will show that the mean value of $\mathcal{T}(j,\underline{y})$ is larger than that of $\mathcal{T}(i,\underline{y})$. Thus, we expect the probability of the error event defined in (14) to be small.

Define $\mathcal{G}_y \triangleq \{\underline{y} \ : \ (1-\eta)M\Gamma \leq \|\underline{y}^c\|_1 \leq (1+\eta)M\Gamma\}$. That is, $\mathcal{G}_y$ denotes the set of all $\underline{y}$'s where the number of tests with negative outcome lie between $[(1-\eta)M\Gamma \ \ (1+\eta)M\Gamma]$. We claim that

$$\text{Pr}(\mathcal{E}) \leq \sum_{\underline{y}\in\mathcal{G}_y} \text{Pr}(\mathcal{E}(\underline{y}))\text{Pr}(\underline{y}) + \text{Pr}(\{\underline{y} \in \mathcal{G}_y^c\}). \quad (15)$$

Let us now analyze $\text{Pr}(\mathcal{E}(\underline{y}))$ for any $\underline{y} \in \mathcal{G}_y$. Let $\underline{y}_L$ be an arbitrary observation vector such that $\|\underline{y}_L^c\|_1 = (1-\eta)M\Gamma$ and $\underline{y}_H$ be such that $\|\underline{y}_H^c\|_1 = (1+\eta)M\Gamma$.[5] For any $i \in S_d$ and for any $j \notin S_d$, define $\mu_i \triangleq \mathbb{E}(\mathcal{T}(i,\underline{y}_H))$, $\mu_j \triangleq \mathbb{E}(\mathcal{T}(j,\underline{y}_L))$, $\sigma_i^2 \triangleq \text{Var}(\mathcal{T}(i,\underline{y}_H))$ and $\sigma_j^2 \triangleq \text{Var}(\mathcal{T}(j,\underline{y}_L))$. We claim that

$$\mu_j = Mp\left(\Gamma^{'} - \psi_{cb}(1-\Gamma^{'})\right)$$
$$\text{and } \mu_i = Mp\left(\gamma_0\Gamma^{''} - \psi_{cb}(1-\Gamma^{''})\frac{1-\gamma_0\Gamma}{1-\Gamma}\right), \quad (16)$$
$$\sigma_j^2 \leq Mp\left(\Gamma^{'} + \psi_{cb}^2(1-\Gamma^{'})\right)$$
$$\text{and } \sigma_i^2 \leq Mp\left(\gamma_0\Gamma^{''} + \psi_{cb}^2(1-\Gamma^{''})\frac{1-\gamma_0\Gamma}{1-\Gamma}\right), \quad (17)$$

where $\Gamma^{'} = (1-\eta)\Gamma$ and $\Gamma^{''} = (1+\eta)\Gamma$. An explanation of the above equations in presented in Appendix B. To simplify (15) further, we present the following proposition:

Proposition 1. *Define $\tau \triangleq \frac{(\mu_j+\mu_i)}{2}$. Let $\epsilon_0 > 0$. Then, for all $\underline{y} \in \mathcal{G}_y$,*

$$Pr(\mathcal{E}(\underline{y})) \leq K\binom{N-K}{L-1}(P_{eh})^{N_0} + KP_{ed}, \quad (18)$$

where $P_{eh} \triangleq Pr\left(\{\mathcal{T}(j,\underline{y}_L) < \tau + \epsilon_0\}\right)$ for any $j \in S_z$ and $P_{ed} \triangleq Pr\left(\{\mathcal{T}(i,\underline{y}_H) > \tau\}\right)$ for any $i \in S_d$.

The proof of Proposition 1 is presented in Section VI-A3. The above definitions of P_{eh} and P_{ed} are unambiguous because the corresponding probabilities are independent of the specific choice of indices j and i, respectively. From Proposition 1, the upper bound on $\text{Pr}(\mathcal{E}(\underline{y}))$ applies to all $\underline{y} \in \mathcal{G}_y$, and thus, from (15), we get

$$\text{Pr}(\mathcal{E}) \leq K\binom{N-K}{L-1}(P_{eh})^{N_0} + KP_{ed} + \text{Pr}(\{\underline{y} \in \mathcal{G}_y^c\}). \quad (19)$$

To analyze $\text{Pr}(\{\underline{y} \in \mathcal{G}_y^c\})$, let us define $Z_l \triangleq \mathbb{I}_{\{\underline{y}(l)=0\}}$ for $l = 1, 2, \ldots, M$. We need to bound the probability that $\sum_{l=1}^M Z_l$ lies outside $[(1-\eta)M\Gamma \ \ (1+\eta)M\Gamma]$. Since the rows of the test matrix are i.i.d., Z_l are also i.i.d. with $\mathbb{E}(Z_l) = \Gamma$ (see Lemma 1). Thus, by using the multiplicative form of the Chernoff bound [37], [38][6], we get

$$\text{Pr}(\{\underline{y} \in \mathcal{G}_y^c\}) \leq 2\exp\left(-M\eta^2\frac{\Gamma}{3}\right). \quad (20)$$

Our next task is to bound P_{eh} and P_{ed} defined in the above proposition. For a given $\underline{y}$ and any k, since $\mathcal{T}(k,\underline{y})$ is a sum of M independent random variables (see Lemma 1), each bounded by $\max(1,\psi_{cb})$, we can use Bernstein's inequality [37] (also see Appendix G) to bound the probability of their deviation from their mean values. The choice of $\psi_{cb} = \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$ ensures that $\psi_{cb} < 1$.[7] Thus, for any $i \in S_d$, with $\delta_0 \triangleq \tau - \mu_i = \frac{\mu_j-\mu_i}{2}$,

$$P_{ed} = \text{Pr}\left(\mathcal{T}(i,\underline{y}_H) > \tau\right)$$
$$= \text{Pr}\left(\mathcal{T}(i,\underline{y}_H) > \mu_i + \delta_0\right) \leq \exp\left(-\frac{\delta_0^2}{2\sigma_i^2 + \frac{2}{3}\delta_0}\right). \quad (21)$$

Similarly, for any $j \in S_z$, we choose $\epsilon_0 = \frac{\mu_j-\tau}{2} = \frac{\mu_j-\mu_i}{4}$, and get

$$P_{eh} = \text{Pr}\left(\mathcal{T}(j,\underline{y}_L) < \tau + \epsilon_0\right)$$
$$= \text{Pr}\left(\mathcal{T}(j,\underline{y}_L) < \mu_j - \epsilon_0\right) \leq \exp\left(-\frac{\epsilon_0^2}{2\sigma_j^2 + \frac{2}{3}\epsilon_0}\right). \quad (22)$$

We now proceed separately for each algorithm to arrive at the final results. Before that, by choosing $p = \frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$, $\left[1 - \frac{(1-u)\alpha}{K}\right]^K \geq \exp\left(-2\alpha(1-u)\right) = e^{-2}$. This follows from the fact that for $0 < b < 1$, $(1-b) \leq e^{-b} \leq 1 - \frac{b}{2}$. Thus, $(1-q)e^{-1} \geq \Gamma \geq (1-q)e^{-2}$. We also note that $\gamma_0 < 1$ for any $u < 0.5$ and for all $K > 1$.

1) Proof for *RoAl*

For **RoAl**, $\psi_{cb} = 0$. From (16) and (17), $\mu_j - \mu_i = Mp\Gamma\left((1-\gamma_0) - \eta(1+\gamma_0)\right)$. Thus, for $\eta = \beta\frac{1-\gamma_0}{1+\gamma_0}$ for some $0 < \beta < 1$, $\mu_j - \mu_i = (1-\beta)(1-\gamma_0)Mp\Gamma > 0$. Recall, $\delta_0 = \frac{\mu_j-\mu_i}{2}$ and $\epsilon_0 = \frac{\mu_j-\mu_i}{4}$. Since $\sigma_j^2 \leq Mp(1-\eta)\Gamma$ and $\sigma_i^2 \leq Mp(1+\eta)\gamma_0\Gamma$, we have

$$2\sigma_i^2 + (2/3)\delta_0 < Mp\Gamma\left(2(1+\eta)\gamma_0 + [(1-\eta) - (1+\eta)\gamma_0]/3\right) < 2Mp\Gamma. \quad (23)$$

Similarly,

$$2\sigma_j^2 + (2/3)\epsilon_0 < Mp\Gamma\left(2(1-\eta) + [(1-\eta) - (1+\eta)\gamma_0]/6\right) < 3Mp\Gamma. \quad (24)$$

[5] Ideally, we should work with $\underline{y}_L$ and $\underline{y}_H$ such that $\|\underline{y}_L^c\|_1 = \lceil(1-\eta)M\Gamma\rceil$ and $\|\underline{y}_H^c\|_1 = \lfloor(1+\eta)M\Gamma\rfloor$. As will become clear in the subsequent analysis, the remainder terms due to the floor and ceiling operations are small and asymptotically vanish. Hence, for clarity of exposition, we present our analysis assuming that $(1+\eta)M\Gamma$ and $(1-\eta)M\Gamma$ are integers.

[6] For ease of reference, we have stated it in Appendix G.

[7] Since $u < 0.5$, it follows that $\gamma_0\Gamma < 0.5$.

Thus, from (21) and (22), we have

$$P_{ed} \leq \exp\left(-\frac{Mp\Gamma(1-\beta)^2(1-\gamma_0)^2}{8}\right)$$
$$\text{and } P_{eh} \leq \exp\left(-\frac{Mp\Gamma(1-\beta)^2(1-\gamma_0)^2}{48}\right). \quad (25)$$

Thus, choosing $p = \frac{1}{(1-u)K}$ and noting that $\Gamma \geq e^{-2}(1-q)$, from (19) and (20) we get,

$$\mathbb{P}(\mathcal{E}) \leq \exp\left[-\frac{M(1-\gamma_0)^2(1-q)N_0}{C_{a1}K(1-u)} + \log\left(K\binom{N-K}{L-1}\right)\right] + 2\exp\left[-\frac{M\beta^2(1-\gamma_0)^2\Gamma}{3(1+\gamma_0)^2}\right] + \exp\left[-\frac{M(1-\gamma_0)^2(1-q)}{C_{a2}K(1-u)} + \log K\right], \quad (26)$$

with $C_{a1} = \frac{48e^2}{(1-\beta)^2}$ and $C_{a2} = \frac{8e^2}{(1-\beta)^2}$. Thus, if M is chosen as specified in (3), with the constants C_{a1}, C_{a2} chosen as above, then the error probability is upper bounded by $\exp\left(-c_0 \log\left[K\binom{N-K}{L-1}\right]\right) + \exp(-c_0 \log K)$ + $2\exp(-c_0' K \log K)$, where $c_0' \geq \frac{C_{a2}(1-u)}{(1-\gamma_0)^2(1-q)} \frac{\beta^2(1-\gamma_0)^2\Gamma}{3(1+\gamma_0)^2} > \frac{\beta^2}{3(1-\beta)^2}$.

*2) Proof for **CoAl***

We first note that with $\psi_{cb} = \psi_0$, where $\psi_0 \triangleq \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$, we have (see Appendix B)

$$\mu_j - \mu_i = Mp\Gamma(1+\psi_0)\left[(1-\gamma_0) - \eta\left(1+\gamma_0\frac{1-\gamma_0\Gamma}{1-\Gamma}\right)\right]. \quad (27)$$

Thus, for $\eta = \beta\frac{1-\gamma_0}{1+\gamma_0\frac{1-\gamma_0\Gamma}{1-\Gamma}}$ for any $0 < \beta < 1$, $\mu_j - \mu_i = Mp\Gamma(1+\psi_0)(1-\gamma_0)(1-\beta) > 0$. With $\psi_{cb} = \psi_0$, we have $\sigma_i^2 \leq Mp\gamma_0\Gamma\left(1+\eta+\psi_0\frac{1-\Gamma''}{1-\Gamma}\right) \leq 2Mp\gamma_0\Gamma(1+\psi_0)$. Also, we note that $\psi_0 < 1$. Thus, $2\sigma_i^2 + (2/3)\delta_0 < Mp\Gamma(1+\psi_0)\,(4\gamma_0 + (1-\beta)(1-\gamma_0)/3) \leq 4Mp\Gamma(1+\psi_0)$. Thus, from (21), we get

$$P_{ed} \leq \exp\left(-\frac{Mp\Gamma(1+\psi_0)(1-\beta)^2(1-\gamma_0)^2}{32}\right). \quad (28)$$

With ψ_0 as above, $\sigma_j^2 \leq Mp\Gamma(1+\psi_0)(1+\gamma_0)$ and thus, $2\sigma_j^2 + (2/3)\epsilon_0 < Mp\Gamma(1+\psi_0)\left(2+2\gamma_0+\frac{(1-\gamma_0)}{6}\right) < 4Mp\Gamma(1+\psi_0)$, since $1+\psi_0 = \frac{1}{1-\gamma_0\Gamma}$. Thus, from (22), we get

$$P_{eh} \leq \exp\left(-\frac{Mp\Gamma(1+\psi_0)(1-\beta)^2(1-\gamma_0)^2}{64}\right). \quad (29)$$

The next steps follow exactly as for **RoAl**. Hence, if M is chosen as specified in (3), with the constant C_{a1} and C_{a2} chosen as $\frac{64e^2}{(1-\beta)^2}$ and $\frac{32e^2}{(1-\beta)^2}$, respectively, then the error probability remains smaller than $\exp\left(-c_0 \log\left[K\binom{N-K}{L-1}\right]\right) + \exp(-c_0 \log K) + \exp(-c_0 \log K) + 2\exp(-c_0' K \log K)$.

3) Proof of Proposition 1

For $i \in S_d$, define $\mathcal{H}_i(\underline{y}) \triangleq \{\mathcal{T}(i,\underline{y}) \leq \tau\}$. The error event in (14) is a subset of the right hand side in the following equation:

$$\mathcal{E}(\underline{y}) \subset \bigcup_{i\in S_d}\left(\left\{\bigcup_{S_z\in\mathcal{S}_z}\bigcap_{j\in S_z}(\mathcal{E}_{ij}(\underline{y})\cap\mathcal{H}_i(\underline{y}))\right\}\cup\overline{\mathcal{H}_i(\underline{y})}\right), \quad (30)$$

where $\mathcal{E}_{ij}(\underline{y}) \triangleq \{\mathcal{T}(j,\underline{y}) \leq \mathcal{T}(i,\underline{y})\}$ for any $i \in S_d$ and $j \in S_z$. In the above, we have used the fact that, for any two sets A and B, $A \subset \{A\cap B\}\cup\overline{B}$. Further, using monotonicity properties, we have

$$\{\mathcal{E}_{ij}(\underline{y})\cap\mathcal{H}_i(\underline{y})\} \subset \{\mathcal{T}(j,\underline{y}) \leq \tau\} \subset \{\mathcal{T}(j,\underline{y}) < \tau+\epsilon_0\}, \quad (31)$$

where $\epsilon_0 > 0$ is any constant.

We note that given $\underline{y}$, $\mathcal{T}(l,\underline{y})$ for any item l (see (2)), can be represented as a difference of sums of i.i.d. random variables. The number of variables involved in each sum depends only on the number of zeros in $\underline{y}$, which is restricted to $[(1-\eta)M\Gamma \;\; (1+\eta)M\Gamma]$ for $\underline{y} \in \mathcal{G}_y$ due to the way the set $\mathcal{G}_y$ was chosen. Thus, for any $i \in S_d$, $j \notin S_d$ and for all $\underline{y} \in \mathcal{G}_y$,

$$\Pr(\{\mathcal{T}(j,\underline{y}) < \tau+\epsilon_0\}) \leq \Pr(\{\mathcal{T}(j,\underline{y}_L) < \tau+\epsilon_0\})$$
$$\text{and } \Pr(\{\mathcal{T}(i,\underline{y}) > \tau\}) \leq \Pr(\{\mathcal{T}(i,\underline{y}_H) > \tau\}). \quad (32)$$

Further, we note that given $\underline{y}$, $\mathcal{T}(j_1,\underline{y})$ is independent of $\mathcal{T}(j_2,\underline{y})$ for any $j_1, j_2 \in S_z$, $j_1 \neq j_2$ as these can be represented as a function of only $\underline{x}_{j_1}$ and $\underline{x}_{j_2}$, respectively, and $\underline{y}$ does not depend upon $\underline{x}_{j_1}$ and $\underline{x}_{j_2}$ (see Lemma 1). Using this observation, the claim in the proposition now follows from (30), (31) and (32) by accounting for the cardinalities of different sets involved in the union bounding.

4) Proof of Corollary 1

For the uniform case, we use the union bound over all possible choices of the defective set. The proof of the corollary follows the same steps as the proof of Theorem 1; the only difference comes on account of the additional union bounding that has to be done to account for all possible choices of the defective set. Here, we discuss the multiplicative factors that have to be included because of this additional union bound. Let $\mathcal{S}_d$ denote the set of all possible defective sets. Note that $|\mathcal{S}_d| = \binom{N}{K}$. First, from (30) in the proof of Proposition 1, for a given $\underline{y}$,

$$\mathcal{E}(\underline{y}) \subset \left\{\bigcup_{S_d\in\mathcal{S}_d}\bigcup_{i\in S_d}\bigcup_{S_z\in\mathcal{S}_z}\bigcap_{j\in S_z}\left(\mathcal{E}_{ij}(\underline{y})\cap\mathcal{H}_i(\underline{y})\right)\right\} \bigcup\left\{\bigcup_{S_d\in\mathcal{S}_d}\bigcup_{i\in S_d}\overline{\mathcal{H}_i}(\underline{y})\right\}, \quad (33)$$

The first and second term above correspond to the first and second term in (19), respectively. Thus, for the first term in (19), an additional multiplicative factor of $\binom{N}{K}$ is needed to account for all possible defective sets. For the second term, observe that

$$\bigcup_{S_d\in\mathcal{S}_d}\bigcup_{i\in S_d}\overline{\mathcal{H}_i} \subseteq \bigcup_{i\in[N]}\overline{\mathcal{H}_i}. \quad (34)$$

Thus, for the second term in (19), the multiplicative factor of K in (18) gets replaced by a factor of N, and no additional combinatorial multiplicative factors are needed. Similarly, an additional multiplicative factor of $\binom{N}{K}$ is needed for the third term in (19). We also note a change in the third term in (26): $2\exp\left[-\frac{M(1-\gamma_0)^2(1-q)}{C_{a3}} + K\log N\right]$, where C_{a3} is some positive constant. Here, we have used the fact that $\log\binom{N}{K} \leq K\log N$. The corollary now follows by noting that C_{a2}' can be chosen as $\max\left\{C_{a2}, \frac{C_{a3}(1+\psi_0)}{1-u}\right\}$.

B. Proof of Theorem 2

Let $\mathbf{X} \in \{0,1\}^{M\times N}$ denote the random test matrix, $\underline{y} \in \{0,1\}^M$ the output of the group test, $Y_z(\underline{y}) \triangleq \{l \in [M] : \underline{y}(l) = 0\}$ with $M_z \triangleq |Y_z(\underline{y})|$, and $Y_p(\underline{y}) \triangleq \{l \in [M] : \underline{y}(l) =$

$1\}$ with $M_p \triangleq |Y_p(\underline{y})|$. Let $\mathbf{X}_z \triangleq \mathbf{X}(Y_z(\underline{y}),:) \in \{0,1\}^{M_z \times N}$ and $\mathbf{X}_p \triangleq \mathbf{X}(Y_p(\underline{y}),:) \in \{0,1\}^{M_p \times N}$. Although $\mathbf{X}_z$ and $\mathbf{X}_p$ depend upon $\underline{y}$, we omit explicitly denoting this dependence to keep the notation light.

For the ease of analysis of the LP described in (5), we work with the following equivalent program:

$$\textbf{(LP0a)} \quad \begin{aligned} \underset{\underline{z}}{\text{minimize}} \quad & \underline{1}_{M_z}^T \mathbf{X}_z \ \underline{z} \\ \text{subject to} \quad & \underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N, \\ & \underline{1}_N^T \underline{z} \geq (N-L). \end{aligned} \tag{35}$$

The above formulation has been arrived at by eliminating the equality constraints and replacing the optimization variable $\underline{z}$ by $(\underline{1}_N - \underline{z})$. Hence, the non-defective subset output by (35) is indexed by the *smallest* L entries in the solution of (LP0a) (as opposed to largest L entries in the solution of (LP0)). We know that strong duality holds for a linear program and that any pair of primal and dual optimal points satisfy the Karush-Kuhn-Tucker (KKT) conditions [39]. Hence, a characterization of the primal solution can be obtained in terms of the dual optimal points by using the KKT conditions. Let $\underline{\lambda}_1, \underline{\lambda}_2 \in \mathbb{R}^N$ and $\nu \in \mathbb{R}$ denote the dual variables associated with the inequality constraints in (LP0a). The KKT conditions for any pair of primal and dual optimal points corresponding to (LP0a) can be written as follows:

$$\underline{1}_{M_z}^T \mathbf{X}_z - \underline{\lambda}_1 + \underline{\lambda}_2 - \nu \underline{1}_N = \underline{0}_N \tag{36}$$

$$\underline{\lambda}_1 \circ \underline{z} = \underline{0}_N; \ \underline{\lambda}_2 \circ (\underline{z} - \underline{1}_N) = \underline{0}_N; \ \nu(\underline{1}_N^T \underline{z} - (N-L)) = 0; \tag{37}$$

$$\underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N; \ \underline{1}_N^T \underline{z} \geq (N-L); \ \underline{\lambda}_1 \succcurlyeq \underline{0}_N; \ \underline{\lambda}_2 \succcurlyeq \underline{0}_N; \ \nu \geq 0; \tag{38}$$

Let $(\underline{z}, \underline{\lambda}_1, \underline{\lambda}_2, \nu)$ be the primal, dual optimal point, i.e., a point satisfying the set of equations (36)-(38). Let S_d denote the set of defective items. Further, let $\hat{S}_L$ denote the index set corresponding to the smallest L entries, and hence the declared set of non-defective items, in the primal solution $\underline{z}$. We first derive a sufficient condition for successful non-defective subset recovery with **RoLpAl**.

Proposition 2. *If $\underline{\lambda}_2(i) > 0 \ \forall \ i \in S_d$, then for a given output $\underline{y}$, $\hat{S}_L \cap S_d = \{\emptyset\}$.*

Proof: See Appendix C.

Define $\theta_0 \triangleq \max_{\{i:\underline{\lambda}_1(i)=0\}} \underline{1}_{M_z}^T \mathbf{X}_z(:,i)$ and $\theta_1 \triangleq \min_{\{i:\underline{\lambda}_1(i)>0\}} \underline{1}_{M_z}^T \mathbf{X}_z(:,i)$. We relate θ_0, θ_1 and ν as follows:

Proposition 3. *The dual optimal variable ν satisfies*

$$\theta_0 \leq \nu < \theta_1.$$

Proof: See Appendix D.

Let $\mathcal{E}(\underline{y})$, $\mathrm{P}(\mathcal{E}(\underline{y}))$, S_z and $\mathcal{S}_z$ be as defined at the beginning of this section. The sufficiency condition presented in proposition 2 for successful non-defective subset recovery, in turn, leads to the following:

Proposition 4. *For a given output $\underline{y}$, the error event associated with **RoLpAl** satisfies*

$$\mathcal{E}(\underline{y}) \subseteq \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:,i) \ \geq \ \underline{1}_{M_z}^T \mathbf{X}_z(:,j), \forall j \in S_z \right\}. \tag{39}$$

Proof: Define $\mathcal{E}_0(i,\underline{y}) \triangleq \{\underline{\lambda}_2(i) = 0\}$. We first note, from (36), that for any $i \in [N]$

$$\underline{\lambda}_2(i) = 0 \implies \underline{1}_{M_z}^T \mathbf{X}_z(:,i) = \underline{\lambda}_1(i) + \nu \geq \nu. \tag{40}$$

From proposition 3 and (40),

$$\mathcal{E}_0(i,\underline{y}) \subseteq \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:,i) \geq \theta_0 \right\}. \tag{41}$$

We note that there exist strictly less than L items for which $\underline{\lambda}_1(i) > 0$; otherwise, from (37), the solution would violate the primal feasibility constraint: $\underline{1}_N^T \underline{z}(i) \geq (N-L)$. Thus, there exist at least $(N-K)-(L-1)$ non-defective items in the set $\{i : \underline{\lambda}_1(i) = 0\}$. From (41), there exists a set S_z of $(N-K)-(L-1)$ non-defective items such that $\left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:,i) \geq \underline{1}_{M_z}^T \mathbf{X}_z(:,j), \forall j \in S_z \right\}$. Taking the union bound over all possible S_z, we get

$$\mathcal{E}_0(i,\underline{y}) \subseteq \bigcup_{S_z \in \mathcal{S}_z} \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:,i) \geq \underline{1}_{M_z}^T \mathbf{X}_z(:,j), \forall j \in S_z \right\}, \tag{42}$$

and (39) now follows since using Proposition 2 we have $\mathcal{E}(\underline{y}) \subseteq \cup_{i \in S_d} \mathcal{E}_0(i,\underline{y})$. ∎

Note that, for a given $\underline{y}$ and i, the quantity $\underline{1}_{M_z}^T \mathbf{X}_z(:,i)$ is the same as the quantity $\mathcal{T}(i,\underline{y})$ with $\psi_{cb} = 0$ as defined in the proof of Theorem 1, and (39) is the same as (14). Thus, following the same analysis as in Section VI-A, if M satisfies (3) with $\psi_0 = 0$, the LP relaxation based algorithm **RoLpAl** succeeds in recovering L non-defective items with probability exceeding $1 - \exp\left(-c_0 \log\left[K\binom{N-K}{L-1}\right]\right) + \exp(-c_0 \log K) + \exp(-c_0 \log K) + 2\exp(-c_0' K \log K)$.

C. Proof for Theorem 3

We use the same notation as in Theorem 2 and analyze an equivalent program that is obtained by replacing $(1 - \underline{z})$ by $\underline{z}$. We note that **LP2** differs from **LP0** only in terms of the objective function, and the constraint set remains the same. Thus, the complimentary slackness and the primal dual feasibility conditions are the same as given in (37) and (38), respectively. The zero gradient condition for **LP2** is given by

$$\underline{1}_{M_z}^T \mathbf{X}_z - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p - \underline{\lambda}_1 + \underline{\lambda}_2 - \nu \underline{1}_N = \underline{0}_N. \tag{43}$$

Let $i \in S_d$, and for a given $\underline{y}$, define $\mathcal{E}_i(\underline{y}) \triangleq \{i \in \hat{S}_L\}$. Note that $\mathcal{E}(\underline{y}) \subseteq \cup_{i \in S_d} \mathcal{E}_i(\underline{y})$. Further, $\mathcal{E}_i(\underline{y}) \subseteq \mathcal{A}_i(\underline{y}) \cup \mathcal{B}_i(\underline{y})$, where $\mathcal{A}_i(\underline{y}) \triangleq \{\lambda_2(i) = 0\}$ and $\mathcal{B}_i(\underline{y}) \triangleq \left\{ \mathcal{E}_i(\underline{y}) \cap \{\lambda_2(i) > 0\} \right\}$. Let us first analyze $\mathcal{B}_i(\underline{y})$. Using similar arguments as in Propositions 2 and 3, it can be shown that,

$$\begin{aligned} \mathcal{B}_i(\underline{y}) &\subseteq \{\nu = 0\} \\ &\subseteq \bigcup_{S_z \in \mathcal{S}_z} \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:,j) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:,j) \leq 0, \forall j \in S_z \right\}, \end{aligned} \tag{44}$$

and

$$\bigcup_{i \in S_d} \mathcal{B}_i(\underline{y}) \subseteq \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \left\{ \mathcal{T}(j,\underline{y}) \leq 0, \forall j \in S_z \right\}. \tag{45}$$

Further, using similar arguments as in the proof of Theorem 2, it can be shown that

$$\begin{aligned} \mathcal{A}_i(\underline{y}) \subseteq \bigcup_{S_z \in \mathcal{S}_z} \Big\{ & \underline{1}_{M_z}^T \mathbf{X}_z(:,i) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:,i) \\ & \geq \underline{1}_{M_z}^T \mathbf{X}_z(:,j) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:,j), \forall j \in S_z \Big\}, \end{aligned}$$

and

$$\bigcup_{i \in S_d} \mathcal{A}_i(\underline{y}) \subseteq \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \left\{ \mathcal{T}(i,\underline{y}) \geq \mathcal{T}(j,\underline{y}), \forall j \in S_z \right\}. \tag{46}$$

We note that (46) is the same as (14) and we analyze it in a manner similar to Theorem 1. In particular, in proposition 1 we assert that: $\left\{ \mathcal{T}(i,\underline{y}) \geq \mathcal{T}(j,\underline{y}), \forall j \in S_z \right\} \subseteq$

$\{\mathcal{T}(i,\underline{y}) > \tau\} \cup \{\mathcal{T}(j,\underline{y}) \leq \tau, \forall j \in S_z\}$ for some τ. We note that $\{\mathcal{T}(j,\underline{y}) \leq 0, \forall j \in S_z\} \subset \{\mathcal{T}(j,\underline{y}) \leq \tau, \forall j \in S_z\}$ for any $\tau > 0$. That is, if $\tau > 0$, then we would have already accounted for the event $\mathcal{B}_i(\underline{y})$ while upper bounding $\mathcal{A}_i(\underline{y})$. It is easy to verify from the parameters chosen in the proof of Theorem 1 that the value of τ is indeed greater than zero. Thus, following the analysis as in Section VI-A, if M satisfies (3) with $\psi_0 \triangleq \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$, the LP relaxation based algorithm **CoLpAl** succeeds in recovering L non-defective items with probability exceeding $1 - \exp\left(-c_0 \log\left[K\binom{N-K}{L-1}\right]\right) + \exp(-c_0 \log K) + \exp(-c_0 \log K) + 2\exp(-c_0' K \log K)$. This concludes the proof.

D. Proof of Theorem 4: Necessary Number of Observations

For the purpose of this proof, recall that P_e was defined in (11). We need to prove that $\lim_{N\to\infty} P_e = 0$ implies the bound on the number of observations as given by (13). Towards that end, we first find, by lower bounding P_e, the conditions on M that will lead to the error probability being bounded away from zero. We consider a genie-aided lower bound, where we assume that the active set is partially known. Let ω denote the index of the defective set, which is distributed uniformly over the set $\mathcal{I}^d$. Suppose $K-j$ items from the defective set are revealed to us, with the $K-j$ items chosen uniformly at random from the K defective items. Define a partition for S_ω as $S_\omega = S^{(j)} \cup S^{(K-j)}$, where $|S^{(j)}| = j$ and $|S^{(K-j)}| = K-j$ and $S^{(j)} \cap S^{(K-j)} = \{\emptyset\}$. Let the set $S^{(K-j)}$ correspond to the set of defective items revealed to us. Conditioned on $S^{(K-j)}$, ω is uniformly distributed over the set of K sized subsets of indices in $\mathcal{I}^d$ containing the known set $S^{(K-j)}$. Now consider $H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)})$:

$$\begin{aligned}
&H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \\
&\quad = H(E|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) + H(\omega|E, \underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \quad (47)\\
&\quad \overset{(a)}{\leq} H_b(P_e) + (1-P_e)H(\omega|E=0, \underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \\
&\qquad + P_e H(\omega|E=1, \underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \quad (48)\\
&\quad \overset{(b)}{\leq} H_b(P_e) + (1-P_e)\log\binom{N-K+j-L}{j} \\
&\qquad + P_e H(\omega|\mathbf{X}, S^{(K-j)}) \quad (49)\\
&\quad \overset{(c)}{\leq} H_b(P_e) + (1-P_e)\log\binom{N-K+j-L}{j} \\
&\qquad + P_e \log\binom{N-K+j}{j}. \quad (50)
\end{aligned}$$

In the above, (a) follows since E is a binary RV and $H(E|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \leq H(E) = H_b(P_e) \leq 1$. Since the entropy of any discrete valued RV is bounded by the logarithm of the alphabet size, the second term in (b) is obtained by considering the cardinality of the remaining number of outcomes conditioned on the outcome of E. When $E=0$, i.e., when there is no error, the number of ways of choosing the set $S^{(j)}$ is $\binom{N-K+j-L}{j}$. The third term in (b) follows since conditioning reduces entropy. We obtain (c) by using the fact that $H(\omega|\mathbf{X}, S^{(K-j)}) = \log\binom{N-K+j}{j}$. Also,

$$\begin{aligned}
&H(\omega, E|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \\
&\quad = H(\omega|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) + H(E|\omega, \underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) \\
&\quad = H(\omega|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}), \quad (51)
\end{aligned}$$

since the decoder output is a function of $(\underline{\mathbf{y}}, \mathbf{X})$ and hence, $H(E|\omega, \underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) = 0$. Thus, we get

$$\begin{aligned}
\log\binom{N-K+j}{j} &= H(\omega|\mathbf{X}, S^{(K-j)}) \\
&= H(\omega|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) + I(\omega; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}) \\
&\overset{(a)}{\leq} H(\omega|\underline{\mathbf{y}}, \mathbf{X}, S^{(K-j)}) + I(\mathbf{X}_{S_\omega}; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}) \quad (52)\\
&\overset{(b)}{\leq} H_b(P_e) + \log\binom{N-K+j-L}{j} \\
&\qquad + P_e\Gamma_l(L,N,K,j) + I(\mathbf{X}_{S^{(j)}}; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}), \quad (53)
\end{aligned}$$

where (a) follows from data-processing inequality [34], and (b) follows from (50), (51) and using the fact that $I(\mathbf{X}_{S_\omega}; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}) = I(\mathbf{X}_{S^{(j)}}; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)})$. Further, we have

$$\begin{aligned}
&I(\mathbf{X}_{S^{(j)}}; \underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}) \\
&\quad = H(\underline{\mathbf{y}}|\mathbf{X}, S^{(K-j)}) - H(\underline{\mathbf{y}}|\mathbf{X}_{S^{(j)}}, \mathbf{X}, S^{(K-j)}) \quad (54)\\
&\quad \overset{(a)}{\leq} H(\underline{\mathbf{y}}|\mathbf{X}_{S^{(K-j)}}) - H(\underline{\mathbf{y}}|\mathbf{X}_{S^{(j)}}, \mathbf{X}_{S^{(K-j)}}) \\
&\quad = I(\mathbf{X}_{S^{(j)}}; \underline{\mathbf{y}}|\mathbf{X}_{S^{(K-j)}}), \quad (55)
\end{aligned}$$

where (a) follows since conditioning reduces entropy and $\underline{\mathbf{y}}$ depends on $(\mathbf{X}_{S^{(j)}}, \mathbf{X}, S^{(K-j)})$ only through $(\mathbf{X}_{S^{(j)}}, \mathbf{X}_{S^{(K-j)}})$. Using basic properties of entropy, mutual information and the i.i.d. assumption across observations, it can be shown that [17]:

$$I(\mathbf{X}_{S^{(j)}}; \underline{\mathbf{y}}|\mathbf{X}_{S^{(K-j)}}) \leq MI(X_{S^{(j)}}; Y|X_{S^{(K-j)}}) = MI^{(j)}. \quad (56)$$

Thus, we get a genie aided lower bound on the probability of error as

$$P_e \geq 1 - \frac{H_b(P_e) + MI^{(j)}}{\Gamma_l(L,N,K,j)} \quad \forall\ j = 1, 2, \ldots, K. \quad (57)$$

This further implies

$$M \geq \frac{(1-P_e)\Gamma_l(L,N,K,j) - H_b(P_e)}{I^{(j)}} \quad \forall\ j = 1, 2, \ldots, K. \quad (58)$$

The above equation holds for all $j = 1, 2, \ldots, K$ and thus, the lower bound on the number of observations follow easily by noting that $H_b(P_e) \to 0$ as $P_e \to 0$. Hence the proof.

VII. Simulations

In this section, we present numerical results illustrating the performance of the algorithms proposed in this work for non-defective subset recovery. We empirically find the number of tests required to achieve a given performance level, which helps to highlight the practical ability of the proposed algorithms to recover a non-defective subset using significantly fewer measurements compared to the indirect method of first identifying the defective subset and then choosing the required number of items from the complement set. The simulation results validate the general theoretical trends, and facilitate a direct comparison of the presented algorithms.

Our setup is as follows. For a given set of operating parameters, i.e., N, K, u, q and M, we choose a defective set $S_d \subset [N]$ randomly such that $|S_d| = K$ and generate the test

output vector y according to (1). We then recover a subset of L non-defective items using the different recovery algorithms, i.e., **RoAl**, **CoAl**, **RoLpAl**, **RoLpAl++** and **CoLpAl**, and compare it with the defective set. The empirical probability of error is set equal to the fraction of the trials for which the recovery was not successful. The recovery is deemed unsuccessful if the output non-defective subset contains at least one defective item. This experiment is repeated for different values of M and L. For each trial, the test matrix $\mathbf{X}$ is generated with random Bernoulli i.i.d. entries, i.e., $\mathbf{X}_{ij} \sim \mathcal{B}(p)$, where $p = 1/K$. Also, for **CoAl** and **CoLpAl**, we set $\psi_{cb} = \psi_{lp} = \frac{\gamma_0 \Gamma}{1-\gamma_0 \Gamma}$. Unless otherwise stated, we set $N = 256$, $K = 16$, $u = 0.05$, $q = 0.1$ and we vary L and M.

Figure 2 shows the variation of the empirical probability of error with the number of tests, for $L = 64$ and $L = 128$. These curves demonstrate the theoretically expected exponential decay of the average error rates with the number of tests. They also show the similarity of the error rate performance of **RoAl** and **RoLpAl**, and the performance improvement offered by **RoLpAl++** at higher values of L. We also note that, as expected, the algorithms that use tests with both positive and negative outcomes perform better than the algorithms that use only tests with negative outcomes.

Figure 3 presents the number of tests M required to achieve a target error rate of 10% as a function of the size of the non-defective subset, L. We note that for small values of L, the algorithms perform similarly, but, in general, **CoAl** and **CoLpAl** are the best performing algorithms across all values of L. We also note that, as argued in Section III-C, **RoLpAl++** performs similar to **RoLpAl** for small values of L and for large values of L the performance of the former is the same as that of **CoLpAl**. Also, as mentioned in Section V, we note the linear increase in M with L, especially for small values of L. We also compare the algorithms proposed in this work with an algorithm that identifies the non-defective items by first identifying the defective items, i.e., we compare the "direct" and "indirect" approach [22] of identifying a non-defective subset. We first employ a defective set recovery algorithm for identifying the defective set and then choose L items uniformly at random from the complement set. This algorithm is referred to as "**InDirAl**" in Figure 3. In particular, we have used "**No-LiPo-**" algorithm [18] for defective set identification. It can be easily seen that the "direct" approach significantly outperforms the "indirect" approach. We also compare against a non-adaptive scheme that tests items one-by-one. The item to be tested in each test is chosen uniformly at random from the population. We choose the top L items tested in all the tests with negative outcomes as the non-defective subset. This algorithm is referred to at "**NA1by1**" (Non-Adaptive 1-by-1) in Figure 3. It is easy to see that the group testing based algorithms significantly outperform the **NA1by1** strategy.

Figure 4 compares the number of tests required to achieve a target error rate of 10% for **CoAl** with the information theoretic lower bound for two different values of K. It can be seen that the empirical performance of **CoAl** is within $O(\log K)$ of the lower bound. The performance of the other algorithms is found to obey a similar behavior.

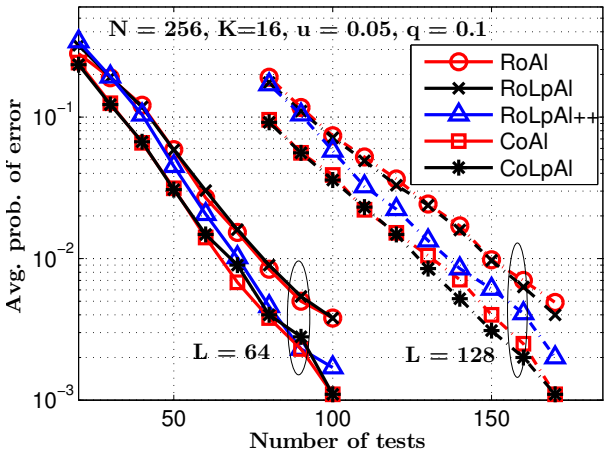


Fig. 2. Average probability of error (APER) *vs.* number of tests M. The APER decays exponentially with M.

TABLE III
ROBUSTNESS OF THE NON-DEFECTIVE SUBSET IDENTIFICATION ALGORITHMS TO UNCERTAINTY IN THE KNOWLEDGE OF K. THE NUMBERS IN THE TABLE ARE $\Delta_M(\hat{K}, K_t)$.

$K_t = 16$, $N = 256$, $L = 128$, $q = 0.1$, $u = 0.05$			
	$\Delta_K = 0.75$	$\Delta_K = 1.5$	$\Delta_K = 2.0$
RoAl	1.13	1.06	1.20
CoAl	1.13	1.04	1.17
RoLpAl	1.09	1.04	1.17
RoLpAl++	1.04	1.00	1.17
CoLpAl	1.11	1.03	1.19

As discussed in Section V, the parameter settings require the knowledge of K. Here, we investigate the sensitivity of the algorithms on the test matrix designed assuming a nominal value of K to mismatches in its value. Let the true number of defective items be K_t. Let $M(\hat{K}, K_t)$ denote the number of tests required to achieve a given error rate when the test is designed with $K = \hat{K}$. Let $\Delta_M(\hat{K}, K_t) \triangleq \frac{M(\hat{K},K_t)}{M(K_t,K_t)}$. Thus, $\Delta_M(\hat{K}, K_t)$ represents the penalty paid compared to the case when the test is designed knowing the number of defective items. Table III shows the empirically computed Δ_M for different values of uncertainty factor $\Delta_K \triangleq \frac{\hat{K}}{K_t}$ for the different algorithms. We see that the algorithms exhibit robustness to the uncertainty in the knowledge of K. For example, even when $\hat{K} = 2K_t$, i.e., $\Delta_k = 2$, we only pay a penalty of approximately 17% for most of the algorithms. Also, as suggested by the analysis of the upper bounds in Section V, the algorithms exhibit asymmetric behavior in terms of robustness and are more robust for $\Delta_k > 1$ compared to when $\Delta_k < 1$.

Figure 5 shows the performance of different algorithms with the variations in the system noise parameters. Again, in agreement with the analysis of the probability of error, the algorithms perform similarly with respect to variations in both the additive and dilution noise.

VIII. CONCLUSIONS

In this work, we have proposed analytically tractable and computationally efficient algorithms for identifying a non-defective subset of a given size in a noisy non-adaptive

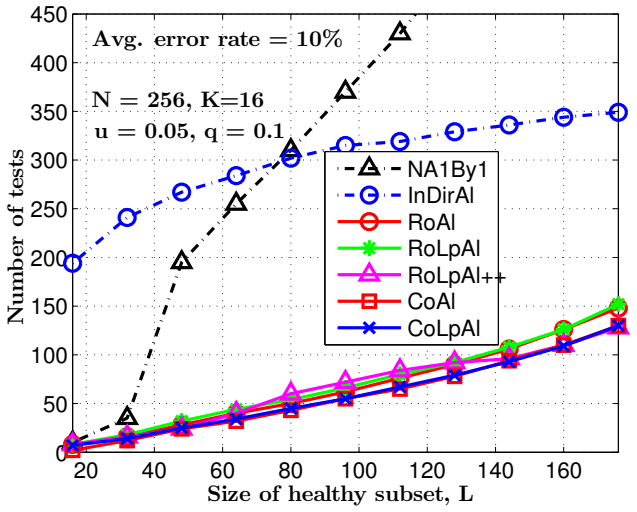


Fig. 3. Number of tests *vs.* size of non-defective subset. Algorithm **CoLpAl** performs the best among the ones considered. The direct approach for finding non-defective items significantly outperforms both the indirect approach ("**InDirAl**"), where defective items are identified first and the non-defective items are subsequently chosen from the complement set [22], as well as the item-by-item testing approach ("**NA1by1**").

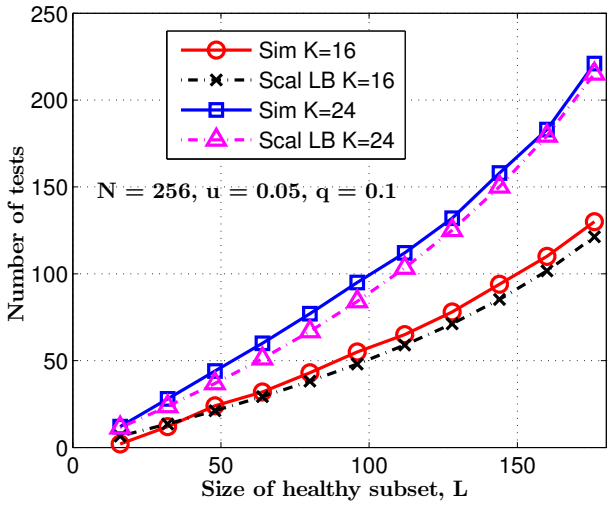
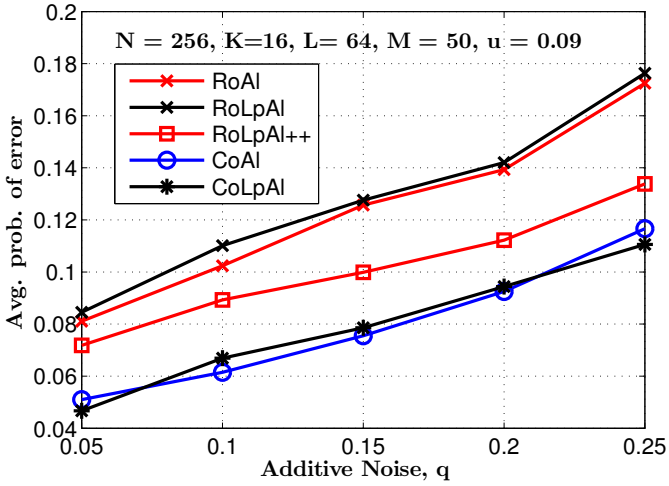
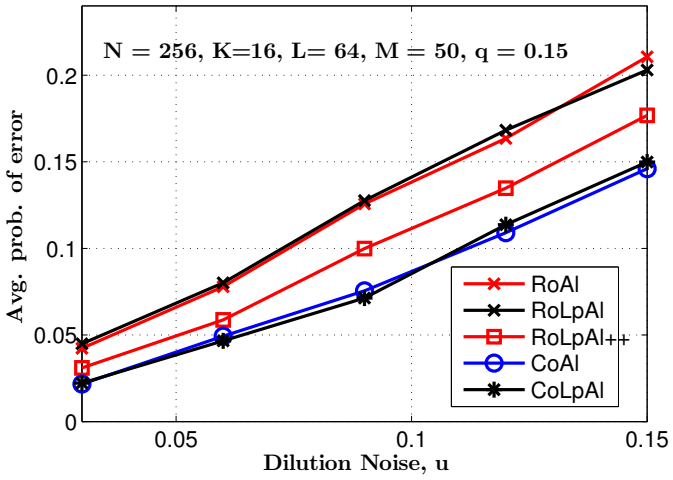


Fig. 4. Comparison of **CoAl** with the *scaled* information theoretic lower bounds. Here, the lower bounds have been scaled by a multiplicative factor of $\log(K)$. The close agreement of the scaled lower bound with the performance of the algorithm shows that **CoAl** is within a $\log(K)$ factor of the lower bounds.

group testing setup. We have derived upper bounds and lower bounds on the number of tests for correct non-defective subset identification, and we have shown that the upper bounds and information theoretic lower bounds are order-wise tight up to a poly-log factor. The algorithms are robust to the uncertainty in the knowledge of system parameters. Also, algorithms that use both positive and negative outcomes, namely **CoAl** and the LP relaxation based **CoLpAl**, gave the best performance for a wide range of values of the size of non-defective subset to be identified, L. In this work, we have considered the randomized pooling strategy. It will be interesting to study deterministic constructions for the purpose of non-defective



(a)



(b)

Fig. 5. Variation of the average probability with (a) additive noise (q) and (b) dilution noise (u).

subset identification; this could be considered in a future extension of this work.

APPENDIX

A. *Proof of Lemma 1*

We note that a test outcome is 0 only if none of the K defective items participate in the test and the output is not corrupted by the additive noise. We get (a) by noting that the probability that an item does not participate in the group test is given by $(1-p)+pu$; (b) follows from (1). For (c), given that $X_{li} = 1$ for any $i \in S_d$, the outcome is 0 only if the i^{th} item does not participate in the test (despite $X_{li} = 1$) *and* none of the remaining $K-1$ defective items participate in the test (either the entry of the test matrix is zero or the item gets diluted out by noise) *and* the test outcome is not corrupted by additive noise. That is, $\mathbb{P}(Y_l = 0 | X_{li} = 1) = u(1-(1-p)u)^{K-1}(1-q) = \gamma_0 \Gamma$. The other part follows similarly. To obtain (d), note that for any $i \in S_d$ and $j \notin S_d$, $\mathbb{P}(Y_l | X_{li}, X_{lj}) = \mathbb{P}(Y_l | X_{li})$. By Bayes' rule and part (b) in this lemma, we get: $\mathbb{P}(X_{li}, X_{lj} | Y_l) =$

$\frac{\mathbb{P}(Y_l|X_{li},X_{lj})}{\mathbb{P}(Y_l)}\mathbb{P}(X_{li})\mathbb{P}(X_{lj}) = \mathbb{P}(X_{li}|Y_l)\mathbb{P}(X_{lj}|Y_l)$. Since the rows of the test matrix $\mathbf{X}$ are i.i.d., part (e) follows by noting that for a given $\underline{y}$ and any item i,

$$\mathcal{T}(i,\underline{y}) = \sum_{j\in\{k\,:\,\underline{y}(k)=0\}} \underline{x}_i(j) - \psi_{cb} \sum_{j\in\{k\,:\,\underline{y}(k)=1\}} \underline{x}_i(j),$$

i.e., it can be represented as a difference of sums of i.i.d. random variables. Hence the proof.

B. Proof of (16), (17) and (27)

Given $\underline{y}$, $\mathcal{T}(i,\underline{y})$ can be written as a linear combination of independent Bernoulli random variables (see Lemma 1). For $\underline{y} = \underline{y}_H$, there will be $M\Gamma^{''}$ zeros and $M(1-\Gamma^{''})$ ones in $\underline{y}$. For any $i \in S_d$, the expressions for μ_i and σ_i^2 now follow from Lemma 1: $\Pr(X_{li}=1|Y_l=0) = p\gamma_0$ and $\Pr(X_{li}=1|Y_l=1) = p\frac{(1-\gamma_0\Gamma)}{1-\Gamma}$ for any $i \in S_d$. Similarly, for $\underline{y} = \underline{y}_L$, the expressions for μ_j and σ_j^2 hold because $\Pr(X_{lj}=1|Y_l=0) = p$ and $\Pr(X_{lj}=1|Y_l=1) = p$ for any $j \in S_z$.

For (27), with $\psi_{cb} = \psi_0 = \frac{\gamma_0\Gamma}{1-\gamma_0\Gamma}$, $\mu_j - \mu_i = Mp\Gamma\left[(1-\eta) - \gamma_0\left((1+\eta) + \frac{1-\Gamma^{'}}{1-\gamma_0\Gamma} - \frac{1-\Gamma^{''}}{1-\Gamma}\right)\right]$, where $\Gamma^{'} = (1-\eta)\Gamma$ and $\Gamma^{''} = (1+\eta)\Gamma$. Note that $(1+\eta) - \frac{1-\Gamma^{''}}{1-\Gamma} = \frac{\eta}{1-\Gamma}$ and $(1-\eta) - \gamma_0\frac{1-\Gamma^{'}}{1-\gamma_0\Gamma} = (1-\eta)(1+\psi) - \frac{\gamma_0}{1-\gamma_0\Gamma}$. Thus, $\mu_j - \mu_i = Mp\Gamma\left((1-\eta)(1+\psi) - \frac{\gamma_0}{1-\gamma_0\Gamma} - \frac{\eta}{1-\Gamma}\right)$. Equation (27) now follows by using the fact that $\frac{1}{1-\gamma_0\Gamma} = 1+\psi_0$.

C. Proof of Proposition 2

We first prove that, for all $i \in \hat{S}_L$, $\underline{\lambda}_2(i) = 0$. The proof is based on contradiction. Suppose there exists $j \in \hat{S}_L$ such that $\underline{\lambda}_2(j) > 0$. This implies, from the complimentary slackness conditions (37), $\underline{z}(j) = 1$ and thus, $\underline{\lambda}_1(j) = 0$. Since this item is in $\hat{S}_L$, it implies that $\underline{1}_N^T\underline{z} > (N-L)$. Hence, $\nu = 0$. From the zero gradient condition in (36), we get $\underline{1}_{M_z}^T\mathbf{X}_z(:,j) = -\underline{\lambda}_2(j) < 0$, which is not possible, as all entries in $\mathbf{X}$ are nonnegative. It then follows that $\forall\ i \in \hat{S}_L$, $\underline{\lambda}_2(i) = 0$. Thus, if $\underline{\lambda}_2(i) > 0\ \forall\ i \in S_d$, then these items cannot belong to the first L entries in the primal solution $\underline{z}$, i.e., $S_d \cap \hat{S}_L = \{\emptyset\}$.

D. Proof of Proposition 3

Suppose $\nu < \theta_0$. Then, by definition of θ_0, there exists i such that $\underline{\lambda}_1(i) = 0$ and $\nu < \underline{1}_{M_z}^T\mathbf{X}_z(:,i)$. Thus, from (36), $\underline{\lambda}_2(i) = \nu - \underline{1}_{M_z}^T\mathbf{X}_z(:,i) < 0$, which violates the dual feasibility conditions (38). Thus, $\nu \geq \theta_0$. Similarly, let $\nu \geq \theta_1$. Then, there exists i such that $\underline{\lambda}_1(i) > 0$ and $\nu \geq \underline{1}_{M_z}^T\mathbf{X}_z(:,i)$. Thus, from (36), $\underline{\lambda}_2(i) = \underline{\lambda}_1(i) + \nu - \underline{1}_{M_z}^T\mathbf{X}_z(:,i) > 0$, which is a contradiction, since, by (37), $\underline{\lambda}_1(i) > 0$ implies $\underline{\lambda}_2(i) = 0$. Thus, $\nu \geq \theta_1$ is not possible.

E. Order-Tight Results for Necessary and Sufficient Number of Tests with Group Testing

In this section, we present a brief sketch of the derivation of the order results for the necessary number of tests presented in Table I. First, $I^{(j)} = H(Y|X_{S^{(K-j)}}) - H(Y|X_{S^{(K-j)}}, X_{S^{(j)}})$ [17], where $H(\cdot|\cdot)$ represents the entropy function [34]. From (1), we have

$$H(Y|X_{S^{(K-j)}}) = \sum_{l=0}^{K-j}\left[\binom{K-j}{l}p^l(1-p)^{K-j-l} H_b\left((1-q)u^l(1-p(1-u))^j\right)\right], \quad (59)$$

$$H(Y|X_{S^{(K-j)}}, X_{S^{(j)}}) = \sum_{i=0}^{K}\left[\binom{K}{i}p^i(1-p)^{K-i}H_b\left((1-q)u^i\right)\right]. \quad (60)$$

We use the results from [35] for bounding the mutual information term. We collect the required results from [35] in the following lemma.

Lemma 2. *Bounds on $I^{(j)}$ [35]: Let $p = \frac{\delta}{K}$. $I^{(j)}$ can be expressed as $I_1^{(j)} + I_2^{(j)}$, where*

$$I_1^{(j)} = \delta e^{-\delta(1-u)}(1-q)\left(u\log u + 1 - u\right)\frac{j}{K} + O\left(\frac{1}{K^2}\right). \quad (61)$$

For the case with $u = 0$ and $q > 0$ we have:

$$I_2^{(j)} = \delta e^{-\delta}\left(\log(\frac{1}{q}) - (1-q)\right)\frac{j}{K} + O\left(\frac{1}{K^2}\right), \quad (62)$$

and for $q = 0$, $u \geq 0$ we have:

$$\delta e^{-\delta}\left((1-u)\left[\log\frac{K}{j\delta(1-u)}\right] - u\right)\frac{j}{K} + O\left(\frac{1}{K^2}\right) \leq I_2^{(j)} \leq \delta e^{-\delta(1-u^2)}\left((1-u)\left[\log\frac{K}{j\delta(1-u)}\right] - u + u^2\right)\frac{j}{K} + O\left(\frac{1}{K^2}\right). \quad (63)$$

Thus, with $\delta = 1$ and large K, neglecting $O(1/K^2)$ terms, (a) For $u = 0$, $q > 0$, we get $I^{(j)} \approx \frac{j}{eK}\log(\frac{1}{q})$; and (b) For $q = 0$, $0 \leq u \leq 0.5$, simplifying further, we get

$$\frac{j}{eK}(1-u)\log\frac{K}{j} \lessapprox I^{(j)} \lessapprox \frac{j}{e^{1/2}K}(1-u)\left(\log\frac{K}{j} + 1\right). \quad (64)$$

In the above, we have used the notation "$\approx$" and "$\lessapprox$" to highlight the fact that $O(\frac{1}{K^2})$ terms have been neglected in the above expressions for $I^{(j)}$. The order results for the lower bounds now follow by first noting that $\max_{1\leq j\leq K}\frac{\Gamma_l(L,N,K,j)}{I^{(j)}} \geq \frac{\Gamma_l(L,N,K,1)}{I^{(1)}}$, and, for the scaling regimes under consideration the combinatorial term, $\Gamma_l(L,N,K,1)$ can be asymptotically bounded as $\lim_{N\to\infty}\Gamma_l(L,N,K,1) \geq \log\frac{1-\beta_0}{1-\alpha_0-\beta_0}$.

F. Affine characterization of the function $\frac{H_b(\alpha)}{1-\alpha}$

Lemma 3. *Let $H_b(\cdot)$ represent the binary entropy function. Then, for $0 < \alpha \leq \alpha_h < 1$, there exist positive constants $c_0, c_1 > 0$, with c_1 depending on α_h, such that*

$$\frac{H_b(\alpha)}{1-\alpha} \leq c_0\alpha + c_1. \quad (65)$$

To establish (65), we note that

$$\begin{aligned}\frac{H_b(\alpha)}{1-\alpha} &= -\frac{\alpha}{1-\alpha}\log(\alpha) - \log(1-\alpha)\\ &= \frac{\alpha}{1-\alpha}\sum_{i=1}^{\infty}\frac{(1-\alpha)^i}{i} + \sum_{i=1}^{\infty}\frac{\alpha^i}{i} \qquad (66)\\ &\leq \alpha\left(1+\frac{(1-\alpha)}{2}+\frac{(1-\alpha)^2}{3}\right)\\ &\quad + \frac{\alpha(1-\alpha)^3}{4}\left(\sum_{i=1}^{\infty}(1-\alpha)^{i-1}\right) \qquad (67)\\ &\quad + \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{3} + \frac{\alpha^4}{4}\left(\sum_{i=1}^{\infty}\alpha^{i-1}\right)\\ &\leq \frac{17}{6}\alpha + \frac{1}{4}\left[(1-\alpha)^3 + \frac{\alpha^4}{1-\alpha}\right] \leq c_0\alpha + c_1,\end{aligned}$$

where $c_0 = 17/6$ and c_1 is obtained by appropriately bounding the second term when $0 \leq \alpha \leq \alpha_h$. In particular, for $\alpha_h \leq 0.5$, $c_1 = 0.25$ will satisfy (65).

G. Chernoff Bounds

Theorem 5. *([38], Ch. 4) Let $X_1, X_2, \ldots, X_n$ be independent $\mathcal{B}(p)$ random variables. Let $X = \sum_{i=1}^{n} X_i$ and let $\mu = \mathbb{E}(X)$. Then, for any $0 < \delta < 1$, the following Chernoff bounds hold:*

$$\mathbb{P}\left(X \geq (1+\delta)\mu\right) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) \qquad (68)$$

$$\mathbb{P}\left(X \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\delta^2\mu}{2}\right). \qquad (69)$$

Theorem 6. *(Bernstein Inequality [37]) Let $X_1, X_2, \ldots, X_n$ be independent real valued random variables, and assume that $|X_i| < c$ with probability one. Let $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}(X)$ and $\sigma = Var(X)$. Then, for any $\delta > 0$, the following hold:*

$$\mathbb{P}\left(X > \mu + \delta\right) \leq \exp\left(-\frac{\delta^2}{2\sigma^2 + \frac{2}{3}c\delta}\right) \qquad (70)$$

$$\mathbb{P}\left(X < \mu - \delta\right) \leq \exp\left(-\frac{\delta^2}{2\sigma^2 + \frac{2}{3}c\delta}\right). \qquad (71)$$

ACKNOWLEDGMENTS

The authors thank Anuva Kulkarni for several interesting discussions in the initial phase of this work. The authors are also deeply grateful to the anonymous reviewers for their detailed feedback, which helped to correct several technical and non-technical errors in the paper.

REFERENCES

[1] A. Sharma and C. R. Murthy, "On finding a subset of non-defective items from a large population using group tests: Recovery algorithms and bounds," in *Proc. ICASSP*, Apr. 2015, pp. 4150–4154.

[2] R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, Dec. 1943.

[3] D. Du and F. Hwang, *Pooling designs and non-adaptive group testing: Important tools for DNA sequencing*, World Scientific, 2006.

[4] M. Sobel and P. A. Groll, "Group testing to eliminate efficiently all defectives in a binomial sample," *Bell System Technical Journal*, vol. 38, no. 5, pp. 1179–1252, 1959.

[5] H. Q. Ngo and D.-Z. Du, "A survey on combinatorial group testing algorithms with applications to DNA library screening," in *Discrete mathematical problems with medical applications*, vol. 55 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pp. 171–182. Amer. Math. Soc., 2000.

[6] A. J. Macula and L. J. Popyack, "A group testing method for finding patterns in data," *Discrete Appl. Math.*, vol. 144, no. 1-2, pp. 149–157, 2004.

[7] D. M. Malioutov and K. R. Varshney, "Exact Rule Learning via Boolean Compressed Sensing," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2013, pp. 775–783.

[8] S. Dash, D. M. Malioutov, and K. R. Varshney, "Screening for learning classification rules via boolean compressed sensing," in *Proc. ICASSP*, 2014, pp. 3360–3364.

[9] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 185–191, Mar 1985.

[10] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically," *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 249–278, Mar. 2005.

[11] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, "Group testing and sparse signal recovery," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, Oct. 2008, pp. 1059–1063.

[12] W. Kautz and R. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inf. Theory*, vol. 10, no. 4, pp. 363–377, 1964.

[13] P. Erdos, P. Frankl, and Z. Furedi, "Families of finite sets in which no set is covered by the union of r others," *Israel Journal of Mathematics*, vol. 51, no. 1-2, pp. 79–89, 1985.

[14] M Ruszinkó, "On the upper bound of the size of the r-cover-free families," *J. Comb. Theory, Ser. A*, vol. 66, no. 2, pp. 302–310, 1994.

[15] A. G. Dyachkov and V. V. Rykov, "Bounds on the length of disjunctive codes," *Problems of Information Transmission*, vol. 18, no. 3, pp. 7–13, 1982.

[16] A. Sebo, "On two random search problems," *Journal of Statistical Planning and Inference*, vol. 11, no. 1, pp. 23–31, Jan. 1985.

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

[18] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-adaptive group testing: Explicit bounds and novel algorithms," *CoRR*, vol. abs/1202.0206, 2012.

[19] D. Cabric, S. M. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz, "A cognitive radio approach for usage of virtual unlicensed spectrum," in *Proc. of 14th IST Mobile Wireless Communications Summit*, 2005.

[20] FCC, "Et docket no. 02-155," *Spectrum policy task force report*, Nov. 2002.

[21] A. Sharma and C. R. Murthy, "Group testing based spectrum hole search for cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3794–3805, Oct. 2014.

[22] A. Sharma and C. R. Murthy, "On finding a subset of healthy individuals from a large population," *CoRR*, vol. abs/1307.8240, 2013.

[23] H. B. Chen and F. K. Hwang, "A survey on nonadaptive group testing algorithms through the angle of decoding," *Journal of Combinatorial Optimization*, vol. 15, no. 1, pp. 49–59, 2008.

[24] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group testing with probabilistic tests: Theory, design and application," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 7057–7067, Oct. 2011.

[25] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.

[26] J. Yoo, Y. Xie, A. Harms, W. U. Bajwa, and R. A. Calderbank, "Finding zeros: Greedy detection of holes," *CoRR*, vol. abs/1303.2048, 2013.

[27] D. M. Malioutov and M. B. Malyutov, "Boolean compressed sensing: LP relaxation for group testing," in *Proc. ICASSP*, 2012, pp. 3305–3308.

[28] A. G. D'yachkov, "Bounds for error probability for a symmetrical model in designing screening experiments," *Problems Inform. Transmission*, vol. 17, pp. 245–263, 1981.

[29] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," in *Int. Symp. Found. Comp. Theory*, 2009, pp. 62–73.

[30] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications - (extended abstract).," in *ICALP (1)*, Luca Aceto, Monika Henzinger, and Jir Sgall, Eds. 2011, vol. 6755 of *Lecture Notes in Computer Science*, pp. 557–568, Springer.

[31] J. Scarlett and V. Cevher, "How little does non-exact recovery help in group testing?," in *Proc. ICASSP*, March 2017, pp. 6090–6094.

[32] M. B. Malyutov, "The separating property of random matrices," *Mat. Zametki*, vol. 23, no. 1, pp. 155–167, 1978.

[33] R. G. Gallager, *Information Theory and Reliable Communication*, John Wiley & Sons, Inc., New York, NY, USA, 1968.

[34] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.

[35] D. Sejdinovic and O. Johnson, "Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction," in *Proc. Allerton Conf. on Commun., Control and Comput.*, 2010, pp. 998–1003.

[36] Y. Nesterov, *Introductory lectures on convex optimization : A basic course*, Kluwer Academic Publ., 2004.
[37] G. Lugosi, *Concentration-of-Measure Inequalities*, Lecture Notes. Available at http://www.econ.upf.edu/ lugosi/anu.pdf, 2006.
[38] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, New York, NY, USA, 2005.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Mar. 2004.

Abhay Sharma received the B.E. (Hons) Electrical and Electronics Engineering degree from BITS, Pilani, India, in 1996, the M.S. degree in Electrical Engineering from Ohio State University, Columbus, USA, in 2000, and the Ph.D. degree in Electrical Communication Engineering from the Indian Institute of Science, Bangalore, India, in 2015. From 2000 to 2005 he worked with Analog Devices Inc., where he designed and implemented physical layer algorithms for cellular communication. From 2006 to 2009 he worked with Allgo Embedded Systems, Bangalore, India, in the area of video signal processing and emerging W-PAN wireless technologies. He worked as a Staff Engineer at Qualcomm, Inc. Bangalore, in 2015, and is currently working as a Member of Technical Staff at the Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bangalore. His research interests include sparse signal processing, wireless communications, statistical signal processing and machine learning.

Chandra R. Murthy (S'03–M'06–SM'11) received the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology Madras, India, in 1998, the M. S. and Ph. D. degrees in Electrical and Computer Engineering from Purdue University and the University of California, San Diego, USA, in 2000 and 2006, respectively. From 2000 to 2002, he worked as an engineer for Qualcomm Inc., where he worked on WCDMA baseband transceiver design and 802.11b baseband receivers. From Aug. 2006 to Aug. 2007, he worked as a staff engineer at Beceem Communications Inc. on advanced receiver architectures for the 802.16e Mobile WiMAX standard. In Sept. 2007, he joined the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India, where he is currently working as an Associate Professor.

His research interests are in the areas of energy harvesting communications, multiuser MIMO systems, and sparse signal recovery techniques applied to wireless communications. His paper won the best paper award in the Communications Track in the National Conference on Communications 2014. He has 45+ journal papers and about 80 conference papers to his credit. He was an associate editor for the IEEE Signal Processing Letters during 2012-16. He is an elected member of the IEEE SPCOM Technical Committee for the years 2014-16, and has been re-elected for the 2017-19 term. He is a past Chair of the IEEE Signal Processing Society, Bangalore Chapter. He is currently serving as an associate editor for the IEEE Transactions on Signal Processing, the Sadhana Journal and as an editor for the IEEE Transactions on Communications.